Health Technology Assessment 2010; Vol. 14: No. 16

Randomised controlled trials for policy interventions: a review of reviews and meta-regression

S Oliver, AM Bagnall, J Thomas,
J Shepherd, A Sowden, I White,
J Dinnes, R Rees, J Colquitt, K Oliver
and Z Garrett

March 2010
DOI: 10.3310/hta14160

Health Technology Assessment
NIHR HTA programme
www.hta.ac.uk

How to obtain copies of this and other HTA programme reports
An electronic version of this title, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (www.hta.ac.uk). A fully searchable DVD is also available (see below).

Printed copies of HTA journal series issues cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our despatch agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per issue and for the rest of the world £3 per issue.

How to order:

– fax (with **credit card details**)
– post (with **credit card details** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you to either print out your order or download a blank order form.

Contact details are as follows:

Synergie UK (HTA Department)
Digital House, The Loddon Centre
Wade Road
Basingstoke
Hants RG24 8QW

Email: orders@hta.ac.uk

Tel: 0845 812 4000 – ask for 'HTA Payment Services'
(out-of-hours answer-phone service)

Fax: 0845 812 4001 – put 'HTA Order' on the fax header

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *University of Southampton* and drawn on a bank with a UK address.

Paying by credit card
You can order using your credit card by phone, fax or post.

Subscriptions

NHS libraries can subscribe free of charge. Public libraries can subscribe at a reduced cost of £100 for each volume (normally comprising 40–50 titles). The commercial subscription rate is £400 per volume (addresses within the UK) and £600 per volume (addresses outside the UK). Please see our website for details. Subscriptions can be purchased only for the current or forthcoming volume.

How do I get a copy of *HTA on DVD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd/index.shtml). *HTA on DVD* is currently free of charge worldwide.

The website also provides information about the HTA programme and lists the membership of the various committees.

Randomised controlled trials for policy interventions: a review of reviews and meta-regression

S Oliver,[1*] AM Bagnall,[2] J Thomas,[1] J Shepherd,[3] A Sowden,[2] I White,[4] J Dinnes,[3] R Rees,[1] J Colquitt,[3] K Oliver[1] and Z Garrett[1]

[1]Social Science Research Unit, Institute of Education, University of London, UK
[2]Centre for Reviews and Dissemination, University of York, UK
[3]Southampton Health Technology Assessments Centre, Wessex Institute for Health Research & Development, University of Southampton, UK
[4]Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge, UK

*Corresponding author

Declared competing interests of authors: none

Published March 2010

DOI: 10.3310/hta14160

This report should be referenced as follows:

Oliver S, Bagnall AM, Thomas J, Shepherd J, Sowden A, White I, *et al.* Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess* 2010;**14**(16).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (*SciSearch*®) and *Current Contents*®/Clinical Medicine.

NIHR Health Technology Assessment programme

The Health Technology Assessment (HTA) programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA programme then commissions the research by competitive tender.

Second, the HTA programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

Criteria for inclusion in the HTA journal series

Reports are published in the HTA journal series if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this issue of the journal was commissioned by the National Coordinating Centre for Research Methodology (NCCRM), and was formally transferred to the HTA programme in April 2007 under the newly established NIHR Methodology Panel. The HTA programme project number is 06/90/22. The contractual start date was in April 2004. The draft report began editorial review in January 2009 and was accepted for publication in March 2009. The commissioning brief was devised by the NCCRM who specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA programme or the Department of Health.

ISSN 1366-5278

Published by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk), on behalf of NETSCC, HTA.

Printed on acid-free paper in the UK by the Charlesworth Group. MR

Abstract

Randomised controlled trials for policy interventions: a review of reviews and meta-regression

S Oliver,[1*] AM Bagnall,[2] J Thomas,[1] J Shepherd,[3] A Sowden,[2] I White,[4] J Dinnes,[3] R Rees,[1] J Colquitt,[3] K Oliver[1] and Z Garrett[1]

[1]Social Science Research Unit, Institute of Education, University of London, UK
[2]Centre for Reviews and Dissemination, University of York, UK
[3]Southampton Health Technology Assessments Centre, Wessex Institute for Health Research & Development, University of Southampton, UK
[4]Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge, UK

*Corresponding author

Objectives: To determine whether randomised controlled trials (RCTs) lead to the same effect size and variance as non-randomised studies (NRSs) of similar policy interventions, and whether these findings can be explained by other factors associated with the interventions or their evaluation.

Data sources: Two RCTs were resampled to compare randomised and non-randomised arms. Comparable field trials were identified from a series of health promotion systematic reviews and a systematic review of transition for youths with disabilities. Previous methodological studies were sought from 14 electronic bibliographic databases (Applied Social Sciences Index and Abstracts, Australian Education Index, British Education Index, CareData, Dissertation Abstracts, EconLIT, Educational Resources Information Centre, International Bibliography of the Sociological Sciences, ISI Proceedings: Social Sciences and Humanities, PAIS International, PsycINFO, SIGLE, Social Science Citation Index, Sociological Abstracts) in June and July 2004. These were supplemented by citation searching for key authors, contacting review authors and searching key internet sites.

Review methods: Analyses of previous resampling studies, replication studies, comparable field studies and meta-epidemiology investigated the relationship between randomisation and effect size of policy interventions. New resampling studies and new analyses of comparable field studies and meta-epidemiology were strengthened by testing pre-specified associations supported by carefully argued hypotheses.

Results: Resampling studies offer no evidence that the absence of randomisation directly influences the effect size of policy interventions in a systematic way. Prior methodological reviews and meta-analyses of existing reviews comparing effects from RCTs and non-randomised controlled trials (nRCTs) suggested that effect sizes from RCTs and nRCTs may indeed differ in some circumstances and that these differences may well be associated with factors confounded with design. No consistent explanations were found for randomisation being associated with changes in effect sizes of policy interventions in field trials.

Conclusions: From the resampling studies we have no evidence that the absence of randomisation directly influences the effect size of policy interventions in a systematic way. At the level of individual studies, non-randomised trials may lead to different effect sizes, but this is unpredictable. Many of the examples reviewed and the new analyses in the current study reveal that randomisation is indeed associated with changes in effect sizes of policy interventions in field trials. Despite extensive analysis, we have identified no consistent explanations for these differences. Researchers mounting new evaluations need to avoid, wherever possible, allocation bias. New policy evaluations should adopt randomised designs wherever possible.

Contents

Glossary and list of abbreviations

Glossary

Controlled before-and-after study A controlled trial in which outcomes are measured before and after exposure to the intervention.

Controlled clinical trial A controlled trial of a clinical intervention in which people are allocated to receive one of two or more interventions, but not randomly. This term is used when we report the work of other authors using the same term.

Non-randomised controlled trial A controlled trial in which people are allocated to receive one of two or more interventions, but not randomly. This term includes controlled trials of clinical and non-clinical interventions.

Non-randomised study A study with a design that does not include randomisation, with or without a control group, e.g. controlled trial, cohort studies, case-controlled studies, surveys.

Randomised controlled trial A study in which people are allocated at random (by chance alone) to receive one of two or more interventions. One of these interventions is the standard of comparison or control.

List of abbreviations

AEI	Australian Education Index
ASSIA	Applied Social Sciences Index and Abstracts
BEI	British Education Index
CBA	controlled before-and-after study
CCT	clinical controlled trial (not randomised)
CDSR	Cochrane Database of Systematic Reviews
CI	confidence interval
CONSORT	Consolidated Standards of Reporting Trials
DARE	Database of Abstracts of Reviews of Effects
df	degrees of freedom
DoPHER	Database of Promoting Health Effectiveness Reviews
EPDS	Edinburgh Postnatal Depression Score
EPPI-Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ERIC	Educational Resources Information Centre
IBSS	International Bibliography of the Sociological Sciences
ICC	intracluster correlation coefficient
MRC	Medical Research Council
MSM	men who have sex with men
nRCT	non-randomised controlled trial
NRS	non-randomised study
PAIS	Public Affairs Information Service
PHSE	Personal, Social and Health Education
RCT	randomised controlled trial
sdUAI	sero-discordant or unknown status unprotected anal intercourse
SIGLE	System for Information on Grey Literature in Europe
SMD	standardised mean difference
SSCI	Social Science Citation Index
TREND	Transparent Reporting of Evaluations with Nonrandomized Designs

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices, in which case the abbreviation is defined in the figure legend or in the notes at the end of the table.

Executive summary

Background

While the randomised controlled trial (RCT) is generally regarded as the design of choice for assessing the effects of health care, within the social sciences there is considerable debate about the relative suitability of RCTs and non-randomised studies (NRSs) for evaluating public policy interventions.

Objectives

To determine whether RCTs provide the same effect size and variance as NRSs of similar policy interventions, and whether these findings can be explained by other associated factors.

Methods

This study employed four approaches:

1. Resampling studies: comparing controlled trials that are identical in all respects other than the use of randomisation by 'breaking' the randomisation in a trial to create smaller non-randomised trials and smaller randomised trials by resampling randomised and non-randomised comparisons from the data.
2. Replication studies: comparing randomised and non-randomised arms of controlled trials mounted simultaneously in the field.
3. Investigating comparable 'field' studies: controlled trials drawn from systematic reviews that include both randomised and non-randomised studies. These include structured narrative reviews and sensitivity analyses within meta-analyses.
4. Meta-epidemiology: investigating associations between randomisation and effect size using a pool of more diverse randomised and non-randomised studies within broadly similar areas. These more diverse studies can be drawn from across reviews addressing different questions, or from broad sections of literature.

This study sought earlier reports of all four approaches and conducted new analyses for three of these approaches (1, 3 and 4 above) across a range of public policy sectors. The new analyses were strengthened by testing pre-specified associations supported by carefully argued hypotheses. Data were drawn from: two RCTs of policy interventions for resampling studies; comparable studies drawn from systematic reviews of health promotion and of transition for youths with disabilities; and a systematic search for prior work. The search strategy comprising free text terms for RCT and non-randomised studies (e.g. non-experimental, pseudorandom, semi-random) was applied to 14 electronic bibliographic databases spanning health, education, social policy and social science in June and July 2004 [Applied Social Sciences Index and Abstracts (ASSIA), Australian Education Index (AEI), British Education Index (BEI), CareData, Dissertation Abstracts, EconLIT, Educational Resources Information Centre (ERIC), International Bibliography of the Sociological Sciences (IBSS), ISI Proceedings: Social Sciences and Humanities, PAIS International (Public Affairs Information Service), PsycINFO, SIGLE (System for Information on Grey Literature in Europe), Social Science Citation Index (SSCI), Sociological Abstracts]. This was supplemented by citation searching for key authors, contacting review authors and searching key internet sites.

For investigating comparable field studies, and the meta-regression, studies were coded for characteristics of the population, policy intervention and evaluation. Differences in effect sizes between studies were investigated using random-effects meta-regression to allow for unexplained heterogeneity between studies as well as the known uncertainty in estimated effect sizes (measured by their standard errors). Associations between different characteristics of the studies and whether or not they employed randomisation were measured using chi-squared tests.

Results

Reviews of methodological studies and empirical reviews

Prior methodological reviews included a review of within-study comparisons of randomised and non-randomised participants, six single meta-analyses and one review of meta-analyses. Between

them these covered interventions for preventing juvenile delinquency, treatment of alcohol abuse, and other psychological, mental health or health-care interventions. These studies investigated whether randomisation influenced effect sizes. Most also investigated the influence of other variables or modifiers of effect such as population, sample size, attrition, intervention, type of control group and publication status. The results suggest that effect sizes from RCTs and non-randomised controlled trials (nRCTs) may indeed differ in some circumstances and that these differences may well be associated with factors confounded with design. Inter-relationships among variables make it difficult to determine the likely impact of any one factor.

A systematic review of meta-analyses of existing reviews comparing effects from RCTs and nRCTs found that the effect sizes were similar in five reviews, dissimilar in eight reviews, and mixed in three. Most reviews appeared to ignore the variability associated with effect size. Considerable variation in the studies pooled within reviews, in terms of population, intervention, outcome and other methodological details, makes it difficult to separate the potential effect of random assignment from the potential effects of all the other variables.

Resampling studies

Re-analysis of data from two trials suggests that nRCTs can give the same answers as RCTs. This was a tightly controlled examination in which the only factor that was different between the RCTs and nRCTs was randomisation.

Comparable 'field' studies and meta-epidemiology

In the examination of trials sampled from systematic reviews we found considerable variation, with RCTs producing smaller effect sizes than nRCTs in systematic reviews conducted at the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) (using within review comparisons and meta-epidemiology) and larger effect sizes than nRCTs in the studies reviewed by Colorado State University (using meta-epidemiology alone).

Investigation of potential confounding factors in the EPPI-Centre reviews suggests that RCTs have smaller effect sizes, even though their sample sizes tend to be smaller with participants allocated individually (both attributes associated to some extent with effect size) and their theoretical frameworks more readily apparent. Other attributes commonly associated with quality were not associated with randomisation or effect size: attrition rates, time to follow-up or quality of reporting.

Conclusions

From the resampling studies we have no evidence that the absence of randomisation directly influences the effect size of policy interventions in a systematic way. At the level of individual studies, non-randomised trials may lead to different effect sizes, but this is unpredictable. Many of the examples reviewed and the new analyses in the current study reveal that randomisation is indeed associated with changes in effect sizes of policy interventions in field trials. Despite extensive analysis, we have identified no consistent explanations for these differences.

Recommendations for research

1. Policy evaluations should adopt randomised designs whenever possible.
2. Policy evaluations should also adopt other standard procedures for minimising bias and conducting high-quality assessment of effects of intervention, particularly blinded allocation of either individuals or groups and the avoidance of small sample sizes.
3. Feasibility studies of randomising geographical areas, communities and regions should be carried out for evaluating policy interventions in a range of sectors, implemented within interventions, communities and across regions.
4. Feasibility studies of blinded allocation should be carried out for policy interventions in a range of sectors, implemented within interventions, communities and across regions.
5. Clear descriptions should be included in systematic reviews of how judgements of equivalence (or otherwise) have been reached when comparing the effects found in randomised and non-randomised studies of policy interventions.
6. Research is required into the reasons for choosing randomisation or not, particularly in the presence and absence of an explicit collective plan of action.

Chapter 1

Policy interventions and their evaluation

The NHS Research and Development Methodology Programme identified the need to investigate the implications of randomised and non-randomised evaluation designs for assessing the effectiveness of policy interventions.

The work of Sacks[1] and the classic paper by Schulz[2] showed that the benefit ascribed to a clinical intervention depends on the methodology used in the study. For instance, the effect size tends to be more pronounced in historically controlled than in randomised controlled trials (RCTs) of the same intervention, and in poorly randomised than in rigorously randomised studies. Concurrently controlled and randomised studies produce more similar results,[3] although the researchers urge caution when interpreting this finding, as the number of studies included in the review was small. While research comparing the effect sizes produced by different study designs is growing in clinical topics, little work has been done with respect to policy/management interventions. These are defined as those interventions that are not confined to an individual practitioner, and include, but are not limited to, health. Examples would include peer-led teaching and health promotion in schools. Non-randomised studies (NRSs) in these areas may be cross-sectional or before and after (i.e. either with or without baseline measurements), and many of the randomised studies may be cluster randomised.

The Research and Development Methodology Programme required the compilation of existing studies that compare findings of randomised and non-randomised studies of policy interventions in order to: (1) analyse effect sizes in which similar interventions have been examined by different methods and (2) extract and summarise the information bearing on the effects of study type and quality of study findings, with the ultimate aim of learning about biases (mean bias and spread of biases) associated with different study types.

Defining policy and intervention

The study required a definition of 'policy interventions' that would facilitate selection of systematic reviews of policy interventions and individual trials of policy interventions. The term 'policy intervention' is used throughout the UK government's policy hub website (www.policyhub.gov.uk/search_result.asp), but without a definition. We have been unable to find a definition of 'policy intervention'. The closest we have found in dictionaries are definitions of 'policy' as:

> a course of action or principle adopted or proposed by a government, party, individual, etc.
>
> Oxford English Dictionary

> a plan of action adopted by an individual or social group
>
> WordNet, a lexical database for the English language (www.cogsci.princeton.edu/cgi-bin/webwn)

> policy (plan) noun [C] a set of ideas or a plan of what to do in particular situations that has been agreed officially by a group of people, a business organization, a government or a political party
>
> Cambridge Advanced Learner's Dictionary

We are not alone in struggling to define 'policy' and 'policy intervention'. In seeking a sound and operational definition of policy intervention, we have referred to the public policy literature and the literature about evaluation and evidence-informed policy/practice.

Jenkins[4] observed that:

> Pursuit of the question 'what is public policy?' leads one down the tangled path towards a definition where many have been before and

> from which few have emerged unscathed. There is, as Lineberry and Masotti (1975) point out, little in the way of a consistent conceptualization of the term 'policy' itself and pages could be, and have been, filled with competing definitions. The problem may be to provide an account that captures the detail and density of the activities embraced by the policy arena. With this detail in mind, it is worth considering the following definition of public policy:
>
> 'a set of interrelated decisions taken by a political actor or group of actors concerning the selection of goals and the means of achieving them within a specified situation, where those decisions should, in principle, be within the power of those actors to achieve' (Roberts 1971)
>
> [This definition] stresses the point that policy is more than a single decision. As Anderson (1975) has argued, 'policy making typically involves a pattern of action extending over time and involving many decisions'.

The US has a strong history of employing controlled trials to evaluate 'social programs'[5] that fall within our understanding of policy interventions. For instance, House[6] cites the dictionary definition of intervention as 'interference that may affect the interests of others'. He goes on to talk about the inherently 'messy' social context within which emerge 'complex disordered events we call interventions' (House, p. 323). House distinguishes between generic development, policy-making and site-specific interventions (p. 325). 'Policymaking interventions' consist of 'establishing rules and guidelines'; in the case of education, Standard Attainment Tests in the UK would be an up to date example of the type of educational intervention that House discusses under this heading.

Some of the literature refers to 'social interventions' and 'policy analysis and evaluation research'. Haveman[7] described social interventions as programmes that, when evaluated, can inform policy, but some 'are' policy. There is extensive literature on the 'War on Poverty–Great Society' developments in the US initiated in 1965, in which the various types of social intervention that represented local changes in policy were evaluated by government mandate and these evaluations were considered directly relevant to government policy.[8]

Defining 'policy intervention'

The focus of our investigation was on evaluations of interventions for public policy or service organisation and management that:

- are intended to serve communities or populations
- require more than the efforts of individual practitioners to apply
- are not a one-to-one service.

We have adapted the definition of public health interventions provided by Rychetnik *et al.*[9] In order to embrace broader public policy, this definition of interventions is paraphrased as:

> a set of actions with a coherent objective to bring about change or produce identifiable outcomes. These include policy, regulatory initiatives, single strategy projects or multi-component programmes. Policy interventions are intended to serve communities or populations. They are distinguished from one-to-one services that are for the benefit of individuals.

These interventions require more than the efforts of individual practitioners to be applied. They may include legislation or regulation; setting of policy or strategy at the level of national or local government, or institutions; the provision or organisation of services; environmental modification; or facilitating lay or public delivered support/education. These interventions may fall within public policy for health, education, social care, welfare, housing, criminal justice, transport and urban renewal.[10]

Another interpretation of the term 'policy intervention' refers to intervening in policy making rather than policy intervention. Devlin *et al.*[11] considered the parameters of policy-making interventions in relation to service user perspectives on HIV policy. To paraphrase them:

> policy interventions seek to influence decision-making … and ensure that policy supports or at least does not impede [services]. These interventions relate therefore to local and national policy makers (within governmental and statutory sectors) and local and national resource allocators (for example government departments and local authorities). They can also involve seeking to influence those people or agencies charged with the production and supply of information to

> support policy development and resource allocation. Therefore, they might also seek to influence applied and academic social researchers, epidemiologists, policy advisors and local public health surveillance personnel (collectively called, the research and policy community).

Examples of interventions that impact on policy-makers and seek to influence their (drafting) policy and legislation might include:

- lobbying government departments, local authorities, research bodies
- taking part in national consultation processes undertaken by policy and lobbying organisations and government
- joining professional associations, research/policy forums
- applying for funding from local authorities or government bodies
- subscribing to information sources of national policy-makers and lobbyists.

This distinction between collectively effecting change through setting policy, and collectively effecting change through implementing prior policy decisions is apparent in the policy analysis literature.[12] Harrison[12] describes policy as a process, rather than simply as an output of a decision, or an input to management. The policy process begins within the arena of political science with setting agendas around problematic issues, and progresses to designing and evaluating efforts to solve these problems. Thus, a 'policy intervention' may be either a method for influencing the policy-making process, or a method for influencing the policy implementation process.

Evaluating public policy interventions

Within the area of public policy there has been wide debate about the suitability of experimental evaluation methods. While it has been suggested that the RCT should be the 'gold standard' and used whenever possible,[13,14] others have argued that evaluating social and policy interventions is a complex task and that RCTs, and experimental designs in general, are not always practical or even desirable.[15,16] Nutbeam[17] suggests that complex multicomponent interventions (e.g. directed at communities or regions using a range of media, delivered in a number of settings) are more likely to be effective in bringing about population health gains than 'single issue' initiatives (e.g. directed at individuals or small groups, using fewer media, delivered in a particular setting), but are much harder to evaluate. For example, it might not be possible to allocate whole communities or regions to study groups randomly, and it is not easy to isolate the effects of competing interventions, thus confounding the results. The World Health Organization commenting on the evaluation of health promotion goes as far as saying:

> The use of randomised control trials to evaluate health promotion initiatives is, in most cases, inappropriate, misleading and unnecessarily expensive.[18]

Oakley *et al.*[19] have summarised the objections to RCTs for evaluating social interventions, arguably a category that includes all policy interventions, as: randomised experiments oversimplify causation, cannot be carried out in complex institutional and other settings or to test complex interventions, ignore the role of theory in understanding intervention effectiveness, are inappropriate in circumstances in which 'blinding' is impossible, are politically unacceptable and too expensive, have been tried and failed, are unethical because valued treatments are withheld from control groups and/or experimental/quantitative research is inherently exploitative, and perfectly good alternatives to RCTs that pose none of these problems exist and should therefore be used instead. These objections focus largely on the science, ethics and feasibility of randomisation. They have led to a dearth of randomised studies in some policy areas, which needs to be taken into account when preparing research syntheses, and to research communities who remain disinclined to mount randomised evaluations. Oakley *et al.*[19] used three recent UK trials of policy interventions (day care for preschool children, social support for disadvantaged families, and peer-led sex education for young people) to consider issues relating to the use of randomisation and suggest some practical strategies for its use in trials of social interventions. Their refutations of the objections to RCTs are supported by an analysis of the relevant theoretical literature.[20]

Indeed, experimental evaluations have long been considered the optimal design for evaluation in some fields of social policy, particularly in the US.[5] Oakley[21] cites examples of experimental policy evaluations that date back as far as the early decades of the twentieth century, and discusses how experimental methods became popular, particularly in the US, between the 1960s and the 1980s to evaluate the effectiveness of public policy:

> This history is conveniently overlooked by those who contend that randomised controlled trials have no place in evaluating social interventions. It shows clearly that prospective experimental studies with random allocation to generate one or more control groups is perfectly possible in social settings. (p. 1239)

More recently there have been key trials that have evaluated the effectiveness of so-called 'complex' health promotion interventions. For example, the North Karelia Youth Program[22] was a large-scale multicomponent intervention evaluated using an RCT involving over 4000 participants, featuring a range of activities including classroom education, media campaigns, changes to nutritional content of school meals, health screening, and health education initiatives in the workplace.

Particularly innovative are experimental evaluations of interventions addressing environmental or structural factors integrating sexual health and employment/economic policy. Examples include a matched controlled trial of the impact of an employment creation programme on teenage pregnancy[23] and a cluster RCT of microcredit schemes for impoverished women to develop increased economic independence, social status, and power within sexual negotiations, thereby reducing HIV transmission.[24]

An analysis of the reasons for not adopting the RCT design concludes that, despite serious practical objections and partial remedies, RCTs are logically and empirically superior to all currently known alternatives.[25] The view that the RCT is inappropriate to test the success of policy interventions is refuted by a bibliometric analysis, which concludes that between 6% and 15% of impact evaluations of childhood interventions in education and justice employ a randomised design.[26]

Our own experience of conducting RCTs supports their use for evaluating social interventions.[19] Our experience of conducting and evaluating systematic reviews reveals their widespread use elsewhere. Of the 75 evaluation studies identified in a recent systematic review of interventions to promote healthy eating and physical activity among young people,[27,28] 31 (41%) used an RCT design, 30 (40%) used a controlled trial (without randomisation), and 14 (19%) used only one study group with outcomes measured before and after the intervention. While the evidence base in this area is likely to comprise a vast range of evaluation designs, the role of experimental evaluation cannot be discounted.

Efforts to consolidate this evidence base have increased, together with a recent surge in production of systematic reviews of the effects of policy interventions.[29,30] Reviews have recently been completed, or are in the process of being completed, in the areas of health (e.g. interventions to improve vaccination coverage),[31] education (e.g. after school programmes)[32] and criminology (e.g. 'Scared straight' interventions to discourage juvenile delinquency).[33]

Randomisation and effect sizes of clinical interventions

The RCT is widely regarded as the design of choice for evaluating the effectiveness of clinical interventions in health care, as it can provide the most internally valid estimate. The main benefit of the RCT is the use of a randomisation procedure that, when properly concealed, ensures that the subjects receiving the treatment and control are equal with respect to all conditions except for receiving the treatment or the control. With sufficient sample sizes, and a truly random generation of the allocation sequence, comparison groups should on average be equal with respect to both known and unknown prognostic factors at baseline.[34] RCTs also have written protocols specifying, and thus standardising, important aspects of participant enrolment, intervention, observation and analysis.[35]

Our knowledge of the importance of certain design features of RCTs has been derived primarily in the field of clinical health-care interventions.[2,36,37] Meta-epidemiological techniques have successfully been used to investigate variations in the results of RCTs of the same intervention according to features of their study design.[38] Substantial numbers of systematic reviews of RCTs have been identified, and results compared between the trials meeting and not meeting various design criteria such as proper randomisation, concealment of allocation and blinding. These comparisons have then been aggregated across the reviews to obtain an estimate of the systematic bias removed by the design feature.[2,37] The results have been shown to be reasonably consistent across clinical fields, providing some evidence that meta-epidemiology may be a reliable investigative technique.[39]

The use of meta-epidemiology has also been extended from the comparison of design features within a particular study design to comparisons between study designs. A recent Health Technology Assessment report reviewed eight such examples:[40] seven considered medical interventions, while one considered psychological interventions. The conclusions of these reviews varied, partly due to variations in their methods and rigour but also because of limitations in the meta-epidemiological methods used. The only robust conclusion that can be drawn is that in some circumstances the results of randomised and non-randomised studies differ, but it cannot be proved that differences are not due to other confounding factors. The key lessons that can be learned from this work are:

- The identification and selection of comparisons of randomised and non-randomised evidence should be systematic. This will not overcome the problem of selective publication of primary studies (if studies with positive results are more likely to be published, regardless of design, meta-epidemiological reviews will find designs showing intervention effects in the same direction if not of similar magnitude) but should at least ensure that all available comparisons are included regardless of whether designs show similar or conflicting results.
- To reduce confounding from factors other than lack of randomisation, randomised and non-randomised studies should be assessed for differences in the participants, interventions and outcomes. The possibility of temporal confounding of study types (NRSs typically being performed prior to the RCTs) should also be assessed.
- The similarity of randomised and non-randomised studies should be assessed for differences in study methods other than allocation. Discrepancies and similarities between study designs could be partly explained by differences in other unevaluated aspects of methodological quality of the RCTs and/or the NRSs, such as blinding or intention-to-treat analysis.
- Sensible, objective criteria should be used to determine differences or equivalence of study findings as these can have a large influence on the conclusions drawn. The amount of data available is also important; for example in one review, for each intervention five RCTs on average were compared with four NRSs. Hence the absence of a statistically significant difference cannot be interpreted as evidence of 'equivalency', and clinically significant differences in treatment effects cannot be excluded.[40]

These previous investigations also suggest that there may be variability in the direction of bias introduced when randomisation is not used. Selection bias is commonly thought of as resulting from the systematic selection of either high or low risk participants to receive an intervention. This would lead to the intervention group being 'heavily weighted by the more severely ill'[41] or alternatively including those least likely to suffer adverse consequences from an intervention (less severely ill). If in fact selection bias arises due to haphazard variations in case-mix, there will be a mixture of under- and overestimates of the treatment effect. The results might all be biased, but not all in the same direction.[40] In these circumstances, an increase in the heterogeneity of treatment effect (beyond that expected by chance) rather than (or as well as) a systematic bias would be expected. Deeks *et al.*[40] suggest that a formal statistical comparison should aim to compare the heterogeneity in treatment effects, and not just the average treatment effects between randomised and non-randomised groups.

Randomisation and effect size of policy interventions

The effects of policy interventions have been assessed through the use of RCTs, nRCTs and other study designs. The choice has been influenced by the relative rigour of the designs and the feasibility in the circumstances of applying prospective designs and random allocation of interventions. The weight given to each of these influences (rigour and feasibility) when embarking on policy evaluations may be driven by philosophy, as much as by research evidence.

Although studies largely from clinical areas have identified detailed design features of rigorous RCTs that reduce systematic bias in estimating effect sizes,[39] meta-epidemiological investigations of medical and psychological interventions concluded that it is less clear what influences the differences in results drawn from randomised and non-randomised studies, as results of NRSs sometimes, but not always, differ from results of randomised studies of the same intervention.[40] There is growing evidence in the meta-analytic literature that even strong quasi-experimental designs assessing

criminology are more likely to report a result in favour of treatment and less likely to report a harmful effect of treatment than randomised studies.[42]

Chapter 2 considers the methodologies appropriate for investigating the extent and possible causes of such differences.

Chapter 2

Methodology: design and data sources

Four approaches have been adopted in previous studies to investigate the relationship between randomisation and effect size of interventions:

1. Comparing controlled trials that are identical in all respects other than the use of randomisation by 'breaking' the randomisation in a trial to create non-randomised trials. These are often called resampling studies.
2. Comparing randomised and non-randomised arms of controlled trials mounted simultaneously in the field. These are replication studies.
3. Comparing similar controlled trials drawn from systematic reviews that include both randomised and non-randomised studies. These include structured narrative reviews and sensitivity analyses within meta-analyses.
4. Investigating associations between randomisation and effect size using a pool of more diverse randomised and non-randomised studies within broadly similar areas. These more diverse studies can be drawn from across reviews addressing different questions, or from broad sections of literature. This is known as meta-epidemiology.

This study sought reports of all four approaches conducted by others, and built on their work by conducting original research with new analyses for three of these approaches (1, 3 and 4 above) across a range of public policy sectors.

The latter two approaches were strengthened in new analyses by testing pre-specified associations supported by carefully argued hypotheses.

Resampling of randomised controlled trials

Resampling studies re-analyse data from RCTs to explore widely used alternatives to randomisation such as: comparing areas, matching areas and adjusting for differences between groups using multivariate analysis. By 'breaking' the randomisation in the trials, this analysis creates non-randomised trials and explores the extent to which established alternatives to randomisation are able to find the same results as the original RCTs.

Because these studies are based on trials that are identical other than the use of randomisation, they explore the direct association between randomisation and effect size without being confounded by other factors that might influence effect size when calculated from similar, but not identical, field trials.

Such studies were sought in a methodological review described in Chapter 4. Two new re-sampling studies are reported in Chapter 7. The data are drawn from two trials of social support for families with young children, one carried out in the UK, and the other in Canada. Both trials span the health and social care sectors.

Replication studies

Replication studies assess the effects of intervention from different comparisons within the same study. In order to investigate the role of randomisation, replication studies compare the effect sizes from randomised and non-randomised comparisons. Such studies were sought in the methodological review described in Chapter 4. We did not have access to data from other replication studies for new analyses.

Comparable field studies

Randomised and non-randomised evaluations drawn from a single review are comparable field studies for addressing the following questions. Do randomised and non-randomised evaluations lead to differences in effect sizes and variance? If so, are these differences due to the randomisation or to other factors associated with randomisation? If differences are due to characteristics of the interventions or their evaluation, is it possible to overcome these difficulties in the design or analysis of evaluations and/or research syntheses whether or not studies are randomised?

Extensive exploratory analyses would be expected to identify some associations between randomisation and other factors, if only by chance. To avoid the risk of identifying chance associations, we tested a limited number of well-argued associations for which hypotheses rested on our understanding of policy interventions, research communities and evaluation methodology, or arose from previous research. By drawing on published literature, we proposed a series of potential confounders, and argued how these are likely to be interrelated (see Chapter 3).

We proposed several possible conclusions for an exploratory investigation of published evaluations:

- There is *no systematic difference* between the effect sizes of RCTs of policy interventions and the effect sizes of non-randomised trials; so *non-randomised trials may be adequate* to evaluate policy interventions.
- The effect sizes of RCTs of policy interventions *are systematically different* from the effect sizes of non-randomised trials; this difference cannot be explained by any other variables in the interventions or their evaluation, so it is assumed that *randomisation is required to control for unidentifiable influences*; and examples of RCTs that are ethically and scientifically sound should be sought to model future evaluations of the effects of policy interventions.
- The effect sizes of RCTs of policy interventions *are systematically different* from the effect sizes of non-randomised trials; however, this difference can be explained by one or more other variables in the evaluation, such as baseline differences, that are amenable to statistical adjustment in order to take into account the difference; in these circumstances, *non-randomised trials with the appropriate corrections may be adequate to evaluate the impact of policy interventions*.
- Randomised trials of policy interventions lead to *systematically different effect sizes* compared with non-randomised trials; where this difference can be explained but not quantified by one or more other variables in the evaluation (but this difference is not amenable to adjustment) the *strength of evidence* to support decisions about policy interventions is *necessarily weaker.* An example may be small single centred randomised trials led by enthusiasts, compared with large multicentre uncontrolled trials attempting to assess the impact as an intervention is implemented more widely.
- The *variance of non-randomised trials is greater* than that of RCTs. Deeks *et al.*[40] have found that, while non-randomised controlled trials (nRCTs) do not differ systematically in their effect sizes, their variance is greater than that of RCTs. This suggests that confidence intervals (CIs) for individual nRCTs should be considered to be larger than stated, which means that *statements of statistical significance should be treated with caution*. If the purely statistical studies show that nRCTs differ from RCTs only in variance (not effect size), we would conclude that nRCTs are biased in ways that cannot be explained simply because they are non-randomised: other biases (e.g. selection or publication bias) are at work which lead us to conclude that nRCTs overstate the statistical significance of their interventions, but not necessarily the size of the effect.

Thus, any investigation of a possible association between randomisation and effect size needs to take into account the similarities or differences of interventions and evaluations in which this association is tested.

Meta-epidemiology

Our study extended the use of meta-epidemiology by Deeks *et al.*[40] to policy interventions within health and other sectors to draw together systematically what is already known about the choice of study design for evaluating policy. Lessons learnt from that meta-epidemiological review were then applied to a meta-epidemiological study of policy evaluations using our own data sets.

Policy interventions

In attempting to distinguish 'policy interventions' from interventions examined in earlier methodological studies, our discussions and searches for relevant literature touched on the following issues:

Prior research

From the outset we were aware of a similar methodological study by Deeks *et al.*[40] This study did not have any inclusion or exclusion criteria regarding type of intervention, other than they had to have 'intended effects'. Two chapters of that report are particularly relevant to our work:

- Chapter 3: a review of eight 'meta-epidemiological' reviews that reviewed comparisons of RCTs and NRSs in which the original authors had specifically set out to examine the similarity/differences in results according to randomisation. Deeks *et al.*[40] reviewed interventions that were almost exclusively therapeutic in nature (i.e. aimed to treat or cure disease), but a handful related to the organisation of care and to educational interventions and were eligible for our study.
- Chapter 5: a review of existing systematic reviews that included randomised and non-randomised studies. This chapter included 'policy interventions' as well as clinical/medical interventions. We have already drawn on the reviews within this chapter to outline the range of 'policy interventions' for this study. The study reported here extends the work of Deeks *et al.*[40] by comparing the results of randomised and non-randomised evidence from a wider range of policy interventions.

Resistance to randomised controlled trials

As the purpose of this methodological study is to resolve questions about how essential RCTs are in areas where they are less readily available, we anticipated finding relevant studies (randomised and non-randomised) in areas where there has been some but not complete resistance to RCTs.[14,43,44] These include circumstances in which:

- it is difficult to stop contamination between intervention group(s) and control(s) (e.g. community wide interventions)
- benefit may derive in part from an individual or group actively seeking to participate in the particular intervention (e.g. peer support provided by patient organisations)
- randomisation is not feasible (e.g. legislation)
- interventions are multicomponent (e.g. a combination of health service initiatives, face-to-face health education in schools and in the community plus mass media).

Within this study we shall explore whether differences other than the presence or absence of randomisation could account for any variation that might be found in results, and comment on the extent to which resistance to RCTs is justified.

Complex interventions and their relationship with policy interventions

Although not directly relating to policy interventions, the Medical Research Council (MRC) framework for the development and evaluation of RCTs for complex interventions to improve health seemed to capture the essence of what we have been discussing. However, the MRC definition of complex interventions includes interventions that may be delivered by individuals (e.g. the different social/educational/treatment aspects of physiotherapy distinguish physiotherapy as a complex intervention) (see Appendix 1). We noted that in other less clinical areas, 'multicomponent' interventions was the term of choice, although this usually included multipractitioners too.

Level of policy making

Discussion distinguished policy interventions at different levels (national, regional, community and institution). These distinctions appeared to translate poorly to policy evaluations at these different levels. In particular it was noted that an evaluation of institution-wide policies may precede or follow national endorsement of a policy; the report may not clearly acknowledge which of these circumstances prevail, and whether it is the former or the latter may make little difference to the methodological challenges of evaluation. When a definition of policy intervention was applied to a set of evaluations (see below) it confirmed the observation above that social interventions as programmes, when evaluated, can inform policy, but that some 'are' policy.[7]

Implementation and its relationship with policy

We envisaged many clinical interventions also being 'policy'; for instance, prescribing aspirin following a heart attack. In order to avoid replicating methodological research in the clinical area, we distinguished between, for example, a trial of aspirin treatment for heart attack (a trial of a clinical intervention) and a trial of methods to encourage greater use of aspirin treatment for heart attack (a trial of a social or educational intervention to increase uptake), and included the latter but not the former. We included other similar interventions such as interventions to increase the

uptake of vaccination or screening. This scope made reviews conducted by the Cochrane Effective Practice and Organisation of Care review group particularly relevant.

Developing operational definitions

Examining prior reviews

A draft definition of policy interventions, and illustrative examples, was developed through several rounds of discussion within the research team. It was refined by two researchers independently applying the emerging criteria and definitions to a set of systematic reviews to judge whether each review would be included or excluded. Twenty of these reviews were sampled from the Health Technology Assessment (HTA) report on evaluating NRSs[40] (this was a subsample of a larger set originally chosen for their relevance on the basis of their titles only). The remaining 20 were selected randomly from the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) Database of Promoting Health Effectiveness Reviews (DoPHER; eppi.ioe.ac.uk).

Different categories of policy intervention can be further distinguished by their details, either as elements of policy setting, or as elements of implementing policies such as legislation or regulation, provision or organisation or services, environmental modification, or facilitating education or support delivered by lay people. Examples are offered below.

Setting of policy/strategies

- Government policy (e.g. policies on vaccination/immunisation/screening; fiscal/economic incentives to participate in sport/physical activity; nutritional policies such as the 'National School Fruit Scheme').
- Local government policy (e.g. provision/sponsorship of community based activities to promote cultural diversity and social cohesion, and to prevent discrimination and violence, such as 'Neighbourhood Renewal Strategies'; community-wide inter-agency strategies to promote health such as 'Health Action Zones').
- Institutional policy (e.g. school-wide strategies to promote mental and emotional health, such as bullying/harassment prevention; curriculum review to prevent disaffection with school/academic studies; health promoting hospitals).

Legislation/regulation

- Environmental health regulations (e.g. waste disposal, pollution/emissions and its impact on health, smoking restrictions).
- Taxation (e.g. on tobacco, alcohol).
- Advertising/sponsorship regulation (e.g. on tobacco products).
- Food standards regulations (e.g. nutritional content of school meals).

Provision/organisation of services

- Education (e.g. increasing access to education through initiatives such as 'Education Action Zones; vocational strategies to aid transition from school to work such as the 'Connexions' service; class sizes; training the trainer cascades).
- Health promotion (e.g. increasing access to, and uptake of, facilities/resources; initiatives to promote health in the workplace; mass media campaigns; community development; social support).
- Health care (e.g. increasing access to, and uptake of, facilities/resources; effective organisation of services; effective promotion, dissemination and uptake of evidence based clinical practice guidelines).
- Social services (e.g. effective organisation of services; effective alliances with health and education sectors).

Environmental modification

- Creation of safer cities (e.g. improved street lighting to prevent crime; traffic calming schemes, cycle paths/helmets, and speed cameras to prevent injuries).
- Urban renewal (e.g. housing improvement programmes to promote better living conditions/health/sanitation/hygiene).

Facilitating lay/public delivered support/education

- Facilitating one-to-one support (e.g. lay birth partners and fathers supporting women in childbirth; enabling carer/family support for chronic illness; peer-delivered counselling in schools).
- Facilitating one-to-group support (e.g. peer-delivered health promotion in schools).
- Facilitating community action (e.g. health promotion delivered by the community).
- Facilitating self-directed activities (e.g. self-management of chronic disease; independent learning).

Some of the examples above reflect current UK intersectoral policy initiatives – attempts to set 'joined-up policy'. Some interventions may span health/education/housing and a number of other sectors. Finally, the categories could be viewed as a hierarchy with legislation providing a context for the setting of policy/strategy, which in turn affects how services are provided and organised, and which may also manifest as changes to the physical environment.

The draft criteria and definitions worked well, and minor revisions were made to improve the inter-rater reliability for the handful of cases for which their relevance was questionable. From this sample:

- The majority of reviews described interventions in health care and health promotion. Reviews in other areas (e.g. education) were in a minority.
- Only around a quarter were included ($n = 11$; 28%). The majority were excluded ($n = 25$; 62.5%), and four (10%) were unclear.
- Most of those included fell into the 'Provision/organisation of services' and the 'Setting of policy/strategies' categories. We found none that had addressed 'legislation/regulation'.
- Many of those excluded were interventions delivered by individual practitioners (e.g. mostly health professionals).
- At least five of those excluded were 'pharmacological' interventions, such as vitamin/mineral supplementation, and in one case smoking cessation aids (e.g. lozenges, chewing gum).
- One type of intervention that may be relevant, but usually involved some professional input, was activity under the broad heading of 'self-management'. An example was one of the reviews for which it was 'unclear' whether or not it was a 'policy intervention'.[45] It was about education for the self-management of asthma. Patients generally received written information (e.g. leaflets), and underwent a short interaction with a health professional (plus on-going consultations to monitor progress), but largely managed their illness on a day-to-day basis by themselves. This concept could be applied in many other contexts (e.g. self-learning/education/independent study initiatives). Perhaps a good example would be the development of policies or initiatives to promote distance learning as a way of encouraging greater access to further or higher education. The concept of self-management/education/help could be considered a policy intervention although, in these circumstances, such interventions will likely incur some professional one-to-one input in order to help people initiate their own activities.

Applying draft criteria to sources of policy interventions

In order to develop more detailed operational definitions, draft criteria were applied by two researchers to abstracts and extracted data of outcome evaluations included in: (i) a map of studies of HIV health promotion for men who have sex with men (MSM); (ii) a map of studies of children and healthy eating; and (iii) a review of the promotion of sexual health/prevention of sexually transmitted diseases among women. Refining the criteria involved successive rounds of independent coding and reflective discussion.

For policy interventions

Policy interventions are those interventions which *establish or modify collective plans for action* so as to have *systematic impact on the public*. These policy interventions operate via institutions (e.g. hospitals, practitioner bodies, schools, public authorities, commercial bodies, patient organisations) and communities (e.g. geographical or social groups, networks, people with shared interests) and do not include personal policies of individuals.

Policy interventions require more than the *authority* of individual practitioners to instigate, more than the *resources* of individual practitioners to implement and more *roles or skills* than those of a single practitioner to implement. Their instigation and implementation depends upon interaction between organised groups of people. Groupings that make policy range in formality, geographic scope and purpose, but examples include local, national and international government, the regulatory bodies for practitioners and industry and governing bodies of institutions such as schools, health-care services and workplaces. Because of the involvement of social units in policy instigation and implementation, policy interventions are often better evaluated for their effectiveness through the allocation and study of social units (e.g. schools, communities, wards), as opposed to individuals.

Thus, a policy intervention required one or more of the following:

- more than the *authority* of individual practitioners to instigate
  - consultants' ward procedures are not policy

interventions because a single consultant has the authority to implement them, neither are teacher-led interventions confined to the classroom and falling within the curriculum
  - hospital wide procedures are policy interventions (e.g. complaints procedures) as are school procedures for engaging parents with pupils' work [e.g. allowing parents to withdraw their children from Personal, Social and Health Education (PHSE) lessons]
- more than the *resources* of individual practitioners to implement
  - interventions delivered largely within the resources of an individual practitioner with no additional costs other than their reasonable time are 'practice interventions' (e.g. prescribing paracetamol for infants with fever)
  - interventions requiring resources beyond the reach of individual practitioners in their conventional roles, such as interventions requiring additional budgets are policy interventions (e.g. widespread advertising for smoking cessation clinics; prescribing discounted access to fitness facilities)
- more *roles/skills* than that of a single practitioner to implement
  - procedures implemented by many practitioners within the remit of their individual professional roles are not policy interventions (e.g. sharing the workload of facilitating parent craft classes)
  - procedures requiring a team of mixed roles are policy interventions (e.g. replacing doctors with nurses; or provision of specialist stroke units)
- but may, nevertheless, be delivered by individual providers to individual recipients when the intention is to *implement a policy*
  - one-to-one treatments are not necessarily policy interventions (e.g. drugs, surgical treatments, counselling, therapy)
  - directives to consistently adopt a particular intervention are interventions to implement policy (e.g. prescribing aspirin following a heart attack, or counselling before and after HIV tests)
- or may be evident by the use of clustered designs to evaluate their effectiveness, where clustering implies a collective decision about the implementation of different policy interventions in different arms of the trial
  - one-to-one interventions readily evaluated by random allocation of individuals to different treatments are largely practice interventions where such RCTs can inform practice decisions
  - higher units of allocation (e.g. practitioner, setting) are largely policy interventions where such RCTs can inform policy decisions.

For subcategories within policy intervention

In general, the categories of policy interventions described above (setting of policy/strategies, legislation/regulation, provision/organisation of services, environmental modification and facilitating lay/public delivered support/education) could be readily applied. The reviewers identified the possible need for expansion of the 'Environmental modification' category to include the modification of school meals. Any computer-based interventions were considered policy interventions on the grounds of the costs and staffing required for computer support, in addition to the teaching staff required for implementing the intervention.

The level(s) at which policy has been enacted

Four categories were developed for policy level: policy for an institution, policy for a community, policy for a region and policy for a nation. In general, it was clear when policy interventions were being implemented 'institution wide', although this did not exclude them being implemented 'institution wide' across a region or a nation. Also, it was often not possible to discern from the report whether the policy had been set nationally, regionally or institutionally, or whether institutions were obliged to adopt national or regional policy. These distinctions may have no discernible effect on the methodology of evaluation.

Applying operational definitions

Overall, the definitions and categories as described above (see Developing operational definitions) have proved possible to apply in a way that is consistent between two reviewers working independently. Limitations to the work done so far include characteristics of the studies appraised – EPPI-Centre reviews tend to focus on policy level interventions as described here, and so the tests done so far can only have limited powers to test the discriminatory powers of these tools for including or excluding policy interventions. However, the distribution of types of policy interventions that appeared in each EPPI-Centre review varied,

and discriminating between types of policy interventions appeared practical.

In developing our data sets we coded as policy evaluations those:

- in which there was an explicit directive/policy for the intervention OR
- beyond the capacity of individual providers (in terms of their roles/skills, resources, or authority).

These inclusion criteria match the focus of the commissioning brief on policy/management interventions which was 'those interventions that are not confined to an individual practitioner … examples would include peer-led teaching and health promotion in schools'.

Explicit directives or policies include named policies such as national government legislation or programmes or, less formally, explicit collective action plans in which non-researchers had been involved in the decision-making. Collective action plans could be explicit either from descriptions of planning processes or from descriptions of the products of their planning processes, such as guidelines sponsored nationally or regionally by professional organisations or charities, or commercially available curricula.

Operating the second inclusion criterion requires a judgement about the roles, skills, resources and authority of individuals. For instance, distributing fruit to children at school would be judged a policy intervention because it would be beyond the authority/resources of an individual practitioner on the grounds that we do not expect teachers to pay for it out of their own pockets and if the school were to have a budget for it, it must also have a policy for it.

In addition to inclusion/exclusion criteria, we anticipated being able to separately identify and analyse evaluations of interventions that operate at different levels (national, regional, community, or institutional level) and in different policy sectors (housing, transport, health, crime and justice, etc.) or across policy sectors.

Data sources

Lipsey[46] argues that the most suitable data sets for investigating the association between randomisation and effect size are either small numbers of evaluations that are nearly identical, except for randomisation, or large numbers of interventions allowing for diversity in the study population and design. In the first instance, any association between randomisation and effect size would be readily apparent. In the second, any association would need to be distinguished from associations of effect size with other variables, such as differences in the populations, interventions, outcomes or evaluation methods. Resampling studies, which draw on the data from individual RCTs, take Lipsey's argument for comparing similar trials a step further.

Resampling studies data

We had access to two trials of policy interventions in which data were suitable for resampling studies.

Trial 1: The Social Support and Family Health Study

This study was an RCT which assessed whether increased postnatal support could influence maternal and child health outcomes.[47]

Two support interventions were set up. The first, the Support Health Visitor intervention, was the offer of 1 year of monthly supportive listening visits, the first to take place when the baby was approximately 10 weeks old. The primary focus for this intervention was on the mother and her needs. The second intervention, using the services of local community support organisations, entailed being assigned to one of eight community groups that offered drop-in sessions, home visiting and/or telephone support for a period of 1 year.

The trial compared maternal and child health outcomes for women who had been offered either of the support interventions with outcomes for control women who received standard services only. The primary outcomes were child injury, maternal smoking and maternal psychological well-being. Secondary outcomes were uptake and cost of health services, household resources, maternal and child health, experience of motherhood and child feeding.

No evidence of impact was found for either intervention on the primary outcomes. The Support Health Visitor intervention was popular with women and was associated with some of the secondary outcomes. Greater emphasis could, in future research, include the social support role of health visitors, developing more culturally sensitive

outcome measures and exploring the role of social support on the delay of subsequent pregnancy.

Trial 2: The effectiveness of home visitation by public health nurses in preventing the recurrence of child physical abuse and neglect

Home visitation by public health nurses is known to be effective in preventing child abuse and neglect.[48] This RCT therefore aimed to investigate if home visitation by public health nurses to disadvantaged first-time mothers was effective in reducing recidivism.

Families with a history of one child being exposed to physical abuse or neglect were assigned to either a control or intervention group. The control group received standard treatment. The intervention group received a programme of home visitation by nurses in addition to the standard treatment. The main outcome was recurrence of child physical abuse and neglect, and analysis was by intention to treat.

At 3 years' follow-up, recurrence of physical abuse did not differ between control and intervention groups, making the intervention ineffective. Although hospital records showed significantly higher recurrence of hospital attendance in the intervention group than in the control group, the authors concluded that this may be due to specific advice from public health nurses. No significant differences were found for secondary outcomes. This suggested that this home-based strategy was not effective, and that much more effort needed to be made towards prevention of child abuse or neglect before it becomes established as a pattern of behaviour in a family.

Similar studies drawn from systematic reviews

We sought readily available systematic review data stored on EPPI-Reviewer, software for storing and analysing data about primary research for inclusion in systematic reviews. This source includes data from reviews of health promotion (conducted by or in collaboration with the EPPI-Centre published 1999–2004) and education (conducted by the EPPI-Centre or by review groups supported by the EPPI-Centre published before June 2004).

Of the 24 education systematic reviews, only seven included RCTs and NRSs; these included policy interventions of Interactive Communication Technology (ICT) for literacy (three reviews), out-of-home integrated care and education, paid adult support in mainstream schools, supporting pupils with emotional and behavioural difficulties in mainstream primary schools, and personal development planning for improving student learning. Between them they included 32 RCTs and 82 NRSs. This was considered too few studies to analyse further considering their diversity.

There were nine systematic reviews conducted by the EPPI-Centre between 1996 and 2004, and one Cochrane review conducted using EPPI-Centre software. These reviews all included both randomised and non-randomised studies. These reviews were of health promotion, with studies often conducted in educational settings. Between them they addressed workplace health promotion, peer-delivered interventions, mental health, physical activity (two reviews), healthy eating (two reviews), cervical cancer and sexual lifestyle; (nine reviews with a total of 206 studies). See Appendix 2 for summaries of these reviews.

Meta-epidemiological data

The studies described above, when combined as 206 controlled trials of health promotion, also provided a suitable data set for meta-epidemiological investigations.

Another data set suitable for this approach was also available from ongoing work by Colorado State University reviewing and synthesising the past 20 years of research and advancements in the area of transition for youths with disabilities (www.ncset.org/publications/viewdesc.asp?id=714). These data from 126 studies are also held on EPPI-Centre software.

Reviews of reviews

For a review level analysis, data were sought from methodological studies and systematic reviews of randomised and non-randomised studies using systematic search strategies, selection criteria and data extraction procedures.

Chapter 3

Hypothetical associations between randomisation and effect sizes of policy interventions

In Chapter 1 we described divergent views on the appropriateness of RCTs for evaluating the effects of policy interventions. Here we build on our experience of reviewing policy interventions, and the relevant literature, to propose how randomisation and effect sizes may be associated in evaluations of policy interventions.

Our original objectives were to determine whether RCTs lead to the same effect size and variance as NRSs of similar policy interventions, and whether these findings can be explained by other factors associated with the interventions or their evaluation. To meet the second of these objectives we proposed a number of variables for which arguments could be mounted, hypothesising links between them, randomisation and effect size. These variables and hypotheses are presented below and summarised in *Table 1*.

Potential confounders associated with participants of the evaluation

The population of a given study may be related to both the design of an evaluation and its effect size.

Baseline characteristics

Groups in nRCTs may differ at baseline for a number of reasons. Recipients of the intervention may have self-selected, or those who declined to participate may have been assigned to the control/comparison group. Alternatively, recruitment may have favoured those most amenable to participation or those in most need, or excluded older people or those with comorbidities. Well-conducted RCTs, with their standardised procedures for recruitment and data analysed according to the intention to treat rather than receipt of the intervention, are more likely to have more equivalence between groups. Non-equivalence at baseline may influence the calculated effect size and variance.

Attrition

Attrition rates may be linked to the quality of the evaluation. Higher attrition may be expected in community and home settings than in organisational settings where it is easier to employ randomisation and good follow-up. High attrition may also be associated with losing a disproportionate number of people who are socially disadvantaged and more resistant to interventions, and hence lead to a misleadingly large effect size.[49,50] For this reason, high attrition may be associated with both lack of randomisation and higher effect sizes. Potential technical solutions in primary research include greater investment in recruiting and retaining participants, perhaps using the NHS number for tracking. Other solutions for primary studies and systematic reviews include adjusting for attrition, assuming those lost to follow-up have poor outcomes.

Potential confounders associated with the intervention

There are arguments for linking the theoretical underpinning of intervention design, public involvement in developing interventions, and the geographical or organisational scope of interventions with the presence or absence of randomisation and effect size.

Theoretical underpinnings

Policy interventions pose serious challenges to evaluations of effectiveness because they may be large and difficult to replicate consistently, and have diffuse boundaries. Evaluation methodologies differ in their responses to such challenges. RCTs, and systematic reviews of RCTs, are the methods of choice employed by public health physicians wishing to elucidate causal effects in variable circumstances. These methodologies emphasise the need to reduce bias by employing randomisation,

preferably with blinded allocation to treatment and outcome measurement, minimising attrition, and analysing according to the intended treatment. Examples include community interventions for preventing smoking in young people,[51] computerised support for prescribing practice[52] and day care for preschool children.[53]

By contrast, many social scientists are known to employ different theories from experimentalists evaluating policy interventions. The relatively new profession of health promotion specialists[21] has favoured 'an approach to evaluation that implicitly acknowledges the need for outcome data but explicitly concentrates on process or illuminative data that helps us understand the nature of that relationship'.[54] The centrepiece of the health promotion paradigm is the concept of empowerment – enabling people to increase control over, and to improve, their own health. Empowerment claims to attribute responsibility to people not for the existence of a problem, but for finding a solution to it. The goal is then 'full and organised community participation and ultimate self-reliance'.[55] This approach is endorsed by Arblaster[56] in a systematic review of the effectiveness of health service interventions aimed at reducing inequalities in health. The review concluded that characteristics of successful interventions specifically aimed at reducing health differentials include ensuring interventions address the expressed or identified needs of the target population, and the involvement of peers in the delivery of interventions. The tradition of community development rests heavily on public involvement, and we expect community-based interventions to include the public more often in identifying the aims of the intervention, and/or participating in its development. Examples include impact evaluations of a large-scale social marketing initiative to encourage fruit and vegetable consumption[57] and of bar-based, peer-led community-level intervention to promote sexual health among gay men.[58] With this understanding we anticipate evaluations of community development to be more theoretically informed from the tradition of social science, and less subject to randomisation. This expectation is supported by a comparison of the CONSORT (Consolidated Standards of Reporting Trials) statement (a checklist and flow chart, to help improve the quality of reports of RCTs) and the TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) statement. The TREND statement, unlike the CONSORT statement, seeks information about theories used in designing behavioural interventions.[36,59]

In summary, the public health approach has tended to emphasise randomisation but not public involvement or community-based approaches, compared with health promotion where the reverse is so; with the expectation within public health that randomisation leads to more conservative estimates of effect size and the expectation within health promotion that public involvement leads to interventions with greater effect sizes.

Similarly, divergent views about appropriate methods for evaluating interventions are found in the areas of social welfare[44] and education[25] where The Centre for Evidence-Based Social Services (www.ex.ac.uk/cebss/introduction.html) and the Evidence-based (www.cemcentre.org/ebeuk/) Education Network UK stand out from many of their British professional colleagues in social welfare and education, respectively, as advocates for randomised evaluation.

Setting and boundaries of the intervention

In the area of public policy, community-wide interventions, or regional/national interventions may pose challenges in being less easily manipulated for the purposes of evaluation than individual or institutionally based interventions. Standardised implementation of interventions may be more difficult across large communities, regions or whole countries than in single organisations and therefore may be less effective or more variable in effectiveness.

Interventions with a broader reach (communities, regions, nations) have more diffuse boundaries than those set within institutions. Randomisation is less often applied to community, regional or national interventions. Clustered trials are more appropriate for these and some organisational level interventions, in order to reduce the likelihood of participants experiencing comparison interventions to which they have not been allocated. However, clustering reduces the power of a trial, so clustered evaluations are less likely to show statistically significant effectiveness. The solution is to increase the size of the trial, yet recruitment can be particularly challenging in community settings. Moreover, attrition may be greater in larger scale interventions, where tracking of individuals is more difficult than within an organisation (see Attrition).

Providers of the intervention

Community development and peer delivery specialists value health promotion theory and process evaluations more than RCTs. These interventions may therefore be found to have less randomisation. Theories underpinning community development and peer delivery anticipate more effective interventions through their greater relevance. Characteristics of interventions effective for reducing health inequalities include community commitment and peer delivery.[56] Examples in the area of smoking cessation include the non-randomised evaluation of the Wessex Healthy School Award where the intervention was delivered by the school community.[60]

In contrast, clinicians design and deliver their own interventions and evaluations, and work in a culture that favours randomisation; for instance, many of the smoking cessations' interventions for pregnant women are delivered by health professionals and evaluated by RCTs.[61]

Also, working in the area of criminology, Lipsey and Wilson[62] have shown a statistical association between randomisation and effect size in 'demonstration' projects in which the researcher had greater control of both the intervention and randomisation. Our data set of health promotion evaluations provides an opportunity to test this association in another area.

Potential confounders associated with outcomes

The design of the evaluation provides a wealth of potential confounders. Among these are the choice of outcome domains (e.g. knowledge, attitudes, behaviour or health) and the choice of outcome measures ('hard' or 'soft'). These are considered below.

Choice of outcome domains

The impact of health education has traditionally been considered in terms of changes in knowledge, attitudes, behaviour and health. Kirkpatrick's[63] hierarchy of outcomes from the policy area of professional training presents the higher level outcomes (health and behaviour) as harder to attain than lower level outcomes (knowledge and attitudes). The choice of outcomes may be strongly influenced by the intervention setting. Other broader health behaviour theories present knowledge and positive attitudes as necessary but not sufficient for improved behaviour and health in many theories of health behaviour.[64] Thus there is support from these two different policy areas of professional training and health behaviour change for the argument that outcomes in the domains of knowledge, attitudes, behaviour and health are successively more difficult to influence and therefore associated with lower effect sizes.

The choice of outcomes may be strongly influenced by the intervention setting. For instance, the measurement of any health outcome may be easier in a clinical setting where randomisation is also more readily acceptable by staff. Following this argument, evaluations with health outcomes are more likely to be associated with patient populations than community populations. For instance, generating evidence about smoking cessation in pregnancy lends itself to short-term health outcomes such as birth weight, gestational age at birth, perinatal mortality, method of delivery, and measures of anxiety, depression and maternal health status in late pregnancy and after birth as seen in a systematic review of 64 RCTs (51 RCTs and six clustered RCTs).[61] Such data can be easily collected in a clinical setting where randomisation is also feasible. Similarly, a review of smoking cessation for hospitalised adults where nine of the 17 included studies (16 RCTs, one quasi-RCT) measured death of the patient as well as abstinence from smoking.[65] Both these reviews included randomised or quasi-randomised trials.

In contrast, a review of community interventions for preventing smoking in young people[51] found that over half the controlled trials were non-randomised, and outcomes were restricted to knowledge about the effects of smoking, attitudes to smoking, intentions to smoke in the future and smoking cessation.

Choice of outcome measures

The choice of particular outcomes or measures may be associated with randomisation, because some outcomes are more feasible in a clinical setting where randomisation is also readily accepted. For instance, both cluster or individual randomisation and the use of clinical outcome measures requiring blood tests may be difficult to impose elsewhere. For example, cotinine tests as markers for smoking cessation may be more common in clinical settings, and provide smaller effect sizes than unconfirmed reports of non-smoking. For instance, only one study out of 17 (6%) in a review of community interventions for preventing smoking

in young people[51] used cotinine measurements to confirm non-smoking. Similarly, a review of community interventions for adults found 10 of 32 studies included serum thiocyanate analysis and one included cotinine analysis to validate smoking status, giving a total of 34% of trials with biochemical validation of smoking cessation.[66] In contrast, in a review of smoking cessation during pregnancy, when women often attend clinics or hospitals, 35 of 47 (74%) included studies with biochemically validated cessation as an outcome.[61] The implications of this are that 'hard' outcomes (in this case, validated smoking cessation) and randomisation are both easier in a clinical setting where clinical tests and access to medical records are easier. A parallel argument might be made in education sector evaluations where self-reported understanding may be less reliable than examination results, so assessing learning within schools may be more objective. In both cases, evaluating interventions within an institution where testing is routine and randomisation is easier leads to more objective findings.

Potential confounders associated with design of the evaluation

The design of the evaluation provides a wealth of potential confounders: sample size; presence of a control group; concealment of allocation; follow-up; the number of clusters; and quality of reporting. Each of these is considered below.

Sample size

Larger sample sizes are more likely to be found in nRCTs or natural experiments that can provide convenience samples on a large scale. Smaller sample sizes, more commonly found in RCTs, are more likely to lead to spurious results, with those reporting positive findings more likely to be published.

Control group

Control groups are always found in RCTs, but only sometimes in NRSs. The lack of a control group may result in a misleading effect size as the study does not take into account possible simultaneous influences.

Blinding

Patients who know that they are on a new, experimental treatment are likely to have an opinion about its efficacy, as are their clinicians or the other study personnel who are measuring responses to therapy. These opinions, whether optimistic or pessimistic, can systematically distort both the other aspects of treatment and the reporting of treatment outcomes, thereby reducing our confidence in the study's results. In addition, unblinded study personnel who are measuring outcomes may provide different interpretations of marginal findings or differential encouragement during performance tests, either one of which can distort their results.[67] Blinding of community interventions is more challenging, and therefore linked with other characteristics of community interventions such as fewer randomised evaluations.

Follow-up

Longer term follow-up may be easier within institutions, where randomisation is also easier. However, long-term follow-up exposes interventions to additional scrutiny in terms of sustainability or maintenance of effect, and may reveal declining effect sizes. Alternatively, long-term follow-up is also associated with greater attrition which in turn is associated with greater effect sizes.

Clustering

Clustered trials may be 'natural experiments' without randomisation, particularly if there are few clusters. Natural experiments may be more likely to have enthusiasts supporting the intervention, and non-enthusiasts supporting the comparisons, and therefore lead to greater effect sizes. This introduces bias from lack of blinding (see above).

Quality of reporting

The quality of reporting of some aspects of evaluation may be associated with researchers' disciplines. In particular, triallists, who more often use randomisation, may also be more likely to report pre- and postintervention data. Natural experiments in particular may face challenges in collecting pre-intervention data because researchers have less influence. They may be invited to evaluate a policy intervention only after implementation has begun (for example, evaluation of Sure Start Plus, a UK Government pilot initiative to support pregnant young women and young parents under 18 years of age).[68] Although the CONSORT and TREND statements both encourage the reporting for baseline data for RCTs and non-randomised designs respectively, the

TABLE 1 *Hypotheses linking randomisation and effect size through 'effect moderators' or 'cofounders'*

Potential confounder	Association with randomisation	Association with effect size	Possible technical solutions
Participants			
Baseline characteristics			
Groups may differ at baseline because: either recipients of the intervention have self-selected or those who declined to participate have been assigned to the control/comparison group; or recruitment favoured those most amenable to participation, or those in most need, or excluded older people or those with comorbidities	nRCTs are more likely to have more heterogeneous populations and non-equivalence between groups	Heterogeneity and non-equivalence at baseline may influence the calculated effect size and variance	Randomisation wherever possible; better matching and assessment of baseline characteristics elsewhere. More pragmatic trials reflecting 'real world' problems that will be more generalisable and more likely to be implemented
Attrition			
Higher attrition may be expected in community and home settings than in organisational settings Higher attrition may be expected in transient populations (e.g. commercial sex workers, asylum seekers, socially excluded people)	It is easier to employ randomisation and have good follow-up for trials set in organisations. Attrition and randomisation can be used as quality markers for trials	High attrition may be associated with losing a disproportionate number of socially disadvantaged people who are more resistant to health promotion/public health initiatives	Greater investment in recruiting and retaining participants – use of NHS number for tracking. Adjusting for attrition, assuming those lost to follow-up have poor outcomes Innovative strategies for managing contact with transient populations (e.g. using 'peer evaluators')
Intervention			
Theoretical underpinnings of the intervention			
Public health triallists value experimental methodologies more than do health promotion specialists, who place more emphasis on involving the community in developing and delivering the intervention	Experimental methodologies in public health are associated with randomisation Health promotion is associated with community development but not randomisation	Rigorous public health trials minimise effect sizes Community development is mounted with the expectation that it will maximise effectiveness	Cross disciplinary research Encourage a results driven culture among social scientists
Public involvement in developing the intervention			
Empowerment theories attribute responsibility to people not for the existence of a problem, but for finding a solution to it	The goal of 'full and organised community participation and ultimate self-reliance'[55] is a feature of social work such as community development and youth work, rather than a feature of public health and randomised experiments	Successful interventions specifically aimed at reducing health differentials include ensuring interventions address the expressed or identified needs of the target population, and the involvement of peers in the delivery of interventions[56]	

continued

TABLE 1 *Hypotheses linking randomisation and effect size through 'effect moderators' or 'cofounders' (continued)*

Potential confounder	Association with randomisation	Association with effect size	Possible technical solutions
Setting and boundaries of the intervention			
Interventions with a broader reach (communities, regions, nations) have more diffuse boundaries than those set within institutions	Randomisation is less often applied to community, regional or national interventions. Clustered trials are more appropriate for these and some organisational level interventions Attrition may be greater in larger scale interventions, where tracking of individuals is more difficult than within an organisation (see Attrition)	Clustering reduces the power of a trial, so clustered evaluations are less likely to show effectiveness Standardised implementation of interventions may be more difficult across large communities, regions or whole countries than in single organisations, and therefore may be less effective	Larger scale cluster trials for policy interventions Greater investment in tracking participants – use NHS number
Provider of the intervention: Community/peer provider			
Community development and peer delivery specialists value health promotion theory and process evaluations more than RCTs	These interventions may therefore be found to have less randomisation	Peers may be seen as more credible sources of information than professionally trained, health educators, and may be particularly helpful in reaching 'at risk' populations	Cross disciplinary research
Clinician			
Many clinicians who design and provide interventions value RCTs	These interventions may therefore be found to have more randomisation	Methodological rigour associated with randomisation is likely to lead to lower effect sizes[69]	
Researcher provider			
Researchers have more control over the intervention and evaluation	Theoretically, the researcher would therefore be better able to randomise	Interventions will be found to be more consistently implemented by enthusiasts, and therefore more effective	
Outcomes			
Choice of outcome domains			
Health outcomes are more readily measured in clinical settings	Clinical settings are more likely to mount RCTs, and have clinical providers, and long-term follow-up (see above)		
Choice of outcome measures			
Clinical outcomes are more commonly found in clinical settings Choice of 'hard' or 'soft' outcomes can be associated with randomisation	If clinicians favour RCTs, clinical outcome measure may be associated with greater randomisation	Clinical outcomes will be found to be more resistant to change than 'softer' outcomes such as reported behaviour	

TABLE 1 *Hypotheses linking randomisation and effect size through 'effect moderators' or 'cofounders' (continued)*

Potential confounder	Association with randomisation	Association with effect size	Possible technical solutions
Evaluation design			
Sample size			
Sample size affects the choice of study design	Larger sample size may be more likely in NRSs	Smaller sample sizes are more likely to give spurious results; of these, those with positive results are more likely to be published	Weight of evidence by sample size
Control group			
Study design is linked with the use of a control group	Control groups are always found in RCTs, but only sometimes in NRSs	Use of a control group leads to smaller effect sizes than uncontrolled evaluations	
Blinding			
Blinding of participants, recruiters, intervention providers and outcome assessors to the intervention allocation	Blinding is easier with randomisation	Poor concealment, more common in nRCTs, will be associated with greater effect sizes	
Follow-up			
Length of follow-up periods is linked with study design	Long follow-up may be easier within institutions, where randomisation is also easier	Long follow-up will be associated with declining effect size	
Clustering			
Clustered trials with few clusters are more likely to be 'natural experiments'	Natural experiments do not include randomisation	Natural experiments may be more likely to have enthusiasts supporting the intervention and non-enthusiasts supporting the comparisons, and therefore lead to greater effect sizes	
Quality of the reporting			
Quality of reporting specific elements of a study is associated with researchers' disciplines	Better reporting (of pre- and postintervention data) will be seen to be associated with triallists who also support randomisation	Reporting of pre- and postintervention data precludes effect sizes inflated by differences between groups	Adjusting for differences between groups, in primary studies and in reviews

CONSORT statement, introduced in 1996, has had much longer to influence the reporting of RCTs than the TREND statement, introduced in 2004, has had to influence non-randomised trials. Thus, a greater proportion of randomised studies than NRSs might be expected to report baseline data.

By influencing the reporting of studies, both statements have been able to influence the quality of the design and conduct of studies, and may be expected to reduce bias and, consequently, effect sizes.

Reporting of pre-and postintervention data precludes effect sizes inflated by differences between groups.

Chapter 4

Review of methodological literature

Introduction

The aim of this chapter is to review what is already known and to set the context for later chapters through identifying, examining and discussing a range of methodological studies that compare randomised and non-randomised designs outside the area of health. We were aware that methodological research on policy interventions had been published in the broader social sciences and wished to see what could be learned from this before conducting our own investigation within health.

A search was conducted to identify studies in areas such as education, psychology and social care, which had investigated the influence of evaluation methodology (principally differences between randomised and non-randomised studies), intervention attributes, characteristics of study populations and a range of other factors on effect sizes. The key characteristics of the studies were tabulated and described (in terms of their aims, methods, effect modifiers investigated, and results and conclusions). Particular attention was paid to the effect modifiers investigated and how these were characterised in order to investigate them further at the level of systematic reviews and primary studies. This chapter reports the results of the review of the methodological literature.

Methods

Although this was not intended to be a systematic review of the literature, we nevertheless adopted standard practices for literature searching, retrieval, extraction and synthesis, when possible.

In terms of inclusion criteria we were interested in capturing empirical studies examining the association between various population, intervention, outcome, and study characteristics and effect size. The principle characteristic of interest was whether or not randomisation had been used to assign participants to study groups. We were also interested in other methodological attributes such as methods for measuring outcomes, follow-up procedures, and sample size, together with any other variables that may be associated with presence or absence of randomisation, as potential 'effect modifiers'. We anticipated capturing a range of types of study including within-study comparisons (i.e. RCTS that include a non-randomised control/comparison group to allow at least one non-experimental estimate of effect); single meta-analyses comparing the results of randomised and non-randomised trials; and meta-reviews (i.e. reviews of meta-analyses which summarise the conclusions from individual meta-analyses with respect to associations between variables and effect sizes).

A search strategy was designed, tested and revised by the information scientist. Searching for studies that describe both randomised and non-randomised evaluation designs is problematic in electronic bibliographic databases, particularly outside the health field. This is due to poor indexing of study designs in the controlled vocabulary (where one exists), the limited search capabilities of many databases, and the lack of abstracts describing the studies. The strategy underwent various revisions to balance sensitivity with specificity before being executed. The final strategy comprised free-text terms for RCTs, and NRSs (e.g. non-experimental, pseudorandom, semi-random) (see Appendix 3).

A number of electronic databases were searched:

- Applied Social Sciences Index and Abstracts (ASSIA)
- Australian Education Index (AEI)
- British Education Index (BEI)
- CareData
- Dissertation Abstracts
- EconLIT
- Educational Resources Information Centre (ERIC)
- International Bibliography of the Sociological Sciences (IBSS)
- ISI Proceedings: Social Sciences and Humanities
- PAIS International (Public Affairs Information Service)

- PsycINFO
- SIGLE (System for Information on Grey Literature in Europe)
- Social Science Citation Index (SSCI)
- Sociological Abstracts.

Further details are given in Appendix 3, for both this and the search reported in Chapter 5.

In addition to this search, citation searching was undertaken to identify publications by authors known to have published widely on the issue of evaluation design and effects. These included William Shadish, Harris Cooper, Larry Hedges, Steven Glazerman and Dan Levvy. Authors of reviews included in Chapter 5 were contacted to identify further methodological publications. Finally, a number of internet sites were searched including the catalogue of the British Library, the Library of Congress, search engines (Copernic and Google) and information gateway sites [OMNI (www.intute.ac.uk/medicine/) and SOSIG (www.intute.ac.uk/socialsciences/)].

Once literature searching was complete, relevant papers were retrieved and sifted to assess their relevance. They were classified according to their study type (i.e. replication studies, single meta-analyses, reviews of meta-analyses, etc.), read and key details extracted and tabulated. They were not assessed for their methodological quality. The purpose was to identify hypotheses for testing with our own data (see Chapters 7–9).

Discussion of methodological literature

Type of studies

After sifting we identified one review of within-study comparisons of randomised and non-randomised participants,[70,71] six single meta-analyses[42,46,72–75] and one review of meta-analyses.[76] These eight studies form the basis of this chapter. Their scope and methods are described in *Table 2*.

We also identified nine single within-study comparisons of randomised and non-randomised participants, seven of which were included in the review of such comparisons by Glazerman *et al.*[70,71] and consequently are not discussed any further (see Methods of analysis). In addition, six small concurrent comparisons of randomised/non-randomised study participants were identified[81–86] and a number of commentary papers were found. These weaker sources of evidence were not tabulated.

Scope of the studies

The scope of the studies varied in terms of the interventions included, the populations included and the methodological focus. The topic areas examined by the eight studies varied. They included interventions for preventing juvenile delinquency,[42,46] treatment of alcohol abuse[73,87] and psychological interventions.[74–76] There was an overlap with the health field, with Heinsman and Shadish[72] including interventions in the mental health and health-care fields.

In terms of the type of interventions evaluated, there was a strong focus on psychotherapy and psychosocial programmes, partly reflecting the specialist interests of the authors, many of whom were common to more than one publication (For example, Heinsman and Shadish, 1996;[72] Shadish, 1997;[88] Shadish, 2000;[75] Lipsey, 2003;[46] Wilson and Lipsey, 2001[76]). Some studies were broad in their inclusion of interventions, such as Weisburd *et al.*,[42] who included interventions aimed at communities, families, schools and labour markets.

In terms of scope, many of the studies had the primary aim of comparing effect sizes between randomised and non-randomised evaluation designs. In addition, some also examined the relationship between other modifiers and effects. For example, Shadish and Ragsdale[74] also investigated the influence of sample size, attrition, type of control group, publication status and a number of other factors. In other studies the focus was broader, with randomisation only one variable of interest among many. For example, in the review by Lipsey[46] [described in further detail in Reviews of meta-analyses of randomised and non-randomised studies ($n = 6$)], page 71 stated that the aim was to 'illustrate the hazards and complexities of investigating moderator variables in meta-analysis'. Wilson and Lipsey[76] in their review of meta-analyses [described further in Synthesis of meta-analyses ($n = 1$)] examined the association between a vast range of population, intervention and study, variables and effect size variance. Shadish *et al.*[75] [described in further detail in Reviews of meta-analyses of randomised and non-randomised studies ($n = 6$)] had a slightly different focus, examining the characteristics of studies judged to be 'clinically representative', in terms of relationship to study design and effect size (see Effect modifiers).

Where random/non-random allocation was the key issue investigated, some studies sought to delineate the influence of different types of NRS on study

outcomes, rather than analysing all NRSs as one homogeneous group. For instance, Weisburd *et al.*[42] classified five types of study. These included (1) correlational studies where an intervention is measured in only one group; (2) studies where a clear temporal sequence can be observed between an intervention and an outcome; (3) studies in which one group receives an intervention, compared with another group that does not; (4) studies in which intervention and comparison groups are compared, with other mediating factors controlled for/or with a matched comparison group; and (5) RCTs. Type (1) was considered non-experimental, type (2) was a 'stronger non-experimental/weaker quasi-experimental', types (3) and (4) were considered quasi-experimental and (5) was an RCT. The authors reported mean effect sizes for each of these types, and also separately compared the quasi-experimental studies (3 or 4) and the higher quality quasi-experimental designs (4) with the RCTs (5).

The rationale and context for the studies varied. For example, Weisburd *et al.*[42] posed the question of whether the type of research design used to evaluate a crime and justice intervention influences its conclusions. Writing from the perspective of the Campbell Collaboration Crime and Justice Coordinating Group, they acknowledged the gold standard status of the RCT for assessing effectiveness, but noted the lack of well-conducted RCTs in the crime and justice area and that systematic reviews that restrict their inclusion criteria to RCTs may be unrealistic. Their central question, therefore, was 'What are the potential shortcomings of including NRSs within systematic reviews?'. Specifically, 'Are they likely to over or under-estimate the effects of interventions?'.

In contrast, Moyer and Finney[87] were interested in the generalisability of RCTs in the field of alcohol treatment. They suggested that participants recruited into RCTs may be more likely to benefit than those who receive treatment under more typical 'real life' conditions, which tend to prevail in NRSs. The study therefore investigated the extent to which RCTs had been used in the field, whether participants differ between the two designs, whether the interventions are implemented differently between the two designs; and whether 'post-treatment' functioning of participants between the designs differ, controlling for differences in participant characteristics and other methodological features.

In summary, while one of the central aims of each study was to examine the influence of randomisation on intervention effects, each did so from a variety of different perspectives.

Methods of analysis

As discussed in Type of studies, three main types of study were used to examine the influence of design and other characteristics on intervention effects.

Reviews of within-study comparisons (n = 1)

The first approach, as employed by only one study, was reviews of within-study comparisons, also referred to as the design replication study. As described earlier, this type of study uses an RCT to evaluate an intervention, the results of which are compared with one or more non-experimental comparison/control groups. The aim is to ascertain how similar the outcomes for the non-experimental groups are to those of the randomised groups of the same intervention. Glazerman *et al.*[70,71] systematically reviewed replication studies in the fields of welfare, job training and employment services. The review sought to ascertain whether non-randomised methods produce similar results to well-designed RCTs, which non-randomised methods were more likely to replicate the outcomes from well-designed RCTs and under which conditions they were likely to perform better, and whether averaging multiple effects from NRSs produced similar results to those obtained by well-designed RCTs.

A study protocol was published on the Campbell Collaboration website (www.campbellcollaboration.org) prior to initiation of the review. To be included, studies had to compare randomised and non-randomised groups from within the same study, and the same intervention in the same sites. Any differences other than the presence or absence of randomisation would confound the results.

Twelve studies relating to nine interventions were included. Quality assessment was performed on each of the studies and was found to be generally good. The analysis explored the analytic techniques used to adjust for differences between the comparison group and the randomised population and how selection bias varies according to the source of the comparison group. For example, the source of the comparison group was coded according to whether it was drawn from a national data set or from the control group of another RCT, or whether members from the same geographic

TABLE 2 *Study methods*

Study	Study aim	Inclusion criteria	Study identification	Analysis
Review of within-study comparisons				
Glazerman, 2003[70,71]	To assess the value of replication studies to assess the ability of NX designs to produce valid impacts of social programs on participants' earnings	RCTs with an additional comparison group to allow at least one NX estimate of programme impact The experimental/NX comparison had to be based on estimates from the same experiment and had to pertain to the same intervention in the same sites	Search not detailed here 12 studies pertaining to nine interventions were included	Design and context variables coded. To estimate the average bias reduction, both bivariate analyses (tabulations) and multivariate analyses (regression) were used. Bivariate analyses were sample size weighted
'Single' meta-analyses				
Shadish, 1996[74]	To provide better estimates (than previous meta-analyses) of the differences between randomised experiments and non-equivalent control group designs, by controlling inclusion and coding and analysing potential moderator variables	Primary studies of marital or family psychotherapy or enrichment, taken from a sample of 100 studies acquired for a previous meta-analysis. Included studies had to compare treatment to control conditions and allow effect sizes to be calculated. Studies in which allocation method was not clear or was haphazard were excluded	Studies were identified from a previous meta-analysis[77] 100 studies were included (64 RCTs) 116 were excluded	Each study coded by effect size and design. Differences between RCTs and NRSs are presented by simple comparison of pooled effect size and variance. Effects of including the 116 excluded studies was examined. Outliers were adjusted for. Effects of potential confounder variables were examined singly and in combination by regression analysis
Weisburd, 2001[42]	To determine whether the type of research design used in a crime and justice study influences the conclusions that are reached	Primary studies of criminal justice interventions that included crime or delinquency as an outcome measure and met minimal methodological requirements. Seven broad areas looked at: communities, families, schools, labour markets, places, policing and criminal justice	Studies identified for National Institute of Justice commissioned 'Maryland Report' were included 308 studies were included (46 RCTs)	Each study coded 1, 0 or –1 according to whether the investigator concluded the intervention had worked, had no detected effect or had backfired (IRR) as presented in the Maryland Report and also according a 'scientific methods scale' developed for the Maryland Report as an indicator of internal validity (SMS) where 1 indicated correlational studies and 5 indicated RCTs Cross-tabulated SMS score with mean IRR to give an indication of intervention outcomes according to study design

TABLE 2 *Study methods (continued)*

Study	Study aim	Inclusion criteria	Study identification	Analysis
Heinsman, 1996[72]	A methodological study using meta-analysis to examine the defining feature of the randomised experiment, random assignment to conditions	Studies that included control conditions and allowed effect sizes to be estimated were selected. Studies in which allocation method was not clear or was haphazard were excluded	Studies included in four existing meta-analyses in health, mental health and education were screened for inclusion 51 RCTs and 47 NRSs were included. Total number excluded not reported	Each study coded by design and effect sizes. Differences between RCTs and NRSs given across all interventions and according to each of the four topic areas both by simple comparison of randomised and non-randomised experiments and by using a regression analysis
Shadish, 2000[75]	To study the relationship of clinical representativeness to outcome and to generalise by extrapolation from that research to a clinically representative target of interest	Studies of psychological therapies in any setting, excluding those that used psychotropic medication or were purely preventive	Studies were sourced from the Shadish and Heinsman[88] meta-analysis that met their stage 1 criteria for clinical representativeness ($n = 41$), 40 studies randomly sampled from each of the same meta-analyses used by Shadish and Heinsman,[88] and nine studies of 'clinic therapy' from Weisz *et al.*[78] Authors refer to time lag bias; most of included studies are pre-1990. More recent studies may have non-significantly smaller effect sizes and may be significantly less clinically representative	Each study coded for effect size (independently/blind to other variable coding), clinical representativeness criteria, treatment characteristics and outcome characteristics. Correlation between effect size and clinical representativeness scores given in a scatter plot, then broken down by source of studies, random or non-random assignment and year of publication. Multiple regression analysis used to predict effect size from coded variables
Lipsey, 2003[46]	To illustrate the hazards and complexities associated with investigating and interpreting confounded moderator variables, by examining the difference in effect sizes associated with randomised vs non-randomised designs	Studies of psychosocial intervention programmes to prevent or reduce juvenile delinquency. Studies had to use a control group design involving random assignment or matching or present pre-intervention data indicating the degree of initial equivalence between the treatment and control groups. Set in an English speaking country and reported 1950 or later	Subset of studies drawn from a previous meta-analysis[79,80] 382 studies included, 51% RCTs (Note: this paper is about confounded moderator variables; the above the papers may be more useful for looking at effects of randomisation)	Each study coded for design and effect size. Analysis presents pooled effect sizes for studies with various combinations of moderator variables, and makes simple comparisons

continued

TABLE 2 *Study methods (continued)*

Study	Study aim	Inclusion criteria	Study identification	Analysis
Moyer, 2002[87]	To compare the participants, methodological features and post-treatment functioning in randomised and non-randomised studies of alcohol treatment (due to concerns about generalisability of findings from RCTs)	Alcohol treatment trials. Only those that randomised by individual were included	Details missing 324 studies included (232 randomised)	Each study coded for methodological features and effect size. Examined the following with regard to randomisation or not: participant selection; participant characteristics; participant pre-treatment characteristics across conditions; types of treatments; methodological features; participant post-treatment functioning
Meta-reviews				
Wilson, 2001[76]	To determine the influence of (study) method features relative to substantive intervention features on observed study outcomes	Meta-analyses of psychological interventions (i.e. treatments whose intention was to induce psychological change, whether emotional, attitudinal, cognitive or behavioural) for which standardised mean difference effect sizes could be estimated	Large number of psychology and sociology databases supported by manual searches Identified 319 meta-analyses, 76 of which contained both RCTs and non-randomised comparative studies	For each meta-analysis: Coded (1) total effect size variance around the grand mean effect size and (2) effect size variance according to selected study features. Determined the proportion of effect size variance associated with the study features of interest (eta-squared) for each meta-analysis Main analysis was the description and comparison of the mean eta-squared values and when appropriate the mean difference product moment correlation coefficient (*r*) indices for different subgroups

IRR, investigator reported result; NX, non-experimental; SMS, scientific methods scale.

area as the randomised population were sampled. The adjustment techniques used were coded as to whether background variables were used as covariates in a regression model, matching methods were used (e.g. propensity scores), pre-intervention measures of the outcome were taken into account, or an econometric sample selection model was used.

Three sets of analyses were performed. Univariate analysis described the range of bias estimates across the studies in terms of annual earnings. Bivariate analyses then explored the influence of source of comparison group and adjustment analytic techniques on bias, expressed in annual earnings. Finally, multivariate regression assessed the independent impact of these variables on bias estimates. The authors acknowledged the relatively small number of constituent studies in the review as limiting the sophistication of their analysis. They therefore describe the results of the regression as being illustrative.

To explore whether averaging multiple effects from NRSs produced similar results to those obtained by well-designed RCTs, an aggregation exercise was conducted. The distribution of the 1150 bias estimates from the included studies was examined to see whether positive and negative bias estimates cancelled each other out.

Reviews of meta-analyses of randomised and non-randomised studies (n = 6)

Six reviews in which meta-analysis was used to explore the influence of selected variables on effects were identified. Few, if any, of these reviews could be considered as 'standard' systematic reviews of effectiveness. Their aim was not necessarily to summarise the effectiveness of interventions in a given area. Rather, they were methodological studies initiated specifically or in part to answer questions regarding the most appropriate and valid study designs to use to evaluate intervention programmes in given disciplines. Most reported highly structured and generally transparent methods for identifying, extracting and analysing primary studies. Although this transparency may fall short of what is currently considered to be accepted standards for synthesising evidence about effectiveness, it was generally adequate for methodological investigations, particularly in the current work where the purpose was to identify hypotheses for testing with new data.

The usual method for identifying primary evaluations included in these meta-analyses was to sample studies from authors' own existing databases of studies, often used in previous meta-analyses. For example, Shadish and Ragsdale[74] included 100 studies of marital and family psychotherapy or enrichment, the majority of which were sourced from their previous meta-analysis published 3 years earlier.

All of the reviews reported inclusion criteria for primary studies, in varying detail. In general, studies had to report specific interventions (e.g. psychotherapy, or psychosocial treatment) for a particular purpose (e.g. preventing juvenile delinquency, or alcoholism). Some reviews specified use of a control/comparison group and reporting of method of allocation to study groups as inclusion criteria. The latter is particularly necessary to determine whether methods purporting to be randomised really were randomised. Primary studies also had to report sufficient data to allow effect sizes to be estimated, a necessary step in testing the influence of randomisation on study outcomes.

The number of constituent studies in the reviews varied from 90 to 382. In general, the number of studies included was relatively large, with three reviews each including over 300 studies. The greater the number of included studies, the more likely that meta-analyses will be a useful technique for examining associations between study design and effect size. The proportion of randomised and non-randomised studies included in the reviews varied. In one study only 15% of studies were randomised.[42] In another the proportion of randomised studies reached 71%.[87]

Statistical procedures varied in complexity across the reviews. Average effect size estimates were used in most cases to express the influence of study and intervention variables on results. Methods for calculating effect sizes included standardised mean differences (SMDs), and *d*-scores. There were, however, a few exceptions. For example, Moyer and Finney[87] noted the limitations of effect size calculations, namely that they exclude studies for which effect sizes cannot be calculated, and they average over a number of different types of outcome to create a single independent effect size for each study, which fails to take into account the 'strength of the competition' with which treatments are compared or limits the studies to those with a standard treatment or a control condition. The authors suggested this is problematic in alcohol treatment trials where 'no-treatment' control conditions are rarely found. Instead, they measured the proportion of participants abstinent from alcohol and the proportion 'improved' (drinking moderately) following treatment.

In terms of analysis, some reviews analysed the results of randomised and non-randomised studies separately, comparing overall effect sizes between the two (e.g. Moyer and Finney,[87] and Weisburd *et al.*[42]). Other reviews also pooled all studies together regardless of design, and examined the effect of randomisation alongside other mediating variables in predictive regression models. For example, Lipsey[46] assessed the effect of potential confounding moderator variables in randomised and non-randomised studies. Firstly, it was speculated that certain study designs may be associated with higher or lower effect sizes. However, this may not necessarily be an artefact of design itself, but may reflect correlation between certain designs with particular types of intervention. To demonstrate this, the number of randomised and non-randomised designs that had been employed to evaluate 'demonstration' programmes (in which the researcher is involved in delivering or planning the intervention) and 'routine practice' programmes (in which interventions are delivered as part of routine services and evaluated externally) was mapped. Not surprisingly, randomised designs were more common in demonstration projects than in practice projects. Secondly, mean effect sizes

were calculated for demonstration and routine practice programmes, and also stratified according to whether a randomised or non-randomised design was used. Finally, a range of other potential confounders were tabulated to demonstrate their relationship with study design and type of evaluation project. Significant associations were plotted in a table, with the effect size differential associated with variation in the moderator.

In another example, Shadish *et al.*[75] analysed the relationship between studies judged clinically representative, moderating variables and effect sizes in a meta-analysis of 90 psychological therapy evaluations. The study was initiated due to a concern that the effect of psychological treatments, as estimated in previous meta-analyses, may be higher than would be found under routine clinical conditions. The aim was therefore to meta-analyse studies meeting a definition of clinical representativeness, taking into account the moderating effect of study design and other variables that might be associated with clinically representative studies.

Studies sourced from their previous meta-analysis (mostly conducted before the 1990s) were coded according to clinical representativeness criteria (e.g. the study population compared with those who might consult psychological services in practice; the intervention was delivered in a routine practice setting); effect size; treatment characteristics (e.g. number and duration of sessions); and study design (e.g. random or non-random allocation). A total of 1324 effect sizes (SMDs) were calculated from the 90 studies (range 1–168, and mean of 14.71 per study) and aggregated using both random and fixed effects. The correlation between effect size and clinical representativeness scores was illustrated in a scatter plot, then broken down by source of studies, random or non-random assignment and year of publication. Multiple regression analysis was then used to predict effect size from coded variables. Through this analysis the authors were able to determine whether non-randomised designs have significantly higher clinical representativeness scores (on the assumption that NRSs are more likely to be carried out in 'practice' settings), and the extent to which this influences effect size.

In summary, these studies have used a range of methods to assess differences in effects between randomised and non-randomised studies. Studies have commonly been sourced from authors' own data sets from their previous meta-analyses. They have been screened against methodological and topic-specific inclusion criteria, coded and in a minority of cases quality assessed. Statistical procedures have varied in complexity, although most studies have used multivariate regression analysis.

Synthesis of meta-analyses (n = 1)

The final approach, as employed by one study,[76] was a pooling of meta-analyses. This approach has been described as 'meta-epidemiology', whereby a substantial number of meta-analyses that contain both randomised and non-randomised evaluations are pooled to estimate differences in effect when randomisation is removed.[40]

Wilson and Lipsey[76] sought to investigate the influence of a range of study methods relative to intervention characteristics, participant characteristics and measurement features on effects. Randomisation was one of a number of attributes examined. Following a search of a number of electronic databases they synthesised 319 meta-analyses of psychological, behavioural and educational interventions, 76 of which contained both RCTs and non-randomised comparative studies. Rather than grouping together randomised and non-randomised studies of similar interventions and populations, a 'lumping' approach was adopted. Thus, studies of differing methods and intervention types were combined in a series of analyses, and the influence of different characteristics were explored.

For each meta-analysis the authors coded the total effect size variance around the grand mean effect size, and the effect size variance according to selected study features (e.g. type of research design, type of intervention, type of outcome measure, participant characteristics). They determined the proportion of effect size variance associated with the study features of interest for each meta-analysis using the eta-squared technique. Eta-squared is the ratio of the between-group sum of squares to the total sum of squares. In terms of study design they estimated variance according to randomised versus non-randomised controlled designs, and comparison group (randomised or non-randomised controlled designs) versus one group pre–post test designs. They also carried out further analyses to examine the direction and strength of any relationships.

Effect modifiers

Effect modifiers investigated

There were differences across the eight studies in the effect modifiers investigated (*Table 3*). In part this reflected variations in the overall aims of each study, with some studying differences in effect:

- according to study design[42]
- according to study design but also taking into account other variables that might be confounded with design[46,72,74,87]
- according to methods (including design) and substantive factors (related to the intervention)[76]
- according to some other variable of primary interest, but also assessing study design.[75]

The review of reviews by Glazerman *et al.*[70,71] did not aim to investigate the influence of effect modifiers, but did investigate differences between randomised and non-randomised methods, and the role of different techniques for comparing randomised and non-randomised groups within a single study.

In general the factors studied across the eight studies could be classified into one of four categories:

1. population characteristics (e.g. age, gender, ethnicity, socioeconomic status, diagnosis)
2. intervention characteristics (e.g. type, intensity, duration, standardisation, implementation)
3. outcome features (e.g. specificity, self-report, outcome interval)
4. design features (e.g. type of design, attrition, sample size, self-selection into study).

The factors falling outside of these categories were publication status[46,72,74] and year of publication.[74,75]

Some studies gave a clear rationale for the variables they selected as potential effect modifiers,[72,74] and other studies gave detailed descriptions of each variable of interest. For example, Wilson and Lipsey[76] within their four broad categories (respondent, treatment, measurement and design) clearly defined individual variables such as (1) age, gender, ethnicity and socioeconomic status of respondents; (2) treatment type, components, and dosage; (3) outcome constructs and how they are measured and source of outcome information; and (4) type of comparison group, design type, methodological quality and sample size.

Effect modifiers identified

In the study in which design was the main factor of interest, the authors concluded that the weaker the design the more likely it was for a study to report a result in favour of intervention and the less likely it was to report a harmful effect.[42] When RCTs were compared with the 'highest quality quasi-experimental' studies (defined as 'studies in which intervention and comparison groups are compared, with other mediating factors controlled for/or with a matched comparison group') the same pattern of results was maintained. The authors themselves note that the studies, although all in crime and justice, are very different and that very few of them examined a specific type of intervention. If the authors had explored the impact of study design on outcomes for specific types of intervention then different findings may have emerged.

The four studies that investigated the impact of study design on effect size but also took into account other variables that might be confounded with design[46,72,74,87] reported conflicting findings. There were even discrepancies within studies when different techniques were used to examine differences in effect sizes.[72]

The study by Lipsey[46] focused on interventions to prevent or reduce juvenile delinquency and reported larger effect sizes from NRSs. However, he went on to conclude that such a finding is valid only if randomised and non-randomised studies are otherwise similar in terms of characteristics (participants, interventions, outcomes and study methods) other than randomisation that might be related to effect size. In further analyses, Lipsey found that a range of moderator variables were associated with effect size, including participant (gender, ethnicity) and intervention characteristics (e.g. intervention type, duration, intensity), and that some methodological variables (e.g. attrition, sample size and duration of intervention) were also related to design. These findings suggest that great care has to be given to disentangling the relationships between moderators and identifying those that have independent relationships with effect sizes.

In contrast, two studies concluded that the effect sizes from RCTs were much larger than from NRSs.[72,74] The interventions investigated were marital and family therapy,[74] scholastic aptitude test coaching, ability grouping in classrooms, pre-surgical education and drug abuse prevention.[72]

However, this finding was influenced by inclusion in the analysis of variables thought to be confounded with method of assignment (such as level of activity of intervention compared with control, pre-test effect size, self versus other selection into conditions). In both studies much of the discrepancy in effect size appeared to be due to design being confounded with other variables and, when the effects of the confounds were removed, the difference in effect size between randomised and non-randomised studies was much smaller.

The final study in this group of four found abstinence rates after intervention for alcohol use disorders to be similar for randomised and non-randomised studies.[87] This effect remained after controlling for differences in features between randomised and non-randomised studies (application of inclusion/exclusion criteria, length of follow-up and follow-up rates).

The meta-epidemiological study in which design and intervention features were the factors of interest included 319 meta-analyses of psychological, behavioural and educational interventions.[76] In the first analysis the authors were interested in estimating the proportion of total variance in observed effect sizes associated with study features. The authors found that different treatment types were associated with the largest proportion of effect size variability and that overall individual study features accounted for between 2% and 8% of effect size variance. Interestingly, the proportion of variance associated with study design was slightly smaller than that associated with most of the substantive features of the intervention, particularly measurement of the outcome. In further analyses the authors investigated both the direction and the strength of the relationship between effect size and individual study features. The authors concluded that there was little difference (on average) between the results from randomised and non-randomised studies, but that this should not be interpreted as evidence for the equivalence of randomised and non-randomised designs. Overall findings from this study suggest that effect sizes are to some extent a function of specific features of study methods and of the intervention itself.

Although the main factor of interest in the study by Shadish *et al.*[75] was to what extent studies evaluating the effects of psychological therapies were 'clinically representative', the authors also compared effect sizes between randomised and non-randomised studies. They found significantly larger effect sizes from randomised studies, which they concluded was due to a bias caused by self-selection into treatment by participants in NRSs (explored using pre-test effect sizes). This misleadingly makes treatment appear less effective in NRSs. The authors explain this by suggesting that clients in the most psychological distress were more likely to self-select into intervention groups, leaving less distressed clients to form the control group. At baseline they tend to score worse than control group clients on measures of distress. Even if treatment is effective, the post-treatment effect size for the intervention group is likely to be relatively small. Interestingly, they found that effect sizes for NRSs where clients self-selected into groups were much lower than NRSs where allocation was conducted by researchers (e.g. alternate methods of allocation, matching of intervention/control participants). This suggests that NRSs that could be considered to be 'quasi-randomised' are more likely to be associated with effects similar to those of RCTs. Finally, the study found that RCTs were less clinically representative than NRSs, although there were a substantial number of clinically representative RCTs. The fact that clinical representativeness is associated with lower effect sizes is probably an artefact of confounds such as self-selection bias in many NRSs that happen to be clinically effective.

The final study reviewed, by Glazerman *et al.*,[70,71] was cautious in its conclusions about differences between within-study randomised and non-randomised estimates of effect. The review suggests that long-standing debates about the appropriateness of non-randomised methods cannot yet be resolved, at least within the area of welfare and employment programmes. Non-randomised methods sometimes came close to replicating results generated by randomised methods, but sometimes they were dramatically different. In terms of the different methods of comparing randomised and non-randomised methods within single studies, the authors found that bias was lower when the comparison group was drawn from within the same evaluation, as opposed to a national data set. The same was found when the control group was locally matched to the treatment population, or drawn as a control group in an evaluation of a similar programme or the same programme at a different study site. In general, statistical adjustments to compensate for non-randomised methods reduced bias, but methods such as regression, propensity score matching or other forms of matching did not differ greatly in terms of bias reduction.

The overall conclusion from this set of nine studies (*Table 4*) is that in some situations the results of randomised and non-randomised studies appear to differ and sometimes they appear similar, but, importantly, these differences may be linked to a range of other features that are confounded with design (see *Table 3*).

This is similar to the conclusion reached by Deeks *et al.*[40] in their investigation of the results of randomised and non-randomised studies evaluating health-care interventions. However, the range of confounders investigated in the studies described here, with the exception of Weisburd *et al.*,[42] appeared more diverse.

Summary and implications

We identified and analysed eight methodological studies, mostly meta-analyses, in the disciplines beyond health. The studies have examined the influence of evaluation methodology, principally randomisation, on study outcomes, and the mediating effect of variables such as type of intervention or participant characteristics. In some cases these studies appear to have been instigated by researchers to improve the way meta-analyses address heterogeneity and mediating variables in the interpretation of effects. In other cases, studies are reported to appeal not just to the research community, but also to policy-makers and practitioners to underpin their decision-making around choice of intervention.

The studies varied in aims and scope, but in general they posed similar questions; notably, whether presence or absence of randomisation in a controlled evaluation significantly influences outcomes. From this, inferences may be made about the degree to which absence of randomisation causes bias. Statistical procedures also varied,

TABLE 3 *Confounders of randomisation and effect size*

Hypothetical confounders (see Chapter 3) of effect size	Confounders indentified in the literature
Participants of the evaluation	
Baseline characteristics	Participant characteristics (gender, ethnicity)[46]
Attrition	Clinical representativeness[88]
	Self selection by participants[72,74]
	Pre-test measures high[72,74]
	Differential attrition[42]
The intervention	
Theoretical underpinnings	Intervention characteristics (e.g. intervention type, duration, intensity)[46]
Setting and boundaries of the intervention	
Providers of the intervention	Level of activity of intervention compared with control[72,74]
	No treatment vs alternative treatment in control group[76]
Outcomes	
Choice of outcome domains	Operationalisation of outcome measures[76]
Choice of outcome measures	Use of researcher-developed outcome measure[76]
Design of the evaluation	
Sample size	Strength of study design[42]
Control group	Methodological variables (e.g. attrition, sample size and duration of intervention)[46]
Blinding	
Follow-up	Comparison group was drawn from within the same evaluation, locally matched or drawn from a national data set[70,71]
Clustering	
Quality of reporting	
	Publication bias[42,72,74]

TABLE 4 *Study results*

Study	Factors investigated	Results	Author conclusions
Review of within-study comparisons			
Glazerman, 2003[70,71]	Source of comparison group (same labour market[b], control group from other site[b], national data set) Statistical method (regression[b], matching[b], selection correction or instrumental variables, none) Type of matching (propensity score: 1 to 1[b], propensity score: 1 to many[b], other matching technique, no matching) Specification test result (not recommended, recommended[b], no test[b]) Quality of background data: regression (poor set of controls, extensive set of controls[b], very extensive set of controls[b], regression not used) Quality of background data: matching (poor set of covariates, extensive set of covariates, very extensive set of covariates[b], matching not used) Quality of background data: overall (used prior earnings[b], did not use prior earnings) Experimental sample size (small: < 500 controls; medium: 500–1500 controls[b]; large: > 1500 controls[b]) Experimental impact finding (programme effective, programme ineffective, indeterminate[b])	Bivariate analysis[b] Multivariate analysis: Estimated six regression models. Found that average bias reduced by similar amount when either regression or matching were used and by a further degree if both used. Baseline measures of the outcome (i.e. pre-programme earnings) were also important, as was use of a control group matched to the same geographic area or labour market. Use of national data sets tended to increase average bias. Using a control group from another site (i.e. the control group from the same RCT in anther area) reduced bias even further, but clearly such control groups will not readily be available to evaluators Finally examined effect to which positive and negative biases could cancel each other out. Concluded that if a sufficiently high number of non-experimental estimators could be found the inference that could be drawn from such evidence might be improved but not in a predictable way	Those who plan and design new studies to evaluate the impacts of training or welfare programmes on participants earnings can use the empirical evidence to improve non-experimental designs, but not to justify their use Also note that the various authors used different standards to assess the size of the bias and, in some cases, reached different conclusions with the same data. Their conclusions should be further probed than was possible here. Also some studies used more realistic replication than others of what would have happened in the absence of randomisation
'Single' meta-analyses			
Weisburd, 2001[42]	Cross-tabulated SMS score with mean IRR to give an indication of intervention outcomes according to study design	RCTs had the lowest mean IRR (0.22, SD 0.70) and correlational studies the highest (0.80, SD 0.42), suggesting a linear inverse relationship between study design and outcome 80% of correlational studies showed positive intervention effects compared with 65% of quasi-experimental studies and 37% of RCTs	Authors believe their findings point to the possibility of an overall positive bias in non-randomised criminal justice studies, although this may be confounded by publication bias or different attrition rates across designs

TABLE 4 *Study results (continued)*

Study	Factors investigated	Results	Author conclusions
Shadish, 1996[74]	Random vs non-random assignment[a] Pre-test effect size[a] Publication status ([a]but randomisation effect robust to publication status) Attrition[a] Use of matching – stratifying Random assignment, no differential attrition vs non-random assignment plus matching[a] Internal vs external control group Self vs other selection of participants Specificity of outcome Sample size Effect size calculation method Use of self-report outcome Treatment standardisation Active vs passive control group	Overall average effect size significantly higher for RCTs than for NRSs, and variance component smaller, although this may not be significant Regression model showed significant but smaller effect of randomisation when combined with internal/external control group, self vs other selection of participants, specificity of outcome, sample size, effect size calculation method, use of self-report outcome, treatment standardisation and active vs passive control group. Effect size was significantly higher with published than with unpublished works, when pre-test effect size is high and when participants do not self-select	Authors suggest that although effect sizes of RCTs were significantly larger than those of NRSs, much of the discrepancy was due to confounding with other variables. When the effects of the confounders were removed the difference was halved. The importance of the finding depends on whether one is discussing meta-analysis or primary studies, how precise an answer is needed and whether some adjustment to the data from studies using non-random assignment is possible. Authors conclude that NRSs may produce acceptable approximations to RCTs under some circumstances but RCTs remain the gold standard
Heinsman, 1996[72]	Self vs other report % differential attrition[a] Specific vs general measure Published vs unpublished Passive vs active control group[a] Exact vs approximate effect size[a] Sample size Pre-test effect size[a] Random vs non-random assignment Standardised treatment vs not Self vs other selection into conditions[a] Use of matching – stratifying or not % total attrition[a] Internal vs external control group	Overall average effect size higher for RCTs than NRSs; however, the size and direction of differences within the four areas varied considerably Regression model showed no effect from method of assignment. Effect size was higher with low differential and total attrition, with passive controls with higher pre-test effect sizes, when the selection mechanism did not involve self-selection of subjects into treatment and with exact effect size computation measures	Authors suggest that if RCTS and NRS were equally well designed they would yield roughly the same effect size

continued

TABLE 4 *Study results (continued)*

Study	Factors investigated	Results	Author conclusions
Shadish, 2000[75]	Clinical representativeness score Year of publication Length of therapy in minutes[a] Total attrition Differential attrition Reactivity scale Outcome specificity[a] Matching Internal control group Passive control group Self-selection Did not use structure Random assignment Unpublished work Adult presenting problem Not brief therapy Behavioural orientation Weeks to post-test	Randomised studies tend to report larger effect sizes than non-randomised studies (ascribed to selection bias in NRSs) Clinical representativeness scores were significantly different across the three sources. NRSs had significantly higher clinical representativeness scores than RCTs. A significant negative correlation was seen between clinical representativeness scores and year of publication Regression analysis found that effect sizes were larger the greater the dose of therapy, when highly specific measures were used, when internal control groups were used, when outcome was measured near the end of therapy, for behaviourally oriented therapies, with more representative clinical structure, for participants without clinically representative mental health problems and when therapy was not limited to a fixed number of sessions	Authors conclude that psychological therapies are robustly effective across conditions that range from research-oriented to clinically representative; previous findings that clinical representativeness leads to lower effect size are probably an artefact of other confounding variables, especially biased self-selection; increased dose of therapy is associated with larger effect sizes; and larger effects are seen in studies using outcome measures closely tailored to treatment goals
Lipsey, 2003[46]	Random vs non-random assignment Research/demonstration vs practice programmes Type of control Attrition Sample size Official records Outcome interval Published Gender Age Ethnicity Adjudicated Prior offences Treatment type Custodial Amount Intensity Implementation problems	Mean effect size for research/demonstration programmes was significantly larger than that for practice programmes within each design (randomised and non-randomised). Within each type of programme, non-randomised designs are associated with larger effects than randomised designs. Smaller effect sizes for RCTs confounded with and accentuated by a difference between the mean effects for research/demonstration and practice programmes Many moderator variables are associated with effect size, some quite strongly. A number of them are also related to type of design or type of programme or both. Effect size differences associated with moderator variables are generally as large as or larger than those associated with type of design	Great care must be taken when interpreting the relationship between a moderator variable and effect sizes in meta-analysis, especially if the relationship appears to have implications for practice or policy

TABLE 4 *Study results (continued)*

Study	Factors investigated	Results	Author conclusions
Moyer, 2002[87]	Random vs non-random assignment Type of treatment Participant characteristics Other methodological features	RCTs were more likely to use recognised diagnostic criteria to characterise participants and to stringently implement treatment and assess outcomes. NRSs were more likely to assess outcomes in higher proportions of participants over longer follow-up periods and to have greater statistical power to detect treatment effects Abstinence and improvement rates following active treatment were similar for the two types of design, even when differences in study features were controlled	The contrasting strengths and weaknesses of randomised and non-randomised studies suggest that they should be considered as complementary forms of treatment evaluation in the alcohol treatment field and perhaps more generally
Petrosino, 2003[26]	Random or potentially random vs non-random assignment	RCTs and potentially randomised trials as a percentage of outcome evaluation studies for each database: ERIC 16%, Criminal Justice Abstracts 19%, MEDLINE 54%, NCCAN 12%, PsycINFO 21%, Sociofile 9%	Randomised studies are used in nearly 70% of childhood interventions in health care but probably in 6–15% of kindergarten to 12th grade interventions in education and juvenile justice
Meta-reviews			
Wilson, 2001[76]	Treatment features: type component intensity/duration Respondent features: age gender ethnicity socioeconomic status diagnosis ability group Measurement features: construct operationalisation source of information researcher-developed measure[a] Design features: comparison group type[a] design type: randomised vs non-randomised[a], comparison group vs pre–post test[a] methodological quality sample size[a]	2% (95% CI 1 to 3) of overall effect size variance was associated with non-random allocation compared with 8% (95% CI 6 to 10) associated with treatment type or 8% (95% CI 2 to 14) with operationalisation of outcome measures Mean linear correlation of random allocation with effect size 0.04[a] (i.e. RCTs yielded slightly higher effect sizes), with a range from –0.60 to 0.77, i.e. both large over- and underestimates were found from NRSs. Correlation was 0.18[a] for use of no treatment vs alternative treatment in control group, or 0.10[a] for use of researcher-developed outcome measure Overall mean effect size difference for random allocation 0.03 compared with 0.13[a] for researcher-developed outcome measure, 0.26[a] for no treatment control vs alternative treatment control, and –0.18[a] for sample size	Randomisation: findings cannot be taken as evidence of equivalence of RCTs and NRS designs, more likely that the selection bias in one NRS is offset by an opposite bias in another such comparison, i.e. neither consistent under- nor overestimate of effect size Found that operationalisation of outcome measures is at least as important as type of design, if not more so State design features and outcome operationalisations are often related to treatment type, duration, respondent characteristics and other substantive features of the interventions

CI, confidence interval; ERIC, Educational Resources Information Centre; IRR, investigator reported result; NCCAN, National Clearinghouse on Child Abuse and Neglect Information (NCCAN Clearinghouse); SD, standard deviation; SMS, scientific methods scale.

a Indicates statistically significant effect ($p < 0.05$) in at least one analysis.

b Indicates categories with lowest level of associated bias compared with other categories per explanatory variable.

although most studies employed predictive meta-regression models to identify associations.

The overall conclusions from this set of eight studies are that in some situations the results of randomised and non-randomised studies appear similar and sometimes they appear to differ. Importantly, these differences may be linked to a range of other features likely to be confounded with design, such as participant and intervention characteristics. Inter-relationships among variables make it difficult to determine the likely impact of any one factor, which is of vital importance when the findings have direct implications for policy or practice. Thus, in terms of providing answers to the questions posed in Chapter 3 about whether randomised and non-randomised studies lead to differences in effect sizes and whether any observed differences are due to randomisation per se or to other factors associated with randomisation, the findings from the methodological literature studied suggest that effect sizes from the two types of study may indeed differ and that these differences may well be associated with factors confounded with design. In the following chapter this issue is further investigated through a systematic review of systematic reviews that have evaluated (via randomised and non-randomised studies) the effects of policy interventions.

Chapter 5

Systematic review of systematic reviews

Aim

To search for, assess and synthesise systematic reviews that have:

- compared the results of policy interventions estimated from randomised and non-randomised studies
- described the methods used by reviewers to identify factors other than the use of randomisation that may have influenced the results of randomised and non-randomised studies

in order to identify any differences in average effect and/or variability between designs, and to identify any variables (confounders/moderators of effect) that might affect the above.

Methods

Selection of systematic reviews

Systematic reviews meeting the following criteria were eligible for inclusion:

- completed or published between 1999 and 2004 (limit applied to try and ensure the inclusion of reviews with up-to-date methods)
- evaluated a policy intervention (see Chapter 1, Defining 'policy intervention')
- included both randomised and non-randomised studies and have estimated intervention effects separately according to design (or provide sufficient data to allow us to do so)
- used quantitative synthesis (meta-analysis).

Reviews which either estimated intervention effects separately according to design but did not quantitatively synthesise the studies or used quantitative synthesis (meta-analysis) but did not estimate intervention effects separately according to design were excluded from the main analysis but are discussed briefly in Additional policy intervention reviews.

A flowchart of the inclusion process is presented in *Figure 1*. Each review potentially meeting inclusion criteria was screened by one reviewer using a predefined electronic form, and checked by a second reviewer. Disagreements were resolved by consensus, with reference to a third reviewer if necessary.

Literature searches and data sources

There are probably even fewer definitive terms available for 'systematic reviews' or equivalent in the wider literature than there are for primary evaluation or trial designs. Ideally, to ensure that the searches retrieved relevant references, the strategy would have included terms for 'review' or at least for 'literature review'.[89] However, attempts at searching using these terms on electronic databases (e.g. MEDLINE) produced an unmanageable number of references.

We therefore decided to search for all available references relating to policy interventions from databases that exclusively contain citations to reviews and systematic reviews. All reviews potentially relating to policy interventions available on the following databases were obtained: Centre for Reviews and Dissemination DARE (Database of Abstracts of Reviews of Effects); the Cochrane Database of Systematic Reviews (CDSR), the Campbell Collaboration's Database, C2 RIPE (Register of C2 Systematic Reviews of Interventions and Policy Evaluation); the EPPI-Centre DoPHER; and the (former) Health Development Agency Evidence Base Database.

A number of test searches were then completed in databases with a focus beyond health to estimate the feasibility of attempting a comprehensive search of the non-health literature for systematic reviews. The searches were restricted to very specific terms for 'systematic review', without using proximity operators, and were further restricted by date range (2003–4).

The following free-text terms and indexed keywords (if available) were used: meta-analysis, meta-analysis, systematic review, systematic overview, collaborative review, integrative research, integrative review, research integration, narrative

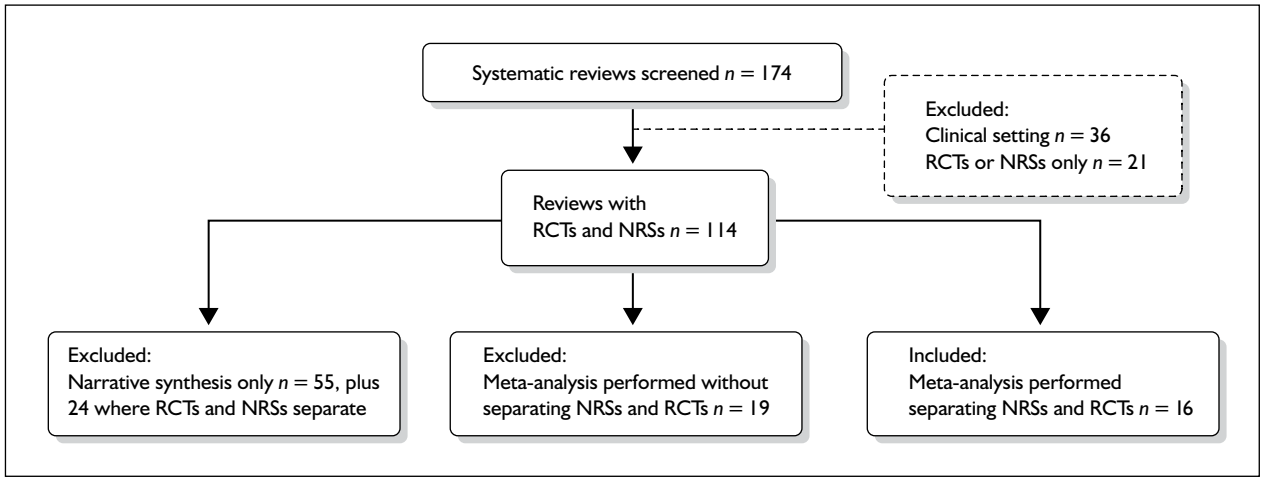

FIGURE 1 *Flow chart of the inclusion process.*

synthesis, evaluation synthesis, meta synthesis, realist synthesis, descriptive synthesis, explanatory synthesis and pool data.

The following databases were searched:

- ASSIA
- BEI
- CareData
- ERIC
- HDA HealthPromis
- PAIS International
- SIGLE
- SSCI
- Sociological Abstracts.

The results of the test searches confirmed that it would not be worthwhile conducting thorough, comprehensive searches of databases with a focus beyond health for systematic reviews. For the year 2003–4 alone, a precise search without 'literature review' identified 2494 records, and a more sensitive search with 'literature review' added identified 45,596 records. Full details of the search strategies can be found in Appendix 3.

Data extraction and analysis

A data extraction form for recording relevant information from each systematic review was designed and piloted. Full systematic reviews were pre-screened independently by two reviewers. Those meeting the inclusion criteria had data extracted by one reviewer and the completed data extraction forms checked against the full paper by a second reviewer. Any disagreements were resolved by consensus or by referral to a third reviewer if necessary.

The following information was extracted from each review:

- details of literature search used to identify studies for inclusion, including databases searched, years and whether details of the search strategy were available
- method of assessing studies for inclusion
- details about quality assessment
- number of studies/participants per design
- method of synthesis used
- results according to study design and other quality features.

Review methods and results were tabulated and discussed narratively. They were first classified into three groups according to the authors' judgement regarding the equivalence or otherwise ('similar', 'not similar' or 'mixed') of the results of RCTs and NRSs. We focused on the following key methodological aspects:

- Whether authors attempted to examine similarities or differences in the following aspects across study designs (or whether they provide sufficient study details to allow us to judge): study populations; interventions used (design and provider of intervention); design of evaluation; outcomes assessed; study dates (if NRSs have largely been conducted before RCTs they may be more likely to show positive effects, i.e. their positive results leading to the RCTs being commissioned in the first

place); and other aspects as recorded by review authors.
- Whether authors tried to assess heterogeneity either across or within study designs, using statistical methods or other approaches to identifying heterogeneity.
- What criteria were used to establish equivalence (or otherwise) of the results of RCTs and NRSs and whether these criteria were sensible and objective.

Details about these items were recorded so that any observed differences in the results of randomised and non-randomised studies could be considered and were not simply attributed to lack of randomisation.

Results

The results of test searches found that thorough, comprehensive searches of non-health databases for systematic reviews would not be worthwhile. Therefore the results reported here are for studies found through health-related databases alone.

Description of reviews

Sixteen reviews met inclusion criteria (*Table 5* and Appendix 4).

Interventions

Eight reviews included children,[90–97] usually in schools, and eight included adults, usually in hospitals.[98–105] None of the included reviews assessed the effects of legislation and only one (the hospital falls prevention programme review)[103] included interventions aimed at modifying the environment. In terms of the scope of the intervention, the majority aimed to make changes within institutional settings, such as hospitals, schools or the workplace. Some of these also aimed to influence the wider community, in terms of multicomponent interventions based in schools, the home and community settings.

Study identification and inclusion

Information on the extent of searching was extracted to help determine whether the identification of included studies within the reviews was likely to be biased. Evaluations of publication bias have noted differences in the frequency of publication of randomised and observational studies.[106] The extent of searching across the included reviews was variable, ranging from two electronic databases plus reference lists[98,103] to more extensive searching in the majority of included reviews. Two of the reviews used specialised databases: Langhorne *et al.*[102] used the Cochrane Stroke Group Specialised Register of Controlled Trials, and Mullen *et al.*[93] used the Prevention Research Synthesis Project database. Four of the 16 reviews restricted inclusion to English language publications.[90,93,94,97] Included studies in other reviews all appeared to be English language. Some reviews specified that studies had to take place in the US, Canada or the UK, so restriction to English language was probably appropriate in these reviews.

Most reviews included fewer studies in the meta-analysis than in the review overall; this is to be expected as not all included studies would report all outcomes of interest. The exception to this would be where a review specified only one or two outcomes of interest and restricted inclusion to studies that reported those outcomes.

Quality assessment

Information on quality assessment within the reviews was extracted to help determine whether validity was assessed appropriately for each study design, i.e. are we getting a true picture of whether the included studies were of good quality? Only three of the reviews did not assess study validity.[91,93,100] In the others, a mixture of checklists, scales and components were used. Jacobs *et al.*[92] only assessed the validity of RCTs.

Synthesis methods

There were two approaches to meta-analysis. The first involved keeping RCTs and NRSs separate throughout the review and meta-analysis process, although the rationale for doing this was rarely explicit. This approach was used in eight of the 16 reviews.[91,92,94,99,101–104] Four of these reviews stated a priori that they would investigate potential moderators of effect.[91,95,96,104]

The alternative approach was to pool all studies in a meta-analysis, and then investigate potential moderators of effect (an aspect of the study that varies from one study to the next) including randomisation on the average estimate of effect. This approach was taken in the remaining eight reviews, and usually involved a large number of studies that varied enormously in terms of intervention, population, outcomes and outcome measures. Outcomes were converted to effect sizes to enable them to be pooled.[90,92–94,96–98,100,101,105,106] Six of these reviews stated which potential moderator variables they would be investigating

TABLE 5 *Summary of review methods (n = 16)*

Method	Category	Total
Intervention type	Rehabilitation/treatment	4
	Hospital policy	2
	Prevention	7
	Health promotion	2
	Criminal justice	1
Scope of implementation	Policy for a nation	0
	Policy for a region	0
	Policy for a community	5
	Policy for an institution	15
Search	More than three electronic sources + reference lists	11
	Three or fewer electronic sources + reference lists	5
Language restriction	English only	4
	No restriction	4
	Not stated	8
Number using quality-related inclusion criteria		1 (NRS only)
Number using quality assessment	Not conducted	3
	Conducted	13 (1 RCT only)
Number of RCTs	Median (IQR; range)	11.5 (3.5–16; 2–144)
	Number not reported	2
Number of NRSs	Median (IQR; range)	14.5 (8–25; 1–174)
	Number not reported	2

IQR, interquartile range.

a priori, either by means of subgroup analysis (Tobler *et al.*,[107] Davis and Gidycz[90] and Mullen *et al.*[93]), regression analysis (Griffith *et al.*[100] and Wilson *et al.*[105]) or using both sensitivity and subgroup analysis (Cambach *et al.*[98]). Wilson *et al.*[96] used sensitivity analyses, and Wilson *et al.*[97] used multiple regression to investigate potential moderators of effects but these were not pre-stated investigations.

In the following sections we discuss the included reviews according to whether the authors judged the results of RCTs and NRSs to be 'similar', 'non-similar' or 'mixed'.

Review results: where authors judged results from RCTs and NRSs to be 'similar' (*n* = 5)

Five reviews are included in this section: Cameron *et al.*,[99] Kwan and Sandercock,[101] Langhorne *et al.*,[102] Tobler *et al.*,[107] and Wilson *et al.*[105] (*Table 6*, Appendices 4.1–4.13).

Method of pooling

Only two reviews had a stated objective to investigate differential effects of randomisation (among other variables).[105,107] Tobler *et al.*[107] aimed to 'empirically confirm that the inclusion of non-randomised pre-test/post-test research designs does not overestimate intervention success', while Wilson *et al.*[105] aimed to investigate the influence of study design on findings. Both used the 'lumping' approach to meta-analysis, pooling all studies and then investigating potential moderators of effect, including randomisation (overall effect is reported for the subset of randomised studies in both reviews). Neither review discussed weaknesses with this approach. Both reviews investigated the magnitude of effect but neither reported assessing the variance associated with it, which ideally should be investigated for RCTs versus NRSs.

TABLE 6 *Summary of review findings*

		Total *n* (%)	Results judged 'similar' *n* (%)	Results judged 'not similar' *n* (%)	Results mixed *n* (%)
Statistical heterogeneity identified by design?	Yes	4 (25)	2 (40)	2 (25)	0
	No	11 (69)	3 (60)	5 (62.5)	3 (100)
	Narrative only	1 (6.25)	0	1 (12.5)	0
Obvious differences between RCTs/NRSs (author or reviewers opinion)	Yes	4 (25)	1 (20)	3 (37.5)	0
	No	12 (75)	4 (80)	5 (62.5)	3 (100)
	Narrative only	0	0	0	0
Sources of heterogeneity investigated?	Population				
	Yes	2 (12.5)	1 (20)	1 (12.5)	0
	No	14 (87.5)	4 (80)	7 (87.5)	3 (100)
	Intervention				
	Yes	4 (25)	2 (40)	2 (25)	0
	No	12 (75)	3 (60)	6 (75)	3 (100)
	Comparator				
	Yes	2 (12.5)	2 (40)	0	0
	No	14 (87.5)	3 (60)	8 (100)	3 (100)
	Outcomes				
	Yes	1 (6.25)	1 (20)	0	0
	No	15 (93.75)	4 (80)	8 (100)	3 (100)
Rationale for pooling approach given?	Yes	6 (37.5)	2 (40)	3 (37.5)	1 (33.3)
	No	7 (44)	2 (40)	4 (50)	1 (33.3)
	Partially?	3 (18.5)	1 (20)	1 (12.5)	1 (33.3)
Criteria to judge equivalence of study results by design given?	Yes	3 (18.5)	0	2 (25)	1 (33.3)
	No	13 (81.5)	5 (100)	6 (75)	2 (66.7)

The other three reviews pooled RCTs and NRSs separately; only one of which[101] provided any form of justification for this approach, stating in the discussion section that 'non-randomised studies are highly susceptible to bias and there is significant statistical heterogeneity between the studies'.

None of the five reviews in this section reported the criteria that were used to judge equivalence between results of RCTs and NRSs (Appendix 4.2), so it is not clear whether equivalence was judged in a systematic or pre-specified way, or, if so, how sensible and objective were the criteria used.

Assessment of heterogeneity

Statistical heterogeneity was tested for separately by randomised and non-randomised study design in three of the five reviews in this section.[99,101,102]

None of the five reviews in this section described whether or not there was clinical and methodological heterogeneity between RCTs and NRSs in terms of participants, interventions, outcomes, methodology (other than randomisation) or any other aspect. Nor did they give sufficient information for us to make a strong judgement, although there did seem to be elements of the populations, interventions, comparators and outcomes measured that differed between randomised and non-randomised designs. As few details were provided, it is difficult to assess whether the RCTs and NRSs were sufficiently similar to allow a comparison to be made (i.e. whether like was being compared with like). The authors did not mention any obvious differences between RCTs and NRSs except for Kwan and Sandercock,[101] who stated that the comparator was poorly described in NRSs. Based on the limited

information presented about the included studies, we noted no obvious systematic differences between RCTs and NRSs.

Comparison of review results and authors' conclusions regarding similarity between RCTs and NRSs

Appendix 4.3 gives the pooled results of the RCTs and NRSs separately, along with the results of any heterogeneity tests performed.

In one of the reviews[102] there were clear differences in the results of RCTs and nRCTs for at least one outcome measure; however, study numbers were extremely small. In this review,[102] two pooled RCTs indicated a significantly increased risk of death but in the single NRS there was a small but not significant drop in risk. The NRS also indicated a significant decrease in the number of patients admitted to hospital but the RCTs showed a non-significant increase. Significant statistical heterogeneity was seen in some outcomes in RCTs but, as there was only one NRS in the meta-analysis, there was no heterogeneity in NRS outcomes. The authors stated that considerable heterogeneity between trials made it difficult to draw specific conclusions.

Kwan and Sandercock[101] found the results of RCTs indicated a trend towards longer hospital stay with the intervention whereas NRSs indicated a shorter stay, but neither effect was statistically significant at the 5% level. Significant statistical heterogeneity was seen for the outcome 'duration of hospital stay' in NRSs but not in RCTs. The authors identified that RCTs and NRSs were showing trends to give answers in opposite directions but that the differences were not statistically significant.

In the review by Tobler *et al.*,[107] effect sizes were given without CIs, so it is difficult to judge whether there were differences between results of RCTs and NRSs. Pooled effect sizes were slightly smaller in the NRSs than in the RCTs. The authors commented that lack of random assignment does not seem to greatly bias results relative to other problems. Mean effect sizes for studies with random assignment and non-random assignment differed by only 0.03. Removing other sources of bias influenced the results far more.

For the remaining two reviews[99,105] results for RCTs and NRSs were very similar in terms of magnitude and direction. In the review by Cameron *et al.*,[99] significant statistical heterogeneity was seen for length of hospital stay and mortality in RCTs but not in NRSs. Results of RCTs and NRSs were judged similar although the authors stated that for some outcomes there was greater heterogeneity between RCTs than between the pooled data from RCTs and that from cohort studies. Wilson *et al.*[105] noted that the difference in results between randomised and non-randomised studies was unremarkable and not statistically significant.

Results of additional heterogeneity investigations

Cameron *et al.*[99] and Kwan and Sandercock[101] did not carry out any further investigations of sources of heterogeneity. Cameron *et al.*[99] state that both the experimental and the control interventions were complex and varied in nature, and propose that differences in case-mix within-study populations may have led to heterogeneity as 'it would be expected that… treatments and programmes targeting those most likely to benefit are most likely to demonstrate effectiveness'. Kwan and Sandercock[101] identified that the definition of the intervention 'care pathway' may have been a source of variation, and further urge readers to be cautious when interpreting results, because of the presence of variation between studies and small numbers of participants.

Langhorne *et al.*[102] carried out a sensitivity analysis to investigate heterogeneity in terms of intervention, trial design and patient follow-up. Details of the sensitivity analysis are not reported, other than that it did not alter the review's conclusions.

Tobler *et al.*[107] and Wilson *et al.*[105] were the only reviews in this group to report the details of any further heterogeneity investigations. Both reviews used the lumping approach to synthesis, and use of random assignment was only one of many covariates investigated as a potential source of heterogeneity.

Tobler *et al.*[107] analysed random assignment as one of many potential moderators of effect across a large meta-analysis (Appendix 4.4). Other potential moderators included design and provider of the intervention, design of the evaluation and population in terms of school grade, special populations and levels of drug use. The authors concluded that removing other potential sources of bias influenced the results of the review far more than removing studies without random assignment to the intervention.

Wilson *et al.*[105] also analysed many potential moderators of effect across a large meta-analysis including intervention design and design of the

evaluation (Appendix 4.4). The authors concluded that positive findings may result from participant characteristics rather than any positive effect of the intervention itself.

Summary

It seems reasonable to conclude that overall we found no evidence for clear systematic differences in results of RCTs and NRSs in the examples reviewed, but there is insufficient evidence on which to base any wider conclusions.

In the reviews included in this section, other potential sources of confounding are suggested to be more important than randomisation, but it could be just chance that the RCTs/NRSs are similar in these examples.

Review results: where authors judged results from RCTs and NRSs to be 'non-similar' (*n* = 8)

Eight reviews are included in this section: Cambach *et al.*,[98] Davis and Gidycz,[90] Griffith *et al.*,[100] Jacobs *et al.*,[92] Mullen *et al.*,[93] Oliver *et al.*,[103] Smedslund *et al.*,[104] and Wilson *et al.*[96] (see *Table 6*, Appendices 4.5–4.7).

Method of pooling

Three reviews[90,93,98] had a stated objective to consider differences in intervention effect between RCTs and NRSs. In each review the use of randomisation was one of several other methodological characteristics investigated.

Five[90,93,96,98,100] used the 'lumping' approach and three[92,103,104] the 'splitting' approach to analysis.

Two reviews described the criteria that were used to judge equivalence between results of RCTs and NRSs. Griffith *et al.*[100] stated that non-overlapping 95% CIs would allow conclusions about the strength of one 'predictor' (including randomisation) in comparison with another to be drawn. Mullen *et al.*[93] used a chi-squared statistic to assess the likelihood of the magnitude of the between-subgroup differences, in terms of both results of effect estimates and of heterogeneity.

Assessment of heterogeneity

Three of the eight reviews included in this section attempted to identify statistical heterogeneity separately for RCTs and NRSs.[92,100,104] Only two reviews[98,100] described any clinical and methodological heterogeneity between RCTs and NRSs in terms of interventions, participants, outcomes and methodology (other than randomisation).

Griffith *et al.*[100] found statistically significant heterogeneity reported among NRSs but not among RCTs. They reported that several NRSs (but not RCTs) involved patients considered to be treatment failures.

Jacobs *et al.*[92] found significant heterogeneity for three outcomes among RCTs but none for NRSs. The authors did not identify any possible causes of heterogeneity, although from our own assessment sample size seemed to be larger in NRSs than in RCTs. Smedslund *et al.*[104] found no significant heterogeneity.

In Cambach *et al.*[98] the NRSs were all undertaken in an outpatient setting, whereas the RCTs were in a mixture of settings. No further sources of within-group heterogeneity for RCTs and NRSs were identified.

None of the remaining reviews in this section[90,93,96,103] discussed clinical or methodological heterogeneity or reported sufficient detail of included studies for us to draw our own conclusions on the similarity or otherwise of the included studies.

Comparison of review results and authors' conclusions regarding similarity between RCTs and NRSs

In five reviews, effect sizes were larger in NRSs than in RCTs.[90,92,93,100,104] Davis and Gidycz[90] concluded that higher mean effect sizes were seen when studies did not use random assignment of participants. Other variables were also associated with increased effect size. The 95% CIs for the average estimates were not reported.

Griffith *et al.*[100] found significant heterogeneity in NRSs but not RCTs, and concluded that studies without random assignment reported 'better outcomes' than those with random assignment. Smedslund *et al.*[104] found no significant heterogeneity for RCTs or NRSs at the one time point at which a heterogeneity test was used. The authors concluded that NRSs showed larger effects than RCTs at all time points but that the RCTs were probably more reliable. In Griffith *et al.*[100] the 95% CIs overlapped by 0.02, and in Smedslund *et al.*[104] CIs also overlapped.

In Mullen *et al.*[93] 95% CIs did not overlap and the authors reported that between-subgroup

differences were not significantly different for random versus non-random assignment. Despite RCTs indicating a significant benefit and NRSs indicating no significant differences between the groups in this review, the authors found that random assignment versus non-random assignment when compared in the stratified subgroup analysis explained only 7.3% of the total heterogeneity.

Jacobs *et al.*[92] found significant heterogeneity for some outcomes in RCTs but no significant heterogeneity for any outcome in NRSs. This review concluded that RCTs found a statistically significant change in two outcomes that was not found in cohort studies.

Two reviews[96,103] found larger effect sizes in RCTs than in NRSs. In Oliver *et al.*,[103] 95% CIs overlapped, and in Wilson *et al.*[96] they did not. Wilson *et al.*[96] found a statistically significant difference in effect between RCTs and NRSs ($p < 0.05$) and concluded that randomised designs gave larger mean effects than non-randomised ones.

In one review,[98] results of RCTs and NRSs were in opposite directions, although it is unclear which direction indicated a positive result. Ninety-five per cent CIs were not reported so we cannot tell if these overlap. The conclusions of Cambach *et al.*[98] and of Oliver *et al.*[103] with regard to similarity between findings of RCTs and NRSs were unclear.

Results of additional heterogeneity investigations

Cambach *et al.*[98] carried out subgroup analyses using variables relating to study participants, interventions and comparators. They reported that outcomes were not significantly heterogeneous with regard to any of the variables investigated, including randomisation, although they also state that methodological quality may have biased the outcomes. Davis and Gidycz[90] carried out subgroup analyses using variables relating to participant age, intervention design and provider, study methodology and publication status. They found that several participant, programme and methodology characteristics were significantly related to effect size, including age, number of intervention sessions, extent of active participation in the intervention and type of outcome measure. Wilson *et al.*[96] carried out subgroup analyses using variables relating to methodology and population, and concluded that study design appeared to be related to observed effects, although inclusion of 'weak' designs did not seem to increase effect sizes. Both intervention and population variables were reported to be moderators of effect size.

Griffith *et al.*[100] investigated the effect of eight potential moderator variables on effect size; these included characteristics of intervention design, plus random assignment. They found that five of the eight variables, including randomisation, had a significant effect. Randomisation was associated with smaller effect sizes. Mullen *et al.*[93] investigated the effect of 13 potential moderator variables on effect size, including participants, aspects of intervention design, and provider and methodology. They found that eight of 13 variables explained more than 5% of the observed heterogeneity, with ethnicity explaining more than two times the total heterogeneity of any other variable.

Jacobs *et al.*[92] did not carry out statistical investigation of heterogeneity. They reported that the discrepancy in results for one outcome between RCTs and NRSs may reflect the differences between study populations in heterogeneity secondary to study design and/or bias. Oliver *et al.*[103] did not carry out statistical investigations of heterogeneity and made no comment regarding potential moderators of effect. Smedslund *et al.*[104] did not carry out statistical investigations of heterogeneity but commented that smoking cessation outcomes were influenced not only by the interventions but also by the settings and organisational context.

Summary

In reviews included in this section there was some evidence of dissimilar results arising from RCTs and NRSs but the CIs of the effect sizes overlapped in many cases. There was no real consideration of other differences between study designs that could contribute to these findings, although other variables were found to impact on overall intervention effects.

Review results: where authors judged results from RCTs and NRSs to be 'mixed' (*n* = 3)

Three reviews are included in this section: Guyatt *et al.*,[91] Thomas *et al.*,[94] and Wilson *et al.*[97] (see *Table 6*, Appendices 4.8–4.10). In these reviews, similarity and differences between results of RCTs and NRSs varied across outcomes.

Method of pooling

Only one review[91] had a stated objective to investigate differential effects of randomisation (among other variables). The 'splitting' approach to synthesis was taken in all three reviews but no justification for the approach was given.

A *z*-score was used to generate a *p*-value related to the null hypothesis that there were no real differences in results from RCTs and NRSs. The other two reviews did not state what criteria were used to judge equivalence of results between RCTs and NRSs so it is difficult to assess whether these were sensible and objective.

Assessment of heterogeneity

None of the reviews attempted to identify any statistical heterogeneity in RCTs and NRSs.

Only Guyatt *et al.*[91] attempted to narratively assess clinical heterogeneity in terms of population, recruitment, intervention and duration of follow-up. No obvious differences were reported between RCTs and NRSs.

There were no obvious differences between RCTs and NRSs expressed by Thomas *et al.*[94] in terms of population, interventions and outcomes from the details provided. Wilson *et al.*[97] did not report on clinical or methodological heterogeneity between included studies, and insufficient detail of included studies was given to enable the reader to judge whether there were obvious differences between RCTs and NRSs with regard to clinical or methodological features.

Comparison of review results and conclusions regarding similarity of results between RCTs and NRSs

Appendix 4.9 gives the pooled results of the RCTs and NRSs separately, along with the results of any heterogeneity tests performed.

In Guyatt *et al.*,[91] no significant effects were seen for any outcome in RCTs; in NRSs significant effects were seen in five of eight outcomes assessed; however, not all were in the same direction. The authors reported that there were statistically significant differences between the findings of RCTs and NRSs for two outcomes. They stated that relying on the results from observational studies would lead to the conclusion that the interventions have a positive effect, while relying on the results of RCTs would lead to the conclusion that the interventions did not have an effect.

In Thomas *et al.*,[94] significant effects were seen more often in RCTs than in NRSs, although these could go in either direction. The conclusions of this review regarding similarity of results between RCTs and NRSs are not clearly stated.

Wilson *et al.*[97] reported that effect size was larger in RCTs than in NRSs for some outcomes (but not significantly so).

Results of additional heterogeneity investigations

Guyatt *et al.*[91] investigated intervention design, methodological variables (random assignment and length of follow-up), gender and year of study as potential moderators of effect. They stated that interventions in RCTs and NRSs were similar in nature and intensity and that the studies were conducted at similar times and had similar lengths of follow-up. They concluded that it is likely that participants who received the intervention in the observational studies were more predisposed to a positive outcome.

Thomas *et al.*[94] investigated inclusion of a physical activity component and overall quality assessment as potential moderators of effect. Results suggested that in some circumstances observed variability in effect size between studies might be explained in part by whether or not the interventions promoted physical activity as well as healthy eating. The investigation of quality assessment as a potential moderator of effect appeared to focus only on randomisation and outcome measurement, and the authors concluded that this did not have a significant effect.

Wilson *et al.*[97] investigated aspects of intervention design and intervention provider, evaluation design and participants as potential moderators of effect. They found that, although there were no significant differences between RCTs and NRSs, other aspects of study design did seem to influence the outcome. Although significant differences in effect size between RCTs and NRSs were initially found, when other potential moderating or confounding variables were accounted for, differences were no longer significant.

Summary

In the three reviews in this section there were no consistent differences in effect size between NRSs and RCTs. Differences that were seen could be accounted for by other potential moderating variables.

Additional policy intervention reviews

Narrative reviews that summarised results separately by study design

Twenty-four systematic reviews contained both RCTs and NRSs and summarised results separately by study design, without using meta-analysis (Appendices 4.11 and 4.12).[108–128,158–161] Not all of the reviews intended to separate RCTs and NRSs; in four reviews it is clear that studies are reported individually because so few were found. Seven reviews stated a priori an intention to separate RCTs and NRSs. In other reviews the separation occurred either because the reviewers deemed it inappropriate to pool studies due to heterogeneity among participants, interventions, outcomes and study designs, or because RCTs are one of a number of study designs classified as 'good quality', or no rationale was given for separation of RCTs and NRSs.

In four reviews, intervention effects appeared stronger in NRSs than in RCTs. The interventions reviewed were: immunisation; health education; feedback and audit; and tobacco sales. Interventions were mostly educational or administrative/legislative procedures aimed at health professionals or shop owners.

In five reviews, intervention effects appeared stronger in RCTs than in NRSs. The interventions reviewed were: service delivery and organisation; early rehabilitation; payments for health professionals (two reviews); rehabilitation; and psychosocial interventions (two reviews). Interventions were broader and aimed at communities, patients or recipients of services rather than at health or other professionals.

This pattern, if it is a pattern, was not seen in reviews that met the inclusion criteria, perhaps because NRSs were more similar to each other and RCTs were more similar to each other, and this is why they were pooled together in meta-analyses, in the included reviews.

In the other 15 reviews in this section, RCTs and NRSs appeared to have similar effects or it was not possible to tell whether or not they were similar.

Meta-analyses that pooled different study designs

Nineteen reviews contained both RCTs and NRSs and pooled the results without separating by study design (Appendix 4.13).[129–147,162] None of the reviews gave a clear rationale for pooling the study designs together. In three reviews,[131,141,146] potential moderators of effect including randomisation were investigated after pooling. Randomisation was not a moderator of effect in one review[131] and, in the other two reviews,[141,146] the results of the investigation with regard to randomisation as a potential moderator of effect were not reported. In three reviews, only randomised and quasi- or pseudo-randomised study designs were included. When selecting reviews for inclusion in this evaluation, we considered studies described as quasi- or pseudo-randomised to be non-randomised. However, a pseudo-, quasi- or non-randomised controlled trial is not as different from a RCT as a non-randomised observational study would be.

Conclusions from additional policy intervention reviews

In reviews that discussed RCTs and NRSs separately but without meta-analysis, and in reviews that pooled both study designs together, it was unusual for review authors to explicitly state their rationale for doing so.

When a rationale was stated for not pooling, it usually related to RCTs being methodologically stronger study designs than NRSs, although other study designs could also be rated as 'strong' and combined with RCTs in some of these reviews (e.g. longitudinal designs in Reeves[125]).

When a rationale was not stated for pooling, it often seemed to be the case that other features of the included studies were expected to bring more heterogeneity to the results of the review than randomisation. Sometimes randomisation was investigated along with other study features as a potential moderator of effect.

Discussion

Inclusion criteria

This investigation of the effects of randomisation in evaluations of policy interventions was not as straightforward as the investigation by Deeks *et al.*[40] for evaluating NRSs in health care. In their investigation many of the included reviews specifically aimed to investigate differences between RCTs and NRSs. If we used similar inclusion criteria to such reviews in the field of policy interventions, we would only have included reviews that were already included in the report by Deeks *et al.*[40] Most of the reviews included in our investigation did not have a stated intention

of investigating differences between RCTs and NRSs. Most did not even have the stated intention of separating RCTs and NRSs in the analysis. Those that did have this intention did not always give a rationale for doing so. When a rationale was stated, it was either in order to separate more methodologically sound study designs (randomisation being one indicator of quality, but often not the only one), or to investigate potential moderators of effect. Potential moderators included randomisation and other features of study design, also aspects of the intervention, participants and outcomes measured.

Searches

Searching by study design is problematic even in MEDLINE: indexing of RCTs in MEDLINE by publication type and medical subject heading has improved in recent years but is still inadequate. Searching for NRSs is much more difficult; there are many study designs that could be classed as non-randomised and there is little definitive terminology. Comprehensive and consistent indexing according to study design is lacking. Databases beyond MEDLINE very often have poor indexing by study design, and these problems of definition become more pronounced in databases that are non-health related. Non-health databases in addition rarely have a thesaurus of keyword terms included; the records often lack an abstract and a number of databases only have rudimentary search capabilities. There are probably even fewer definitive terms available for 'systematic reviews' or equivalent in the non-health literature than there are for trial designs. Ideally, to ensure that the searches retrieved relevant references, the strategy would have included terms for 'review' or at least for 'literature review'. However, attempts at searching using these terms produced an unmanageable number of references.

Results judged similar

Even in the reviews for which authors judged results from RCTs and NRSs to be 'similar', pooled results of NRSs tended to be more positive than RCTs in two of five reviews and more negative in one. Heterogeneity was assessed separately by design in three of the five reviews in this section: there was greater heterogeneity among RCTs in one review and among NRSs in another. Other potential confounders or moderators of effect in reviews in this section were population variables (three reviews) and intervention variables (two reviews). The two reviews that carried out secondary analysis of moderators of effect including randomisation concluded that other potential sources of bias influenced the results of the review more than randomisation.

Results judged not similar

In the eight reviews for which authors judged results were 'not similar' between RCTs and NRSs, six found that NRSs had more positive results than RCTs and two found that RCTs had more positive results than NRSs. In one review it was unclear which was more positive. Only three reviews in this section assessed heterogeneity by design; one found more heterogeneity in NRSs and one found more heterogeneity in RCTs. The third found no significant heterogeneity in either group.

Other potential confounders or moderators of effect in reviews in this section were population variables (four reviews), intervention variables (three reviews) and study design/methodological variables (three reviews). No review found that random assignment had a strong effect on outcomes – population variables seemed to be more important.

Potential confounding variables in the three reviews in which results of RCTs compared with NRSs were judged to be 'mixed' included participant and methodological variables. Wilson *et al.*[97] found that other methodological variables were more likely than random assignment to influence outcome.

One possible reason for other variables influencing outcomes more than study design could be that in the reviews we found, randomised and non-randomised study designs have been used in different types of populations/settings/interventions (i.e. the review authors have not set inclusion criteria restricting these other variables). In theory, if a randomised design is chosen, potential confounding variables should be distributed evenly between groups. In reality we cannot confirm this because we do not have reviews in which randomised and non-randomised designs have been applied to the same population/intervention/setting, so we cannot compare them. There is too much heterogeneity between the included studies to isolate the effect of randomisation. This reflects the broad nature of many systematic reviews of policy interventions compared with reviews of more tightly defined health-care interventions (e.g. pharmacological interventions). And so, while within a review it might appear that other potential moderators of effect have a stronger effect than

randomisation, within a single RCT this would hopefully not be the case. We cannot confirm or refute this based on the work we have done.

The results of this investigation show us that the 'state of the art' in terms of systematic reviews of policy interventions does not yet answer the question of whether RCTs and NRSs are of similar validity in evaluating policy interventions. While it can be argued that RCTs can sometimes be difficult and/or unethical to conduct in certain settings, and that results are not always generalisable to 'real life',[15,16] it can also be argued that NRSs can be subject to so many biases as to make it doubtful whether it is useful to include them in systematic reviews at all.[148] It seems clear that further investigation should be carried out in the form of properly conducted systematic reviews of policy interventions that include both RCTs and NRSs with the pre-stated objective of investigating differences between them. Methodological studies need to be indexed much more comprehensively in electronic databases.

Criteria used to judge equivalence of RCTs and NRSs

As in the study by Deeks *et al.*,[40] the manner in which results were judged to be equivalent between study designs varied between the reviews (*Table 7*). In the majority (13 out of 16) of the included reviews, the criteria used to judge the equivalence of the results were not described. None of the five reviews in which the authors judged the results of RCTs and NRSs to be similar described how it reached such a judgement.

Two of the eight reviews in which the authors judged the results of RCTs and NRSs to be

TABLE 7 *Equivalence criteria used in reviews of RCTs and NRSs*

Review	Equivalence criteria
Results judged similar	
Cameron, 2000[99]	None
Kwan, 2004[101]	None
Langhorne, 1999[102]	None
Tobler, 2000[107]	Not stated
Wilson, 2000[105]	Not stated
Results judged not similar	
Cambach, 1999[98]	None
Davis, 2000[90]	None
Griffith, 2000[100]	Regression analysis examined effects of moderator variables on effect size, giving an estimate of between-groups variance (*Qb*). 95% CIs were calculated; non-overlapping CIs allowed conclusions to be drawn about the strength of one predictor in comparison with another
Jacobs, 2002[92]	None
Mullen, 2002[93]	The contribution of grouping variables to variation in the effect size estimates were examined using (1) the chi-squared statistic with Bonferroni correction (to assess likelihood of differences between subgroups) and (2) between-group heterogeneity (*Qb*), to assess the magnitude of any effect. A substantial contribution by a moderator variable to the overall heterogeneity was defined as ≥5%
Oliver, 2000[103]	None
Smedslund, 2004[104]	Not stated
Wilson, 2001[96]	Not stated
Results judged mixed	
Guyatt, 2000[91]	A *z*-score was used to generate a *p*-value related to the null hypothesis that there were no real differences in results from observational studies and randomised trials
Thomas, 2003[94]	Not stated
Wilson, 2003[97]	Not stated

dissimilar described their criteria for equivalence. One[93] began by using the chi-squared statistic to identify potential moderating variables. Both calculated a measure of variance between groups (*Qb*) which was used to estimate the magnitude (or strength of contribution) of the potential moderator variables to the overall effect size. In one of the reviews[100] 95% CIs were then calculated for each potential moderator variable and non-overlapping CIs were taken to indicate relative strength of moderator or predictor variables. In the other review[93] a contribution of 5% or more by a proposed moderator variable to the overall heterogeneity surrounding the effect size was defined as a substantial contribution.

One of the three reviews in which results were judged to be mixed defined the criteria used to judge equivalence between the groups.[91] This consisted of a *p*-value generated from a *z*-score based on the null hypothesis that there were no differences between results of RCTs and NRSs. This technique, while being less subjective than simply using authors' judgement, does not examine the relative contribution of potential moderator variables other than randomisation to the overall effect size.

Given these findings it seems important to note that sensible and objective criteria to judge equivalence or otherwise of results of RCTs and NRSs should be included and applied in systematic reviews that include both study designs. These should be explicitly defined in the review protocol and, where possible, should use methods that take the effects of potential moderating variables other than randomisation into account.

Conclusions

Considerable variation in the studies pooled within reviews, in terms of population, intervention, outcome and other methodological details, makes it difficult to separate the potential effect of random assignment from the potential effects of all the other variables.

Not only should the magnitude of the pooled effect estimates be compared between RCTs and NRSs, but also the variability associated with them. However, most included reviews did not do this. Most did not state what criteria were used to judge equivalence between findings of RCTs and NRSs.

The existing systematic reviews of policy interventions do not help us to determine whether RCTs and NRSs give similar results when evaluating policy interventions. Further research should be carried out (see below).

Recommendations for research

Systematic reviews should be carried out with the intention of investigating differences in effects of policy interventions between RCTs and NRSs. Sensible and objective criteria that are supported by empirical evidence should be used to judge equivalence or otherwise of results of RCTs and NRSs in these investigations. Not only should the magnitude of the pooled effect estimates be compared between RCTs and NRSs, but also the variability associated with them. Methodological indexing terms should be developed to enable more fruitful searching of health and non-health electronic databases.

Chapter 6

Methods for testing the hypotheses developed in Chapters 3–5

Aims

The aims of this part of the study are: (1) to test our main hypothesis that RCTs produce different results when compared with other study designs and (2) to test whether this finding can be explained by the hypotheses developed in Chapters 3–5. These hypotheses outline possible relationships between various factors which might be associated with the use of randomisation and/or the effect size of evaluations. These factors may therefore explain, confound, strengthen or weaken the conclusions drawn from (1).

We adopted three of the four possible approaches mentioned earlier (see Chapter 2):

- Comparing controlled trials that are identical in all respects other than the use of randomisation, by 'breaking' the randomisation in a trial to create non-randomised trials. This approach uses original primary data from two RCTs of policy interventions in resampling studies.
- Comparing similar controlled trials drawn from systematic reviews that include both randomised and non-randomised studies (i.e. analysing comparable field studies). This approach uses a series of systematic reviews of health promotion interventions conducted by the EPPI-Centre (all EPPI-Centre health promotion reviews available at the time of the current study).
- Investigating associations between randomisation and effect size using a pool of more diverse studies within broadly similar areas. This meta-epidemiological approach uses the pooled data from the EPPI-Centre reviews mentioned above, and data from trials of interventions to support transition from school into adult life reviewed by Colorado State University.

As the methods of analyses for the second and third approach overlap, descriptions of methods and results are combined as a single meta-epidemiological study.

Creating NRSs from RCT data

This part of the study builds on work conducted by Deeks *et al.*[40] and explores the difference between randomised and non-randomised trials in a tightly controlled way through the use of statistics and primary data from RCTs. By generating non-randomised trials from within an RCT we were able to look at differences between randomised and non-randomised trials in a population of studies that only differed by their method of allocation. Thus, in effect, we had a set of trials that we knew to be free of all the confounders that Chapters 3–5 predict might moderate and mediate observed differences between RCTs and nRCTs.

We were fortunate to be given data from two RCTs (see Chapter 2). One evaluated postnatal support and the other evaluated the prevention of child physical abuse and neglect. These data preserved the anonymity of the trial participants, but contained baseline and outcome information for all individuals included in the original trial analyses.

Creating randomised and non-randomised trials

In order to replicate, as far as possible, circumstances that might lead to the creation of non-randomised trials by researchers working 'in the field', we created non-randomised trials based on the area in which participants lived. Each area had a number of participants who received the intervention, and a number who did not. Each area could therefore be considered to be a mini-RCT. We had six such areas in Trial 1 and four in Trial 2 (after combining two small areas). We were then able to create non-randomised comparisons by comparing the people who received the intervention in one area with people who did not in other areas. Thus we were able to create 30 non-randomised trials from Trial 1 and 12 from Trial 2. We used all individuals in the selected areas, in contrast to Deeks *et al.*[40] who drew randomised samples in the selected areas, because our RCTs had far fewer participants than those of Deeks *et al.*

It is important to note that the average intervention effect in the mini-NRSs must equal the average intervention effect in the mini-RCTs. Our focus will therefore be on comparing the standard deviations of the intervention effects in mini-NRSs and in mini-RCTs. Our design does not allow for the possibility that real NRSs might induce bias by (consciously or not) assigning interventions to 'more promising' areas.

Methods for analysis

We had three questions to answer in our re-analysis of the two trials:

1. Do the results of the non-randomised trials differ from those of the randomised trials?
2. Does matching areas on baseline characteristics enable non-randomised trials to approximate the results of the randomised trials?
3. Can adjusting for baseline characteristics in the analysis enable non-randomised trials to approximate the results of the randomised trials?

For the first of the above analyses, we calculated the log odds ratios of the outcomes of interest in all randomised and non-randomised trials. In Trial 1 we had 30 non-randomised and six randomised trials. In Trial 2 we had 12 non-randomised and four randomised trials.

In the second analysis we matched the intervention areas with control areas on baseline characteristics of the study participants. This gave us six non-randomised trials from Trial 1 and four non-randomised trials from Trial 2.

In the third analysis we took each possible comparison to create non-randomised trials, and adjusted for baseline differences in the analysis using logistic regression.

To answer our research questions in each of the above analyses, we tested for differences in the variances of the randomised and non-randomised trials using an *F*-test. Because this test wrongly assumes that all mini-NRSs are independent, the *p*-values produced are likely to be too small.

The results of the first re-analysis can be derived algebraically (see Appendix 5). A key quantity turns out to be a correlation coefficient r, derived by computing the observed log odds in each arm in each area, and forming the correlation between the log odds in the intervention arm and the log odds in the control arm. In the appendix we show that the standard deviation of the NRSs is greater than the standard deviation of the RCTs whenever r is greater than 0. It follows that we can test whether the standard deviation of the NRSs equals the standard deviation of the RCTs by testing whether $r = 0$.

Methods for analysing comparable field studies and meta-epidemiology

The aim of this part of the study was, first, to investigate whether study design influences a study's effect sizes by analysing 'comparable policy evaluations' (i.e. evaluations of similar policies) from sets of studies with randomised and non-randomised study designs selected from systematic reviews of policy interventions. The second aim was to test whether these findings can be explained by the hypotheses developed in Chapters 3–5, which were based on the findings of existing systematic reviews. This part of the study is based on examining differences in the effect sizes of individual studies across nine health promotion reviews.

Identification and description of policy intervention evaluations within reviews

Predetermined inclusion criteria and descriptive codes were applied to studies previously reviewed in depth. Reviewers inspected abstracts (where these were available) and previous coding for each study. Previous coding of the EPPI-Centre data set described outcome evaluations according to a standardised keyword system developed by the EPPI-Centre[149] covering the type of study (e.g. outcome evaluation, survey, case–control study); the country where the study was carried out; the health focus of the study; the study population; and, for reports describing or evaluating interventions, the intervention site, intervention provider and intervention type. (These studies were also sometimes further classified with review-specific codes.) In addition, extracted data described in detail the population, development and delivery of the intervention, research design and methodological attributes, and the type of outcomes measured (when relevant). Where necessary, reviewers referred to the full reports of these evaluations. Standardised coding and extracted data were also available for inspection as part of the Colorado data set.

The inclusion criteria distinguished policy intervention evaluations from evaluations of other kinds of intervention. Descriptive codes for this study classified evaluations according to our typology of policy interventions (setting of policy/strategies, legislation/regulation, provision/organisation of services, environmental modification, facilitating lay/public delivered support/education); the presence or absence of an explicit collective plan of action; the level at which policy was implemented (international, national, regional, community or institution); the attrition rate; and evaluation designs (RCT or other evaluation design) based on Deeks *et al.*'s[40] coding framework.

Initially, reviewers worked separately on a subset of studies in the reviews so as to quickly assess the likely availability of policy intervention evaluations and to see how easy/useful it was to apply the inclusion criteria and descriptive codes. This subset was chosen so that it spanned a range of social and organisational settings. The reviewers' independent responses were compared, discrepancies discussed, and amendments made relating either to descriptions of individual studies or to the definitions of the terms describing policy interventions. Ultimately, screening and coding for each outcome evaluations in the reviews was carried out independently by two reviewers, with discrepancies resolved through consensus.

The results were tabulated to describe the balance of policy intervention evaluations and other intervention evaluations found in each review, and the type of policy interventions (setting of policy/strategies, legislation/regulation, provision/organisation of services, environmental modification, facilitating lay/public delivered support/education), the level at which they operate (national, regional, community or institutional), and whether they were evaluated with an RCT or another design.

Analysis of each study

EPPI-Centre reviews

We calculated measures of effect for all studies. Given that many of the outcomes used different scales and different combinations of continuous and dichotomous data, we selected the SMD as being the only measure that would enable us to compare and combine results.[150] Our software, EPPI-Reviewer, can calculate this quantity. To accommodate the inconsistent and incomplete reporting of quantitative data from controlled trials, EPPI-Reviewer has been adapted to calculate measures of effect from a minimum of available data. EPPI-Reviewer can also compute appropriate measures of effect and standard errors from cluster-randomised trials comparing groups of individuals (e.g. classes or schools).

Outcomes were classified as being in one of four 'outcome domains': knowledge, attitudes, behaviour and health state. For studies that reported more than one outcome per domain, we included in our analysis only the outcome that was most commonly reported across all studies in that review. Thus each study had up to four outcomes calculated, though many did not report outcomes in all our domains.

Many authors did not report enough information to calculate an effect size. Some simply reported that 'there were no significant differences between the groups' and did not supply numeric information. In these cases we assumed a SMD of zero; the standard error was calculated because for the SMD it depends only on the sample sizes. A sensitivity analysis was conducted to examine the impact this had on our results by comparing the results obtained with and without the studies in question. Only 12 outcomes out of 376 fell into this category, and the sensitivity analysis revealed no differences in results as a result of this.

The intracluster correlation coefficient (ICC) for cluster trials was often not reported. We assumed an ICC of 0.02 for these trials and adjusted standard errors accordingly by inflating them by the square root of the design effect. All reported ICCs were in the range of 0.01 and 0.02, as were the ICCs we were able to calculate from primary data, so we felt confident in assuming this value. We were able to calculate an ICC for one sexual health study (0.01), and this value was assumed for the other trials in that review.

Colorado reviews

All the studies in the Colorado data set that met our inclusion criteria already had SMDs calculated. Unlike outcomes in the EPPI-Centre data set, only one effect size had been calculated for each study, and these outcomes were not split into different domains and are mostly concerned with social and education skills such as comprehension and communication.

Analysis combining studies: potential confounders

Any differences between the effect sizes of randomised and non-randomised studies may be explained by other variables that are also related to effect size. For instance, 'hard outcomes' provided by clinical data may be more easily obtained in trials set in clinical establishments where randomisation is also more acceptable to the community of researchers and practitioners. In contrast, 'soft outcomes' such as self-reported behaviour may be more optimistic and more commonly relied upon on in community settings where randomisation is less acceptable to the practitioners and researchers.

The first stage of analysis was therefore to test for associations between randomisation and any attributes of policy interventions or their evaluations where a theoretical argument may be mounted, or for associations that have been shown in other empirical studies, including results reported in Chapters 3–5. In order to explore these associations between randomisation and other attributes, we cross-tabulated the attribute of interest against study design and tested for statistical relationships using the chi-squared test. (Statistical tests for this analysis were carried out in EPPI-Reviewer.) For some dimensions, e.g. intervention provider, the studies could have more than one attribute. In order to avoid recounting any studies that had more than one attribute in the chi-squared test, the 'count' for each study was calculated as one divided by the number of times the study appeared in the test. So, for a study with two different intervention providers, the value of the study was 0.5 in the two cells in which it appeared.

Analysis combining studies: comparing effect sizes of randomised and non-randomised studies

The second stage of analysis investigated the differences between the observed effect sizes of randomised and non-randomised studies within the same systematic review. Because we had data from several systematic reviews, we also wanted to investigate whether or not any differences between the observed effect sizes of randomised and non-randomised studies were consistent across systematic reviews. The analysis was based on the estimated effect size for each study in each outcome domain and allowed for random error in those estimated effect sizes, as expressed by the standard errors. We have controlled for the variation that might be introduced by longer term follow-up by always selecting the outcomes measured as soon as possible after the intervention.

Model for one review

Our model for the estimated intervention effects incorporates the following features:

- the overall intervention effect may be different in each review
- NRSs and RCTs may differ systematically
- the systematic difference between NRSs and RCTs may be different in each review
- random error in each estimated intervention effect is captured by its standard error, but there may be additional heterogeneity between intervention effects
- this additional heterogeneity between intervention effects may differ in magnitude between RCTs and NRSs.

Formally, our model is:

$$y_{ij} = \delta_i + b_i t_{ij} + u_{ij} + e_{ij} \quad (1)$$

where y_{ij} is the estimated intervention effect in the jth study in the ith review; δ_i is the average true intervention effect in the ith review; b_i is the average difference between RCTs and NRSs in the ith review – the 'bias term'; t_{ij} is 0 if the jth study in the ith review is a RCT and 1 if it is an nRCT; u_{ij} is a study-specific random effect that has mean 0 and standard deviation σ_{iR} if the study is a RCT and σ_{iN} if the study is an NRS; and e_{ij} is random error that has mean 0 and standard deviation equal to the calculated standard error s_{ij}.

Model (1) was fitted separately for each review and for each outcome domain (in the EPPI-Centre reviews). Estimation was carried out in two ways. Firstly, we fitted separate random-effects meta-analysis models to the RCTs and to the NRSs, and we estimated b_i as the difference in estimated intervention effects, with squared standard error equal to the sum of the squared standard errors of the separate estimated intervention effects. Secondly, we used meta-regression,[150] which additionally assumed equal variances ($\sigma_{iR} = \sigma_{iN}$). We also explored whether the data were consistent with equality of variances using a likelihood ratio test between a single meta-regression model and a pair of meta-regression models.

In this model, our main interest is in whether b_i is zero or not. However, there is typically substantial

uncertainty about a single b_i, and it is therefore useful to combine the b_i in a second stage model.

Model for all reviews

In the second stage model, the bias term b_i is assumed to follow a normal distribution across reviews:

$$b_i \sim N(\beta, \Phi^2) \quad (2)$$

where β expresses the average bias of nRCTs (presupposes common direction of estimated intervention effects) and Φ expresses review-specific bias.

If β and Φ are both zero then NRSs do not differ systematically from RCTs in any policy area. If β is non-zero but Φ is zero then NRSs differ systematically from RCTs and the difference is *the same* across different policy areas. If Φ is non-zero then NRSs differ systematically from RCTs and the difference is *different* across different policy areas (so that the difference is likely to be small or zero in some policy areas but not in others).

Model (2) was fitted separately for each outcome domain (in the EPPI-Centre reviews). The model was fitted by applying a standard random-effects meta-analysis model to the estimates of b_i and its standard error from fitting Model (1).

To explore which other study factors are associated with estimated intervention effects, we repeated the above analysis with t_{ij} redefined as each other study factor in turn.

Meta-confounding

As different study characteristics were likely to be correlated, we estimated independent effects by combining statistically significant ($p < 0.05$) variables from the above univariate analysis in a multivariate meta-regression.[150] These variables covered public involvement, settings and boundaries of the intervention, intervention provider (clinician) and reporting quality. This analysis allowed for us to investigate the possibility of meta-confounding while keeping the number of tests to a reasonable number in order to minimise the chances of false positive results. The aim of this analysis was to discover whether these other factors strengthen, weaken or otherwise change the result of testing our main hypothesis. We acknowledge, however, that this approach risks excluding potentially relevant interactions as it was based on a subset of the possible range of variables that could have been included. All statistical tests in this analysis were carried out using the metareg command in STATA.

Data

Given that some studies appeared in more than one review (for example, studies concerned with children and physical activity sometimes also had a component on healthy eating) and there was a danger that some studies could appear in the analysis twice, we organised the reviews into chronological order and excluded studies from reviews if they had already appeared in an earlier review. *Table 8* reports the number of policy interventions in each review and the number of non-overlapping policy interventions included in the analysis, and *Tables 9–21* show characteristics of the studies in the EPPI-Centre reviews. A sensitivity analysis that included all studies in all reviews showed us that, even though the number of studies appearing in some reviews was greatly diminished, our results were unaffected by this decision. There was no overlap between the studies in the EPPI-Centre reviews and the Colorado data set, so no action needed to be taken between data sets.

TABLE 8 *Number of studies analysed across whole set of EPPI-Centre reviews*

Review	Total number of policy interventions in this review	Number in this analysis
1. Workplace health promotion[152]	46	46
2. Peer-delivered health promotion[153]	47	47
3. Preventing cervical cancer[154]	29	26
4. Young people: physical activity[28]	13	12
5. Young people: healthy eating[27]	22	9
6. Young people: mental health[155]	4	4
7. Children: physical activity[156]	19	16
8. Children: healthy eating[94]	30	26
9. Men who have sex with men[157]	10	10

TABLE 9 *Types of study (n = 176)*

Type of study	Number
RCT	97
nRCT	79

TABLE 10 *Whether or not interventions were based on an explicit theoretical model*

	No	Yes	Total
RCT	27	70	97
nRCT	41	38	79
Total	68	108	176

TABLE 11 *Whether interventions were based on explicit public involvement*

	Explicit public involvement	No explicit public involvement/not stated	Total
RCT	12	85	97
nRCT	14	65	79
Total	26	150	176

TABLE 12 *Identification of aims*

	Aims indentified by target population	Aims: not stated/unclear/other than target population	Total
RCT	4	93	97
nRCT	4	75	79
Total	8	168	176

TABLE 13 *Were lay people involved in developing the intervention?*

	Yes	No	Total
RCT	24	73	97
nRCT	26	53	79
Total	50	126	176

TABLE 14 *Intervention site (not mutually exclusive categories)*

	Community	Institution	Total
RCT	31	82	113
nRCT	24	68	92
Total	55	150	205

TABLE 15 *Intervention provider (not mutually exclusive categories)*

	Community	Lay	Researcher	Practitioner	Total
RCT	12	44	8	58	122
nRCT	14	39	11	51	115
Total	26	83	19	109	237

TABLE 16 *Choice of measurement tool (not mutually exclusive categories)*

	Clinical test	Non-clinical test	Total
RCT	24	93	117
nRCT	10	79	89
Total	34	172	206

TABLE 17 *Choice of outcome measures (not mutually exclusive categories)*

	Clinical risk factor/health problem or state	Other outcome	Total
RCT	26	95	121
nRCT	14	79	93
Total	40	174	214

TABLE 18 *Intervention provider (not mutually exclusive categories)*

	RCT	nRCT	Total
Not stated	11	10	21
Unclear	6	1	7
Not relevant (e.g. mass media)	1	2	3
Community	2	5	7
Community worker	6	5	11
Counsellor	3	2	5
Health professional (specify)	16	12	28
Health promotion/education practitioner	16	7	23
Lay therapist	0	1	1
Parent	7	7	14
Peer (specify)	33	30	63
Psychologist	4	1	5
Researcher	7	9	16
Residential worker	0	0	0
Social worker	1	3	4
Teacher/lecturer	34	33	67
Other (specify)	14	14	28
Total	161	142	303

TABLE 19 *Was the allocation to intervention and control/comparison groups performed blind?*

	RCT	nRCT	Total
Not relevant (study not a trial)	0	0	0
Not stated	88	39	127
Unclear (please specify)	3	3	6
Yes	5	1	6
No	1	36	37
Total	97	79	176

TABLE 20 *Were participants aware which group they were in for the evaluation?*

	RCT	nRCT	Total
Not relevant (study not a trial)	0	0	0
Not stated	76	51	127
Unclear	9	15	24
Yes	10	11	21
No	3	2	5
Total	98	79	177

TABLE 21 *Was outcome measurement performed blind?*

	RCT	nRCT	Total
Not relevant (study not a trial)	0	0	0
Not stated	84	59	143
Unclear	3	11	14
Yes	10	1	11
No	2	8	10
Total	99	79	178

Chapter 7

Results: testing our main hypothesis that RCTs are the same as NRSs

Hypothesised relationships between randomisation, effect size and potential moderators, or confounders were translated into null hypotheses for empirical testing. This chapter reports the results of testing our principal null hypothesis that, based on previous research,[40] there is no detectable difference in effect size between RCTs and NRSs, but that the variance of NRSs is greater than that for RCTs.

This hypothesis is tested with data from the two reconstructed RCTs, the nine EPPI-Centre reviews (separately and pooled) and the Colorado data set of policy evaluations reviewed in depth.

Results from creating randomised and non-randomised trials from two RCTs

Using the Social Support and Family Health Trial (Trial 1),[47] we created six RCTs, based on the participants in six areas, and 30 nRCTs (by comparing the intervention groups in each area with the control groups from every other area). There were 731 participants in this trial: 367 in the intervention group and 364 in the control group. The same technique in Trial 2 gave us 12 nRCTs and four RCTs. There were 160 participants in this trial: 88 in the intention group and 72 in the control group.

The three main outcomes in Trial 1 were smoking, depression [Edinburgh Postnatal Depression Score (EPDS) score 12] and whether or not the child had had an accident in the last year. All were binary outcomes, so we were able to use the same methods for all three. We calculated odds ratios, comparing the odds, for example, of smoking in the intervention group with the odds of smoking in the control group. We then plotted the results obtained from the RCTs and nRCTs on dotplots *(Figure 2*).

We then compared the variances of the RCTs and nRCTs. As can be seen from *Table 22*, the variances of the log odds ratios were also very similar between the different types of studies. Using the STATA sdtest, we found that none of the differences were statistically significant with p-values ranging from 0.6807 for smoking to 0.9998 for maternal depression. We therefore do not have any evidence to support our hypothesis that the variance of the effect sizes of nRCTs can be expected to be greater than that for RCTs.

The standard deviations are surprisingly similar for smoking, EPDS and accidents. This is explained by the small values of the correlations between intervention and control arm log odds in different areas, which are 0.32, 0.13 and 0.32 respectively (see the formulae in Appendix 5).

Analysis of the primary outcomes in Trial 2,[48] 'neglect' and 'physical abuse', yielded similar results (*Figure 3*). The tests for variance were also not significant ($p = 0.4$ and $p = 0.5$).

We also tested two further ways of comparing randomised with non-randomised trials: matching areas and adjusting on baseline variables. We matched each of the intervention groups in the six areas in the Social Support Study against a control group from another area on three variables: lone parenthood, type of housing and ethnic group.

Areas in the home visitation trial (Trial 2)[48] were matched on a measure of deprivation developed for the trial. *Figure 4* shows the dotplot for this analysis. As before, there were no significant differences in variance using the STATA sdtest with p-values of 0.89 (smoking), 0.85 (EPDS) and 0.97 (accidents). *Figure 5* shows the dotplot for this analysis in Trial 2. For the outcome 'physical abuse' there was no statistical difference between the two types of studies ($p = 0.40$). However, for neglect, nRCTs showed significantly smaller log odds ratios than the RCTs ($p = 0.018$).

Our final analysis consisted of using logistic regression to adjust the results of the trials according to the same baseline characteristics on which we used to match areas in the previous

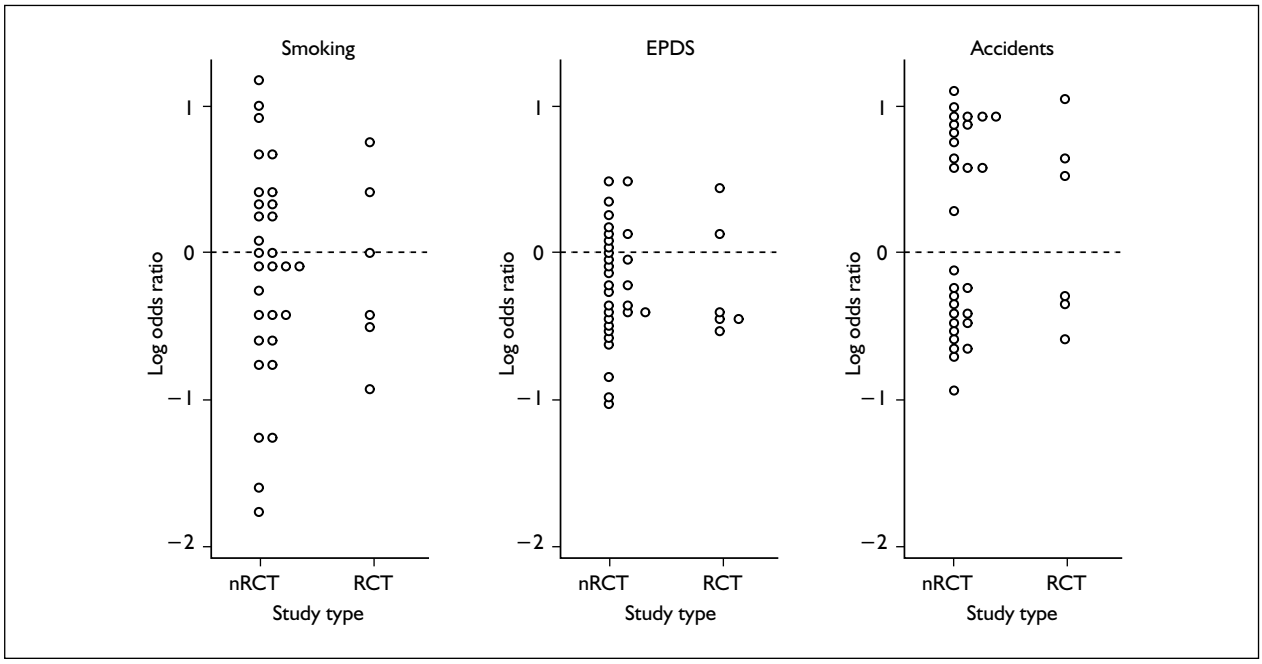

FIGURE 2 *Trial 1 dotplots for the three main outcomes. EPDS, Edinburgh Postnatal Depression Score.*

TABLE 22 *Standard deviations of log odds ratio for different outcomes in Trial 1*

Trial type	Smoking	EPDS	Accidents
RCT	0.632	0.397	0.663
nRCT	0.732	0.397	0.674
Test of equality	$p = 0.53$	$p = 0.81$	$p = 0.53$

EPDS, Edinburgh Postnatal Depression Score.

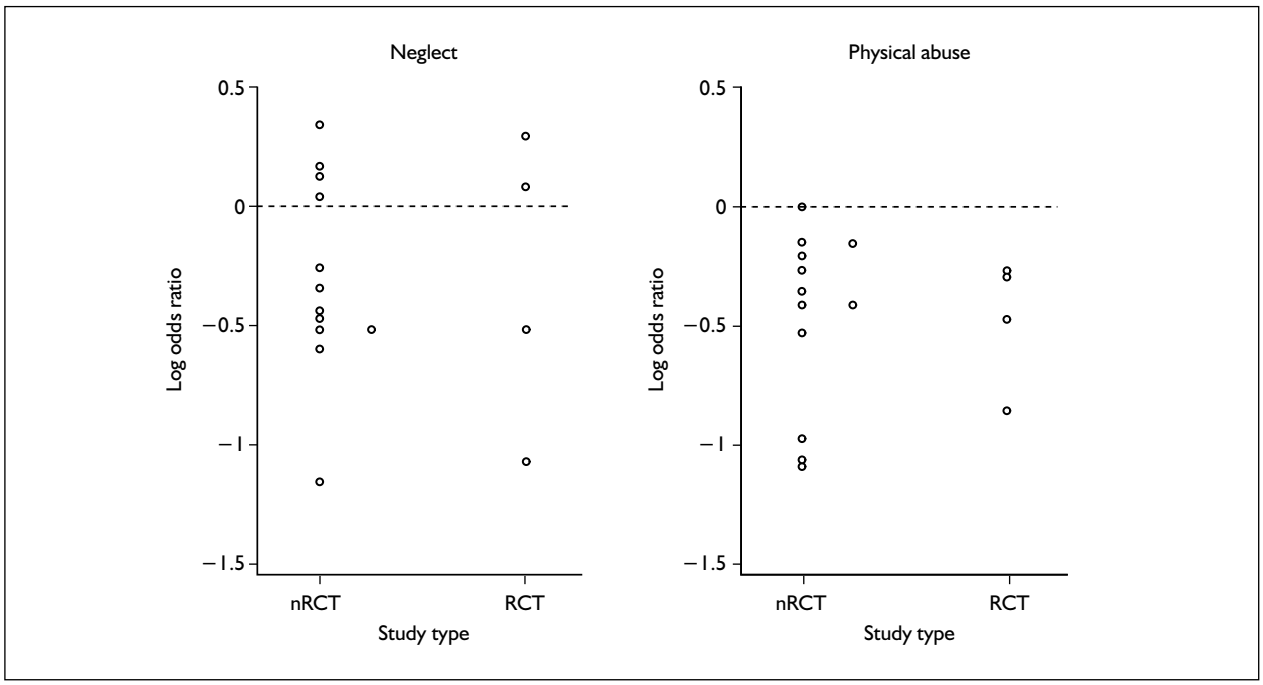

FIGURE 3 *Trial 2 dotplots for the two main outcomes.*

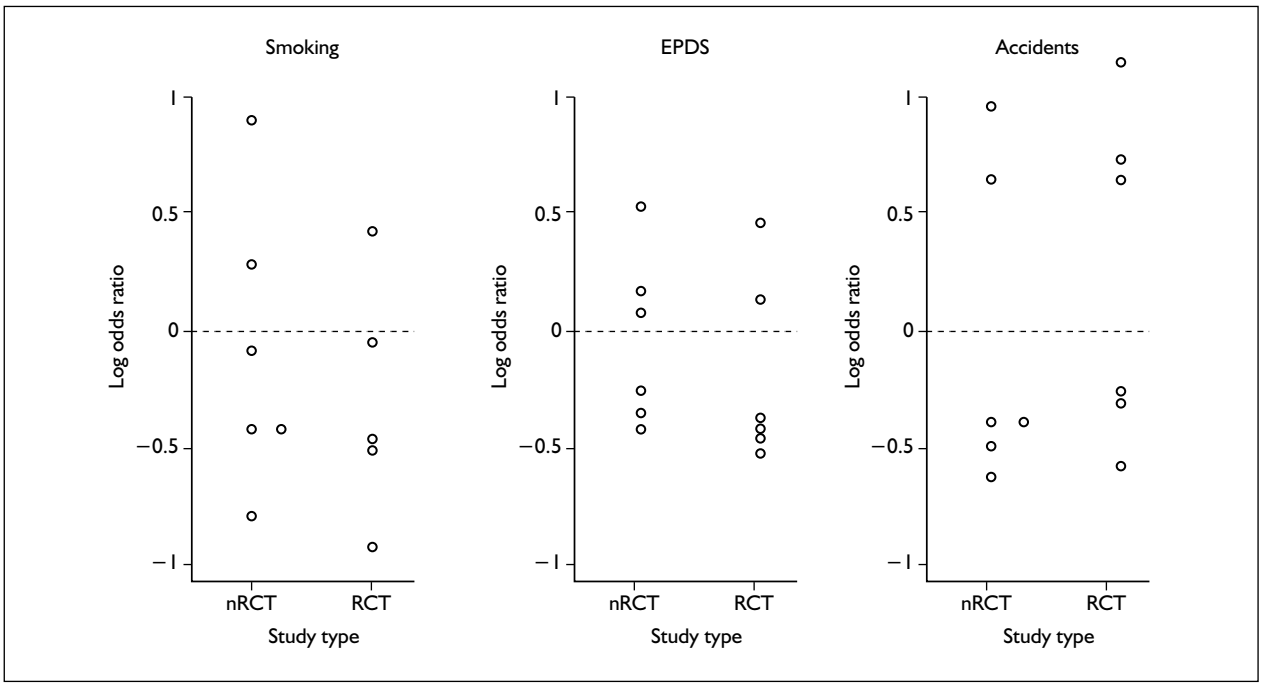

FIGURE 4 *Trial 1 dotplots for the three main outcomes. EPDS, Edinburgh Postnatal Depression Score.*

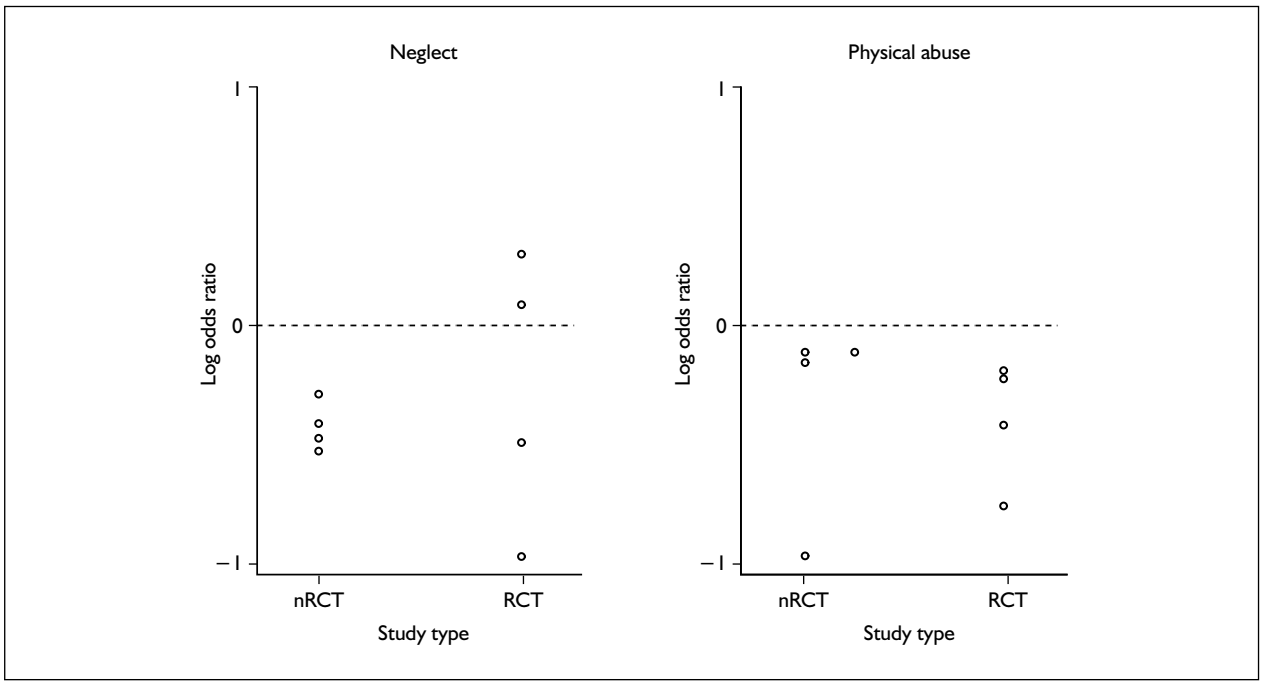

FIGURE 5 *Trial 2 dotplots for the two main outcomes.*

analysis. We used the same methods as described above to generate non-randomised comparisons and compared 30 non-randomised trials with six RCTs in Trial 1 and 12 non-randomised trials with four RCTs in Trial 2.

Dotplots for this analysis in Trial 1 are shown in *Figure 6* and for Trial 2 in *Figure 7*. Again, tests for differences in variance showed no significant differences between the two types of trial.

Results from the EPPI-Centre reviews

Presented here are the findings that resulted from examining the data in the EPPI-Centre reviews to answer our main research question: do RCTs have the same effect sizes as nRCTs? As Deeks *et al.*[40] found, the answer to this question has two aspects: (1) the overall average size of effect and (2) the range of different effect sizes covered by the different study types (their variance). As one of the inclusion criteria for the EPPI-Centre reviews was the presence of a control group, the comparison made here is between RCTs and nRCTs.

Because variance is an important part of our question, we fitted the model described in Chapter 6 separately for RCTs and nRCTs. First, we conducted a random-effects meta-analysis (STATA: metan) separately for the RCTs and nRCTs in each review in each outcome domain (knowledge, attitudes, behaviour, health state). This gave us two overall effect sizes and standard errors for each review in each outcome domain. We then calculated the bias term b as the difference between the RCTs and nRCTs in each review with a standard error calculated as $se = \sqrt{se_1^2 + se_2^2}$ where se_1 and se_2 are the standard errors of the overall effect sizes for the RCTs and nRCTs. This gave us a bias term (the difference between the effect sizes of RCTs and nRCTs) and a standard error in each outcome domain for each review. The final stage in this analysis was to combine the bias terms for each review in a random-effects meta-analysis (STATA: metan). The direction of effect from this analysis tells us if, overall, RCTs have larger or smaller effect sizes than nRCTs, and each point on the forest plots below represents the results from one review.

The pooled effect size of –0.28 (95% CI 0.64 to 0.09) indicates that the nRCTs have bigger effect sizes than the RCTs, but this result is not statistically significant ($p = 0.14$). However, there is a high degree of heterogeneity between the reviews [$Q = 24.86$, degrees of freedom (df) = 7, $p < 0.001$]. Taken together, these results suggest that the nRCTs have bigger effect sizes than the RCTs in

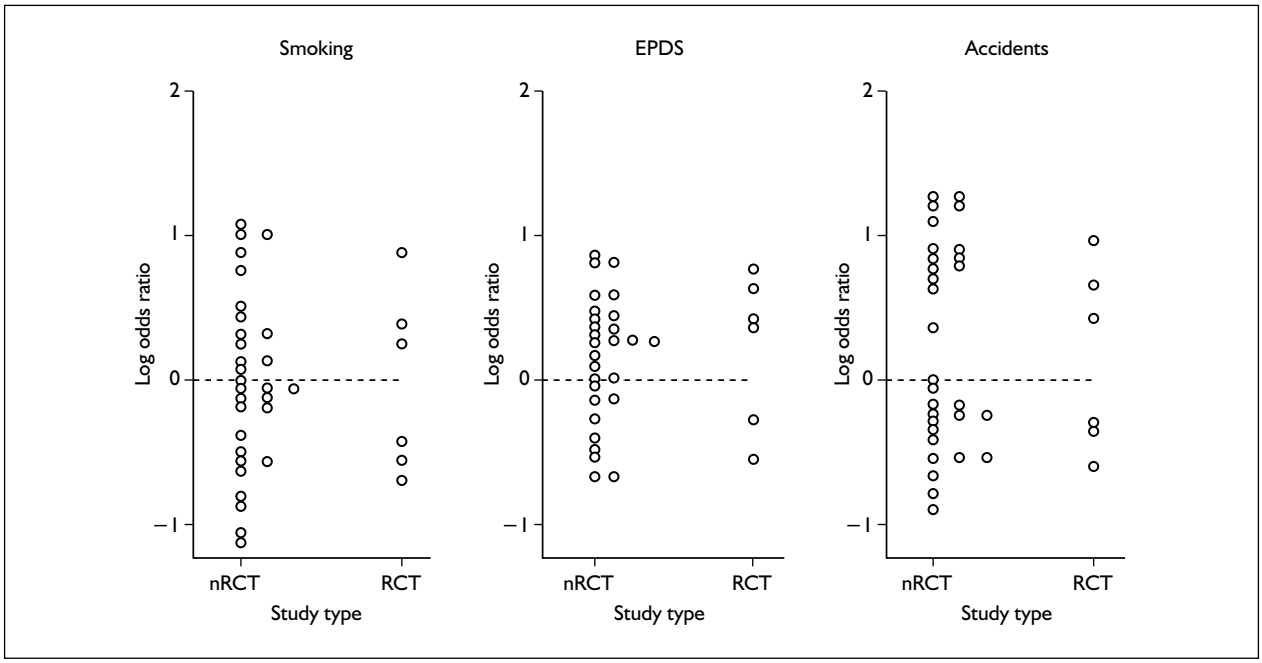

FIGURE 6 *Trial 1 dotplots for the three main outcomes. EPDS, Edinburgh Postnatal Depression Score.*

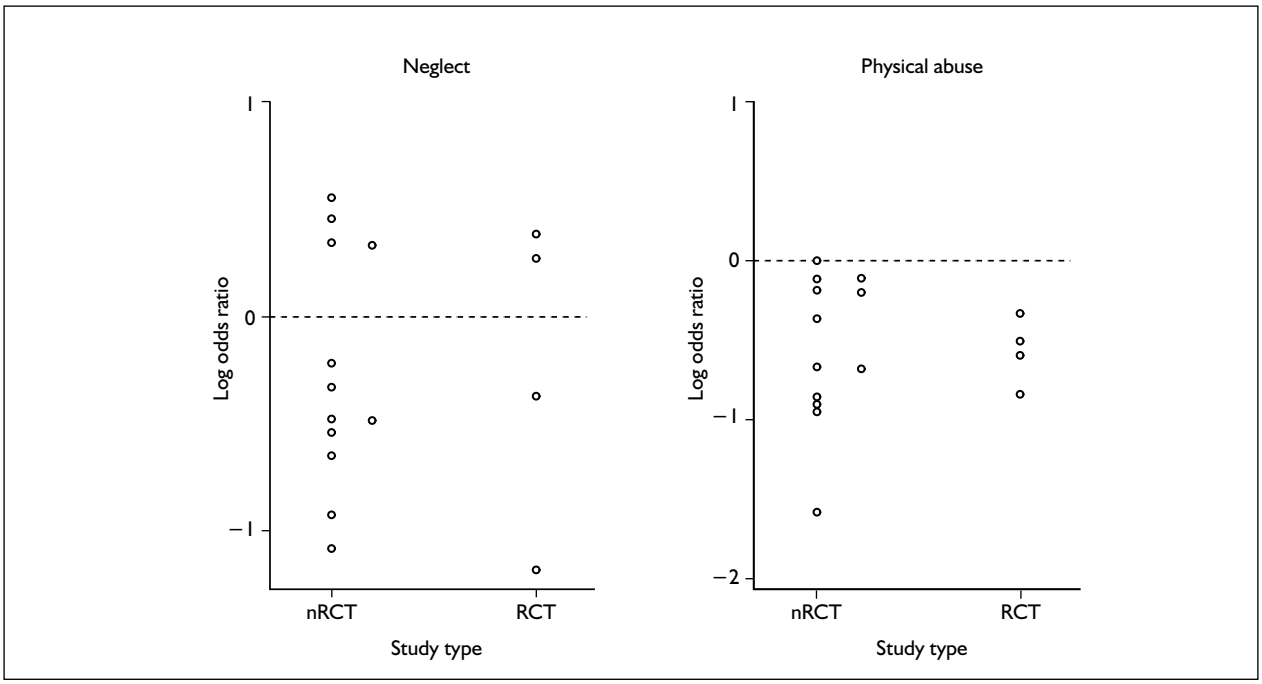

FIGURE 7 *Trial 2 dotplots for the three main outcomes.*

some reviews (e.g. reviews 3[154] and 6[155]) but not in others, and they may even have smaller effect sizes than the RCTs in some reviews (e.g. review 7[156]).

The pooled effect size of –0.166 (95% CI –0.319 to 0.012) indicates that the nRCTs have bigger effect sizes than the RCTs, and this result is statistically significant ($p = 0.034$). The results in this outcome domain are more homogeneous than knowledge ($Q = 6.45$, df = 6, $p = 0.37$).

The pooled effect size of –0.111 (95% CI –0.199 to –0.023) is statistically significant, indicating that for behaviour nRCTs have bigger effect sizes than RCTs. The results are also homogeneous in this outcome domain ($Q = 5.46$, df = 7, $p = 0.60$).

The pooled effect size for health state is not statistically significant –0.084 (95% CI –0.234 to 0.066) and indicates a very slightly larger value for

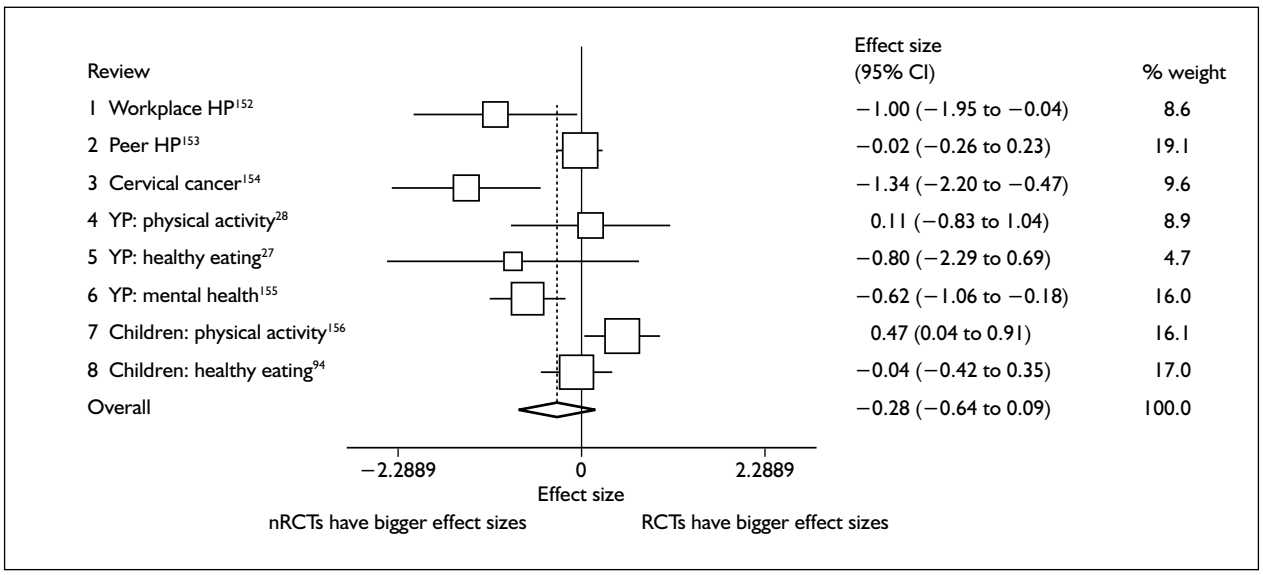

FIGURE 8 *Forest plot of outcome domain: knowledge. HP, health promotion; YP, young people.*

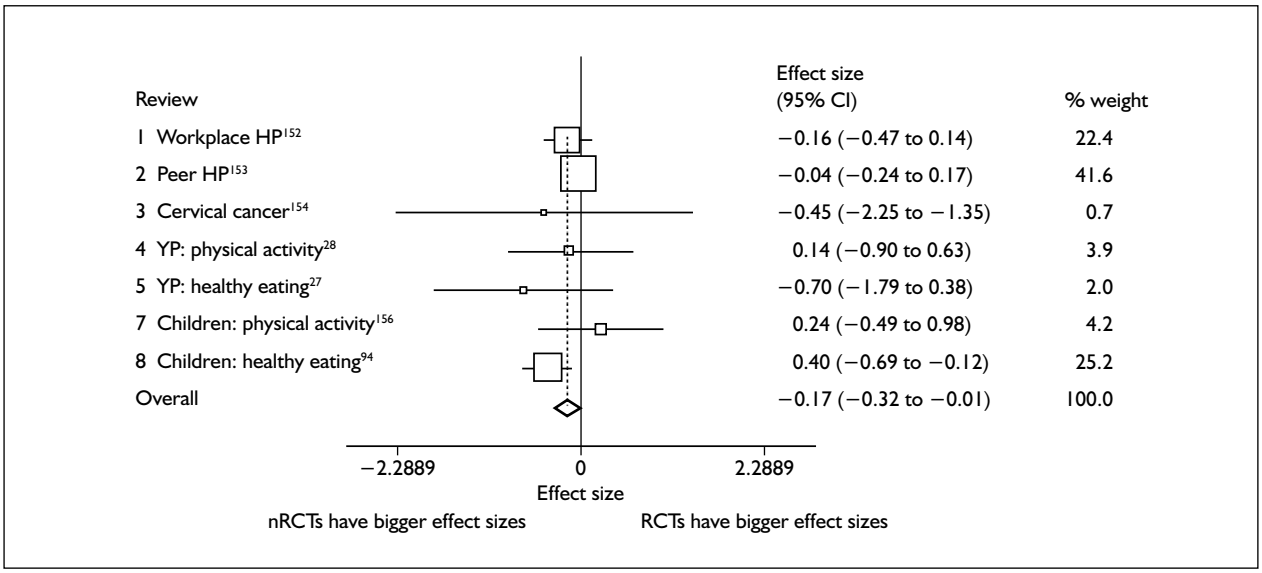

FIGURE 9 *Forest plot of outcome domain: attitudes. HP, health promotion; YP, young people.*

the nRCTs than for the RCTs. The results are not as homogeneous as those for attitudes and behaviour ($Q = 9.88$, df = 6, $p = 0.13$).

We also ran the same analyses presented above using standard meta-regression, which assumes equal variances. The result of this analysis was very similar to the above analysis, suggesting that the variances are not so different that we cannot proceed to a multivariate meta-regression in the next chapter.

Results from the Colorado studies

We followed exactly the same methods for the Colorado studies as for the studies in the EPPI-Centre reviews. Given that this data set of 126 studies is regarded as being a single albeit broad review, we did not need to calculate separate effects for each review or separate effects for different outcome domains. However, the Colorado data set does contain a wider variety of study designs

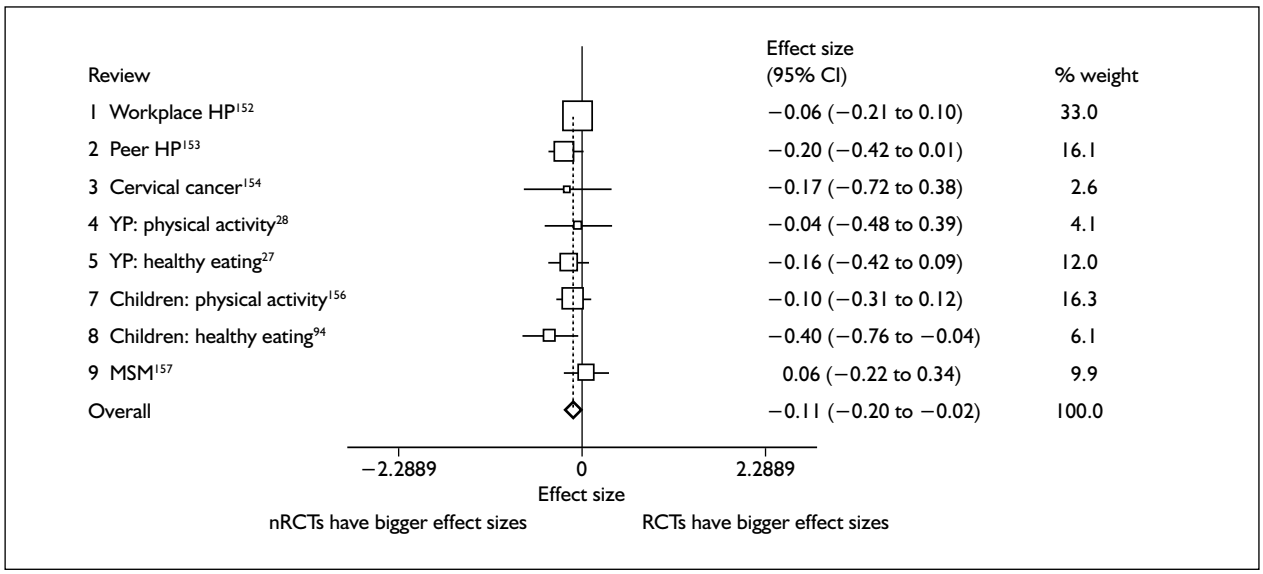

FIGURE 10 *Forest plot of outcome domain: behaviour. HP, health promotion; MSM, men who have sex with men; YP, young people.*

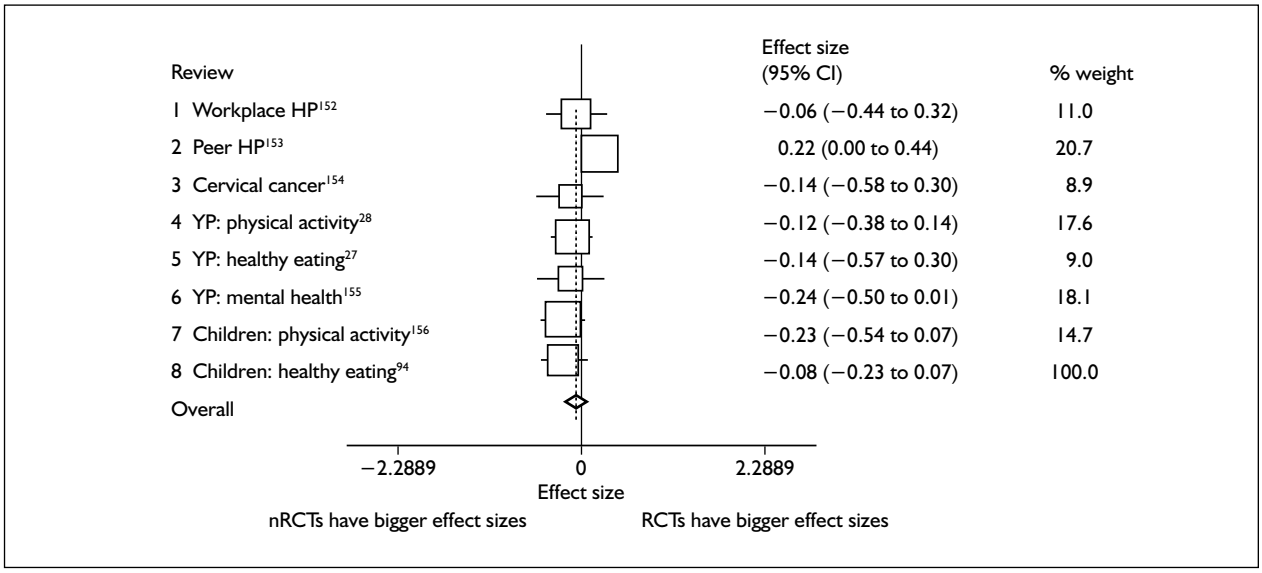

FIGURE 11 *Forest plot of outcome domain: health state. HP, health promotion; YP, young people.*

than the EPPI-Centre data, so we are able to compare RCTs with nRCTs (as above) and also with experiments without control groups (e.g. before-and-after studies). Three studies were found to have extremely large effect sizes: SMDs of approximately five or more. These studies had a disproportionate effect on the analyses and were therefore excluded as effect sizes of this magnitude are extremely rare and implausible.

We found different results in the Colorado studies than in the EPPI-Centre studies. Here, RCTs were found to have much larger effect sizes than non-randomised trials, by a statistically significant 0.368 (95% CI 0.134 to 0.603) of a standard deviation. However, there was no significant difference between the RCTs and the non-controlled studies, 0.044 (95% CI –0.134 to 0.222), although the direction of effect is for RCTs to have slightly smaller effect sizes.

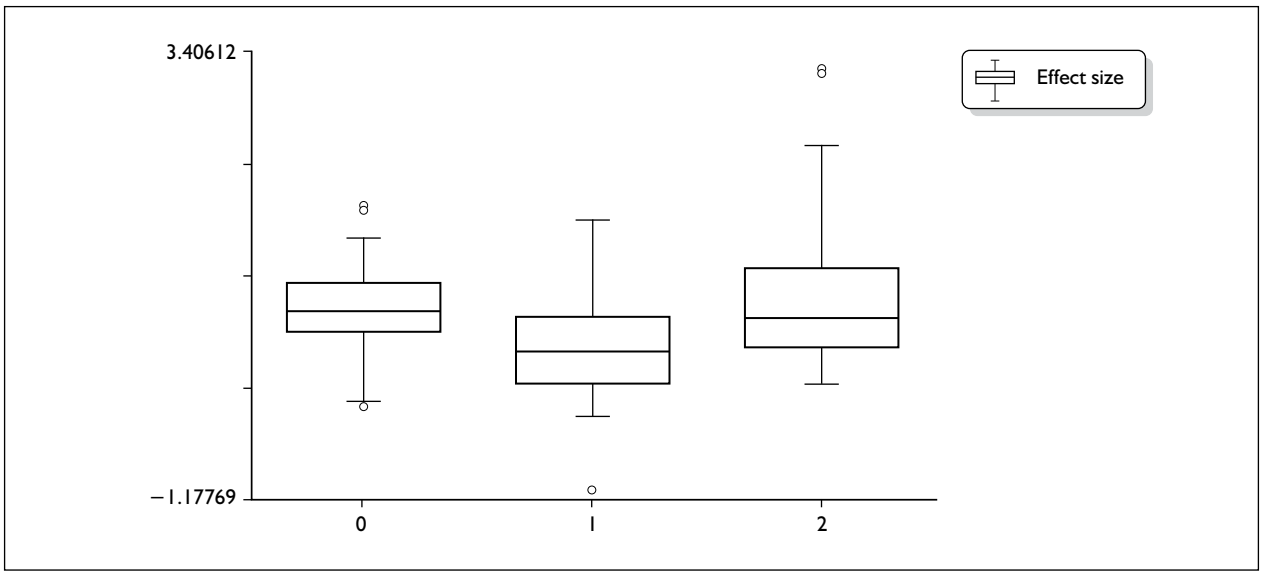

FIGURE 12 *Distribution of effect sizes for different study designs. Distributions of the effect sizes of the different studies: 0 = RCT, 1 = nRCT, 2 = experiments without control groups.*

Testing for variance (STATA: sdtest), we find that the RCTs have a smaller, but not significantly different, variance than the nRCTs ($p = 0.12$), and that RCTs have a much smaller variance than the non-controlled studies ($p = 0.0006$). We also tested for variance by running the regression using traditional meta-regression (STATA: metareg) and obtained very similar results to the above. This suggests that, while the test for different variances suggests there is a difference, we should treat this result cautiously, especially as the test is very sensitive to non-normality, and the interquartile ranges of the study types are not all that dissimilar.

Conclusion

The results from this part of the study give mixed answers to our principal research question. The statistical exercise involving the re-analysis of data from two trials suggests that nRCTs can give the same effect sizes as RCTs. This was a tightly controlled examination in which the only factor that was different between the RCTs and nRCTs was randomisation. However, we could by chance have chosen two trials in whichever area was not an important predictor of outcome, so generalisation from these two trials is difficult.

In the examination of trials sampled from systematic reviews we found considerable variation with RCTs having smaller effect sizes than non-randomised controlled studies in the EPPI-Centre reviews, and larger effect sizes than non-randomised controlled studies in the Colorado studies. The EPPI-Centre and Colorado data sets are very different, however, and we shall explore some of the possible reasons for these different results in Chapter 8. These findings show that NRSs can differ systematically from RCTs, but that the direction and existence of the difference can differ across policy areas.

Chapter 8

Results: testing the hypotheses developed in Chapters 3–5

Earlier chapters have highlighted that it is difficult to state definitively whether RCTs produce different effect sizes to other study designs as this depends on the circumstances. The previous chapter sought to establish whether differences between types of study were discernible in their results using three types of data: nRCTs constructed artificially from RCTs, trials sampled from systematic reviews and trials spanning broad sections of policy sector literature. The re-analysis of trials showed that, in two situations in which the selection of an non-randomised control group was genuinely unbiased, the results of the nRCTs were very similar to those of the RCTs. However, when we moved on to examine real trials in the field, we found that, in one data set (RCTs and nRCTs of health promotion policy interventions) for some outcomes, RCTs had smaller effect sizes, whereas in another data set (RCTs and NRSs of transition policy interventions) we found the opposite.

The aim of this chapter is to explore some of these contradictions and attempt to unpick the reasons why we, and previous studies, have found conflicting results. We shall test the hypotheses, developed in previous chapters, in order to see whether RCTs produce different effect sizes to nRCTs because they are used in different circumstances, with different types of interventions and with different participants. These hypotheses will also serve to explore differences in the findings between the EPPI-Centre and Colorado data sets. Some differences between the data sets are apparent from the outset. The average size of the samples is very different. In the EPPI-Centre studies, the average sample size for RCTs is 990 and for nRCTs it is 535 – not a statistically significant difference when the standard deviations are taken into consideration ($p = 0.27$). In the Colorado studies, on the other hand, RCTs have a mean sample size of 35 and nRCTs a mean of 84. This difference is statistically significant ($p = 0.03$). The non-control group studies are also larger than the RCTs with a mean sample size of 74 ($p = 0.03$).

The methods we use in this chapter were described in detail in Chapter 6, Methods for analysing comparable field studies and meta-epidemiology. Briefly, in order to test the association between different study characteristics (such as the theoretical underpinnings of the intervention) and the use of random assignment, we cross-tabulated study type against the characteristic in question and carried out chi-squared tests. We then tested the same characteristic to see whether theoretical framework, for instance, was associated with larger effect sizes by using the two stage model described in Chapter 6, Methods for analysing comparable field studies and meta-epidemiology, which preserves distinct variances for the different characteristics under investigation. Finally, if a characteristic was found to be associated with statistically significant differences in effect sizes, the characteristic was entered into a multivariate meta-regression in order to see whether it strengthened or weakened the findings presented in Chapter 7.

We present the findings below, following a reminder of the hypotheses being tested in each case. *p*-values are reported for the chi-squared tests and SMDs with CIs for the univariate regression.

Participants

Baseline characteristics

We considered baseline characteristics to be an important variable to explore because of the possible impact of differences in groups on the intervention and evaluation. Groups may differ at baseline because: recipients of the intervention have self-selected or those who declined to participate have been assigned to the control/comparison group; or recruitment favoured those most amenable to participation or those in most need, or excluded older people or those with multiple disadvantages (comorbidities in health, multidimensional identities in social research). Non-randomised controlled trials are more likely to have more heterogeneous populations and non-equivalence between groups. Heterogeneity and non-equivalence at baseline may influence the calculated effect size and variance.

We found that nRCTs were far more likely to report that they had non-equivalent groups at baseline than RCTs in both the EPPI-Centre data set ($p = 0.0002$) and among the Colorado studies ($p < 0.001$). However, in neither the EPPI-Centre nor the Colorado studies did this difference translate into an association with effect size.

Attrition

Higher attrition may be expected in community and home settings than in organisational settings and may be expected in transient populations (e.g. commercial sex workers, asylum seekers and socially excluded people). It is easier to employ randomisation and have good follow-up for trials carried out in organisations. High attrition may be associated with losing a disproportionate number of socially disadvantaged people who are more resistant to health promotion/public health initiatives.

We did not find that attrition rates were associated with study type in either of our data sets and, using meta-regression, we also found that different attrition rates were not associated with different effect sizes.

Intervention

Theoretical underpinnings of the intervention

Logically, interventions underpinned by theory should be more effective. The lack of theoretically based policy interventions has been noted in the fields of changing professional practice and suggested as an explanation for the lack of effective interventions.[151]

In another field, we understand that public health triallists value experimental methodologies more than do health promotion specialists, who place more emphasis on involving the community in developing and delivering the intervention, and we expected to find that experimental methodologies in public health are associated with randomisation.[21] We also expected to find that health promotion is associated with community development but not randomisation.

However, we found in the EPPI-Centre data set that RCTs were more likely to have stated or recognisable theories than nRCTs ($p = 0.0011$), and there were no significant differences among the Colorado studies ($p = 0.6930$). The presence, or absence, of a theoretical framework was not associated with effect size in either data set. The differences between the results for the Colorado and EPPI-Centre studies may be connected with different data extraction questions. The EPPI-Centre question allows reviewers to infer the theoretical framework, whereas the Colorado question asks whether the authors stated what their framework was.

Public involvement in developing the evaluation

Empowerment theories attribute responsibility to people not for the existence of a problem, but for finding a solution to it. The goal of 'full and organised community participation and ultimate self-reliance'[55] is a feature of social work such as community development and youth work, rather than a feature of public health and randomised experiments. Successful interventions specifically aimed at reducing health differentials include ensuring interventions address the expressed or identified needs of the target population and the involvement of peers in the delivery of interventions.[56]

Because of its different data extraction strategy, the EPPI-Centre data set had more relevant information in this area. We found no association between the people who identified the aims of the intervention and whether or not random assignment was employed, or between this variable and the effect size reported by studies ($p = 0.7660$).

We also found no relationship between the use of needs assessments and whether or not a study employed randomisation ($p = 0.3198$); and for outcome domain knowledge, attitudes and health state, the use of needs assessments also had no relationship with effect size. However, for behaviour outcomes (the outcome with the most data) we found that, by 0.171 of a standard deviation (95% CI 0.049 to 0.293), interventions based on needs assessments did worse than those which were not. The addition of interventions based on needs assessments to the multivariate meta-regression did not change our findings with regard to whether RCTs have different effect sizes to nRCTs.

We had a similar result when examining the issue of whether or not lay people were involved in developing the intervention. Approximately the same proportions of randomised and non-randomised studies had involved lay people in

developing their interventions and those studies which did not involve lay people had better results by 0.210 of a standard deviation (95% CI 0.036 to 0.383) for the behaviour outcome domain. The other domains did not show any significant difference. Including this variable in the multivariate meta-regression appeared both to strengthen the importance of study type and lay involvement in the model.

As part of our coding for policy interventions, we also collected data on whether or not interventions described specific collective plans of action to achieve the intervention's goals and also on whether the intervention involved the facilitation of lay/public delivered support or education. Non-randomised controlled trials were significantly more likely to have explicit action plans ($p = 0.0416$) among the EPPI-Centre studies, but were not associated with effect size. There were insufficient data available in the Colorado data set to make a judgement about explicit action plans. There was also no association between lay/public support and type of study or between lay/public support and the effect size of the intervention.

Setting and boundaries of the intervention

Interventions with a broader reach (communities, regions, nations) have more diffuse boundaries than those set within institutions. We expected to see randomisation applied less often to community, regional or national interventions. Clustered trials are more appropriate for these and some organisational level interventions. Attrition may be greater in larger scale interventions, where tracking of individuals is more difficult than within an organisation (see Attrition). Clustering reduces the power of a trial, so clustered evaluations are less likely to show effectiveness. Standardised implementation of interventions may be more difficult across large communities, regions or whole countries than in single organisations and therefore may be less effective.

We collected data on whether there was an explicit formal record of the policy (at institutional, community or regional level), the level at which policy was being enacted (international, national, regional, institutional, community) and whether an intervention was delivered to recipients individually, but we did not find any relationship between these factors and study design or randomisation.

We did find in the EPPI-Centre studies, however, that only nRCTs allocated people by region, with RCTs much more likely to allocate individuals. The relationship between study design and unit of allocation was significant ($p = 0.0117$) and, when comparing individual assignment with assignment by group, studies that allocated by group had smaller effects on attitudes than those allocating by individuals by 0.281 of a standard deviation (95% CI 0.035 to 0.526). The Colorado studies and the other outcome domains in the EPPI-Centre data set showed no significant differences.

The question on unit of allocation was categorised into individuals, family, group/class (e.g. tutor group), institution, community and region. We then carried out a meta-regression on this variable and found that, for attitudes ($p = 0.012$) and behaviour ($p = 0.033$), the size of the allocation unit was negatively correlated with effect size – i.e. that the larger units of allocation had smaller effect sizes [by –0.081 (95% CI –0.144 to –0.018) and –0.051 (95% CI –0.099 to –0.004) respectively]. We then added type of study into the regression finding that this strengthened the size and statistical significance of the associations – both for allocation unit and study type.

We also looked at differences between interventions sited in institutions and those located in the community. RCTs and nRCTs used both settings as much as one another and, possibly looking at the unit of allocation issue from another angle, we also found that community interventions had smaller effect sizes than institutional interventions by –0.159 of a standard deviation (95% CI –0.273 to 0.046).

Provider of the intervention: community/peer provider/ practitioner

Community development and peer-delivery specialists value health promotion theory and process evaluations more than RCTs so these interventions may be found to have less randomisation. Theories underpinning community development and peer-delivery anticipate more effective interventions through their greater relevance, and there is empirical evidence to support this.[56]

To explore the influence of different types of people delivering or providing interventions, we categorised the intervention providers into

community, lay, researcher and practitioner providers. Both RCTs and nRCTs used the same ranges and ratios of providers. Community providers had a significantly worse impact on knowledge (SMD = –0.404, $p = 0.003$) and behaviour (SMD = –0.247, $p = 0.000$), and the direction of the other two outcome domains was also negative. Lay providers, often peers, had more mixed results: better than the other providers in changing health states (SMD = 0.276, $p = 0.000$), similar for influencing attitudes and behaviour, but worse for knowledge (SMD = –0.236, $p = 0.024$). There were no significant results in either direction for practitioner providers.

Researcher provider

Researchers have more control over the intervention and evaluation and so, theoretically, the researcher would therefore be better able to randomise. Interventions will be found to be more consistently implemented by enthusiasts, and therefore be more effective.

This factor was explored using the same categorisation as 'practitioner' (see above). Judging by effect sizes, researchers appear to be better providers for influencing behavioural outcomes (SMD = 0.22, $p = 0.045$) and about the same as other providers for the other influencing outcome domains.

Outcomes

Choice of outcome domains

Health outcomes are more readily measured in clinical settings; clinical settings are more likely to mount RCTs, and have clinical providers and long-term follow-up. With regard to clinical outcomes being more resistant to change, the health state domain has the lowest overall effect size of 0.123 (95% CI 0.060 to 0.185), compared with 0.251 (95% CI 0.201 to 0.302) for behaviour, 0.306 (95% CI 0.193 to 0.418) for attitudes and 0.449 (95% CI 0.356 to 0.543) for knowledge. Meta-regression also suggests this ordering of outcomes is significant ($p = 0.000$).This ordering of effect size in relation to domains supports the hierarchy of outcomes proposed by Kirkpatrick[63] and Munro *et al.*[64]

Choice of outcome measures

Clinical outcomes are more commonly found in clinical settings. The choice of 'hard' or 'soft' outcomes can be associated with randomisation. If clinicians favour RCTs, clinical outcome measure may be associated with greater randomisation. Clinical outcomes will be found to be more resistant to change than 'softer' outcomes such as reported behaviour

In terms of the choice of outcome measure, there is some suggestion in the EPPI-Centre data set that RCTs use clinical tests more than nRCTs, but this is not quite statistically significant ($p = 0.0849$).

Evaluation design

Sample size

Sample size may be related to the choice of study design. Logically, larger sample sizes may be more likely in nRCTs and smaller sample sizes are more likely to give spurious results; of these, those with positive results are more likely to be published.

As suggested in the above hypotheses, the larger units of allocation provide smaller effect sizes. However, when we regress sample size and effect size in the EPPI-Centre data set, we find no association – although the 'direction' is that smaller samples have larger effect sizes ($p = 0.221$). The Colorado data set has the same characteristics, with a suggestion of smaller effect sizes in the larger studies which does not quite reach statistical significance ($p = 0.058$). When the two data sets are combined, we have a set of studies with a much larger spread of sample sizes, and the association between sample size and effect size is more pronounced ($p = 0.03$).

Control group

Control groups are always found in RCTs, but only sometimes in NRSs. We expected to find that the use of a control group leads to smaller effect sizes than are found in uncontrolled evaluations.

As the EPPI-Centre data set did not contain any studies without control groups, this analysis could only be conducted with the Colorado studies. A simple comparison of studies with control groups against studies without does not identify any significant differences. However, combining RCTs with nRCTs conceals differences within the studies that have control groups. The Colorado data set has large effect sizes for RCTs, medium effect sizes for nRCTs and large effect sizes for the non-control group studies. This means that we might expect differences between the nRCTs and the non-control

group studies but not between the RCTs and the non-control group studies. If the above hypothesis is correct – that uncontrolled evaluations have larger effect sizes (and the comparison between nRCTs and non-control group studies is consistent with this: $p = 0.014$) – then the question that remains is why do the RCTs in the Colorado data set have such large effect sizes?

Blinding

Blinding of participants, recruiters, intervention providers and outcome assessors to the intervention allocation is easier for some interventions with randomisation. Poor concealment, common in nRCTs, will be associated with greater effect sizes.

We assessed blinding in three ways in the EPPI-Centre data set: allocation concealment, participant awareness and outcome measurement. RCTs were significantly more likely to use allocation concealment ($p = 0.000$) and blinded outcome measurement ($p = 0.002$) than nRCTs. However, nRCTs were equally likely to conceal allocation from the intervention participants. These results are based on small numbers of RCTs and nRCTs (fewer than 40). As our regression analyses are by outcome domain within each review, there were insufficient studies to carry this out to assess blinding.

The Colorado data set did not record concealment of allocation.

Follow-up

Length of follow-up periods is linked with study design; long follow-up may be easier within institutions, where randomisation is also easier. We expected that long follow-up would be associated with declining effect size.

Despite the expectation that longer follow-ups would lead to smaller effect sizes, we did not find any evidence of this in either data set. There was also no association with effect size.

Clustering

Clustered trials with few clusters are more likely to be 'natural experiments'. Natural experiments do not include randomisation. Natural experiments are more likely to lack blinding and to have enthusiasts supporting the intervention and non-enthusiasts supporting the comparisons, and therefore lead to greater effect sizes. Testing this hypothesis was beyond the capacity of this project.

Quality of the reporting

Quality of reporting specific elements of a study is associated with researchers' disciplines. Better reporting (of pre- and postintervention data) may be seen to be associated with triallists who also support randomisation.[72] Reporting of pre- and postintervention data precludes effect sizes inflated by differences between groups.

We collected data on a range of aspects of reporting quality in the EPPI-Centre data set: pre-intervention information on sociodemographic variables; pre-intervention data on outcome variables; names of measurement tools; postintervention data on outcome variables; whether there were any obvious shortcomings in the numerical reporting; and whether the study was replicable based on the report. None of the categorical answers was found to be associated with study type: RCTs seem to be as well (or as poorly) reported as nRCTs. Some statistically significant results were found when relating the above factors with effect size, but no obvious pattern emerges:

- Studies that provided full information on pre-intervention sociodemographic variables had higher effect sizes for knowledge: 0.291 of a standard deviation (95% CI 0.030 to 0.551); this variable was not significant in uni- or multivariate meta-regression.
- Studies without obvious shortcomings in their reporting have better results for knowledge by 0.434 of a standard deviation (95% CI 0.261 to 0.607), but no difference for other outcome domains; when combined in a multivariate meta-regression with study type, this variable was significant ($p = 0.001$) and moved study type from $p = 0.183$ to $p = 0.053$.
- Studies giving enough information or providing a further source of information on evaluation design are not as effective as those that do not at changing people's knowledge by 0.433 of a standard deviation (95% CI 0.046 to 0.821); this variable was not significant in uni- or multivariate meta-regression.
- Studies that do not give sufficient information to ensure that the content of the intervention is replicable do better in the attitudes domain by 0.559 of a standard deviation (95% CI 0.371 to 0.747) than those that do give sufficient information. When combined in a multivariate meta-regression with study type, this variable

TABLE 23 *Summary of the results of the univariate (unadjusted) and multivariate (adjusted) regressions. A negative value indicates that RCTs had smaller effect sizes than nRCTs*

	Unadjusted (95% CI)	Adjusted (95% CI)
Knowledge	–0.275 (–0.641 to 0.091)	–0.269 (–0.465 to –0.073)
Attitudes	–0.166 (–0.319 to –0.012)	–0.165 (–0.369 to 0.040)
Behaviour	–0.111 (–0.199 to –0.023)	–0.192 (–0.330 to –0.053)
Health state	–0.084 (–0.234 to 0.066)	0.052 (–0.149 to 0.254)

was significant ($p = 0.000$) and moved study type from $p = 0.059$ to $p = 0.163$.

We examined the Colorado data set on two aspects of reporting quality: replicability and the naming of measurement tools. Neither of these issues was seen to be associated with type of study and, again, we had a scattering of statistically significant results without any clear pattern:

- The meta-regression comparing those studies that rated highly on replicability with those that rated poorly suggests that effect size decreases as replicability reduces, but the result is not quite statistically significant at $p = 0.052$.
- Studies that named their measurement tools did significantly worse than those that did not by 0.3319 of a standard deviation (95% CI 0.5283 to 0.1355).

The multivariate regression

After testing each of the above factors in turn for associations with effect size, we placed those that were significant in the univariate analysis of EPPI-Centre studies into a multivariate model to explore independent effects. In order to avoid confounding by review, review was included in the model as a fixed effect. The factors that we explored were: public involvement, settings and boundaries, intervention provider (clinician) and reporting quality. When all factors are placed in the meta-regression, many lose statistical significance. *Table 23* records how the regression affects our overall hypothesis.

Knowledge

The result of exploring the outcome domain knowledge with relation to study type suggested that there was a non-significant effect in favour of smaller effects for RCTs: –0.275 (95% CI –0.641 to 0.091). After taking all the above factors into account in the multivariate regression, the meta-regression now suggests that there is a significant effect in favour of smaller effects for RCTs: –0.269 (95% CI –0.465 to –0.073), $p = 0.007$.

Attitudes

The previous analysis suggested that RCTs had significantly smaller effect sizes: –0.166 (95% CI –0.319 to 0.012). After taking all the above factors into account, the meta-regression suggests that the direction and quantity of the effect is the same, but it is no longer significant ($p = 0.115$).

Behaviour

The previous analysis suggested that RCTs had significantly smaller effects: –0.111 (95% CI–0.199 to –0.023). The meta-regression suggests that the amount by which nRCTs overstate their effects is slightly larger: –0.192 (95% CI–0.330 to –0.053), and the statistical significance of this has increased ($p = 0.007$).

Health state

The previous analysis suggested that there was very little difference between RCTs and nRCTs: –0.084 (95% CI –0.234 to 0.066). The multivariate meta-regression confirms this ($p = 0.611$) with the direction of effect now marginally in favour of larger effects in the RCTs.

Conclusion from EPPI-Centre data

RCTs have statistically significantly smaller effect sizes than nRCTs for behavioural outcomes – and the indications are that this holds true for attitudes and knowledge too. In spite of taking many possible confounding factors into account, the type of study still explains some differences in the observed effect sizes in this data set.

Chapter 9
Discussion

Summary of findings

In two particular cases, trials that are identical in all respects except randomisation (constructed from resampling randomised and non-randomised comparisons from RCT data) led to similar effect sizes, but sometimes with greater variance in the absence of randomisation. In the field, however, effect sizes can differ, yet extensive empirical investigations fail to predict the direction of these differences or the circumstances in which they happen.

We found randomisation to be associated with greater equivalence of groups at baseline, explicit theoretical underpinning of interventions in one data set but not the other, allocation of individuals (rather than groups), allocation concealment and blinded outcome measurement. We found randomisation to be negatively associated with reporting of specific collective plans of action to achieve the intervention's goals. We found no association between randomisation and individual or group interventions, public involvement, institutional or community settings, type of intervention provider, attrition and quality of reporting.

Strengths and weaknesses of study methods

This study employed well-established research methods both for assessing what is already known about randomisation and effect size, and for analysing direct and indirect relationships between randomisation and effect size.

To assess what was already known, we systematically sought and analysed prior studies incorporating the strengths of systematic review methodology (systematic searches and reviewers working independently to analyse each study). Electronic searches were limited by the poor indexing of methodological studies, and the value of meta-analyses that included randomised and non-randomised studies without the explicit aim of comparing the two. In these circumstances, exhaustive searching was not possible; however, systematic electronic searches were complemented by approaches to key methodologists in the area and by searching the World Wide Web. The broad range of methods and contexts of the studies identified was a challenge to assessing their methodological quality. In the absence of clear quality criteria spanning the full range of studies, we chose not to rely on their methods and findings but to use these to design our own original analyses which took into account the strengths and weaknesses of earlier work.

We built on this earlier evidence (see Chapters 4 and 5), and on our understanding of policy evaluation (see Chapter 3), to construct tightly defined predetermined hypotheses for testing with our own data. We hypothesised first that randomisation would lead to differences in effect size and, second, that these differences might be mediated by a number of confounders. We adopted two key methods (resampling studies within single RCTs and meta-regression within reviews) to triangulate the findings of the two approaches. For each of these approaches we used two data sets: the first being data generated by a trial and systematic reviews conducted at the Social Science Research Unit, in the UK, where we were very familiar with the definitions and their application in earlier analyses; and another data set, where we relied on other people's data and their definitions in a Canadian trial and international studies reviewed by American researchers.

For the resampling studies we chose to limit our resampling to comparing groups of data that could reasonably be expected to arise from sampling decisions in the field, rather than calculating numerous effect sizes from resampling thousands of times, as Deeks *et al.*[40] had done.

For analysing review data, we overcame shortcomings of previous studies and went to greater lengths to compare like with like. We nested results within original reviews within which each study shared a set of desired outcomes. This approach minimised differences other than randomisation, compared with other meta-epidemiological studies that relied on a more diverse set of studies. The EPPI-Centre data set

includes a fairly narrow set of study types (RCTs and nRCTs), in which the non-randomised studies would have employed the same methods if only randomisation had been applied. In order to compare like with like as much as possible, we did not calculate an 'average' outcome for each study, but used up to four domains of outcomes per study.

Reviews of previous studies have reported analyses that relied on different ratios of randomised and non-randomised studies. In one study only 15% of studies were randomised.[42] In another, the proportion of randomised studies reached 71%.[87] Our EPPI-Centre data study had fairly similar numbers (97 RCTs and 79 nRCTs), making comparisons of studies easier. Our Colorado data set was less balanced, with 16 non-randomised studies and 46 randomised or quasi-randomised studies of policy interventions.

With this data we conducted a very fine-grained analysis, first within reviews, then across reviews to thoroughly investigate factors that may confound the relationship between randomisation and effect size.

Although assumptions about bias and directions of bias arising from various sources have been made (e.g. theoretical underpinnings of intervention and many of the other non-medical hypotheses from Chapter 3), according to our research this is the first time those assumptions have been empirically tested.

Findings from different data sets

Using the EPPI-Centre data of health promotion evaluations, we found that non-randomised trials resulted in larger effect sizes than randomised trials (statistically significant for two out of four outcomes, and direction of effect the same in the other two). In comparison, using the Colorado data, randomised trials and studies without control groups both resulted in larger effect sizes than non-randomised trials (statistically significant across the pooled outcomes). These differences may be explained by differences in the data sets.

The EPPI-Centre data were a rich source of controlled before-and-after studies ($n = 50$) and non-randomised trials ($n = 23$). Many of these ($n = 47$ and $n = 20$ respectively) were clustered studies where the intervention and control were allocated to groups rather than individuals, as in 'natural experiments' of policy interventions. Many RCTs ($n = 75$) were also clustered in classes, institutions or communities. The outcome domains were matched for comparisons. The design features of the included studies, and the opportunity to match outcome domains, made this data set particularly appropriate for fair comparisons of effect sizes with and without randomisation.

In comparison, the Colorado data set included fewer studies. These studies were more diverse in their designs and had a smaller number of RCTs and nRCTs for comparison. There was no opportunity to compare matched outcomes, as effect sizes were only available for one outcome per study. These differences meant there was less opportunity to compare like with like. The unexpectedly large effect sizes resulting from RCTs may be explained by the small size of the studies: small studies are more likely to produce spurious results, and publication bias leads to greater publication of studies with positive findings. Another explanation may be the different nature of the interventions in this data set. A high proportion of the interventions in the Colorado data were based on information and communications technology (e.g. computers, captioned television, videoconferences). Half of the RCTs evaluated computer-based interventions, whereas approximately one quarter of the non-randomised trials or studies without control groups did the same. In summary, the Colorado data set was dominated by small-scale computer-based studies, and was therefore very different from the natural experiments of large-scale policy interventions, or comparable randomised evaluations found in the EPPI-Centre data.

There was a lack of useable data about 'blinding' of allocation in either data set. There was very little blinding of allocation in randomised or non-randomised studies in the EPPI-Centre data set: such low numbers preclude further investigation. The data available in the Colorado data set refer to blinding of the participants and not blinding of allocation. As blinding can influence effect size,[69] this lack of data is frustrating.

Early in our study we excluded reviews by education review groups allied to the EPPI-Centre as suitable sources of policy interventions because few of these reviews included both randomised and

non-randomised studies. On reflection, they may have been a poor source of 'natural experiments' because three of the eight eligible reviews were of computer-based interventions.

All in all, the EPPI-Centre data proved much more suitable for comparative analysis of randomised and non-randomised studies, so we are more confident of our conclusions from this data, that NRSs of policy interventions inflate effect sizes in the field in ways that we cannot fully explain.

Weaknesses in the Colorado data and the excluded education reviews mean that our conclusions are restricted to health promotion policy, and that we have no corroborating evidence from the main stream education sector or the social services sector.

Comparison with other studies

Our systematic review of empirical comparisons (see Chapter 5) of randomised and non-randomised evaluations of policy interventions revealed inconsistent relationships between randomisation and effect size: randomisation was associated with similar, dissimilar and variable effect sizes in different studies. As these studies did not aim to explore the causes of these differences, they offer little illumination other than to confirm that the design of evaluations is important in assessing effects of policy interventions.

Even methodological studies which did aim to investigate the role of randomisation in assessing effects of policy interventions were inconsistent in their conclusions (see Chapter 4). Our conclusion that RCTs lead to smaller effect sizes than nRCTs is supported by investigations of juvenile delinquency[42,46] and psychological interventions.[76] However, RCTs lead to larger effect sizes of marital and family therapy[74] and scholastic aptitude test coaching, ability grouping in classrooms, pre-surgical education and drug abuse prevention.[72] As in our own study, these differences could not be fully explained by differences in populations, interventions or evaluations.

Our study has shown that carefully designed meta-epidemiological studies can help us to understand bias resulting from study design and that they are particularly powerful alongside other techniques. The re-analyses of data from RCTs shows us what a population of randomised and non-randomised trials might look like; meta-epidemiology enables us to examine actual populations. Without the knowledge gained from the re-analyses of trials, we would be starting from a weaker reference point when exploring whether RCTs and nRCTs have different results in the field. Knowing that non-randomised trials with unbiased control groups do not differ in effect size (on the whole) to RCTs, we can be sure that any differences observed are due either to experimenter bias (arising from the non-randomisation) or to the different types of study being used to evaluate different interventions and/or different populations and/or different outcomes. (The meta-regression was then able to test these different possible confounders.)

The results of our re-analyses of trial data come to much the same conclusions as Deeks *et al.*,[40] with both studies finding that the size of effect did not differ between study types, as a whole. However, Deeks *et al.* found an increase in variance among their nRCTs, whereas we did not find any statistically significant difference. This may be due to the fact that Deeks *et al*. constructed their comparison groups from different regions and cities, while our comparisons were much closer, geographically, some being within the same London borough. This might lead us to recommend that, if a study cannot employ randomisation, selecting nearby areas will offer a better comparison than, for example, 'similar' areas in another city. However, this may limit the generalisability of the findings. Also, when critically appraising nRCTs, the closeness of the comparison areas might be something to bear in mind.

Conclusions

Randomisation does not, according to our reconstructed RCTs and nRCTs, directly influence the effect size of interventions as a whole. Yet, while the many examples reviewed and the new analyses in the current study reveal that randomisation is indeed associated with changes in effect sizes in trials of policy interventions, these differences can lead to larger effect sizes in some cases and smaller ones in others; their direction is difficult to predict. Despite extensive analysis testing of many predefined hypotheses that might have explained this difference, we have failed to identify consistent explanations for these differences. We have tested the possibility that the type of participants,

interventions, selection and measurement of outcomes, and evaluation design might account for the observed differences, but have not found a more consistent predictor of the effect size of interventions than whether or not the evaluation employed random assignment.

Two possibilities could explain the different conclusions arising from the meta-regression and our re-analyses of trials. First, our sample of nRCTs may be biased, possibly because nRCTs are less likely to be published than RCTs when results are not 'exciting'. Second, the nRCTs of policy interventions in the field may have larger effect sizes because of conscious or unconscious experimenter bias when control groups are selected: interventions may be allocated to enthusiastic institutions. As we are unlikely to come to a closer explanation of possible differences than this, decision-makers need to treat the results of nRCTs with caution. Researchers mounting new evaluations need to avoid, wherever possible, allocation bias.

Our study identified 45 evaluations of policy interventions where institutions were allocated randomly to intervention or comparison groups. Such RCTs must be the preferred design for cautious assessment of effects given the feasibility of randomising institutions, and the lower effect sizes of randomised studies. Fewer studies allocated communities or regions, randomly or not, to evaluate the effects of policy interventions.

Recommendations for research to evaluate the effects of policy interventions

1. Policy evaluations should adopt randomised designs wherever possible.
2. Policy evaluations should also adopt other standard procedures for minimising bias and conducting high-quality assessment of effects of intervention, particularly blinded allocation of either individuals or groups, and the avoidance of small sample sizes.
3. Feasibility studies of randomising geographical areas, communities and regions should be carried out for evaluating policy interventions in a range of sectors, implemented within interventions, communities and across regions.
4. Feasibility studies of blinded allocation should be carried out for policy interventions in a range of sectors, implemented within interventions, communities and across regions.
5. Clear descriptions should be included in systematic reviews of how judgements of equivalence (or otherwise) have been reached when comparing the effects found in randomised and non-randomised studies of policy interventions.
6. Research is required into the reasons for choosing randomisation or not, particularly in the presence and absence of an explicit collective plan of action.

Acknowledgements

We are very grateful to our Advisory Group, Professors Doug Altman, Jos Kleijnen and Ann Oakley, and Angela Harden, for being generous with their time and providing valuable guidance throughout the study.

Thanks are also due to Steven Duffy for information support and Zarnie Khadjesari for screening and data extraction of reviews.

Lastly, we would like to thank the anonymous reviewers for their careful attention to the report.

Contribution of authors

JS initiated the conceptual framework for policy interventions. This was refined through further reference to the policy literature, team discussions (SO, JS, AMB, AS, JD, JT, RR, JC), and reflective application of definitions to successive data sets (JS, JC, SO, RR, ZG, KO). SO reviewed the background literature on policy evaluation to inform the hypotheses for empirical testing. The review of methodological studies was designed, conducted and reported by AMB, AS, JD and JS. The review of empirical studies was designed, conducted and reported by AMB, AS, JD, JS and JC. The resampling studies were designed by JT and IW, and conducted by JT, IW and KO. The field studies analyses were designed by JT, SO, IW and RR, conducted by JT, SO, IW, RR, ZG and KO, and reported by JT, SO, IW and KO. Conclusions were drawn and recommendations made by SO, AMB, JT, AS, JD, JS, IW and JC. SO led the study and takes responsibility for the integrity of the work as a whole.

References

1. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–40.

2. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.

3. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**:1–154.

4. Jenkins B. Policy analysis: models and approaches. In: Hill M, editor. *The policy process*: *A reader.* London: Harvester Wheatsheaf; 1993.

5. Oakley A. *Experiments in Knowing*: *gender and method in the social sciences.* Oxford: Polity Press; 2000.

6. House ER, Mathison S. Educational intervention. In: Seidman E, editor. *Handbook of social intervention.* Beverly Hills: Sage; 1983. pp. 323–37.

7. Haveman RH. *Poverty policy and poverty research*: *The great society and the social sciences.* Madison: University of Wisconsin Press; 1987.

8. Nathan RP. *Social science in government*: *uses and misuses.* New York: Basic Books; 1988.

9. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 2002;**56**:119–27.

10. Davies HTO, Nutley S, Smith PC. *What works? Evidence-based policy and practice in public services.* Bristol: The Policy Press; 2000.

11. Devlin W, Keogh P, Nutland W, Weatherburn P. *The field guide*: *Applying making it count to health promotion activity with homosexually active men.* London: Terrence Higgins Trust; 2003.

12. Harrison S. Policy analysis. In: Fulop N, Allen P, Clarke A, Black N, editors. *Studying the organization and delivery of health services.* London: Routledge; 2001. pp. 90–106.

13. Macintyre SP, Petticrew M. Good intentions and received wisdom are not enough. *J Epidemiol Community Health* 2000;**54**:802–3.

14. Oakley A, Fullerton D, Holland J, Arnold S, France-Dawson M, Kelley P, *et al*. Sexual health education interventions for young people: a methodological review. *BMJ* 1995;**310**:158–62.

15. Nutbeam D. Oakley's case for using randomised controlled trials is misleading. *BMJ* 1999;**318**:944–5.

16. Speller V, Learmonth A, Harrison D. The search for evidence of effective health promotion. *BMJ* 1997;**315**:361–3.

17. Nutbeam D. *Assessing the effectiveness of public health interventions, oral presentation, evidence into practice*: *Challenges and opportunities for UK Public Health.* London: The Royal College of Physicians; 2001.

18. WHO European Working Group on Health Promotion Evaluation. *Health promotion evaluation*: *Recommendation to policymakers.* Copenhagen: WHO Regional Office for Europe; 1998.

19. Oakley A, Strange V, Toroyan T, Wiggins M, Roberts I, Stephenson J, *et al*. Using random allocation to evaluate social interventions: three recent UK examples. *Annals* 2003;**589**:170–589.

20. Cook TD, Payne MR. Objecting to the objections to using random assignment in educational research. In: Mosteller F, Boruch R, editors. *Evidence matters*: *randomized trials in education research.* Washington, DC: Brookings Institution; 2002.

21. Oakley A. Experimentation in social science: the case of health promotion. *Social Sciences in Health* 1998;**4**:73–89.

22. Vartiainen E, Tossavainen K, Puska P. North Karelia youth programmes. In: Puska P, Tuomilehto J, Nissinen A, V, editors. *The North Karelia Project. 20 year results and experiences.* Helsinki: The National Public Health Institute; 1995. pp. 289–310.

23. Olsen JJ, Farkas G. Employment opportunity can decrease adolescent childbearing within the underclass. *Eval Program Plann* 1991;**14**:27–34.

24. Bonell CP, Hargreaves JR, Strange V, Pronyk PM, Porter JDH. Should structural interventions be evaluated using RCTs? The case of HIV prevention. *Soc Sci Med* 2006;**63**:1135–42.

25. Cook TD. Why have educational evaluators chosen not to do randomized experiments? *Ann Am Acad Pol Soc Sci* 2003;**589**:114–49.

26. Petrosino A. Estimates of randomized controlled trials across six areas of childhood intervention: a bibliometric analysis. *Ann Am Acad Pol Soc Sci* 2003;**589**:190–202.

27. Shepherd J, Harden A, Rees R, Brunton G, Garcia J, Oliver S, *et al*. *Young people and healthy eating*: *a systematic review of research on barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit; 2001.

28. Rees R, Harden A, Shepherd J, Brunton G, Oliver S, Oakley A. *Young people and physical activity*: *a systematic review of research on barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit; 2001.

29. Davies P, Boruch R. Does for public policy what Cochrane does for health [comment]. *BMJ* 2001;**323**:294–5.

30. Elbourne D, Oakley A, Gough D. EPPI Centre reviews will aim to disseminate systematic reviews in education. *BMJ* 2001;**323**:1252.

31. Briss PA, Rodewald LE, Hinman AR, Shefer AM, Strikas RA, Bernier RR, *et al*. Reviews of evidence regarding interventions to improve vaccination coverage in children, adolescents, and adults: The Task Force on Community Preventive Services. *Am J Prev Med* 2000;**18**:97–140.

32. Zief SG, Lauver S, Maynard R. *Impacts of after-school programs on student outcomes*. Campbell Systematic Reviews, Issue 3; 2006. URL: www.campbellcollaboration.org/library.php (accessed 29 October 2009).

33. Petrosino A, Turpin-Petrosino C, Buehler J. *'Scared Straight' and other juvenille awareness programmes for preventing juvenille delinquency. The Campbell Collaboration*: *C2 Reviews of Interventions, and Policy Evaluations (C2-RIPE).* Philadelphia: Campbell Collaboration; 2002.

34. D'Agostino RB, Kwan H. Measuring effectiveness: What to expect without a randomized control group. *Med Care* 1995;**33**:95–105.

35. Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. *J Clin Epidemiol* 1999;**52**:487–97.

36. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94.

37. Moher D, Pham B, Jones A, Cook DJ, Moher M, Tugwell P, *et al*. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses. *Lancet* 1998;**352**:609–13.

38. Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in meta-epidemiological research. *Stat Med* 2002;**21**:1513–24.

39. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Davey Smith G, Altman D, editors. *Systematic reviews in health care*: *meta-analysis in context.* London: BMJ Books; 2001.

40. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**:1–173.

41. Bradford Hill A. *A short textbook of medical statistics.* London: Hodder & Stoughton; 1977.

42. Weisburd D, Lum CM, Petrosino A. Does research design affect study outcomes in criminal justice? *Ann Am Acad Polit Soc Sci* 2001;**578**:50–70.

43. Ross D, Wight D. The role of randomised controlled trials in assessing sexual health interventions. In: Imrie J, Stephenson JM, Bonell C, editors. *Effective sexual health interventions*: *Issues in experimental evaluation.* Oxford: Oxford University Press; 2003.

44. MacDonald G. Social work: beyond control. In: Maynard A, Chalmers I, editors. *Non-random reflections on health services research.* London: BMJ Publishing Groups; 2007.

45. Gibson PG, Powell H, Wilson A, Abramson MJ, Haywood P, Bauman A, *et al*. *Self-management education and regular practitioner review for adults with asthma*. The Cochrane Database of Systematic Reviews, Issue 3; 2002. Report No.: CD001117. DOI: 10.1002/14651858.CD001117.

46. Lipsey MW. Those confounded moderators in meta-analysis: good, bad, and ugly. *Ann Am Acad Pol Soc Sci* 2003;**587**:69–81.

47. Wiggins M, Oakley A, Roberts I, Turner H, Rajan L, Austerberry H, *et al*. *The Social Support and Family Health Study*: *a randomised controlled trial and economic*

evaluation of two alternative forms of postnatal supoprt for mothers living in disadvantaged inner-city areas. Health Technology Assessment NSH R&D HTA Programme; 2004.

48. MacMillan HL, Thomas BH, Jamieson E, Walsh CA, Boyle MH, Shannon HS, *et al*. Effectiveness of home visitation by public-health nurses in prevention of the recurrence of child physical abuse and neglect: a randomised controlled trial. *Lancet* 2005;**365**:1786–93.

49. Janson SL, Alioto ME, Boushey HA. Attrition and retention of ethnically diverse subjects in a multicenter randomized controlled research trial. *Contr Clin Trials* 2001;**22**:236S–43S.

50. Zebracki K, Drotar D, Kirchner HL, Schluchter M, Redline S, Kercsmar C, *et al*. Predicting attrition in a pediatric asthma intervention study. *J Pediatr Psychol* 2003;**28**:519–28.

51. Sowden A, Stead LF. *Community interventions for preventing smoking in young people*. The Cochrane Database of Systematic Reviews, Issue 1; 2003. Report No.: CD001291. DOI: 10.1002/14651858. CD001291.

52. Walton RT, Gierl C, Yudkin P, Mistry H, Vessey MP, Fox J, *et al*. Evaluation of computer support for prescribing (CAPSULE) using simulated cases. *BMJ* 1997;**315**:791–5.

53. Zoritch B, Roberts I, Oakley A. *Day care for pre-school children*. Cochrane Database of Systematic Reviews 2000, Issue 3. Art. No.: CD000564. DOI: 10.1002/14651858.CD000564.

54. Macdonald G, Davies J. Reflection and vision. Proving and improving the promotion of health. In: Davies J, Macdonald G, editors. *Quality, evidence and effectiveness in health promotion*: *Striving for certainties.* London: Routledge; 1998.

55. Yeo M. Toward an ethic of empowerment for health promotion. *Health Promot Int* 1993;**8**:225.

56. Arblaster L, Lambert M, Entwistle V, Forster M, Fullerton D, Sheldon T, *et al*. A systematic review of the effectiveness of health service interventions aimed at reducing inequalities in health. *J Health Serv Res Pol* 1996;**1**:93–103.

57. Foerster SB, Gregson J, Beall DL, Hudes M, Magnuson H, Livingston S, *et al*. The California children's 5 a Day Power Play! Campaign: evaluation of a large-scale social marketing initiative. *Fam Community Health* 1998;**21**:46–64.

58. Flowers P, Hart G, Williamson L, Frankis J, Der G. Does bar-based, peer-led sexual health promotion have a community-level effect amongst gay men in Scotland? *Int J STD AIDS* 2002;**13**:102–8.

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

60. Moon AM, Mullee MA, Rogers L, Thompson RL, Speller V, Roderick P. Helping schools to become health-promoting environments – An evaluation of the Wessex Healthy Schools Award. *Health Promot Int* 1999;**14**:111–22.

61. Lumley J, Chamberlain C, Dowswell T, Oliver S, Oakley L, Watson L. *Interventions for promoting smoking cessation during pregnancy.* The Cochrane Database of Systematic Reviews, Issue 3; 2009. Art No: CD001055. DOI: 10.1002/14651858. CD001055.pub3.

62. Lipsey MW, Wilson DB. The way in which intervention studies have 'personality' and why it is important to meta-analysis. *Eval Health Prof* 2001;**24**:236–54.

63. Kirkpatrick DL. Evaluation of training. In: Craig R, Mittel I, editors. *Training and Development Handbook.* New York: McGraw Hill; 1967. pp. 87–112.

64. Munro S, Lewin S, Swart T, Volmink J. A review of health behaviour theories: how useful are these for developing interventions to promote long-term medication adherence for TB and HIV/AIDS? *BMC Public Health* 2007;**7**:104.

65. Rigotti NA, Munafo MR, Stead LF. *Interventions for smoking cessation in hospitalised patients.* The Cochrane Database of Systematic Reviews, Issue 3; 2007. Art. No.: CD001837. DOI: 10.1002/14651858. CD001837.pub2.

66. Secker-Walker RH, Gnich W, Platt S, Lancaster T. *Community interventions for reducing smoking among adults.* The Cochrane Database of Systematic Reviews, Issue 2; 2002. Art. No.: CD001745. DOI: 10.1002/14651858.CD001745.

67. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. Are the results of the study valid? *JAMA* 1993;**270**:2598–601.

68. Wiggins M, Rosato M, Austerberry H, Sawtell M, Oliver S. *Sure Start Plus National Evaluation*: *Final Report.* London: Social Science Research Unit, Institute of Education, University of London, 2005.

69. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**:1185–90.

70. Glazerman S, Levy DM, Myers D. *Nonexperimental replications of social experiments*: *a systematic review. Interim Report/Discussion Paper.* Washington, DC: Mathematica Policy Research, Inc.; 2002.

71. Glazerman S, Levy DM, Myers D. Nonexperimental versus experimental estimates of earnings impacts. *Ann Am Acad Pol Soc Sci* 2003;**589**:63–93.

72. Heinsman DT, Shadish WR. Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments? *Psychol Methods* 1996;**1**:154–69.

73. Moyer A, Finney JW, Swearingen CE. Methodological characteristics and quality of alcohol treatment outcome studies, 1970–98: an expanded evaluation. *Addiction* 2002;**97**:253–63.

74. Shadish WR, Ragsdale K. Random versus nonrandom assignment in controlled experiments: do you get the same answer? *J Consult Clin Psychol* 1996;**64**:1290–305.

75. Shadish WR, Navarro AM, Matt GE, Phillips G. The effects of psychological therapies under clinically representative conditions: a meta-analysis. *Psychol Bull* 2000;**126**:512–29.

76. Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol Methods* 2001;**6**:413–29.

77. Shadish WR, Montgomery LM., Wilson P, Wilson MR, Bright I, Okwumabua T. Effects of family and marital psychotherapies: a meta-analysis. *J Consult Clin Psychol* 1993:**61**:992–1002.

78. Weisz JR, Weiss B, Han SS, Granger DA, Morton T. Effects of psychotherapy with children and adolescents revisited: a meta-analysis for clinicians. *J Consult Clin Psychol* 1995;**55**:542–549.

79. Lipsey MW. Juvenile delinquency treatment: a meta-analysis inquiry into the variability of effects. In: Cook TD, Cooper H, Cordray DS, editors. *Meta-analysis for explanation. A casebook*. New York, NY: Russell Sage.

80. Lipsey MW, Wilson DB. Effective interventions for serious juvenile offenders: a synthesis of research. In: Loeber R, Farrington DP, editors. *Serious and violent juvenile offenders: risk factors and successful interventions*. Thousand Oaks, CA: Sage.

81. Aiken LS, West SG, Schwalm DE, Carroll JL, Hsiung S. Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: efficacy of a university-level remedial writing program. *Eval Rev* 1998;**22**:207–44.

82. McKay JR, Alterman AI, McLellan AT, Snider EC. Effect of random versus nonrandom assignment in a comparison of inpatient and day hospital rehabilitation for male alcoholics. *J Consult Clin Psychol* 1995;**63**:70–8.

83. McKay JR, Alterman AI, McLellan AT, Boardman CR, Mulvaney FD, O'Brien CP. Random versus nonrandom assignment in the evaluation of treatment for cocaine abusers. *J Consult Clin Psychol* 1998;**66**:697–701.

84. Vermillion JE. *Regression artifacts in nonequivalent control group designs*: *an empirical investigation of bias in 'ANCOVA' and 'matching' designs.* Annual Meeting of the American Educational Research Association. Boston, MA; 1980. pp. 1–24.

85. Westerberg VS, Miller WR, Tonigan JS. Comparison of outcomes for clients in randomized versus open trials of treatment for alcohol use disorders. *J Stud Alcohol* 2000;**61**:720–7.

86. de C Williams AC, Nicholas MK, Richardson PH, Pither CE, Fernandes J. Generalizing from a controlled trial: the effects of patient preference versus randomization on the outcome of inpatient versus outpatient chronic pain management. *Pain* 1999;**83**:57–65.

87. Moyer A, Finney JW. Randomized versus nonrandomized studies of alcohol treatment: participants, methodological features and posttreatment functioning. *J Stud Alcohol* 2002;**63**:542–50.

88. Shadish WR, Heinsman DT. Experiments versus quasi-experiments: do you get the same answer? In: Bukoski WJ, editor. *Meta-analysis of drug abuse prevention programs.* Washington, DC: DHHS Superintendent of Documents; 1997. pp. 147–64.

89. Kahn EB, Ramsey LT, Brownson RC, Heath GW, Howze EH, Powell KE, *et al*. The effectiveness of interventions to increase physical activity: a systematic review. *Am J Prev Med* 2002;**22**(4 Suppl.):73–107.

90. Davis MK, Gidycz CA. Child sexual abuse prevention programs: a meta-analysis. *J Clin Child Psychol* 2000;**29**:257–65.

91. Guyatt GH, DiCenso A, Farewell V, Willan A, Griffith L. Randomized trials versus observational studies in adolescent pregnancy prevention. *J Clin Epidemiol* 2000;**53**:167–74.

92. Jacobs SE, Sokol J, Ohlsson A. The newborn individualized developmental care and assessment program is not supported by meta-analyses of the data. *J Pediatr* 2002;**140**:699–706.

93. Mullen PD, Ramirez G, Strouse D, Hedges LV, Sogolow E. Meta-analysis of the effects of behavioral HIV prevention interventions on the sexual risk behavior of sexually experienced adolescents in controlled studies in the United States. *J Acquir Immune Defic Syndr* 2002;**30**(Suppl. 1):S94–S105.

94. Thomas J, Sutcliffe K, Harden A, Oakley A, Oliver S, Rees R, *et al*. *Children and healthy eating*: *a systematic review of barriers and facilitators.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London; 2003.

95. Tobler NS, Stratton HH. Effectiveness of school-based drug prevention programs: a meta-analysis of the research. *J Prim Prev* 1997;**18**:71–128.

96. Wilson DB, Gottfredson DC, Najaka SS. School-based prevention of problem behaviors: a meta-analysis. *J Quant Criminol* 2001;**17**:247–72.

97. Wilson SJ, Lipsey MW, Derzon JH. The effects of school-based intervention programs on aggressive behavior: a meta-analysis. *J Consult Clin Psychol* 2003;**71**:136–49.

98. Cambach W, Wagenaar RC, Koelman TW, van Keimpema AR, Kemper HC. The long-term effects of pulmonary rehabilitation in patients with asthma and chronic obstructive pulmonary disease: a research synthesis. *Arch Phys Med Rehabil* 1999;**80**:103–11.

99. Cameron I, Crotty M, Currie C, Finnegan T, Gillespie L, Gillespie W, *et al*. Geriatric rehabilitation following fractures in older people: a systematic review. *Health Technol Assess* 2000;**4**:1–111.

100. Griffith JD, Rowan-Szal GA, Roark RR, Simpson DD. Contingency management in outpatient methadone treatment: a meta-analysis. *Drug Alcohol Depend* 2000;**58**:55–66.

101. Kwan J, Sandercock P. *In-hospital care pathways for stroke.* The Cochrane Database of Systematic Reviews, Issue 4; 2004. Art. No.: CD002924. DOI: 10.1002/14651858.CD002924.pub2.

102. Langhorne P, Dennis MS, Kalra L, Shepperd S, Wade DT, Wolfe CDA. *Services for helping acute stroke patients avoid hospital admission.* The Cochrane Database of Systematic Reviews, Issue 3; 1999. Art. No.: CD000444. DOI: 10.1002/14651858.CD000444.

103. Oliver D, Hopper AH, Seed P. Do hospital falls prevention programs work? A systematic review. *J Am Geriatr Soc* 2000;**48**:1679–89.

104. Smedslund G, Fisher KJ, Boles SM, Lichtenstein E. The effectiveness of workplace smoking cessation programmes: a meta-analysis of recent studies. *Tob Control* 2004;**13**:197–204.

105. Wilson DB, Gallagher CA, MacKenzie DL. A meta-analysis of corrections-based education, vocation and work programs for adult offenders. *Journal of Research in Crime and Delinquency* 2000;**37**:347–68.

106. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997;**9**:15–21.

107. Tobler NS, Roona MR, Ochshorn P, Marshall DG, Streke AV, Stackpole KM. School-based adolescent drug prevention programs: 1998 meta-analysis. *J Prim Prev* 2000;**20**:275–336.

108. Bekker H, Thornton JG, Airey CM, Connelly JB, Hewison J, Robinson MB, *et al*. Informed decision making: an annotated bibliography and systematic review. *Health Technol Assess* 1999;**3**:1–156.

109. Bordley WC, Chelminski A, Margolis PA, Kraus R, Szilagyi PG, Vann JJ. The effect of audit and feedback on immunization delivery: a systematic review. *Am J Prev Med* 2000;**18**:343–50.

110. Buller DB, Borland R. Skin cancer prevention for children: a critical review. *Health Educ Behav* 1999;**26**:317–43.

111. Burns T, Knapp M, Catty J, Healey A, Henderson J, Watt H, *et al*. Home treatment for mental health problems: a systematic review. *Health Technol Assess* 2001;**5**:1–139.

112. Chesnut RM, Carney N, Maynard H, Patterson P, Clay Mann N, Helfand M. *Rehabilitation for traumatic brain injury.* Rockville, MD, US: Agency for Health Care Policy and Research; 1999.

113. DiGuiseppi C, Higgins JPT. *Interventions for promoting smoke alarm ownership and function.* The Cochrane Database of Systematic Reviews, Issue 2; 2001. Art. No.: CD002246. DOI: 10.1002/14651858.CD002246.

114. DiGuiseppi C, Higgins JPT. Systematic review of controlled trials of interventions to promote smoke alarms. *Arch Dis Child* 2000;**82**:341–8.

115. Draper B. The effectiveness of old age psychiatry services. *Int J Geriatr Psychiatr* 2000;**15**:687–703.

116. Emmons KM, Wong M, Hammond SK, Velicer WF, Fava JL, Monroe AD, *et al*. Intervention and policy issues related to children's exposure to environmental tobacco smoke. *Prev Med* 2001;**32**:321–31.

117. Fairbank L, O'Meara S, Renfrew MJ, Woolridge M, Sowden AJ, Lister-Sharp D. A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding. *Health Technol Assess* 2000;**4**:1–171.

118. Giuffrida A, Gosden T, Forland F, Kristiansen IS, Sergison M, Leese B, *et al*. *Target payments in primary care*: *effects on professional practice and health care outcomes.* The Cochrane Database of Systematic Reviews, Issue 4; 1999. Art. No.: CD000531. DOI: 10.1002/14651858.CD000531.

119. Gosden T, Forland F, Kristiansen IS, Sutton M, Leese B, Giuffrida A, *et al*. *Capitation, salary, fee-for-service and mixed systems of payment*: *effects on the behaviour of primary care physicians.* The Cochrane Database of Systematic Reviews, Issue 3; 2000. Art. No.: CD002215. DOI: 10.1002/14651858. CD002215.

120. Hulscher M, Wensing M, van der Weijden T, Grol R. *Interventions to implement prevention in primary care.* The Cochrane Database of Systematic Reviews, Issue 2; 2001. Art. No.: CD000362. DOI: 10.1002/14651858.CD000362.pub2.

121. Hutt R, Rosen R, McCauley J. *Case-managing long-term conditions*: *what impact does it have in the treatment of older people?* London: King's Fund; 2004.

122. Karjalainen K, Malmivaara A, van Tulder M, Roine R, Jauhiainen M, Hurri H, *et al*. *Multidisciplinary biopsychosocial rehabilitation for neck and shoulder pain among working age adults.* The Cochrane Database of Systematic Reviews, Issue 2; 2003. Art. No.: CD002194. DOI: 10.1002/14651858.CD002194.

123. Meads C, Gold L, Burls A, Jobanputra P. *In-patient versus out-patient care for eating disorders.* Birmingham: University of Birmingham, Department of Public Health and Epidemiology; 1999.

124. Pusey H, Richards D. A systematic review of the effectiveness of psychosocial interventions for carers of people with dementia. *Aging Ment Health* 2001;**5**:107–19.

125. Reeves S. A systematic review of the effects of interprofessional education on staff involved in the care of adults with mental health problems. *J Psychiatr Ment Health Nurs* 2001;**8**:533–42.

126. Dobbins M, Beyers J. *The effectiveness of community-based heart health projects: a systematic overview update.* Ontario: Ontario Ministry of Health, Region of Hamilton-Wentworth, Social and Public Health Services Division; 1999.

127. Scott JT, Harmsen M, Prictor MJ, Sowden AJ, Watt I. *Interventions for improving communication with children and adolescents about their cancer.* The Cochrane Database of Systematic Reviews, Issue 3; 2003. Art. No.: CD002969. DOI: 10.1002/14651858.CD002969.

128. Stead LF, Lancaster T. *Interventions for preventing tobacco sales to minors.* The Cochrane Database of Systematic Reviews, Issue 1; 2005. Art. No.: CD001497. DOI: 10.1002/14651858.CD001497. pub2.

129. Allaby M, Forman K, Touch S, Chilcott J. *The use of routine antenatal anti-D prophylaxis for rhesus negative women.* Sheffield: University of Sheffield, Trent Institute for Health Services research; 1999.

130. Cuijpers P. Peer-led and adult-led school drug prevention: a meta-analytic comparison. *J Drug Educ* 2002;**32**:107–19.

131. Dusseldorp E, Van Elderen T, Maes S, Meulman J, Kraaij V. A meta-analysis of psychoeducational programs for coronary heart disease patients. *Health Psychol* 1999;**18**:506–19.

132. Elkan R, Kendrick D, Hewitt M, Robinson JJ, Tolley K, Blair M, *et al*. The effectiveness of domiciliary health visiting: a systematic review of international studies and a selective review of the British literature. *Health Technol Assess* 2000;**4**:1–339.

133. Gruen RL, Weeramanthri TS, Knight SE, Bailie RS. *Specialist outreach clinics in primary care and rural hospital settings.* The Cochrane Database of Systematic Reviews, Issue 4; 2003. Art. No.: CD003798. DOI: 10.1002/14651858.CD003798. pub2.

134. Higginson IJ, Finlay I, Goodwin DM, Cook AM, Hood K, Edwards AG, *et al*. Do hospital-based palliative teams improve care for patients or families at the end of life? *J Pain Symptom Manag* 2002;**23**:96–106.

135. Hyde CJ, Robert IE, Sinclair AJ. The effects of supporting discharge from hospital to home in older people. *Age Ageing* 2000;**29**:271–9.

136. Johnson WD, Hedges LV, Diaz RM. *Interventions to modify sexual risk behaviors for preventing HIV infection in men who have sex with men.* The Cochrane Database of Systematic Reviews, Issue 4; 2003. Art. No.: CD001230. DOI: 10.1002/14651858. CD001230.

137. Kendrick D, Elkan R, Hewitt M, Dewey M, Blair M, Robinson J, *et al*. Does home visiting

improve parenting and the quality of the home environment? A systematic review and meta analysis. *Arch Dis Child* 2000;**82**:443–51.

138. Legler J, Meissner HI, Coyne C, Breen N, Chollette V, Rimer BK. The effectiveness of interventions to promote mammography among women with historically lower rates of screening. *Canc Epidemiol Biomarkers Prev* 2002;**11**:59–71.

139. Monninkhof EM, van der Valk PDLP, van der Palen J, van Herwaarden CLA, Partidge MR, Walters EH, *et al*. *Self-management education for chronic obstructive pulmonary disease.* The Cochrane Database of Systematic Reviews, Issue 4; 2002. Art. No.: CD002990. DOI: 10.1002/14651858.CD002990.

140. Parker G, Bhakta P, Katbamna S, Lovett C, Paisley S, Parker S, *et al*. Best place of care for older people after acute and during subacute illness: a systematic review. *J Health Serv Res Pol* 2000;**5**:176–89.

141. Posavac EJ, Kattapong KR, Dew DE. Peer-based interventions to influence health-related behaviors and attitudes: a meta-analysis. *Psychol Rep* 1999;**85**:1179–94.

142. Prendergast ML, Podus D, Chang E. Program factors and treatment outcomes in drug dependence treatment: an examination using meta-analysis. *Subst Use Misuse* 2000;**35**:1931–65.

143. Szilagyi P, Vann J, Bordley C, Chelminski A, Kraus R, Margolis P, *et al*. *Interventions aimed at improving immunization rates.* The Cochrane Database of Systematic Reviews, Issue 4; 2002. Art No.: CD003941. DOI: 10.1002/14651858.CD003941.

144. Thomson OMA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. *Educational outreach visits*: *effects on professional practice and health care outcomes.* The Cochrane Database of Systematic Reviews, Issue 4; 1997. Art No.: CD000409. DOI: 10.1002/14651858.CD000409.

145. Yabroff KR, O'Malley A, Mangan P, Mandelblatt J. Inreach and outreach interventions to improve mammography use. *J Am Med Wom Assoc* 2001;**56**:166–73.

146. Yin T, Zhou Q, Bashford C. Burden on family members. Caring for frail elderly: a meta-analysis of interventions. *Nurs Res* 2002;**51**:199–208.

147. Ziguras SJ, Stuart GW. A meta-analysis of the effectiveness of mental health case management over 20 years. *Psychiatr Serv* 2000;**51**:1410–21.

148. Reeves BC, MacLehose RR, Harvey IM, Sheldon TA, Russell IT, Black AMS. Comparisons of effect size estimates derived from randomised and non-randomised studies. In: Black N, editor. *Health Services Research Methods*: *a guide to best practice.* London: BMJ Publishing Group; 1998. pp. 73–85.

149. Peersman G, Oliver S. *EPPI Centre Keywording Strategy*: *Data Collection for the Bibliomap Database.* London: EPPI Centre, Social Science Research Unit; 1997.

150. White IR, Thomas J. Standardised mean differences in individually-randomised and cluster-randomised trials, with applications to meta-analysis. *Clin Trials* 2005;**2**:141–51.

151. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A. Making psychological theory useful for implementing evidence based practice: a consensus approach. *Qual Saf Health Care* 2005;**14**:26–33.

152. Harden A, Peersman G, Oliver S, Mauthner M, Oakley A. A systematic review of the effectiveness of health promotion interventions in the workplace. *Occup Med* 1999;**49**:1–9.

153. Harden A, Oakley A, Oliver S. Peer-delivered health promotion for young people: a systematic review of different study designs. *Health Educ J* 2001;**60**:339–53.

154. Sheperd J, Weston R, Peersman G, Napuli IZ. *Interventions for encouraging sexual lifestyles and behaviours intended to prevent cervical cancer.* 1994. Art. No.: CD001035. DOI: 10.1002/14651858.CD001035.

155. Harden A, Rees R, Shepherd J, Brunton G, Oliver S, Oakley A, *et al*. *Young people and mental health*: *a systematic review of research on barriers and facilitators.* London: EPPI Centre; 2001.

156. Brunton G, Harden A, Rees R, Kavanagh J, Oliver S, Oakley A. *Children and physical activity*: *a systematic review of barriers and facilitators.* London: EPPI Centre, Social Science Research Unit, Institute of Education; 2003.

157. Oakley A, Oliver S, Peersman G, Mauthner M. *Review of effectiveness of health promotion interventions for men who have sex with men.* London: Social Science Research Unit, Institute of Education, University of London; 1996.

158. Klassen TP, MacKay JM, Moher D, Walker A, Jones AL. Community-based injury prevention interventions. *Future Child* 2000;**10**:83–110.

159. Linton SJ, van Tulder MW. Preventive interventions for back and neck pain problems: what is the evidence? *Spine* 2001;**26**:778–87.

160. Weigand JV, Gerson LW. Preventive care in the emergency department: should emergency departments institute a falls prevention program for elder patients? A systematic review. *Acad Emerg Med* 2001;**8**:823–6.

161. Weir RP. *Rehabilitation of cerebrovascular disorder (stroke): early discharge and support: a critical review of the literature*. Christchurch: New Zealand Health Technology Assessment (NZHTA); 1999.

162. Thompson O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J. *Continuing education meetings and workshops: effects on professional practice and health care outcomes (review)*. The Cochrane Database of Systematic Reviews; 2001. Art.No.: CD003030. DOI: 10.1002/14651858.CD003030.

Appendix 1
Complex interventions

From the MRC 2000. A framework for development and evaluation of RCTs for complex interventions to improve health.

What is a complex intervention?

Complex interventions are built up from a number of components, which may act both independently and interdependently. The components usually include behaviours, parameters of behaviours (e.g. frequency, timing) and methods of organising and delivering those behaviours [e.g. type(s) of practitioner, setting and location]. It is not easy to define precisely the 'active ingredients' of a complex intervention. For example, although research suggests that stroke units work, what, exactly, is a stroke unit? What are the active ingredients that make it work? The physical setup? The mix of care providers? The skills of the providers? The technologies available? The organisational arrangements?

Health services have to evaluate a wide array of existing and newly proposed complex packages, so that the service can learn what is effective about any given intervention so that it can be more widely applied throughout the service. Some complex interventions are intended as improvements in the form of direct interventions at the level of *individual patient care*; for example, a novel form of cognitive behavioural therapy. Other interventions, although ultimately intended to improve patient care, are actually delivered in the form of an *organisational or service modification*; for example, the introduction of a physiotherapist or Parkinson's disease nurse into primary care services. A third type of complex intervention is further removed again from individual patient care, although ultimately intended just as much to impact there, when an intervention is *targeted on the health professional*; for example, educational interventions in the form of treatment guidelines, protocols or decision-aids. Finally, many complex interventions are delivered *at a population level*; for example, in the form of media-delivered health promotion campaigns.

Appendix 2

In-house review abstracts

Effectiveness of interventions in the workplace: a review

This systematic review aimed to evaluate the potential for using the workplace as a setting for improving adult health.[152] This was seen as an important potential argument owing to the large number of people it would be possible to access, the probable high levels of participations and peer support, and the likely low level of attrition.

Most studies identified were targeted at individuals with varying degrees of environmental modification. No clear evidence of effectiveness was found for type of intervention, topic of intervention or interventions delivered by a particular category of people. Trends of effectiveness were found for comprehensive programmes that combined screening and risk assessment with a range of education programmes, and/or environmental changes. However, these were found in few studies, so replicability cannot be relied upon. Least effective were the weight-loss programmes combining education and financial incentives. No conclusive evidence was found for the effectiveness of peer support.

Due to the general low level of methodological evaluation, data supporting workplace site interventions are not definitive. However, suggestions were made for future research. These included the suggestion that employees should be involved at all levels in the planning and implementation of the activity, the intervention should be supported by top management should be tailor-made to the characteristics of the group. Finally, the quality of reporting should be higher; evaluation in particular should be included in the interventions and a range of outcome measures should be included.

A review of the effectiveness and appropriateness of peer-delivered health promotion interventions for young people

This systematic review synthesised evidence to examine the claim that the peer-delivered approach is a more appropriate an effective method of promoting young people's health (aged 11–24 years) than more traditional approaches.[153]

The most common focus for the outcome evaluations was drugs (including alcohol and smoking), and for the process evaluations it was sexual health. Of the methodologically sound interventions most found clear effectiveness for behavioural outcomes. Studies comparing peers and teachers found equivocal results with regard to effectiveness.

Overall, the review found some evidence to support the effectiveness of peer-delivered health promotion for young people. More than half of the sound studies showed a positive effect at least on behavioural outcome. However, this may be in part because of the scarcity of sounds studies, and lack of good reporting. This report does not encourage peer-delivered health promotion, because there is relatively little sound evidence to support this intuitively appealing idea.

Some suggestions made for future research included implementing health promotion on the basis of a thorough assessment of both self-defined health needs and young people's views on what would be most effective.

Interventions for encouraging sexual lifestyles and behaviours intended to prevent cervical cancer

This review aimed to determine the effectiveness of health education interventions to promote sexual risk reduction behaviours among women in order to reduce transmission of human papillomavirus.[154] Studies were included if they evaluated educational interventions targeting women only, and measured the impact on either behavioural or clinical outcomes. This was the first review to address cervical cancer prevention in terms of sexual behaviour risk reduction.

All of these included outcome evaluations had the primary aim of preventing HIV and other sexually transmitted diseases rather than preventing cervical cancer. Each of the methodologically sound studies showed a statistically significant positive effect on sexual risk reduction, typically with increased use of condoms for vaginal intercourse.

This positive effect was found to be sustained up to 3 months after intervention.

The review drew the following conclusions: that educational interventions targeting socially and economically disadvantaged women including sexual negotiation skill development encouraged at least short-term sexual risk reduction behaviour. This has the potential to reduce the transmission of human papillomavirus and thus possibly reduce the incidence of cervical carcinoma. Health education interventions in which factual information was presented alongside skill development and motivation building was found to achieve short-term increases in reported condom use for vaginal intercourse.

Suggestions for further research included the need for greater attention to gender and culture issues, for interventions need to be sensitive to local culture and context in order to enable women to identify with the health education messages, and that interventions that address power imbalances in relationships are essential for successful implementation. It was also suggested that longer term interventions may show greater effects.

Young people and mental health: a systematic review of research on barriers and facilitators

This systematic review aimed to synthesise evidence on barriers and facilitators of good mental health in young people (aged 11–21 years).[155] A search of systematic reviews was carried out. The findings from this suggested that interventions to promote self-esteem, and those to prevent suicide, were limited in their effectiveness. Some effective interventions were those that addressed young people's concerns about teachers, parental divorce, bereavement and peer rejection.

The in-depth review of outcome evaluations also revealed limited effects for the promotion of self-esteem. Interventions targeting depression also showed no long-term effects, despite some short-term increases in knowledge about symptoms. The in-depth review of young people's views revealed some unexpected findings: that mental health tended to be equated with mental illness and thus not be relevant to the respondents; that young people had surprisingly sophisticated understandings of coping strategies and had wide concerns about mental health. The findings also suggested how irrelevant many health promotion materials are to young people's worries.

The synthesis found some major gaps in the research, including interventions that address young people's concerns about workload, academic achievement, future unemployment, violence and bullying, and physical appearance (among others). These were suggested as possible future directions for research.

Young people and physical activity: a systematic review of research on barriers and facilitators

This systematic review aimed to synthesise the evidence assessing barriers to and facilitators of physical activity among young people (aged 11–16 years), especially those pertaining to socially excluded groups.[28] This was deemed an important subpopulation as particularly low levels of physical activity, which have been linked with other health-damaging behaviours, are found within this group.

The in-depth review, following a wide mapping stage, comprised 12 outcome evaluations which assessed the effect on health behaviour of interventions aimed at a community/society level. Most were delivered in a school or other educational setting by teachers.

A review of views' studies was also conducted. Some important barriers to physical activity identified were those related to the self and other people (e.g. incompetence, self-consciousness), practical and material resources/circumstances such as lack of money, and the school (negative physical education teachers). Facilitators identified included activity being good for losing weight, and parental support. Many suggestions were made about how to increase levels of activity by young people, like making activities more affordable and emphasising their 'fun' side.

These views and a cross-studies synthesis identified many matches between concerns and evaluations. A need was identified for greater concentration on non-traditional activities, such as aerobics, and evaluations specifically targeting young women.

Young people and healthy eating: a systematic review of research on barriers and facilitators

This systematic review aimed to synthesise the evidence from outcome evaluations of interventions, and from the views of young people (aged 11–21 years) to inform readers on barriers to and facilitators of healthy eating.[27]

Most evaluated interventions were carried out in school, so potentially a large proportion of socially excluded young people were missed. This subpopulation was specifically addressed by less than a quarter of identified studies.

The in-depth review focused on interventions aiming to make a change at a community or society level. Most were based in primary or secondary school settings and were delivered by teachers. Some school-based settings were found to be effective at increasing knowledge and improving health behaviour.

The views' study identified the following barriers to healthy eating: costs and wide availability of/preference for fast food. Facilitators were given as a reduction in costs of healthy food, better availability of healthy food and family support, among others.

Most of the gaps identified in the research by the synthesis were found to have been addressed. However, a need for better nutrition information was suggested, and in general more rigorous research in this area is needed.

Children and physical activity: a systematic review of barriers and facilitators

This systematic review aimed to describe the number, types and quality attributes of existing research studies on the barriers to and facilitators of physical activity among children aged 4–10 years.[156] Both a views' study and an evaluation of relevant interventions were carried out.

The interventions investigated were extremely diverse, making it difficult to assess patterns of effectiveness. Most were school-based, all involved parents to varying degrees, but some aimed to tackle sedentary behaviour, and others to increase participation in physical activity.

Studies on children's views about physical activity were scarce. The five that were found focused on barriers to physical activity for children. Twenty distinct barriers were identified, which followed three main themes: preferences and priorities (e.g. a preference for doing other things), family life and parental support, and restricted access to opportunities for participation. Fourteen facilitators to physical activity were also identified, which clustered around the following themes: aspects of physical activity that children value, family life and parental support, and greater access to opportunities for participating in physical activity.

The synthesis found that few health promotion evaluations targeted physical activity outside the physical education lesson, and that children's views rarely informed the development of interventions. These were identified as possible directions for future research.

Children and healthy eating: a systematic review of barriers and facilitators

This systematic review aimed to find out what was known about barriers to and facilitators of healthy eating in children aged 4–10 years.[94] The review was the first systematic review to rigorously integrate the findings from a meta-analysis with a qualitative systematic review.

The in-depth review focused on the barriers to and facilitators of children's consumption of fruit and vegetables. Studies that measured fruit and vegetable outcomes were included, as were those studies that examined children's own perspectives on food and eating.

A substantial amount of research was identified. The relevant outcome evaluations were largely school-based, and often combined learning about the health benefits of fruit and vegetables with hands-on experience. Some interventions also included types of environmental modification, and some targeted multiple outcomes (for instance, body mass index, knowledge, fat intake, physical activity.) These types of interventions were found to have a small but significant positive effect. The larger effect sizes were associated with targeted interventions for parents with risk factors for cardiovascular disease, and with those that focused purely on trying to increase fruit and vegetable increase, rather than 'diluting' the effect by promoting, for example, physical activity.

The main messages from this section were that promoting health eating can be an integral part of a school curriculum, and that effective implementation requires skill, time and support from a wide range of people.

The views' studies comprised eight studies involving children aged 5–11 years and their mothers. From these analyses, the following issues were identified: children do not see it as their role to be interested in health; children do not see health-related messages as relevant or

credible; children understand fruit, vegetables and confectionary very differently; children want to exercise choice over their food, eating is valued as a social occasion, and children note the distinction between advice proffered and observed adult behaviour. The synthesis of views and outcome evaluations revealed that opportunities for developing interventions to increase children's consumption of fruit and vegetables included branding fruit and vegetables as tasty, rather than healthy, and making health messages credible for children.

Suggestions for further research included the need to target interventions towards socially excluded groups and reducing health inequality in general.

HIV health promotion and men who have sex with men: a systematic review of research relevant to the development and implementation of effective and appropriate interventions

This review aims to systematically pull together findings from studies of MSM's views and integrate them with findings from effectiveness studies.[157] The views' studies focused especially on young men (aged 16–25 years), men who sell sex to other men and HIV-positive men. The outcome evaluations all had a comparison or control group, discussed interventions delivered during or after 1996 and measured as the outcome of most importance sero-discordant or unknown status unprotected anal intercourse (sdUAI).

The meta-analysis of the outcome evaluations suggested that counselling or workshops based on cognitive–behavioural techniques for MSM at high risk appear to be more effective at reducing the number of men reporting sdUAI than standard counselling. However, this effect was only found where men were recruited from clinic attendance lists, as opposed to adverts or outreach. The narrative synthesis found no evidence for any effect of interventions targeting sdUAI, although none of the evaluations reported knowledge/awareness/attitudes at sufficient quality to be able to assess these effects. Peer-delivered community-based interventions appeared to have no effect on sdUAI.

Suggestions for future research included supporting future interventions such as counselling based on cognitive–behavioural techniques, and that further rigorously conducted and reported primary and secondary research was required on views of all groups of MSM.

Appendix 3
Search strategies

Literature search to identify methodological studies comparing RCTs with nRCTs beyond health (see Chapter 4)

It was felt that, although the research team had an extensive knowledge of the existing methodological literature comparing the results of RCTs with nRCTs in health, not enough was known about similar methodological studies in the non-health field (social care, education, criminal justice, housing, etc.). An attempt was therefore made to search non-health databases to identify this methodological literature.

Database searches

Although the searches appeared to be fairly straightforward methodological research comparing two types of study design, a number of problems were encountered.

In the first instance, searching by study design is problematic. Indexing of RCTs in MEDLINE by publication type and medical subject heading has improved in recent years, allowing for more accurate searching, but there is still room for improvement. Searching for NRSs is more difficult. There are many study designs that could be classed as non-randomised, and there is as yet little definitive terminology. It is difficult for the indexer to be sure what type of study design has been used, and so comprehensive and consistent indexing according to study design is lacking. These complications are a major issue in MEDLINE; databases beyond MEDLINE often have poorer indexing by study design. These problems of definition and indexing become more pronounced in databases that are non-health related. As well as poor indexing, there is rarely a thesaurus of keyword terms included, the records often lack an abstract, and a number of databases have only rudimentary search capabilities.

Initial attempts at searching for a combination of terms (indexed and free-text) for RCT with nRCT proved unsuccessful. The number of records retrieved was unwieldy, and the percentage of useful studies within these results was very small.

A third search facet was introduced for 'outcomes' (bias, effect size, overestimated results, etc.) in an attempt to refine the strategy. This helped a little but not significantly.

The final search strategy used few terms for RCT (and did not include broader terms for 'study design/methodology'), reduced the non-random terms by rejecting actual study design types (observational, longitudinal, cohort study, cross-sectional, case series, etc.) and used simply 'non-random' and equivalent terms (non-experimental, pseudorandom, semi-random, etc.) alongside a precise 'outcomes' facet, where fewer search terms were used or aligned using restrictive proximity operators.

It was decided that a more sensitive search using 'non-random' study design names, and broader terms for research design should also be completed. This set of results could be referred to in the event that little of relevance was identified in the precise search results.

The following databases were searched:

- ASSIA
- AEI
- BEI
- CareData
- Dissertation Abstracts
- EconLIT
- ERIC
- IBSS
- ISI Proceedings: Social Sciences and Humanities
- PAIS International
- PsycINFO
- SIGLE
- SSCI
- Sociological Abstracts

Additional searches

Citation searching was undertaken in SSCI for authors known to the research team. Internet searches were also undertaken. It was expected that there would be more success looking at potentially relevant sites associated with research methodology and prominent sites in the non-

health area. In most cases any procedural guides identified referred to broader analysis of research, recommending the type of research design to be used, and debating whether or not to use randomisation. The public catalogues of the British Library and the Library of Congress were searched briefly. Finally, a search of the Internet using general search engines (Copernic and Google) and the information gateway (SOSIG; Social Science Information Gateway) was undertaken. These searches were fairly restricted, as any attempt at combining RCT terms with nRCT terms inevitably found health- and clinical-related sites and studies.

Literature search to identify systematic reviews with RCTs and nRCTs in the non-health-care setting (see Chapter 5)

It was decided that it would be easier to retrieve all of the references available on review specific databases rather than attempt any searches for reviews with both RCTs and nRCTs included. All of the reviews available on the following databases were obtained: DARE, CDSR, DoPHER and the Health Development Agency Evidence Base database.

There are probably even less definitive terms available for 'systematic reviews' or equivalent in the non-health literature than there are for trial designs. Ideally, to ensure that the searches retrieved relevant references, the strategy would have included terms for 'review' or at least for 'literature review'. However, attempts at searching using these terms produced an unmanageable number of references.

A number of test searches were completed in broader non-health databases to estimate the feasibility of attempting a comprehensive search of the non-health literature for systematic reviews. The searches were restricted to specific terms for 'systematic review', without using proximity operators, and were further restricted by date range (2003–4).

The following free-text terms and indexed keywords (if available) were used: meta analysis, metaanalysis, systematic review, systematic overview, collaborative review, integrative research, integrative review, research integration, narrative synthesis, evaluation synthesis, meta synthesis, realist synthesis, descriptive synthesis, explanatory synthesis, pool data.

The following databases were searched:

- ASSIA
- BEI
- CareData
- ERIC
- HDA HealthPromis
- PAIS International
- SIGLE
- SSCI
- Sociological Abstracts

The results of the test searches confirmed that it would not be worthwhile conducting thorough, comprehensive searches of non-health databases for systematic reviews.

Searches

The databases searched are listed below with the dates searched and the number of records retrieved. The search strategies used in ASSIA have been listed in full. Full details of the other search strategies used are available from the Centre for Reviews and Dissemination at the University of York.

Methodological studies (see Chapter 4)

Precise methodological searches

ASSIA: Cambridge Scientific Abstracts (CSA). 1987–2004. 28 June 2004

The ASSIA search covered the date range 1987–2004. The search identified 96 records.

((KW = (experiment* group*) or (true experiment*) or (experiment* study) or (experiment* trial*) or (experiment* control*) or (random* or RCT* or GRT*) or DE = (randomized controlled trials)) and ((KW = (non-experiment* or quasiexperiment* or quasi-experiment*) or (quasi-random* or quasirandom* or nonexperiment*) or (quasi-random* or semi-random* or semirandom*) or (non-random* or nonrandom* or pseudo-random)) and ((KW = (error* or confound*) or (overexaggerate* or validity or variable*) or (inaccurate or inaccuracy or exaggerate*) or (reliable or reliability or accuracy) or (disparity or discrepanc* or accurate) or (overestimate* or underestimate* or deviat*) or (outcome* or compar* or estimat*) or (bias* or size* or reliability))

AEI: Dialog. 1976–2004/3. 29 June 2004

The AEI search covered the date range 1976 to March 2004. The search identified six records.

BEI: Dialog. 1976–2004/3. 29 June 2004

The BEI search covered the date range 1976 to March 2004. The search identified 33 records.

CareData. Internet. 2 July 2004

www.elsc.org.uk/caredata/caredata.htm

CareData produced 25 unique records. The search interface available for CareData does not allow for sophisticated search strategies. A first attempt to search using the keyword 'Research methods' identified over 1500 references. A second attempt to combine terms and search in the abstract field only retrieved eight references. Finally, a search for 'random*' was undertaken in the abstract field and the resultant 469 references browsed for any of potential usefulness.

Dissertation Abstracts. Internet. 2002–4. 2 July 2004

wwwlib.global.umi.com/dissertations/

Dissertation Abstracts was searched using simple phrase searches for 'random' and 'non-random' and identified no records.

EconLIT: Ovid WebSPIRS. 1969–2004/5. 2 July 2004

The EconLIT search covered the date range 1969 to May 2004. The search identified seven records.

ERIC: Dialog. 1966–2004/3. 28 June 2004

The ERIC search covered the date range 1966 to March 2004. The search identified 93 records.

IBSS: BIDS. 1951–2004. 2 July 2004

The IBSS search covered the date range 1951–2004. The search identified 13 records.

ISI Proceedings: Social Sciences and Humanities. ISI Web of Knowledge. 1990–2004/6. 30 June 2004

The Social Sciences and Humanities Proceedings search covered the date range 1990 to June 2004. The search identified 48 records.

PAIS International: Ovid WebSPIRS. 1972–2004/4. 2 July 2004

The PAIS search covered the date range 1972 to April 2004. The search identified one record.

PsycINFO: BIDS. Internet. 1872–2004/6. 30 June 2004

The PsycINFO search covered the date range 1872 to June 2004. The search identified 33 records.

SIGLE: Ovid WebSPIRS. 1980–2003/12. 2 July 2004

The SIGLE search covered the date range 1980–2003. Zero records were identified.

SSCI: Web of Science. 1981–2004/6. 29 June 2004

The SSCI search covered the date range 1981 to June 2004. The search identified 107 records.

Sociological Abstracts: WebSPIRS. 1963–2004/6. 2 July 2004

The Sociological Abstracts search covered the date range 1963 to June 2004. The search identified 99 records.

Sensitive methodological searches

The sensitive searches sometimes used two strategies separately. This enabled the use of more nRCT terms in one strategy, and the broader 'research design' terms in combination with a more precise 'outcomes' facet.

ASSIA: CSA. 1987–2004. 5 July 2004

The ASSIA searches covered the date range 1987–2004. The searches identified 371 records. Two search strategies were devised: one used broader terms for nRCTs and combined the RCT, nRCT and outcomes facets; the second strategy combined 'research design' terms with a more precise 'outcomes' facet. The results were then combined and duplicate records removed.

((KW = (study design*) or (study type*) or (study method*) or (trial* design*) or (trial* type*) or (trial* method*) or (experiment* design*) or (experiment* type*) or (experiment* method*) or (research design*) or (research type*) or (research method*)) and ((KW = (validity within 2 efficac*) or (validity within 2 size*) or (validity within 2 therap*) or (validity within 2 benefit*) or (validity within 2 effect*) or (validity within 2 impact*) or (validity within 2 outcome*) or (validity within 2 treatment*) or (validity within 2 finding*) or (validity within 2 evidence*) or (validity within 2 harm*) or (validity within 2 bias*) or (validity within 2 error*) or (validity within 2 result*) or (compar* within 2 efficac*) or (compar* within 2 size*) or (compar* within 2 therap*) or (compar* within 2 benefit*) or

(compar* within 2 effect*) or (compar* within 2 impact*) or (compar* within 2 outcome*) or (compar* within 2 treatment*) or (compar* within 2 finding*) or (compar* within 2 evidence*) or (compar* within 2 harm*) or (evaluat* within 2 efficac*) or (evaluat* within 2 size*) or (evaluat* within 2 therap*) or (evaluat* within 2 benefit*) or (evaluat* within 2 effect*) or (evaluat* within 2 impact*) or (evaluat* within 2 outcome*) or (evaluat* within 2 treatment*) or (evaluat* within 2 finding*) or (evaluat* within 2 evidence*) or (evaluat* within 2 harm*) or (evaluat* within 2 bias*) or (evaluat* within 2 error*) or (evaluat* within 2 result*) or (exaggerate* within 2 efficac*) or (exaggerate* within 2 size*) or (exaggerate* within 2 therap*) or (exaggerate* within 2 benefit*) or (exaggerate* within 2 effect*) or (exaggerate* within 2 impact*) or (exaggerate* within 2 outcome*) or (exaggerate* within 2 treatment*) or (exaggerate* within 2 finding*) or (exaggerate* within 2 evidence*) or (exaggerate* within 2 harm*) or (inaccura* within 2 efficac*) or (inaccura* within 2 size*) or (inaccura* within 2 impact*) or (inaccura* within 2 outcome*) or (inaccura* within 2 treatment*) or (inaccura* within 2 finding*) or (inaccura* within 2 evidence*) or (inaccura* within 2 harm*) or (inaccura* within 2 bias*) or (inaccura* within 2 error*) or (inaccura* within 2 result*) or (accura* within 2 finding*) or (accura* within 2 evidence*) or (accura* within 2 harm*) or (accura* within 2 efficac*) or (accura* within 2 size*) or (accura* within 2 therap*) or (accura* within 2 benefit*) or (accura* within 2 effect*) or (accura* within 2 impact*) or (accura* within 2 outcome*) or (accura* within 2 treatment*) or (accura* within 2 bias*) or (accura* within 2 error*) or (accura* within 2 result*) or (reliab* within 2 efficac*) or (reliab* within 2 size*) or (reliab* within 2 therap*) or (reliab* within 2 benefit*) or (reliab* within 2 effect*) or (reliab* within 2 impact*) or (reliab* within 2 outcome*) or (reliab* within 2 treatment*) or (reliab* within 2 finding*) or (reliab* within 2 evidence*) or (reliab* within 2 harm*) or (reliab* within 2 bias*) or (reliab* within 2 error*) or (reliab* within 2 result*) or (difference* within 2 efficac*) or (difference* within 2 size*) or (difference* within 2 therap*) or (difference* within 2 benefit*) or (difference* within 2 effect*) or (difference* within 2 impact*) or (difference* within 2 outcome*) or (difference* within 2 treatment*) or (difference* within 2 finding*) or (difference* within 2 evidence*) or (difference* within 2 harm*) or (difference* within 2 bias*) or (difference* within 2 error*) or (difference* within 2 result*) or (underestimate* within 2 efficac* within 2 therap* within 2 impact* within 2 finding* within 2 bias*) or (underestimate* within 2 size* within 2 benefit* within 2 outcome* within 2 evidence* within 2 error*) or (* within 2 effect* within 2 treatment* within 2 harm* within 2 result*) or (underestimate* within 2 therap* within 2 impact* within 2 finding* within 2 bias*) or (underestimate* within 2 benefit* within 2 outcome* within 2 evidence* within 2 error*) or (underestimate* within 2 effect* within 2 treatment* within 2 harm* within 2 result*) or (underestimate* within 2 impact* within 2 finding* within 2 bias*) or (underestimate* within 2 outcome* within 2 evidence* within 2 error*) or (underestimate* within 2 treatment* within 2 harm* within 2 result*) or (underestimate* within 2 finding* within 2 bias*) or (underestimate* within 2 evidence* within 2 error*) or (underestimate* within 2 harm* within 2 result*) or (underestimate* within 2 bias*) or (underestimate* within 2 error*) or (underestimate* within 2 result*) or (overestimate* within 2 efficac* within 2 therap* within 2 impact*) or (overestimate* within 2 size* within 2 benefit* within 2 outcome* within 2 evidence* within 2 error*) or (overestimate* within 2 therap* within 2 impact*) or (overestimate* within 2 benefit* within 2 outcome* within 2 evidence* within 2 error*) or (overestimate* within 2 effect* within 2 treatment* within 2 harm* within 2 result*) or (overestimate* within 2 impact*) or (overestimate* within 2 outcome* within 2 evidence* within 2 error*) or (overestimate* within 2 treatment* within 2 harm* within 2 result*) or (overestimate* within 2 finding*) or (overestimate* within 2 evidence* within 2 error*) or (overestimate* within 2 harm* within 2 result*) or (overestimate* within 2 bias*) or (overestimate* within 2 error*) or (overestimate* within 2 result*) or (estimate* within 2 bias*) or (estimate* within 2 error*) or (compar* within 2 random*) or (estimate* within 2 result*) or (estimate* within 2 finding*) or (estimate* within 2 evidence*) or (estimate* within 2 harm*) or (estimate* within 2 impact*) or (estimate* within 2 outcome*) or (estimate* within 2 treatment*) or (estimate* within 2 therap*) or (estimate* within 2 benefit*) or (estimate* within 2 effect*) or (estimate* within 2 efficac*) or (estimate* within 2 size*))

((KW = (true experiment*) or (random* or RCT* or GRT*) or DE = (randomized controlled trials)) and ((KW = (matched pair*) or (paired comparison) or (single case) or (single-case) or (time series) or time-series) or (cross sectional) or (cross-sectional) or (cross over*) or (cross-over*)

or (case stud*) or (case-stud*) or (case control*) or (case-control*) or (historic* control*) or (retrospective or prospective) or (observational or cohort or longitudinal) or (pre post) or (pre-post) or (prepost) or (post test) or (post-test) or (posttest) or (pre test) or (pre-test) or (pretest) or (natural experiment*) or (quasi experiment*) or (quasi-experiment*) or (quasiexperiment*) or (non experiment*) or (non-experiment*) or (nonexperiment*) or (without within 3 random*) or (pseudo random*) or (pseudo-random*) or (pseudorandom*) or (semi random*) or (semi-random*) or (semirandom*) or (quasi random*) or (quasi-random*) or (quasirandom*) or (non random*) or (non-random*) or (nonrandom*) or DE = (cross-sectional studies) or (retrospective studies) or (case studies) or (case controlled studies) or (historical analysis) or (prospective controlled trials) or (observational research) or (cohort analysis) or (longitudinal studies)) and ((KW = confound* or (treatment within 2 effect*) or (treatment within 2 estimate*) or (treatment within 2 size*) or (variable*) or (sample size*) or (statistically significan*) or (error* or (measurement*) or (validity) or (exaggerate*) or (overexaggerate*) or (accuracy) or (inaccurate) or (inaccuracy) or (reliability) or (reliable) or (accurate) or (discrepanc*) or (disparity) or (deviat*) or (underestimate*) or (estimate*) or (difference*) or (characteristic*) or (estimate*) or (overestimate*) or (compar* within 3 evidence) or (compar* within 3 effect*) or (bias*) or (outcome*) or (effect* size*) or DE = (confounding factors) or (outcomes) or (estimates) or (bias) or (effect size) or (reliability))

AEI: Dialog. 1976–2004/3. 5 July 2004
The AEI searches covered the date range 1976 to March 2004. The search identified 69 records.

BEI: Dialog. 1976–2004/3. 5 July 2004
The BEI search covered the date range 1976 to March 2004. The search identified 73 records. The original 'precise' search strategy for BEI was quite sensitive, and so only the 'research design' strategy was required for the 'sensitive' searches.

CareData. Internet. 5 July 2004
www.elsc.org.uk/caredata/caredata.htm

The original 'precise' search of CareData involved a simple sensitive single word search, and so a further 'sensitive' search was not required.

Dissertation Abstracts. Internet. 2002–4. 5 July 2004

wwwlib.global.umi.com/dissertations/

The original 'precise' search of Dissertation Abstracts involved a simple sensitive single word search, and so a further 'sensitive' search was not required.

EconLIT: Ovid WebSPIRS. 1969–2004/5. 5 July 2004
The EconLIT search covered the date range 1969 to May 2004. The search identified 99 records.

ERIC: Dialog. 1966–2004/3. 5 July 2004
The ERIC search covered the date range 1966 to March 2004. The search identified 1507 records.

IBSS: BIDS. 1951–2004. 5 July 2004
The IBSS searches covered the date range 1951–2004. The search identified 263 records.

ISI Proceedings: Social Sciences and Humanities. ISI Web of Knowledge. 1990–2004/6. 5th July 2004
The Social Sciences and Humanities Proceedings searches covered the date range 1990 to June 2004. The search identified 462 records.

PAIS International**: Ovid WebSPIRS. 1972–2004/4. 5 July 2004**
The PAIS search covered the date range 1972 to April 2004. The search identified seven records.

PsycINFO: BIDS. 1872–2004/6. 5 July 2004
The PsycINFO search covered the date range 1872 to June 2004. The search identified 1661 records.

SIGLE: Ovid WebSPIRS. 1980–2003/12. 5 July 2004
The SIGLE search covered the date range 1980 to December 2003. The search identified 16 records.

SSCI: Web of Science. 1981–2004/6. 5 July 2004.
The SSCI search covered the date range 1981 to June 2004. The search identified 181 records.

Sociological Abstracts: WebSPIRS. 1963–2004/6. 5 July 2004.
The Sociological Abstracts search covered the date range 1963 to June 2004. The search identified 231 records.

Internet searches

The following internet sites were searched.

electronic Library for Social Care (eLSC)
7 July 2004.
www.elsc.org.uk/

Social Care Institute for Excellence (SCIE)
7 July 2004.
www.scie.org.uk/

Centre for Evidence-Based Social Services
University of Exeter. 7 July 2004.
www.ex.ac.uk/cebss/

SOSIG
7 July 2004.
www.sosig.ac.uk/

Regard. Economic and Social Research Council (ESRC)
7 July 2004.
www.regard.ac.uk/regard/home/index_html?

The Campbell Collaboration
7 July 2004.
www.campbellcollaboration.org/

British Library Public Catalogue
7 July 2004.
blpc.bl.uk/

Library of Congress Online Catalog
7 July 2004.
www.loc.gov/

Copernic (meta-search engine)
7 July 2004.
www.copernic.com

Google (general search engine)
7 July 2004.
www.google.com/

Systematic review test searches (see Chapter 5)

ASSIA: CSA. 2003–2004. 20 July 2004

The ASSIA search covered the date range 2003–4. The precise search without 'literature review' identified 303 records. With 'literature review' the search identified 4387. (Without date limits the precise search identified 1885.)

(Pool* data) or (Realist synth*) or (Descriptive synth*) or (Explanatory synth*) or (Narrative synth*) or (Evaluation synth*) or (Meta synth*) or (Integrative research) or (Integrative review*) or (Research integration) or (Collaborative review*) or (Collaborative review*) or (Systematic review*) or (Systematic overview*) or (Meta analy*) or Metaanly* or Metanaly* or KW = ((systematic reviews) or (meta analysis))

BEI: Dialog. 2003–4/3. 20 July 2004

The BEI search covered the date range 2003 to March 2004. The precise search without 'literature review' identified 10 records. With 'literature review' added the search identified 85. (Without date limits the precise search identified 54.)

CareData. Internet. 20 July 2004

www.elsc.org.uk/caredata/caredata.htm

The CareData search covered the date range 2003 to March 2004. The searches were conducted using the 'abstract' field option. The precise search without 'literature review' identified 44 records. With 'literature review' added the search identified 330. (Without date limits the precise search identified 120.)

ERIC: Dialog. 2003–2004/3. 20 July 2004

The ERIC search covered the date range 2003 to March 2004. The precise search without 'literature review' identified 57 records. With 'literature review' added the search identified 21,288. (Without date limits the precise search identified 2214.)

PAIS International: Ovid WebSPIRS. 2003–2004/4. 20 July 2004

The PAIS search covered the date range 2003 to April 2004. The precise search without 'literature review' identified one record. With 'literature review' added the search identified 158. (Without date limits the precise search identified 55.)

SIGLE: Ovid WebSPIRS. 2003–2003/12. 20th July 2004

The SIGLE search covered the date range 2003 to December 2003. The precise search without 'literature review' identified 15 records. With 'literature review' added the search identified 843. (Without date limits the precise search identified 281.)

SSCI: Web of Science. 2003–2004/6. 20th July 2004

The SSCI search covered the date range 2003 to June 2004. The precise search without 'literature review' identified 2032 records. With 'literature review' added the search identified 12,566. (Without date limits the precise search identified 9725.)

Sociological Abstracts: WebSPIRS. 2003–2004/6. 20th July 2004

The Sociological Abstracts search covered the date range 2003 to June 2004. The precise search without 'literature review' identified 32 records. With 'literature review' added the search identified 5939. (Without date limits the precise search identified 521.)

Appendix 4

Data for systematic review of systematic reviews (see Chapter 5)

Appendix 4.1: Description of included reviews

Review	Type of intervention	Description of intervention	Target group	Search strategy
Cambach, 1999[98]	Policy for an institution Policy for a community	Pulmonary rehabilitation in patients with asthma and chronic obstructive pulmonary disease	Patients with asthma and/or chronic obstructive pulmonary disease older than 18 years	MEDLINE Current contents Reference lists
Cameron, 2000[99]	Policy for an institution	Geriatric rehabilitation following fractures in older people	Patients aged 65 years or older with any fracture of the lower limbs, pelvis, upper limbs or spine which required hospital care either as an inpatient or in ambulatory care	MEDLINE EMBASE CINAHL Personal reference collections Reference lists Personal contact/contact with authors
Davis, 2000[90]	Policy for an institution	Child sexual abuse prevention programs (in schools)	Children aged 3 –13 years	MEDLINE PsycLIT/PsycINFO Dissertation abstracts online ERIC HealthStar Reference lists Hand searching
Griffith, 2000[100]	Policy for a community	Contingency management in outpatient methadone treatment	Patients receiving outpatient methadone treatment. Mean age 34	MEDLINE PsycLIT/PsycINFO SSCI Science Citation Index Reference lists Personal contact/contact with authors Hand searching
Jacobs, 2002[92]	Policy for an institution	The newborn individualized developmental care and assessment programme	Preterm infants < 37 weeks gestation or < 2500 g at birth	MEDLINE PsycLIT/PsycINFO Cochrane Library (CCTR/Central) EMBASE CINAHL Reference lists Personal contact/contact with authors
Kwan, 2004[101]	Policy for an institution	In-hospital care pathways for stroke	Patients who had been admitted to hospital with a new neurological deficit consistent with a clinical diagnosis of stroke	MEDLINE Cochrane Library (CCTR/Central) EMBASE CINAHL Index to Scientific and Technical Proceedings (ISTP) HealthSTAR Cochrane Stroke Group Specialised Trials Register Reference lists Personal contact/contact with authors Hand searching

Language restrictions	Validity assessment tool	Number of studies per design Number in review/number in meta-analysis		Heterogeneity identified by design?	
		RCTs	NRSs	Statistical	Clinical/method
English, Dutch or German included only	Checklist	*n* = 14/*n* = 6	*n* = 4/*n* = 1	No	No
No language restrictions	Scale (nine-item methodological quality score was devised by the authors)	*n* = 14/2–4 studies per outcome	*n* = 27/1–5 studies per outcome All cohort (concurrent and historical) studies	Yes	No
English language only	Components	Not reported	Not reported	No	No
Not stated	Quality not assessed	*n* = 17/*n* = 17	*n* = 13/*n* = 13	Yes	Yes (none)
No language restrictions	Quality of cohort studies not assessed RCTs: Jadad Scale Checklist Criteria developed by the Neonatal Cochrane review group	*n* = 5/1–4 studies per outcome	*n* = 3/1–4 studies per outcome	Yes	No
Not stated	Did not use preprinted 'selection' forms (presume they mean quality assessment forms) or an overall scoring system but noted important aspects of methodological quality	*n* = 3/*n* = 2 (for one outcome only)	*n* = 7/1–4 depending on outcome	Yes forest plots only	Yes

Review	Type of intervention	Description of intervention	Target group	Search strategy
Langhorne, 1999[102]	Policy for an institution	Services for helping acute stroke patients avoid hospital admission	Stroke patients with a clinical definition of stroke (focal neurological deficit caused by cerebrovascular disease). No specific limit on stroke severity or on the duration between stroke and recruitment into a trial	Cochrane Stroke Group Specialised Register of Controlled Trials Personal contact/contact with authors
Mullen, 2002[93]	Policy for an institution	Behavioural HIV prevention interventions for sexually experienced adolescents in the US	Adolescents of middle or high school age (13–19 years) in the US	Prevention Research Synthesis project database – a cumulative database constructed using manual searches, contacts with experts and searches of databases to obtain published and unpublished reports through 1998 relevant to HIV prevention
Oliver, 2000[103]	Policy for an institution	Hospital fall prevention programmes	Inclusion criteria 'hospital setting' Taken from settings including geriatric, rehabilitation, psychiatry, neurology, general medicine, orthopaedic, oncology, surgery	MEDLINE CINAHL Reference lists
Guyatt, 2000[91]	Policy for an institution Policy for a community	Interventions for adolescent pregnancy prevention	Adolescents aged 18 years or less. Studies had to have been conducted in the US, Australia, New Zealand, the UK, Europe (excluding Eastern Europe) or Scandinavia	MEDLINE PsycLIT/PsycINFO EMBASE Dissertation abstracts online ERIC Popline CINAHL Sociological Abstracts CATLINE (CATalog onLINE) Conference Papers Index NTIS (National Technical Information Services) Reference lists Personal contact/contact with authors Hand searching

Language restrictions	Validity assessment tool	Number of studies per design Number in review/number in meta-analysis		Heterogeneity identified by design?	
		RCTs	NRSs	Statistical	Clinical/method
Not stated	Components: no specific quality assessment tool reported	*n* = 3/1 or 2 per outcome	*n* = 1/*n* = 1	Yes	No
English language only	Quality not assessed Non-randomised studies met eligibility criteria only if they included pre-test measures and either reported no baseline differences between study groups or controlled statistically for such differences	*n* = 10 according to Table 1, Table 3 suggests 9/*n* = 10 according to Table 1, Table 3 suggests 9	*n* = 10 according to Table 1, Table 3 suggests 11/*n* = 6 according to Table 1, 7 according to Table 3	No	No
Not stated	Components: the authors do not state that quality was assessed but discuss aspects of quality at the end of the results section	*n* = 2/*n* = 2	*n* = 19/*n* = 8	No	No
No language restrictions	Quality not assessed	*n* = 13/females: between 7 and 9 depending on outcome; males: 3 or 4 depending on outcome	*n* = 17/females: 6 or 11 depending on outcome; males: between 2 and 6 depending on outcome	No	No

Review	Type of intervention	Description of intervention	Target group	Search strategy
Smedslund, 2004[104]	Policy for an institution	Workplace smoking cessation programmes	Smokers in the workplace	MEDLINE PsycLIT/PsycINFO SSCI Dissertation abstracts online ERIC Sociological Abstracts ABI/Inform BRS Combined Health Information Database Occupational Health and Safety Database Smoking and Health Database Reference lists
Thomas, 2003[94]	Policy for an institution Policy for a community	Barriers and facilitators to healthy eating in children	Children whose average age was between 4 and 10 years	MEDLINE PsycLIT/PsycINFO Cochrane Library (CCTR/Central) also DARE and CDSR Bibliomap HealthPromis (HEA/HDA) SSCI EMBASE ERIC CINAHL PrevRev Cochrane Heart Group internal trials register Reference lists Personal contact/contact with authors Hand searching
Tobler, 2000[107]	Policy for an institution Policy for a community	School-based adolescent drug prevention programmes	All members of the student body, which may have included but did not specifically target high risk youth. Involved school grades 6 –12	Electronic databases not reported Reference lists Personal contact/contact with authors
Wilson, 2000[105]	Policy for an institution	Corrections-based education, vocation and work programmes for adult offenders	Convicted adults or persons identified by the criminal justice system (court) and placed in a prison or jail, or diverted to another corrections-based programme, such as probation	PsycLIT/PsycINFO Dissertation abstracts online ERIC Sociological Abstracts Criminal Justice Periodical Index NCJRS Social SciSearch Social Sciences Abstracts Reference lists Personal contact/contact with authors

Language restrictions	Validity assessment tool	Number of studies per design Number in review/number in meta-analysis		Heterogeneity identified by design?	
		RCTs	NRSs	Statistical	Clinical/method
Not stated	Components	*n* = unclear (states 9 in text, 8 in Table 1, and 10 in Table 3)/*n* appears to be 8	*n* = 11 according to Table 1, 9 according to Table 3/*n* appears to be 10	Yes	No
English language only	Components	*n* = 17/*n* = 2–7 depending on outcome	*n* = 16/*n* = 3–6 depending on outcome	No	No
Not stated	Components: coded but not reported	*n* = 144/*n* = unclear – 144 in Table 6, 141 in text p. 320	*n* = 63/*n* = unclear – 63 in Table 6, 66 in text p. 320	No	No
Not stated	Components	*n* = 3/*n* = 3	*n* = 30/*n* = 30	No	No

Review	Type of intervention	Description of intervention	Target group	Search strategy
Wilson, 2001[96]	Policy for an institution	School-based prevention of problem behaviours	General student population. Some restricted to student population identified as high risk for problem behaviours or delinquency	PsycLIT/PsycINFO ERIC Sociological Abstracts (listed as examples) Personal collections Reference lists
Wilson, 2003[97]	Policy for an institution	School-based intervention programmes for aggressive behaviour	Preschool–12th grade children	MEDLINE PsycLIT/PsycINFO ERIC Dissertation Abstracts International US Government Printing Office Publications National Criminal Justice Reference Service Reference lists Hand searching

Language restrictions	Validity assessment tool	Number of studies per design Number in review/number in meta-analysis		Heterogeneity identified by design?	
		RCTs	NRSs	Statistical	Clinical/method
Not stated	Scale: 5-point Scientific Methods Score – informed by answers to the method rigour items Components	*n* = 42/*n* = 42 (comparisons not studies)	*n* = 174/*n* = 174 (comparisons not studies)	No	No
English language only	Components: study quality not specified, but coded information for each study that described the methods and procedures, including details of design, measures and attrition	179 groups (NOT studies)/not reported	343 groups (NOT studies)/not reported	No	No

Appendix 4.2: Similarity of RCTs and NRSs, where authors judged results to be SIMILAR

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?		
Review	**Population**	**Intervention**	**Comparator**
Cameron, 2000[99]	None		
Kwan, 2004[101]	1/3 RCTs included only patients with ischaemic stroke as opposed to all stroke 5/7 NRSs included only patients with ischaemic stroke None of RCTs reported major differences in observed baseline characteristics between groups although in some studies only limited details were given (e.g. for subtype of stroke, pre-stroke disability or handicap) For the NRSs, baseline characteristics were reported to be similar between groups in two studies, different in certain aspects (race, gender, % haemorrhagic stroke) in four studies, and not reported in one study	RCTs – interventions well described – common elements of care included: involvement of multiple disciplines setting of predefined patient goals and therapeutic activities regular multidisciplinary team meetings NRSs – interventions less well described – common elements of care: involvement of multiple disciplines care planning with specific care protocol RCTs – care pathways were computer generated (1) paper format (1) or not reported (1) NRSs – care pathways were paper format (5) or not reported (2) RCTs – pathways were for stroke rehabilitation (2 studies) or acute care and stroke rehabilitation (1 study) NRSs – pathways were for acute stroke (5 studies) or acute care and stroke rehabilitation (2 studies)	RCTs – control group care poorly defined – described as multidisciplinary care with regular team meetings to discuss patients progress (two studies) NRSs – very poorly described in all studies The paper states that that the comparator was poorly described in NRSs
Langhorne, 1999[102] Country: UK Services for helping acute stroke patients avoid hospital admission	Not clear – studies recruited patients from various sources	Interventions varied from prevention of admission to hospital to early discharge from hospital to community. Varied between all studies regardless of study designs Publication dates: the one included NRS was published at least 10 years before the three included RCTs	NRSs compared area with access to home care stroke team to one without access. Is not clear whether the eligible patients would all come from the community and if so what treatment options were available. For the RCTs the control group care could include inpatient care but not in all cases

Outcomes	Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
	Not stated	None	Yes (Yes)
Only 8/24 (5/21 in meta-analysis) outcomes were reported in both RCTs and NRSs Outcomes only reported in RCTs were: patient and carer satisfaction, dead or dependent at end of follow-up and quality of life Outcomes only reported in NRSs were: death in hospital, death in hospital or discharge to institutional care, complications including pneumonia, urinary tract infection, deep vein thrombosis, dehydration, fluid and electrolyte imbalance, seizures, skin breakdown, falls or fractures, myocardial infarction, first or second CT scan, carotid duplex study, and electrocardiography	(from Discussion) 'non-randomised studies are highly susceptible to bias and there is significant statistical heterogeneity between the studies'	None	RCTs show a trend towards longer stay with the intervention, while NRSs show shorter stay with intervention (But text says significantly shorter and meta-analysis says non-significantly shorter)
	Not stated	None	Unclear – in discussion states that the trials are 'characterised by considerable heterogeneity which makes it difficult to draw specific conclusions' (Unclear – death: trend towards higher mortality in intervention group within the RCTs)

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?		
Review	**Population**	**Intervention**	**Comparator**
Tobler, 2000[107] Country: US Universal school-based drug prevention programmes	Not possible to tell, insufficient data		
Wilson, 2000[105] Country: US Correction-based education on future offending behaviour of adult criminals	Insufficient detail given to judge		

Outcomes	Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
	To empirically confirm that the inclusion of non-randomised pre-test/post-test research designs does not overestimate intervention success and to eliminate other sources of bias, a subset of high quality evaluations was selected. These were randomised, and also met other criteria	Not stated	The lack of random assignment does not seem to greatly bias the studies, relative to other problems. The means for random assignment vs non-random assignment in the large set with problematic variations differ by 0.03. An intermediate set with problematic evaluations removed (e.g. those with cross-sectional research, fewer than 3 hours of intervention etc.) compared for results on random vs non-random assignment differ by 0.06. Removing other sources of bias influences the results far more than does lack of random assignment. (Unclear CIs not reported)
	Randomised and non-randomised studies were pooled. Study type (randomised vs non-randomised) was investigated as an influence on findings, and whether poorer quality studies were driving the positive findings	Not stated	For randomised vs non-randomised studies, the authors state that the difference is unremarkable and statistically non-significant [Little difference between odds ratios for randomised and non-randomised comparisons. CIs are not reported, but the authors note that there were no statistically significant differences under an inverse variance-weighted random-effects model. Note small number of randomised comparisons (3) compared with number of non-randomised comparisons (50)]

Appendix 4.3 Results of RCTs and NRSs, where authors judged results to be SIMILAR

Review	Outcomes with at least one RCT and NRS	Results of RCTs	Heterogeneity test
Cameron, 2000[99] Country: Australia Programmes of care following acute management of fractures in older people	Length of hospital stay, GORU vs orthopaedic unit, WMD	(*n* = 3) 1.631 (95% CI –27.98 to 31.25)	χ^2 26.26, df = 2[a]
	Residential status (return home), GORU vs orthopaedic unit, OR:	(*n* = 4) 1.36 (95% CI 0.86 to 2.13)	χ^2 5.04, df = 3[a]
	Residential status (return home), GHFP vs orthopaedic unit, OR:	(*n* = 2) 2.06 (95% CI 1.08 to 3.93)	χ^2 0.32, df = 1
	Mortality (death by 1 year), GORU vs orthopaedic unit, OR:	(*n* = 4) 0.92 (95% CI 0.57 to 1.48)	χ^2 5.00, df = 3[a]
	Mortality (death by 1 year), GHFP vs orthopaedic unit, OR:	(*n* = 2) 0.85 (95% CI 0.48 to 1.51)	χ^2 5.00, df = 1[a]
	Mortality (death by 1 year), ESD vs orthopaedic unit, OR:	(*n* = 1) 1.01 (95% CI 0.37 to 2.81)	χ^2 0.00, df = 0
Kwan, 2004[101] Country: UK Care pathways vs standard medical care in acute stroke	Duration of hospital stay, WMD	3.99 (95% CI –0.29 to 8.27)	χ^2 0.15, df = 1, *p* = 0.70
	Death by end of follow-up		Insufficient studies
	Discharged to institutional care		Insufficient studies
	Discharged to home		Insufficient studies
	Re-admission or emergency department attendance		Insufficient studies
Langhorne, 1999[102] Country: UK Services for helping acute stroke patients avoid hospital admission	Death: OR	7.76 (95% CI 1.65 to 36.57)	(*n* = 2) χ^2 0.00, df 1, *p* = 0.97
	Death or institutional care: OR	3.07 (95% CI 0.76 to 12.43)	(*n* = 2) χ^2 0.49, df = 1, *p* = 0.48
	Death or dependency: OR	1.91 (95% CI 0.58 to 6.24)	N/A
	Activities of daily living: WMD	–1.07 (95% CI –2.85 to 0.71)	(*n* = 2): χ^2 0.11, df = 1, *p* = 0.74
	Extended ADL: WMD	–0.020 (95% CI –8.24 to 7.84)	N/A
	Subjective health status – patient: WMD	–4.30 (95% CI –9.98 to 1.38)	(*n* = 2): χ^2 0.96, df = 1, *p* = 0.33
	Subjective health status – carer: WMD	–1.42 (95% CI –3.83 to 0.99)	N/A
	Number of patients admitted to hospital: OR	3.99 (95% CI 0.56 to 28.40)	N/A
	Length of total hospital stay: WMD	8.00 (95% CI –11.70 to 27.70)	N/A
Tobler, 2000[107] Country: US School-based drug prevention programmes	RCTs: full set, with problematic evaluations[b]	*n* = 141, WES = 0.19	Not assessed separately for NRSs and RCTs
	Intermediate set (*n* = 139), with problematic evaluations removed[b]	*n* = 94, WES = 0.20	
Wilson, 2000[105] Country: US Correction-based education and future offending behaviour of adult criminals	Programme comparison contrasts	(*n* = 3) OR 1.50 (95% CI not reported)	Not assessed

ADL, Activities of daily living; ESD, early supported discharge; GHFP, geriatric hip fracture programme; GORU, geriatric orthopaedic rehabilitation unit; N/A, not applicable; OR, odds ratio; WES, weighted effect size; WMD, weighted mean difference.

a Indicates statistically significant.

b Problematic evaluations are those with cross-sectional research, fewer than 3 hours of intervention, etc.

Results of NRSs	Heterogeneity test	Author comment
(*n* = 1) –0.900 (95% CI –4.549 to 2.749)	χ^2 0.00, df = 0	Authors state that for some outcomes (e.g. length of stay in evaluation of GORUs) there is greater heterogeneity between RCTs than between the pooled data from RCTs and that from cohort studies, but data not presented
(*n* = 3) 0.85 (95% CI 0.24 to 2.98)	χ^2 28.86, df = 2[a]	
(*n* = 2) 1.89 (95% CI 1.10 to 3.24)	χ^2 0.31, df = 1	
(*n* = 3) 1.44 (95% CI 1.00 to 2.08)	χ^2 1.20, df = 2	
(*n* = 2) 1.18 (95% CI 0.47 to 2.93)	χ^2 0.00, df = 1	
(*n* = 5) 0.93 (95% CI 0.65 to 1.33)	χ^2 1.62, df = 4	
–2.08 (95% CI –4.36, 0.20)	(*n* = 2): χ^2 3.10, df = 1, *p* = 0.08[a]	Authors state that the definition of 'care pathway' may have been a source of variation and urge readers to be cautious when interpreting results, owing to presence of variation between studies and small numbers of participants
	Not assessed[a]	
	(*n* = 4): χ^2 6.16, df = 3, *p* = 0.1039	
	χ^2 7.40, df = 3, *p* = 0.060[a]	
	Not assessed[a]	
0.85 (95% CI 0.65, 1.11)	NRSs: not reported/not applicable/not assessed – only one NRS in meta–analyses	Sensitivity analysis carried out to accommodate variations in trial design, intervention and patient follow-up. Authors state that 'all the conclusions reported above are not altered when sensitivity analyses are carried out to accommodate variations in trial design, intervention, and patient follow up'
0.89 (95% CI 0.68, 1.16)		
0.88 (95% CI 0.65 to 1.18)		
0.40 (95% CI –0.52 to 1.32)		
0.10 (95% CI –2.06 to 2.26)		
N/A		
–0.60 (95% CI –3.16 to 1.96)		
0.57 (95% CI 0.42 to 0.78)		
4.10 (95% CI –2.02 to 10.22)		
n = 66, WES = 0.16	Not assessed separately for NRSs and RCTs	
n = 45, WES = 0.14		
(*n* = 50): OR 1.53 (95% CI not reported)	Not assessed	
Not statistically significantly different under an inverse variance-weighted random-effects model		

Appendix 4.4: Results of any investigation of heterogeneity (where RCTs/NRSs judged SIMILAR)

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Tobler, 2000[107]		Categorical analysis: WES, SE *Population:* Sample size, large 0.10, 0.01; small 0.18, 0.02 School grade, elementary 0.07, 0.01; junior high 0.11, 0.01 Special populations, yes 0.13, 0.02; no 0.11, 0.01 Levels of drug use, non-users 0.17, 0.01; experimental users 0.12, 0.02; abusers 0.25, 0.03 ***Intervention*** *Type of programme:* Knowledge only 0.07, (95% CI –0.02 to 0.16), 0.059 Affective only 0.05 (95% CI –0.04 to 0.14), 0.192 Decisions/values/attitudes 0.11 (95% CI 0.04 to 0.18), 0.002 Knowledge plus affective 0.03 (95% CI –0.02 to 0.08), 0.119 Drug Abuse Resistance Education-type 0.05 (95% CI 0.00 to 0.10), 0.018 Social influences 0.12 (95% CI 0.10 to 0.15), 0.000 Comprehensive life skills 0.17 (95% CI 0.13 to 0.21), 0.000 System-wide change 0.27 (95% CI 0.21 to 0.33), 0.00 *Targeted drug:* Tobacco 0.16, 0.01; alcohol 0.10, 0.01; substance abuse 0.08, 0.01; other 0.11, 0.02 *Leader:* Teachers 0.10, 0.01; peer leaders 0.17, 0.01; clinicians 0.24, 0.03; others 0.08, 0.01 *Intensity:* ≤ 10 hours 0.11, 0.01; 11–30 hours 0.12, 0.01; ≥ 31 hours 0.15, 0.03 ***Comparator*** Type of control group, no treatment 0.11, 0.01; standard health class 0.12, 0.01 Attrition, acceptable 0.16, 0.01; unacceptable 0.09, 0.01

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Wilson, 2000[105] Country: US Correction-based education and future offending behaviour of adult criminals	***Programme type*** Adult basic education/general equivalency diploma $p = 0.03$ Postsecondary education $p = 0.03$ Vocational training $p = 0.14$ Correctional work/industries $p = 0.79$ ***Methodological variables*** Attrition, general $p = 0.01$ Attrition, differential $p = 0.01$ Qualitative method score $p = 0.03$ Qualitative method score squared $p = 0.01$ Reincarceration as outcome $p = 0.01$ Individually, the method variables were not significant predictors of odds ratio variability. But in a multivariate model they emerge as significant predictors. As a block, the programme type variables explained a significant proportion of effect size variability [$Q(4) = 13.20$, $p = 0.01$] after controlling for method variables	***Programme type*** Adult basic education and general equivalency diploma: ($n = 14$) OR 1.44 (95% CI 1.15 to 1.82) Postsecondary education: ($n = 13$) OR 1.74 (95% CI 1.36 to 2.22) Vocational training: ($n = 17$) OR 1.55 (95% CI 1.18 to 1.86) Correctional work/industries: ($n = 4$) OR 1.48 (95% CI 0.92 to 2.17) Multicomponent/other: ($n = 5$) OR 1.33 (95% CI 0.89 to 1.98) ***Method features present/absent*** Subject level matching (post hoc): yes ($n = 14$) OR 1.54, no ($n = 39$) OR 1.53 Performed covariate-adjusted analyses: yes ($n = 12$) OR 1.28, no ($n = 41$) OR 1.64 Attrition, overall: yes ($n = 10$) OR 1.84, no ($n = 43$) OR 1.47 Attrition, differential: yes ($n = 4$) OR 1.37, no ($n = 49$) OR 1.54 Used statistical significance testing: yes (n = 30) OR 1.47, no ($n = 23$) OR 1.61 Reincarceration as outcome: yes ($n = 35$) OR 1.59, no ($n = 18$) OR 1.43 No statistically significant differences under an inverse variance-weighted random-effects model ***Qualitative method score*** Experimental, non-compromised ($n = 2$) OR 1.55 Quasi-experimental, good statistical controls ($n = 4$) OR 1.36 Quasi-experimental, poor and/or no statistical controls ($n = 16$) OR 2.25 Quasi-experimental, clear lack of comparability ($n = 31$) OR 1.33 There is a significant quadratic trend in odds ratios across the four method score categories

OR, odds ratio; SE, standard error; WES, weighted effect size.

Appendix 4.5: Similarity of RCTs and NRSs, where authors judged results to be NOT SIMILAR

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?			
Review	**Population**	**Intervention**	**Comparator**	**Outcomes**
Cambach, 1999[98] Country: the Netherlands Pulmonary rehabilitation in patients with asthma and chronic obstructive pulmonary disease		The four NRSs were all outpatient settings, whereas the RCTs were a mixture of outpatient, inpatient, home-based and physiotherapy practice-based settings		
Davis, 2000[90] Country: US School-based child sexual abuse prevention programmes	Cannot tell			
Griffith, 2000[100] Country: US Contingency management (system of incentives and disincentives) for reducing illicit drug use during treatment	Several studies in which non-random assignment was used involved patients considered to be treatment failures	No obvious differences (but insufficient study details reported to be able to tell)	No obvious differences (but insufficient study details reported to be able to tell)	No obvious differences (but insufficient study details reported to be able to tell)

Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
Randomised vs non-randomised design was stated a priori as a potential source of heterogeneity Appears that all study designs were pooled and subgroup analyses then carried out, including elements of study design (randomised vs non-randomised controlled trial) to examine influence on effect size	None	Unclear (No)
No rationale was stated. All studies were included in a meta-analysis, and methodological quality and effect size were assessed by using multiple regression; random assignment was included	None	Authors state that subset analysis found higher mean effect sizes when studies, did not use random assignment of participants and used a waitlist control instead of an unrelated alternate programme that would control for amount of experimenter contact. Also, studies that used pre-tests to examine initial control group and experiment group equivalence found higher effect sizes than those using only post tests. Studies with more items on outcome measures found higher effect sizes. [No, effect size larger in NRS. With one NRS removed, the difference in effect sizes is reduced (0.725 vs 0.826)]
Randomised and non-randomised studies were pooled (does not specify what constitutes non-randomised). Eight moderators were examined, including assignment of participants No rationale stated as to why assignment was investigated, although authors state that non-randomised studies may involve greater staff expectations and greater baseline levels of use	'The effects of the moderator variables were examined by regressing moderator variables on the effect size, which yields an estimate of between-groups variance (*Qb*). In comparing the relative strength of levels within moderators, 95% CIs were calculated. Non-overlapping CIs allow conclusions about the strength of one predictor in comparison to another while limiting the overall error rate to 5%'	No (No, states that studies employing non-random assignment reported better outcomes than those employing random assignment)

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?			
Review	**Population**	**Intervention**	**Comparator**	**Outcomes**
Jacobs, 2002[92] Country: Canada Objective: The Newborn Individualised Developmental Care and assessment Program compared with conventional care for improving long-term neurodevelopmental outcomes in pre-term and/or low birth weight infants	Insufficient detail to assess any obvious differences in study design, but sample sizes tended to be larger in the cohort studies			
Mullen, 2002[93] Country: US Interventions for sexual risk behaviours for HIV among sexually experienced adolescents in the US	No obvious differences			
Oliver, 2000[103] Falls prevention in hospitals		Interventions appeared to be different		

Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
Separate meta-analyses were carried out for RCTs and for cohort studies: no rationale reported	None	RCTs and non RCTs were similar on 11 outcomes. Cohort studies found a statistically significant change that was not found by RCTs for two outcomes RCTs found a statistically significant change that was not found by cohort studies in two studies The authors state in the discussion that 'there was a discrepancy in results for the duration of mechanical ventilation and supplemental oxygen for RCTs vs cohort studies, which may reflect true differences between study populations or heterogeneity secondary to study design and/or bias'. However, the result section states ' both RCTs and cohort studies reported a significant reduction in requirement for supplemental oxygen'
Included randomised and non-randomised studies in the meta-analysis. Then carried out stratified analyses (including RCTs vs NRSs) to examine variation in size of effects	Two criteria were used to evaluate the contribution of the grouping variables for explaining the variation in the estimation of the summary odds ratios Assessed the likelihood of the magnitude of the between-subgroup differences using a chi-squared statistic Q; degrees of freedom, number of subgroups minus 1. Because 12 stratified comparisons were performed, a Bonferroni correction was used Estimated the magnitude of the contribution as represented by the percentage of the total heterogeneity, Q, explained by the between-group heterogeneity, Q_b. Defined substantial contribution of the variance explained by a stratification variable as ≥ 5%, the larger this percentage, the larger the difference between the subgroups than the differences within subgroups	No, the RCTs indicate a significant effect of the intervention, while the nRCTs indicate no significant effect (Between-subgroup differences were not significantly different for random vs non-random assignment: *Q* statistic 2.40 (1 df), $p = 0.12$. Assignment (random vs non-random) explained 7.3% of the total heterogeneity. Assignment was one of eight variables making a substantial contribution of the variance (meeting the criterion of explaining more than 5% of the total heterogeneity)
No rationale given	None	No (Unclear)

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?			
Review	**Population**	**Intervention**	**Comparator**	**Outcomes**
Smedslund, 2004[104] Country: Norway Worksite smoking cessation programmes	Insufficient detail given			
Wilson, 2001[96] Country: US School-based prevention of crime, substance use, dropout/ non-attendance and other conduct problems	Insufficient detail given			

Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
Randomised and non-randomised trials were analysed separately. No rationale was given as to why	Not stated	Unclear. Authors state that at all three follow-up points, the non-randomised studies showed larger effects and that 'the randomised results are probably closer to the truth, as the nonrandomised studies are probably overestimating the effects' (Yes. Effect sizes for NRSs are larger than RCTs, but they are in the same direction and confidence intervals overlap)
Carried out a meta-analysis of both randomised and non-randomised studies. This was then investigated as an explanatory source of variation in effect size	Not stated	No, $p \leq 0.05$ (statistically significant difference between NRSs and RCTs). 'It is interesting to note that randomised designs yielded larger mean effects than the nonrandomised designs' (No, smaller effect size in NRSs than RCTs, although direction of effect was the same. Confidence intervals do not overlap)

Appendix 4.6: Results of RCTs and NRSs, where authors judged results to be NOT SIMILAR

Review	Outcomes with at least one RCT and NRS	Results of RCTs
Cambach, 1999[98] Country: the Netherlands Pulmonary rehabilitation in patients with asthma and chronic obstructive pulmonary disease	Endurance time	($n = 6$) Effect size 1.4 (95% CI not reported), $p < 0.0001$
Davis, 2000[90] Country: US School-based child sexual abuse prevention programmes	Overall effect	With Nemerofsky 1994: D = 0.725 Without Nemerofsky 1994: D = 0.725
Griffith, 2000[100] Country: US Contingency management (system of incentives and disincentives) for reducing illicit drug use during treatment	Overall effect	$r = 0.22$ (95% CI 0.16 to 0.28)
Jacobs, 2002[92] Country: Canada Objective: The Newborn Individualised Developmental Care and assessment Program (NIDCAP) compared with conventional care for improving long-term neurodevelopmental outcomes in pre-term and/or low birth weight infants	Intraventricular haemorrhage (any)	RR 0.52 (95% CI 0.13 to 2.05), RD –0.14 (95% CI –0.42 to 0.14)
	Intraventricular haemorrhage (severe)	RR 0.39 (95% CI 0.15 to 1.00), RD –0.13 (95% CI –0.26 to –0.01)
	Patent ductus arteriosus	RR 0.95 (95% CI 0.65 to 1.41), RD –0.02 (95% CI –0.20 to 0.16)
	Necrotising enterocolitis	RR 0.21 (95% CI 0.01 to 4.10), RD –0.07 (95% CI –0.19 to 0.06)
	Retinopathy of prematurity (any)	RR 0.84 (95% CI 0.64 to 1.12), RD –0.09 (95% CI –0.24 to 0.06)
	Retinopathy of prematurity (severe)	RR 0.75 (95% CI 0.41 to 1.37), RD –0.07 (95% CI –0.21 to 0.07)
	Pneumothorax	RR 0.15 (95% CI 0.02 to 1.13), RD –0.28 (95% CI –0.52 to –0.05)
	Chronic lung disease (28 days)	RR 1.08 (95% CI 0.84 to 1.39), RD 0.04 (95% CI –0.09 to 0.16)
	Chronic lung disease (36 weeks)	RR 0.26 (95% CI 0.00 to 25.40), RD –0.39 (95% CI –1.04 to 0.26)
	Neurodevelopment 9–12 months (cognitive)	WMD 16.58 (95% CI 9.33 to 23.82)
	Neurodevelopment 9–12 months (motor)	WMD 9.24 (95% CI 0.68 to 17.81)
	Duration (days) ventilation	WMD –25.70 (95% CI –43.94 to –7.46)
	Duration (days) supplemental oxygen	WMD –41.06 (95% CI –65.29 to –16.83)
	Duration (days) hospitalization	WMD –18.38 (95% CI –44.13 to 7.37)
	Weight gain (g/day)	WMD 3.24 (95% CI 0.57 to 5.92)
	Days to full oral feeds	MD –44.90 (95% CI –86.12 to –3.68)
	Gestation at discharge	WMD –0.41(95% CI –1.28 to 0.47)

Heterogeneity test	Results of NRSs	Heterogeneity test	Author comment
Not reported	(*n* = 1) Effect size –1.7 (95% CI not reported), *p* = 0.003	Not reported	
Not reported	With Nemerofsky 1994: D = 1.131 Without Nemerofsky 1994: D = 0.826	Not reported	
Not statistically significant Qw = 25.63, *p* = not statistically significant	*r* = 0.36 (95% CI 0.26 to 0.46)	Statistically significant Qw = 85.56, *p* < 0.001	
Statistically significant Values not reported, but significant heterogeneity for the following outcomes meant that the random effect model was used: Intraventricular haemorrhage (any) Retinopathy of prematurity (any) Chronic lung disease (at 36 weeks)	RR 1.20 (95% CI 0.71 to 2.01), RD 0.05 (95% CI –0.08 to 0.18) RR 0.93 (95% CI 0.17 to 5.17), RD –0.001 (95% CI –0.05 to 0.05) RR 0.71 (95% CI 0.35 to 1.42), RD –0.07 (95% CI –0.21 to 0.07) RR 0.48 (95% CI 0.13 to 1.85), RD –0.05 (95% CI –0.14 to 0.04) RR 5.16 (95% CI 1.58 to 16.84), RD 0.21 (95% CI 0.08 to 0.33) RR 0.97 (95% CI 0.06 to 15.14), RD –0.001 (95% CI –0.05 to 0.04) RR 0.65 (95% CI 0.11 to 3.73), RD –0.02 (95% CI –0.09 to 0.05) RR 0.50 (95% CI 0.30 to 0.83), RD –0.17 (95% CI –0.28 to –0.05) RR 0.65 (95% CI 0.11 to 3.73), RD –0.02 (95% CI –0.09 to 0.05) MD 45.24 (95% CI 35.92 to 54.56) MD 13.70 (95% CI –0.34 to 27.74) WMD –4.67 (95% CI –10.85 to 1.51) WMD –6.64 (95% CI –12.73 to –0.55) WMD –4.83 (95% CI –12.82 to 3.16) MD –3.46 (95% CI –6.69 to –0.23) WMD –14.73 (95% CI –23.45 to –6.02) WMD –0.13 (95% CI –1.14 to 0.88)	Not statistically significant Values not reported, but random-effects model not used for any outcome, therefore it can be assumed that there was no significant statistical heterogeneity	

Review	Outcomes with at least one RCT and NRS	Results of RCTs
Mullen, 2002[93] Country: US Interventions for sexual risk behaviours for HIV among sexually experienced adolescents in the US	Composite behavioural risk variable	OR 0.58 (95% CI 0.42 to 0.81), $p < 0.01$
Oliver, 2000[103] Falls prevention in hospitals	Falls	RR 1.0 (95% CI 0.6 to 1.68)
Smedslund, 2004[104] Country: Norway Worksite smoking cessation programmes	6-month follow-up 12-month follow-up More than 12-month follow-up	OR 1.74 (95% CI 1.26 to 2.40) OR 1.36 (95% CI 1.09 to 1.69) OR 1.26 (95% CI 0.86 to 1.88)
Wilson 2001[96] Country: US School-based prevention of crime, substance use, dropout/non-attendance and other conduct problems	Overall effect size *d*	(42 comparisons) 0.25 (95% CI 0.17 to 0.33)

MD, mean difference; OR, odds ratio; RD, risk difference; RR, relative risk; WMD, weighted mean difference.

Heterogeneity test	Results of NRSs	Heterogeneity test	Author comment
Not reported	OR 0.75 (95% CI 0.48 to 1.16), $p = 0.20$	Not reported	
Not reported	RR 0.76 (95% CI 0.65 to 0.88)	Not reported	
Chi-squared test showed homogeneity at 6-month follow-up	OR 4.65 (95% CI 1.92 to 11.28) OR 2.58 (95% CI 1.37 to 4.86) OR 1.38 (95% CI 0.80 to 2.39)	Chi-squared test showed homogeneity at 6-month follow-up	
Not reported	(174 comparisons) 0.08 (95% CI 0.05 to 0.10)	Not reported	

Appendix 4.7: Results of any investigation of heterogeneity (where RCTs/NRSs judged NOT SIMILAR)

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Cambach, 1999[98]		***Population*** COPD ($n=5$); effect size 1.0, $p<0.0001$ COPD and asthma ($n=2$); effect size 1.8, $p<0.0001$ Within-group heterogeneity $p<0.0001$ Between-group heterogeneity $p=0.009$ FEV1 < 50th percentile ($n=4$); effect size 0.9 $p<0.0001$ FEV1 > 50th percentile ($n=3$); effect size 1.4 $p<0.0001$ Within-group heterogeneity $p<0.0001$ Between-group heterogeneity $p=0.10$ ***Intervention*** Community-based ($n=1$); effect size 2.2, $p<0.0001$ Inpatient/outpatient ($n=6$); effect size 1.0 $p<0.0001$ Within-group heterogeneity $p<0.0001$ Between-group heterogeneity $p=0.001$ ***Comparator*** No intervention in control group ($n=4$); effect size 1.7, $p<0.0001$ Intervention in control group ($n=3$); effect size 0.8, $p<0.0001$ Within-group heterogeneity $p<0.0001$ Between-group heterogeneity $p=0.0008$
Davis, 2000[90]	Type of outcome measure accounted for 16.2% of the variance in effect size, $F(3, 69) = 13.337$, $p<0.001$ Mode of presentation (written material, puppet shows, etc.) did not significantly predict effect size, although the use of puppet shows was a significant predictor in the presence of other variables Qualification of instructor did not significantly predict effect size	Average effect sizes (D) with and without Nemerofsky 1994: ***Population*** Mean age 3–5 years 2.143, 0.937 Mean age 5.01–8 years 1.243, 1.243 Mean age 8.01–12 years 0.770, 0.770 ***Intervention*** Interviewer status: uninformed 0.655, 0.655 Interviewer status: informed 2.020, 0.967 Design: pre and post tests 1.249, 0.886 Design: post tests only 0.601, 0.601

Review	Results of statistical investigations of heterogeneity	
	Regression analysis	**Subgroup analysis**
		Item number < 20 0.862, 0.862
		Item number > 20 1.455, 0.619
		Published 1.005, 0.632
		Not published 1.252, 1.252
		1 session 0.598, 0.598
		2 or 3 sessions 0.663, 0.663
		> 3 sessions 2.153, 1.536
		Active physical participation 1.202, 1.202
		Active verbal participation 1.226, 0.657
		No participation 0.453, 0.453
		Behavioural outcome measure 1.191, 1.191
		Questionnaire outcome measure 1.138, 0.838
		Interview outcome measure 0.617, 0.617
		Did not use written material 0.878, 0.878
		Used written material 1.296, 0.709
		Did not use role play or drama 1.226, 0.688
		Used role play or drama 0.910, 0.910
		Did not use discussion or lecture 0.718, 0.718
		Used discussion or lecture 1.094, 0.821
		Did not use video or film 1.183, 0.704
		Used video or film 0.923, 0.923
		Did not use puppet show 1.084, 0.805
		Used puppet show 0.898, 0.898
		Did not use behavioural training 1.024, 0.663
		Used behavioural training 1.210, 1.210
		Did not use dolls 1.081, 0.817
		Used dolls 0.741, 0.741
		Comparator
		Alternate programme 0.727, 0.727
		Wait list control 1.198, 0.825

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Griffith, 2000[100]	Intervention characteristics Methadone dosage Number of urine specimens per week Type of incentive Targeted CM drug Time to reinforcement delivery Duration of CM Methadone take home Voucher Type of reinforcer Methadone increase/decrease	
Mullen, 2002[93]		***Population*** Age, *Q* 2.42 (1 df) $p = 0.12$, 7.4% of total heterogeneity: < 16 years ($n = 12$) OR 0.71 (95% CI 0.52 to 0.96), $p = 0.03$ ≥ 16 years or high school ($n = 4$) OR 0.47 (95% CI 0.30 to 0.73), $p < 0.01$ Sexual experience, *Q* 4.44 (2 df) $p = 13.5$, 13.5% of total heterogeneity: 8–25% ($n = 4$) OR 0.86 (95% CI 0.48 to 1.52), $p = 0.60$ 33–66% ($n = 7$) OR 0.54 (95% CI 0.42 to 0.69), $p < 0.001$ 83–100% ($n = 5$) OR 0.74 (95% CI 0.42 to 1.31), $p = 0.31$ 100% single ethnic group, *Q* 10.88 (1 df) $p < 0.001$, 33.3% of total heterogeneity: No ($n = 11$) OR 0.80 (95% CI 0.59 to 1.092), $p = 0.16$ Yes ($n = 5$) OR 0.46 (95% CI 0.35 to 0.59), $p < 0.001$ ***Intervention*** Time to first follow-up, *Q* 3.12 (2 df) $p = 0.21$, 9.5% of total heterogeneity: ≤ 1 month ($n = 5$) OR 0.51 (95% CI 0.26 to 1.00), $p = 0.05$ 2–3 months ($n = 6$) OR 0.69 (95% CI 0.48 to 1.00), $p = 0.05$ ≥ 4 months ($n = 5$) OR 0.72 (95% CI 0.44 to 1.19), $p = 0.20$ Intensity, *Q* 2.68 (1 df) $p = 0.10$, 8.5% of total heterogeneity: < 6 sessions ($n = 6$) OR 0.57 (95% CI 0.38 to 0.85), $p = 0.01$ ≥ 6 sessions ($n = 9$) OR 0.74 (95% CI 0.52 to 1.06), $p = 0.10$

Review	Results of statistical investigations of heterogeneity	
	Regression analysis	Subgroup analysis
		Content: perceived risk enhancement, *Q* 0.54 (1 df) *p* = 0.46, 1.7% of total heterogeneity: No (*n* = 8) OR 0.62 (95% CI 0.39 to 1.00), *p* = 0.05 Yes (*n* = 8) OR 0.65 (95% CI 0.47 to 0.91), *p* < 0.01 Content: technical skills with practice, *Q* < 0.001 (1 df) *p* = 1.00, < 0.01% of total heterogeneity: No (*n* = 10) OR 0.64 (95% CI 0.44 to 0.94), *p* = 0.02 Yes (*n* = 6) OR 0.65 (95% CI 0.43 to 0.98), *p* = 0.04 Content: personal skills with practice, *Q* 2.50 (1 df) *p* = 0.11, 7.6% of total heterogeneity: No (*n* = 11) OR 0.70 (95% CI 0.52 to 0.96), *p* = 0.03 Yes (*n* = 5) OR 0.55 (95% CI 0.33 to 0.93), *p* = 0.02 Content: interpersonal skills with practice, *Q* 0.14 (1 df) *p* = 0.71, 0.4% of total heterogeneity: No (*n* = 7) OR 0.65 (95% CI 0.42 to 0.99), *p* = 0.05 Yes (*n* = 9) OR 0.65 (95% CI 0.45 to 0.94), *p* = 0.02 Content: interpersonal skills without practice, *Q* 0.02 (1 df) *p* = 0.88, 0.1% of total heterogeneity: No (*n* = 10) OR 0.66 (95% CI 0.48 to 0.90), *p* = 0.01 Yes (*n* = 6) OR 0.61 (95% CI 0.34 to 1.10), *p* = 0.10 Setting, *Q* 2.98 (1 df) *p* = 0.08, 9.1% of total heterogeneity: Out of class (*n* = 10) OR 0.57 (95% CI 0.40 to 0.81), *p* < 0.01 In class (*n* = 6) OR 0.81 (95% CI 0.54 to 1.23), *p* = 0.32 ***Comparator*** Comparison group treatment, *Q* 0.99 (1 df) *p* = 0.32, 3.0% total heterogeneity: None (*n* = 8) OR 0.66 (95% CI 0.47 to 0.94), *p* = 0.02 Some (*n* = 8) OR 0.65 (95% CI 0.42 to 1.00), *p* = 0.05

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Wilson, 2001[96]	Mixed effects regression analysis: few individual method or sample characteristics are statistically significant predictors of effect size. A substantial proportion of effect size variability was accounted for in the regression models for delinquency (0.33) and dropout/ non-attendance (0.18)	***Outcomes*** Delinquency 40 comparisons 0.04 (95% CI –0.03 to 0.11) Alcohol/drug use 80 comparisons 0.05 (95% CI 0.01 to 0.09) Dropout/non-attendance 39 comparisons 0.16 (95% CI 0.05 to 0.27) Other problem behaviours 73 comparisons 0.17 (95% CI 0.09 to 0.25) ***Methods*** Observed pre-test differences Yes (50 comparisons) 0.11 (95% CI 0.05 to 0.17) No (166 comparisons) 0.09 (95% CI 0.06 to 0.12) Unit of assignment ($p \leq 0.05$) Students (69 comparisons) 0.18 (95% CI 0.12 to 0.24) Not students (147 comparisons) 0.07 (95% CI 0.04 to 0.10) ***Unit of analysis*** Students (196 comparisons) 0.10 (95% CI 0.07 to 0.13) Not students (20 comparisons) 0.08 (95% CI 0.01 to 0.16) Careful selection of measure ($p \leq 0.05$) Yes (40 comparisons) 0.16 (95% CI 0.09 to 0.23) No (176 comparisons) 0.08 (95% CI 0.06 to 0.11)

Review	Results of statistical investigations of heterogeneity	
	Regression analysis	Subgroup analysis
		Overall method rating
		Serious weaknesses (16 comparisons) 0.07 (95% CI –0.02 to 0.16)
		Moderate weaknesses (25 comparisons) 0.03 (95% CI –0.04 to 0.10)
		Some weaknesses/ some strengths (47 comparisons) 0.10 (95% CI 0.04 to 0.16)
		Moderate strengths (104 comparisons) 0.10 (95% CI 0.07 to 0.14)
		Rigorous (24 comparisons) 0.16 (95% CI 0.6 to 0.26)
		Sample
		School grades
		Early elementary (19 comparisons) 0.05 (95% CI –0.06 to 0.16)
		Late elementary (56 comparisons) 0.05 (95% CI 0.00 to 0.11)
		Middle/ junior high school (68 comparisons) 0.09 (95% CI 0.04 to 0.13)
		Senior high school (32 comparisons) 0.14 (95% CI 0.06 to 0.22)
		Level of criminal involvement ($p \leq 0.05$)
		General school population (155 comparisons) 0.07 (95% CI 0.04 to 0.10)
		High-risk population (61 comparisons) 0.20 (95% CI 0.14 to 0.21)

CM, contingency management; COPD, chronic obstructive pulmonary disease; FEV1, forced expiratory volume in 1 second; OR, odds ratio.

Appendix 4.8: Similarity of RCTs and NRSs, where authors judged results to be MIXED

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?			
Review	**Population**	**Intervention**	**Comparator**	**Outcomes**
Guyatt, 2000[91] Country: Canada Interventions to prevent adolescent pregnancy	No	No	No	No
Thomas, 2003[94] Country: UK Barriers to and facilitators of healthy eating	No obvious differences			

Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
Authors discuss differences between randomised and non-randomised studies and stated a priori to explore study design as a possible determinant of outcome	A *z*-score was used to generate a *p*-value related to the null hypothesis that there were no real differences in results from observational studies and randomised trials	No. Observational studies give significant result for six of eight outcomes: initiation of intercourse and responsible sexual behaviour (males and females) and pregnancy and birth control use (females), where randomised trials do not (No. Statistically significant differences observed between randomised and non-randomised designs for: females: initiation of intercourse ($p = 0.01$), pregnancy ($p = 0.02$); males: no significant differences were found) Authors state that relying on observational studies leads to the conclusion that the interventions have a positive effect, while relying on results from RCTs leads to the conclusion that the evidence does not support a positive effect on any outcome
No rationale stated	No rationale stated	Yes Children's knowledge of fruit and vegetables: 'no significant differential effect size, with negligible heterogeneity ($Q = 0.146$, df = 1, $p = 0.702$) being explained by this subdivision of studies' Children's self efficacy: increased effect size for RCTs compared with CTs, but the analogue to the ANOVA does not show that significant heterogeneity was explained by this subgroup analysis ($Q = 1.292$, df = 1, $p = 0.257$) so no conclusions can be drawn Children's consumption of vegetables: heterogeneity not explained by subdivision by randomisation ($Q = 0.801$, df = 1, $p = 0.371$). Consumption of fruit and vegetables: 'little difference between the two groups of studies… no significant difference is explained by this analysis ($Q = 0.061$, df = 1, $p = 0.805$), with significant heterogeneity still remaining' No Children's consumption of fruit: very little effect reported by the seven RCTs, whereas the three CTs increased fruit consumption by nearly two thirds of a serving per day. Heterogeneity explained by subgroup analysis is significant at $p < 0.1$ ($Q = 3.738$, df = 1, $p = 0.0523$), leaving less residual heterogeneity within the groups ($Q = 9.398$, df = 8, $p = 0.31$). However, numbers are small for definitive conclusions [Yes Knowledge of fruit and vegetables Consumption of fruit and vegetables No Self-efficacy – greater effect in RCTs Consumption of fruit – greater effect in NRSs Consumption of vegetables – greater effect in NRSs Unclear Preferences for fruit and vegetables (small number of studies)]

	Obvious differences between RCTs and NRSs, in reviewers'/authors' opinion?			
Review	**Population**	**Intervention**	**Comparator**	**Outcomes**
Wilson, 2003[97] Country: US Objective: school-based intervention programmes for preventing or reducing aggressive behaviour	Insufficient detail given			

ANOVA, analysis of variance; CT, computed tomography.

Rationale for pooling RCTs and NRSs separately	Criteria used to judge equivalence	Authors' conclusions regarding similarity of RCTs and NRSs (reviewer's opinion)
Randomised, quasi-randomised (meaning non-randomised controlled study) and before–after (pre–post test) studies all pooled in the same analysis Study design was investigated as a moderator variable in regression analysis	Not stated	Yes 'The final reduced regression model accounted for 28% of the variance in re-post-test change effect sizes for subject samples receiving intervention. Among the variables relating to study method and procedure, two variables representing the different study designs were included in the model, randomized designs and one-group designs. One-group designs were associated with larger pre-post-test effect sizes (in comparison to nonrandomized designs, but the randomized design variable was not significant. That is, with the other variables in the model held constant, randomized designs did not produce different effect sizes than nonrandomized designs. The nonrandomized designs produce results for both observed and equated effect estimates that generally agree with the randomized results for social competence training with a cognitive-behavioural component' No 'However, the nonrandomized designs yield lower effect estimates than the randomized designs for social competence training without a cognitive-behavioural component and higher estimates for behavioural strategies, creating inconsistent patterns for each' (Yes for some outcomes No for some outcomes)

Appendix 4.9: Results of RCTs and NRSs, where authors judged results to be MIXED

Review	Outcomes with at least one RCT and NRS	Results of RCTs	Heterogeneity test
Guyatt, 2000[91]	*Females:*		Not investigated
Country: Canada	Initiation of intercourse	Pooled OR 1.09 (95% CI 0.90 to 1.32)	
Interventions to prevent adolescent pregnancy	Pregnancy	Pooled OR 1.08 (95% CI 0.91 to 1.27)	
	Responsible sexual behaviour	Pooled OR 1.01 (95% CI 0.75 to 1.36)	
	Birth control use	Pooled OR 0.99 (95% CI 0.64 to 1.54)	
	Males:		
	Initiation of intercourse	Pooled OR 0.81 (95% CI 0.35 to 1.90)	
	Pregnancy	Pooled OR 0.97 (95% CI 0.62 to 1.51)	
	Responsible sexual behaviour	Pooled OR 0.94 (95% CI 0.55 to 1.60)	
	Birth control use	Pooled OR 0.91 (95% CI 0.71 to 1.18)	
Thomas, 2003[94]	Children's knowledge of fruit and vegetables	Effect size 0.68	Not reported
Country: UK	Children's self-efficacy	Effect size 0.12 (95% CI 0.02 to 0.22)	
Barriers to and facilitators of healthy eating	Children's consumption of fruit	Effect size 0.09 (95% CI 0.02 to 0.16)	
	Children's consumption of vegetables	Effect size 0.18 (95% CI 0.11 to 0.26)	
	Consumption of fruit and vegetables	Effect size 0.16 (95% CI 0.08 to 0.24)	
Wilson, 2003[97]	Overall effect size	0.32 ($p < 0.05$)	Not reported
Country: US	ES – Control (*n*), ES – Intervention (*n*)	Focal randomised studies effect size 0.31 ($p < 0.05$)	
Objective: school-based intervention programmes for preventing or reducing aggressive behaviour	*Difference observed, equated:*		
	Social competence, no cognitive behavioural	–0.02 (15), 0.33 (15), 0.35, 0.30	
	Social competence, cognitive behavioural	0.01 (26), 0.37 (26) 0.36, 0.24	
	Behavioural, classroom management	0.25 (11), 0.43 (11), 0.18, 0.08	
	Therapy, counselling	–0.04 (11), 0.25 (11), 0.29, 0.34	
	Multimodal	–0.02 (1), –0.15 (1), –0.13, –0.04	

ES, effect size; OR, odds ratio.

Results of NRSs	Heterogeneity test	Author comment
	Not investigated	
Pooled OR 0.64 (95% CI 0.44 to 0.93)		
Pooled OR 0.74 (95% CI 0.56 to 0.98)		
Pooled OR 1.27 (95% CI 1.10 to 1.46)		
Pooled OR 1.38 (95% CI 1.18 to 1.60)		
Pooled OR 0.71 (95% CI 0.52 to 0.98)		
Pooled OR 0.85 (95% CI 0.68 to 1.06)		
Pooled OR 1.21 (95% CI 1.04 to 1.42)		
Pooled OR 0.82 (95% CI 0.35 to 1.91)		
Effect size 0.63	Not reported	
Effect size 0.09 (95% CI 0.00 to 0.16)		
Effect size 0.38 (95% CI 0.09 to 0.67)		
Effect size 0.27 (95% CI 0.10 to 0.43)		
Effect size 0.14 (95% CI 0.00 to 0.27)		
0.16 ($p < 0.05$)	Not reported	Heterogeneity for all studies was statistically significant 'using fixed effects q tests, we found that none of the ES distributions represented in the means were homogeneous'
Focal non-randomised effect size 0.16 ($p < 0.05$)		
–0.00 (37), 0.07 (37), 0.07, 0.15		
–0.14 (17), 0.09 (17), 0.23, 0.34		
–0.19 (7), 0.24 (7), 0.43, 0.37		
0.14 (5), 0.43 (5), 0.29, 0.34		
–0.01 (16), 0.03 (16), 0.04, 0.20		

Appendix 4.10: Results of any investigation of heterogeneity (where RCTs/NRSs judged MIXED)

	Results of statistical investigations of heterogeneity	
Review	**Regression analysis**	**Subgroup analysis**
Thomas, 2003[94]		***Intervention***
		Consumption of vegetables:
		Did not include a physical activity component, effect size 0.25
		Did contain a physical activity component, effect size 0.15
		Consumption of fruit and vegetables:
		Did not include a physical activity component, effect size 0.25 (95% CI 0.15 to 0.35)
		Did contain a physical activity component, effect size 0.06 (95% CI –0.03 to 0.16). Heterogeneity explained by subdivision is significant ($Q = 7.346$, df = 1, $p = 0.007$)
		Children's knowledge of fruit and vegetables:
		Overall quality assessment, studies rated 'high' effect size 0.72, studies rated medium effect size 0.67. Heterogeneity explained by separation $Q = 1.306$, df = 1, $p = 0.254$
Wilson, 2003[97]		***Population***
		Methods and procedure:
		Randomised design $p = 0.087$
		One-group design $p = 0.000$
		Sample size $p = 0.015$
		Attrition $p = 0.47$
		Measurement characteristics $p = 0.102$
		Treatment test overlap $p = 0.001$

Review	Results of statistical investigations of heterogeneity	
	Regression analysis	**Subgroup analysis**
		Subject characteristics:
		High risk $p = 0.000$
		Age (years, linear) $p = 0.106$
		Age (years, curvilinear) $p = 0.062$
		Age range $p = 0.073$
		Percentage male $p = 0.877$
		White ethnicity $p = 0.789$
		Minority ethnicity $p = 0.545$
		Intervention
		General programme attributes:
		Implementation quality $p = 0.000$
		Programme intensity $p = 0.045$
		One-on-one format $p = 0.057$
		Lay-persons deliver programme $p = 0.001$
		Teachers deliver programme $p = 0.014$

Appendix 4.11: Table of systematic reviews of policy interventions reporting narrative syntheses of both RCTs and NRSs

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Bekker, 1999[108] 366 RCTs 211 NRSs	Interventions affecting informed patient decision-making (e.g. use of a decision aid, offers of free transportation, etc.)	Policy for a community/ institution	Any patient making a health decision	Authors provide an annotated bibliography of 450 studies, only five of which were reviewed in-depth – all of which were RCTs	'no meaningful quantitative meta analysis have been applied as the reported information was too diverse' (p. 9) The five 'good' studies reported in-depth 'used an RCT with low risk of bias', reported 'informed' decision-making variables and referred to a theory
Bordley, 2000[109] 5 RCTs 10 nRCTs	Audit and feedback to improve immunisation delivery	Policy for an institution	Health-care professionals	2/5 RCTs and 3/10 NRSs were conducted in academic centres. No other obvious differences, quite varied within both study types	'after considering the myriad differences among trials (e.g. varying study designs, participants, settings, co-interventions and outcome measures), we concluded that quantitative synthesis or meta analysis was not justifiable' Initially used rigorous inclusion criteria of the Cochrane Collaboration, then broadened inclusion criteria. Because of not having concurrent control groups, simple before-and-after studies and ITS studies are limited by their inability to exclude the role of secular trends or other unknown confounders
Buller, 1999[110] 15 RCTs 10 NRSs	Sun protection programmes for children under 14 years	Health promotion provision/ organisation	Children or their parents or adult caregivers	No obvious differences, but difficult to tell without reference to primary studies	Authors state it was important to classify evaluations on the quality of design as higher quality designs are those that reduce the possibility of rival hypotheses by eliminating threats to internal validity. They note two features: randomisation and comparison of changes in treated units to untreated controls
Burns, 2001[111] 56 RCTs 35 nRCTs (compromised RCTs or NRSs)	Home treatment for mental health problems	Policy for a community/ institution Health-care service/ provision	Patients with mental health problems, mostly psychotic disorders	Studies appear broadly comparable (see Table 1 on page 8, there was only one statistical difference between them – date of publication $p = 0.05$). See also Table 2, for subanalysis of year of publication according to type of control: studies with inpatient controls (both RCTs and NRSs) had median publication year of 1986 and studies with community controls (both RCTs and NRSs) had median publication year of 1994 'Studies of all designs were more likely to study varieties of ACT and ICM than other service models, but RCTs were significantly more likely than NRSs to test these models' (p. 10)	Clinical and methodological heterogeneity, and limited availability of data. Extensive discussion provided on page 67 Two meta-analyses performed: RCTs RCTs + NRSs But no 'NRS only' pooling. Authors state that this was justified in view of the methodological rigour that RCTs offer; however, the distinction between compromised and uncompromised RCTs may be unreliable (owing to poor reporting). (p. 11)

DOI: 10.3310/hta14160

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Chesnut, 1999[112] 11 RCTs 23 NRSs (from evidence tables at back of report)	Rehabilitation methods at various phases in the course of recovery from traumatic brain injury in adults: early rehab in acute care setting; intensity of acute inpatient rehab; cognitive rehab; supported employment; case management	Policy for an institution/ community Health-care organisation/ provision	Adults with traumatic brain injury	Cannot see any systematic differences. May need further examination (incomplete evidence tables at back of report, hard to compare study designs)	No rationale given, although there is discussion of relative pros and cons of different study designs (pp. 32–34). Different study designs were identified for different questions, e.g. for questions 1 and 2 only observational and quasi-experimental studies were available; for question 3 RCTs were available
DiGuiseppi, 2000[113] 13 RCTs 13 nRCTs	Promoting residential smoke alarm ownership	Policy for a community	Community-dwelling individuals	RCTs and nRCTs appear broadly similar in terms of intervention. Most of the RCTs recruited families with babies or small children; most of the NRSs were more generally community based	Meta-analysis to combine odds ratios between intervention and control groups, using a random-effects model Only RCTs were pooled. nRCTs assessed qualitatively. Rationale not explained
Dobbins, 1999[126] RCTs: number unclear NRSs: number unclear Total 13 Note: the criteria 'study design' recorded as 'moderate quality' for all. Overall quality rating 'strong' for 3, 'moderate' for 7, 'weak' for 3	Community heart–health interventions – a wide range	Policy for a community	Communities, both urban and rural. Most targeted 'whole population', some targeted men and some women	Not possible to tell	No clear rationale but 'the [validity] criteria of study design and blinding did not discriminate between 'strong' and 'weak' projects as all projects were rated 'moderate' for these criteria. It may be that these criteria are not applicable when assessing community-based research projects'

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Draper, 2000[115] 8 NRS 9 RCTs	Geriatric acute psychiatry services. Different models of community/liaison psychiatry	Policy for an institution/ community Health-care service delivery Organisation	Elderly patients with or without psychiatric problems	Three RCTs examined psychiatric liaison services for elderly patients who were not being treated for mental disorders; all NRSs included only patients with mental disorders. More dementia in NRS group	None given, apart from 'it is claimed that the best evidence of effectiveness of old age psychiatry services comes from well designed RCTs'
Emmons, 2001[116] 3 RCTs 2 NRS	Interventions aimed at reducing children's exposure to environmental tobacco smoke	Policy for an institution/ community	Parents/ Families	The RCTs were conducted in families with children who were asthmatic or were patients; the NRSs were conducted in all families. Probably more motivation in the RCTs	None given
Fairbank, 2000[117] 14 RCTs 16 nRCTs 29 before–after studies	Any type of intervention designed to promote the uptake of breastfeeding	Policy for a community Health promotion provision/ organisation	Pregnant women, new mothers, women who may breastfeed in the future, and people linked to any of these women	No RCTs evaluated media campaigns or multifaceted interventions. Is hard to spot any other obvious differences without reference to the primary studies	'Owing to the heterogeneity of interventions, settings, participants, outcomes measures and comparison groups, a meta-analysis was not considered to be appropriate. Relative risks… have been estimated for individual RCTs and nRCTs, and forest plots have been presented'
Giuffrida, 2004[118] 1 RCT 1 NRS	The effects of target payments on the behaviour of primary care physicians	Policy for an institution Health-care delivery/ organisation	Primary care physicians	The RCT was performed in the US, while the NRS was conducted with GP practices in Scotland. There appear to be variations in case-mix, with the former study immunising mainly elderly patients, and the latter infants	'Study results were not statistically pooled as there was heterogeneity in the content, design and outcomes of the included studies and there were only two studies'. No stated intent to separate RCTs and NRSs
Gosden, 2004[119] 2 RCTs 2 NRSs	Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians	Policy for an institution Health-care delivery/ organisation	Primary care health professionals	Does not appear to be any obvious differences, other than that the two RCTs recruited paediatricians/counted children for capitation purposes and the NRSs were of general practice	'Study results were not statistically pooled as there was heterogeneity in the content, design and outcomes of the included studies'. No stated intent to separate RCTs and NRSs

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Hulscher, 2001[120] 37 RCTs 18 NRSs	Interventions aimed at primary care clinicians to improve the delivery of preventive services	Policy for an institution Health-care provision/ organisation	Primary care health professionals	Unclear. The studies in general were quite heterogeneous (e.g. in terms of intervention, outcomes, etc.). No obvious differences between the randomised and non-randomised studies	Authors planned to pool studies within homogeneous groups including high methodological quality vs moderate methodological quality (RCTs vs NRSs), but this was not possible owing to differences in interventions, outcome measures used, targeted preventive activities and study settings. An extensive discussion of this is presented in the 'Discussion' section
Hutt, 2004[121] 14 RCTs 5 NRSs	Impact of case management targeted primarily at older people on use of health services including hospital admissions, lengths of stay and use of emergency departments	Policy for community	Older people with chronic illness	Nothing obvious – quite a range of interventions, sample sizes and outcomes across all included studies	RCTs presented first because they provide the highest quality of evidence
Karjalainen, 2003[122] 1 RCT 1 NRS	Multidisciplinary biopsychosocial rehabilitation for neck and shoulder pain among working age adults	Policy for an institution/ community	Adults with neck and shoulder pain	Patients in one study were hospitalised, and therefore may have been in more severe pain. Their intervention was delivered by a clinical psychologist, whereas in the other study the intervention was delivered in a number of community and workplace settings	No rationale stated, although both studies are analysed separately in Meta view using mean differences and 95% CI
Klassen, 2000[158] 6 RCTs 22 NRSs	Injury prevention interventions for children and young people (e.g. education, fiscal incentives, legislation, environmental change, community activities)	Policy for an institution/ community	Children and young people (0–19 years)	A range of injury prevention interventions and outcomes were assessed. No obvious trends between RCTs and NRSs	Rationale for separating RCTs and NRSs not stated explicitly. Authors state that 'RCTs are the study design most likely to provide unbiased estimates of the impact of interventions the review included non-randomised comparison group studies because most studies of community interventions in a previous review did not use a randomised design and logistics of randomisation would be complicated in community based injury prevention programmes'

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Linton, 2001[159] 20 RCTs 7 CCTs	Interventions to prevent back or neck problems, or the development of long-term back or neck problems	Institution (workplace)/ community	Primary prevention: healthy patients not seeking treatment. Secondary prevention: patients with pre-existing back or neck problems	None observed but may need further examination	'RCTs are methodologically stronger than CCTs'
Meads, 1999[123] 1 RCT 11 NRSs	Inpatient compared to outpatient care for eating disorders	Policy for an institution Health-care provision/ organisation	People with eating disorders	Only one RCT included. No obvious differences but difficult to tell	States RCTs preferred and more weight would be given to RCT evidence
Pusey , 2001[124] 18 RCTs 12 NRSs	Psychosocial interventions for carers of people with dementia	Policy for a community	Carers of people with dementia	Variety of interventions and outcomes within both groups. No obvious systematic differences between RCTs and NRSs	Heterogeneous nature of interventions and outcome measures. A previous systematic review only included RCTs; nRCTs were therefore included as if well conducted, they may offer comparable evidence
Reeves, 2001[125] 1 RCT 18 NRSs	Interprofessional education for staff caring for adults with mental health problems	Policy for an institution	Health/Social carers of adults with mental health problems	Only one RCT was included in the review, and the authors question whether it can be described as randomised	No rationale stated, except that that possibly randomised study was one of the more rigorous research designs in comparison with the other included studies
Scott, 2003[127] 3 RCTs 6 NRSs	Any intervention designed to enhance communication between health professionals or others (for example parents or school teachers) and children and/ or adolescents with cancer about their disease, its treatment and their implications	Policy for an institution/ community Health-care provision/ organisation	Children and/ or adolescents diagnosed with cancer	Differences between all studies in aims, interventions, populations and outcomes, but no obvious differences between RCTs and NRSs	No rationale stated

Review details	Description of intervention	Policy level	Target group	Obvious differences between RCTs and NRSs (reviewers' judgement)	Rationale given by authors for not pooling
Stead, 2005[128] 7 RCTs 26 NRSs (including one RCT classed as an NRS by authors)	Interventions for preventing tobacco sales to minors	Legislation/ regulation Policy for a community	Minors defined by the legal age limit in the communities studied	No obvious differences but difficult to tell as interventions and outcomes differed between all studies	Authors chose narrative, rather than quantitative, synthesis because they expected heterogeneity in the study designs, type of interventions and outcomes measured. They state that greater weight was given to controlled studies
Weigand, 2001[160] 1 RCT 2 case-finding studies	Fall prevention programmes	Policy for an institution Health-care provision/ organisation	Emergency department patients > 64 years	No details. The RCT was about preventing falls in patients who had already experienced a fall; the two case-finding studies were about identifying patients at risk of falling	No rationale stated
Weir, 1999[161] 11 RCTs 5 NRSs (+3 cost studies)	Early discharge and community support in the management of patients following a stroke	Policy for a community/ institution	Patients who have had a stroke, and their caregivers	No, but not many details of intervention given so cannot be sure. No obvious differences in sample size or length of follow-up	No rationale stated for not pooling study designs. States that although RCTs are usually best able to reduce effects of bias and confounding, the most important determinant of study validity is the rigour applied to its design and analysis and not necessarily the type of study design used. Also, certain study designs are more appropriate for particular issues

CCT, clinical controlled trial.

Appendix 4.12: Results of systematic reviews of policy interventions reporting narrative syntheses of both RCTs and NRSs

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Bekker, 1999[108]	Authors report that the five RCTs reviewed in-depth are of disparate populations making disparate decisions, and show the importance of the context and social influences on individual's decision-making	Not reported for NRS	(Reviewer) Cannot tell (Author) Not stated	No
Bordley, 2000[109]	Children: one RCT found no significant effects Adults: three of four RCTs reported positive effects	Children: four NRSs showed a significant effect Adults: five of six NRSs reported positive effects	(Reviewer) NRSs tended to have large % increases in immunisation rates (e.g. + 50% in one study). RCTs had modest increases in comparison to controls (typically up to 20% difference). But comparison was difficult as none of the NRS had control groups (Author) Not stated	Probably not. Absolute changes in immunisation rates presented for all study groups, in terms of percentages. Probably would need the *n*'s to be able to re-analyse
Buller, 1999[110]	5 categories of studies: Short duration presentations – improve knowledge, weaker effect on attitudes and sun protection behaviour Multi-unit presentations – NRS not summarised separately Peer-education – improved knowledge Parent and caregiver – improve sun protection for children Community wide – increased proportion of children with sun protection and sunscreen	Short duration studies – improve knowledge and in some cases attitudes and sun protection behaviour Multi-unit presentations – NRSs not summarised separately Peer-education – improved knowledge Parent and caregiver – improve sun protection for children Community wide – improved sun protection and reduced sun-burning	(Authors) Do not compare results of RCTs and NRS (Reviewer) Within the five categories of studies, results from RCTs and NRSs tended to be in the same direction	No
Burns, 2001[111]	Inpatient control studies: eight RCTs mean difference between experimental and control service outcome 'reduction in hospital days' was 6 days per patient per month (p. 59) Community control studies: 29 RCTs mean difference in hospital days was 0.5 days per month	Inpatient-control studies 'When non RCTs were combined with RCTs the difference found across studies of any duration was 3 days per patient per month: about half that found when only RCTs were analysed Community control studies: three NRSs were included with the RCTs and 'made little difference'	(Reviewer) Maybe a slightly stronger effect seen in RCTs for inpatient control studies (Authors) 'it was not a primary aim of the review to test the robustness of NRSs compared to RCTs. The small number of prospective NRSs with well-matched patient groups made this impossible. There were relatively few NRSs with appropriate data to add to the analyses and their inclusion usually had little impact on the results. Where they did affect the results it was often in lessening their significance or reducing the magnitude of the finding'	Yes. Means (SDs) given for RCTs and non RCTs (see pages 59/60) Regression analysis also performed stratified according to RCT/nRCT exploring effect of duration of follow-up and year of publication

DOI: 10.3310/hta14160

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Chesnut, 1999[112]	In studies of cognitive rehabilitation, RCTs supported use of aids to memory	In studies of cognitive rehabilitation, trials had 'mixed results'. Weak evidence that early rehabilitation during acute admission reduces rehabilitation length of stay. Prospective observational studies support use of supported employment. Mixed results for case management	(Authors) No (Reviewers) No	No (not all studies listed in data extraction tables)
DiGuiseppi, 2000[113]	Smoke alarm ownership at follow-up appeared somewhat more likely in the intervention group (OR = 1.26; 95% CI 0.87 to 1.82). Similarly modest positive, statistically non-significant effects on functioning smoke alarms, and on new acquisitions of smoke alarms and functioning smoke alarms, were found. Assumptions of all positive or all negative outcomes resulted in similar or even smaller effect estimates	Five completed NRSs reported smoke alarm outcomes. Two involved safety advice during routine child health surveillance. Results of one of these were similar to those from randomised trials reported. The other trial reported modest effects from free smoke alarms but none from counselling alone. Two trials of community wide education reported no effects on alarm ownership or installation, but, in a third trial, installation of free smoke alarms increased the prevalence of functioning smoke alarms by 19%	(Reviewer) No statistically significant effects were seen in RCTs, and only modest effects in NRSs (Author) 'Evidence from randomised controlled trials indicates that, in general, counselling or education to promote smoke alarms is likely to have only a modest effect, if any, on smoke alarm ownership, function, or acquisition' 'Although observational studies support a substantial beneficial effect of smoke alarm ownership on fire-related injuries there were no data from randomised controlled trials on the effects of counselling or education on fire-related injuries'	Yes: ownership/installation rates provided in tables for NRSs (numbers/percentages). RCTs were pooled anyway
Dobbins, 1999[126]	N/A	N/A	Cannot tell	No (cannot tell which studies were RCTs and which NRSs)
Draper, 2000[115]	Of three trials in people not referred for mental disorders, two showed a positive effect. Of six trials in people with mental disorders, four showed a positive effect	In all studies, improvement was seen in depressed patients In three studies, patients with dementia did not improve or only improved slightly	(Author) 'In summary, the majority of RCTs and audits of acute treatment outcomes indicate that community old age psychiatry services are effective' (Reviewer) NRSs seem to show a more consistent positive effect for depressed patients	No

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Emmons, 2001[116]	Mixed results. One RCT reported significant effects on self-reported exposure reduction, the other two reported no difference in outcomes	Mixed results. One NRS reported no significant differences in outcome. The other reported reductions in exposure	(Reviewer) Cannot tell, seem quite similar (Author) Not reported	No
Fairbank, 2000[117]	Health education: eight RCTs reporting primary outcome – three significantly in favour, four non-significant trend in favour, one unclear Health sector initiative interventions (HSIs) (general): one RCT – significant increase HSI (US Women Infants and Children Programme): two RCTs, both significantly in favour	Health education: four NRSs – three of four reporting primary outcome showed no statistically significant differences between groups. Three before–after studies showed no statistically significant differences HSIs (general): two of three NRSs significant, one of two before–after studies significant HSI (US Women Infants and Children Programme): two of three NRSs significant. Five before–after studies 'supported findings from the RCTs and nRCTs' (none statistically significant)	(Reviewer) The results for RCTs of health education seem to be more in favour of the intervention than the results of NRSs (Author) Not reported	Yes (forest plots presented without pooling)
Giuffrida, 2004[118]	The PCPs receiving the target payment had an influenza vaccination rate 5.9% higher than the control group (the relative difference with the control group was 9.4%), but the difference was not statistically significant. The study also reported that the change in influenza vaccination rate from baseline was larger in the intervention group. The difference in absolute change from baseline between intervention and control group was 6.8% (9.4% was the relative percentage change) and was statistically significant	Reported an improvement in primary and pre-school immunisation rates after the introduction of the target payment remuneration system in the Grampian region. For primary immunisations the proportion of general practices immunising at least 95% and 90% of their eligible populations improved by 50% and 20% respectively	(Reviewer) Cannot tell. Difficult to compare results as the NRS was an interrupted time series and does not have a control/comparison group (Author) 'The studies showed positive effects following the interventions, but the improvements were, in most of the cases, statistically non-significant, which may be in part due to the low power of the studies'	No

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Gosden, 2000[119]	*Capitation payment vs FFS:* One RCT showed no effect on PCP visits/ contacts, decreased referrals to specialists and hospitals in the capitation group only, increased visits to health and emergency departments, decreased hospitalisation, and higher expenditure *Salary payment vs FFS:* One RCT showed no difference in patient visits apart from emergency visits which were higher in salaried than FFS PCPs. For salaried PCPs, the number of patients enrolled and access to PCP was higher, continuity of care was lower	*Capitation payment vs FFS:* One NRS showed increased visits/ contacts, provision of diagnostic and curative services, and reduced prescription renewals and referrals to specialists and hospitals *Mixed capitation systems vs FFS:* One NRS showed no difference between groups	(Author) Not reported (Reviewer) No obvious systematic differences in results; RCTs seem to report a wider range of outcomes then NRSs	No (some data are reported in tables, according to different outcome measures, but never for > 1 RCT or > 1 nRCT within each outcome measure)
Hutt, 2004[121]	2 (of 14) large RCTs reported significant reductions in hospital use	Three of five other studies reported significant reductions in hospital use	(Reviewer) NRS more likely to show a reduction in hospital use (Author) Not reported	No (no data extraction tables)
Hulscher, 2001[120]	5/8 RCTs evaluating information transfer vs no intervention found a significant effect 1 RCT of learning through social influence vs no intervention found a significant effect 9/9 RCTs of physician reminders vs no intervention found a significant effect 12/14 RCTs of multifaceted interventions vs no intervention found a significant effect 2/2 RCTs of learning through social influence vs educational materials found a significant effect 1 RCT found no significant effect of feedback 1 RCT showed a small effect of organisational vs 'other' interventions 6/7 randomised comparisons showed a significant effect of multifaceted vs other interventions 5/6 RCTs showed a significant effect of multifaceted interventions vs group education 2/2 RCTs showed a significant effect of multifaceted interventions vs physician reminders	2/3 NRSs of feedback vs no intervention found a significant effect 1/2 NRSs of physician reminders vs no intervention found a significant effect 1 NRS of organisational interventions vs no intervention found a significant effect 7/10 NRSs of multifaceted interventions vs no intervention found a significant effect 1 NRS found no significant effect of physician reminders 1 NRS showed a significant effect of organisational vs 'other' interventions 3/9 non-randomised comparisons of multifaceted vs other interventions showed a negative effect, two showed no effect and four showed a positive effect	(Reviewer) There are numerous comparisons, many of which stratified according to randomised and non-randomised evidence. There is no obvious trend between them, although possibly slightly stronger positive effects are seen in RCTs than NRSs (Author) Not reported	Possibly (pre/post test means provided in tables. Need to examine further to see if data from NRS are available in the tables – currently it is not very clear)

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Karjalainen, 2003[122]	In the RCT there was no significant difference between the two groups in any of the assessed outcomes apart from the cost of the rehabilitation programme	In the NRS, effects of the multidisciplinary treatment programme did not differ from traditional care in any of the outcomes assessed at 12- and 24-month follow-up	(Reviewer) Yes (Author) Yes	There is only one RCT and one NRS but there is sufficient detail to pool them using SMDs
Klassen, 2000[158]	1/2 RCTs showed a significant effect for bicycle injury prevention 1/1 RCT showed a significant effect for motor vehicle restraint 0/2 RCTs showed a significant effect for pedestrian injury prevention 1/1 RCT demonstrated improved knowledge for alcohol adolescent use and vehicle safety interventions	8/10 NRSs showed a significant effect for bicycle injury prevention 3/4 NRSs showed a significant effect for motor vehicle restraint 2/2 NRSs showed a significant effect for pedestrian injury prevention 1/4 NRSs showed a significant effect on injury prevention 2/2 NRSs demonstrated improved knowledge for alcohol adolescent use and vehicle safety interventions	(Reviewer) Results similar apart from pedestrian injury prevention, where only NRSs showed significant effects (Author) Not reported	Possibly, for some outcomes
Linton, 2001[159]	Only one of nine RCTs reported significant positive effect of back schools All four RCTs of lumbar supports were negative 5/6 RCTs reported positive effects of exercises	3 of 5 CCTs reported positive effect of back schools Both CCTs of lumbar supports were positive No CCTs of exercise	(Reviewer) No (Author) No	No
Meads, 1999[129]	RCT – no significant differences	NRSs – wide variation, no trend seen	(Reviewer) No (Author) No	Yes
Pusey, 2001[124]	11 of 18 RCTs showed either no difference between groups or a difference in only one of many variables measured	5 of 12 NRSs showed either no difference between groups or a difference in one of many variables measured. One NRS showed a negative effect of the intervention	(Reviewer) Similar, perhaps a wider range of effect in the NRSs than in the RCTs (but hard to tell without meta-analysis) (Author) A detailed critique is provided of the studies, but no comment is made about how effects vary according to design	No – not enough data provided in tables

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Reeves, 2001[125]	Outcomes were only reported relating to staff satisfaction (actual results not reported). States 'all papers report positive outcomes'	All papers report positive outcomes	The author discusses the results in relation to 'good', 'acceptable' and 'poor quality' studies. The one RCT falls into the 'good' quality category, along with some 'longitudinal' designs. It is not possible to discern any meaningful differences between results from randomised and non-randomised evidence	No
Scott, 2003[127]	CD-ROM, one RCT, improved feelings of control (compared with book) but not understanding of events School reintegration programme, one RCT, reduced anxiety Self-care coping intervention, one RCT, no significant differences	Computer-assisted instructional programme, one NRS, improved knowledge School reintegration programme, one NRS, some significant improvements Group therapy, two NRSs, no significant difference Planned play and story telling, one NRS, no significant differences Art therapy, one NRS, statistics not reported but more 'good responders' with intervention	(Reviewer) Of the RCTs and NRS that investigate similar interventions, both report mixed but similar results depending on outcomes assessed Authors do not compare results of RCTs and NRSs but state 'findings are difficult to interpret and summarise due to (a) inherent problems with the design of individual studies, and (b) the heterogeneity of their aims, study designs, interventions and controls, patient populations, outcomes, assessment instruments, and methods of analysis'	No
Stead, 2005[128]	Decreased sales: three RCTs found significant reduction in sales to minors, three found no significant difference Ease of access: one RCT that assessed ease of access found that purchase attempts decreased in the intervention communities and increased in the control communities Tobacco use: one RCT that assessed tobacco use found a lower rate of increase in intervention communities	Decreased sales: two NRSs reported significant reduction in sales to minors, all others showed no significant difference Ease of access: five of six NRSs that assessed ease of access found it was reduced following the intervention Tobacco use: five of seven NRSs reported lower prevalence of tobacco use following intervention	(Reviewer) For the outcome of decreased tobacco sales, the RCTs seem to have more positive results than the NRSs (taking a vote-counting approach) Authors do not compare results of NRSs and RCTs	No

Review	Results – RCTs	Results – NRSs	Results judged equivalent (reviewer/ author)	Sufficient detail for re-analysis?
Weigand, 2001[160]	A structured interdisciplinary approach significantly reduced number of falls	Case-finding studies showed it was possible to identify patients at risk of falls	No comment from author (Reviewer) The RCT and two case-finding studies had different aims and outcomes, therefore cannot be compared	No
Weir, 1999[161]	2/5 RCTs supported occupational therapy services in domiciliary setting Effect on caregiver – one of two RCTs showed a significant effect	Six NRSs of specific community-based interventions produced inconsistent results Effect on caregiver – one of two NRS showed a significant effect	*Community based interventions* (Reviewer) No (Author) No – conclude that greater credence should be placed on the RCTs. NRSs produced conflicting results and should be interpreted with caution *Effect on caregiver:* (Reviewer) Equivalent (some evidence but not conclusive) (Author) Not reported	No

CCT, controlled clinical trial; FFS, fee for service; N/A, not applicable; OR, odds ratio; PCP, primary care physician; SD, standard deviation.

DOI: 10.3310/hta14160

Appendix 4.13: Summary table of policy intervention reviews that pooled randomised and non-randomised designs in a meta-analysis

Review details	Description of intervention	Policy level	Target group	Rationale given by authors for pooling	Obvious differences between RCTs and NRSs (reviewers' judgement)	Sufficient detail for re-analysis? (yes/no)
Kendrick, 2000[137] 14 RCTs 3 quasi-experimental/ non-randomised studies	Home visiting to improve parenting and the quality of the home environment	Health promotion provision/ organisation	Mother and child	All studies that reported 'home observation for measurement of the environment inventory' scores (mean and standard deviation or *p*-value) were included in the meta-analysis. Analysis was repeated restricting to RCTs and to studies with a quality score of 0.5 or above No rationale was given by the authors	In two of the three NRSs, families were referred for interventions or had psychosocial problems	No
Elkan, 2000[132] 13 RCTs 2 quasi-experimental studies	Home visiting programmes that offer health promotion and preventive care	Health promotion provision/ organisation	Older people living at home	Meta-analysis was performed on RCTs only; NRSs were not included in the analysis. No rationale for pooling is given	No obvious differences	No
Posavac, 1999[141] 14 RCTs 22 NRSs	Peer-based interventions designed to change attitudes, behaviours, knowledge and skills in ways thought to improve health	Health promotion/ education	Any population	None given. Authors do note that 'given the large samples in many of these studies, even very small differences are statistically significant… therefore a more important issue to consider is whether the effect of a programme, given its cost, is of practical or substantive significance'. Also 'these studies differed greatly from each other; the focus of the meta-analysis was on the effectiveness of peer-based health education, not on a particular form of peer-based interventions'. Randomisation was one of the variables coded but was not used to investigate heterogeneity	No obvious differences (no details presented so cannot tell)	No

Review details	Description of intervention	Policy level	Target group	Rationale given by authors for pooling	Obvious differences between RCTs and NRSs (reviewers' judgement)	Sufficient detail for re-analysis? (yes/no)
Dusseldorp, 1999[131] 28 RCTs 9 quasi-experimental studies	Psychoeducational programmes for cardiac rehabilitation	Policy for an institution	Coronary heart disease patients	Not given 'Quasi-experiments were included only when samples where stratified or matched pairwise, or when a certain time period was used as an assignment rule for patients from the same hospital'. Authors' view: in both RCTs and NRSs, population effect size estimates will be more accurate when pre-test differences are taken into account. The review explores moderating effect of study features: randomisation was not a moderator of treatment success	No RCTs were conducted on health education and exercise training evaluations. Otherwise, no obvious differences	No
Gruen, 2003[133] 5 RCTs 2 CBAs 2 ITSs	Specialist outreach clinics in primary care and rural hospital settings	Health-care provision/ organisation	Patients eligible for specialist care, primary health-care practitioners and specialists	Not given 'Only studies that were similar in terms of setting, intervention and outcome assessment were subjected to statistical meta-analysis'. Only three studies were pooled, all were RCTs. Should we exclude it then?	Differences between all studies, no obvious differences between RCTs and NRSs	Yes (small number of studies for each outcome)
Higginson, 2002[134] 1 RCT 1 prospective study with comparison group 11 retrospective/ observational/cross-sectional	Hospital-based palliative care teams	Institutional policy	Patients with progressive life-threatening illness and their families, carers or close friends	Not given	No obvious differences in sample size, setting or intervention. No further data presented	Effect sizes for individual studies presented. CIs not given
Hyde, 2000[135] 7 RCTs (but no details of randomisation process in 5 of these) 2 quasi-randomised	Supporting discharge from hospital to home	Institutional policy	Older people	Not given But only included randomised or quasi-randomised trials	No obvious differences	Yes (forest plot with data presented).
Johnson, 2003[136] 12 RCTs 1 quasi-experimental study	Modifying sexual risk behaviours for preventing HIV infection	Policy for a community	Men who have sex with men	Not given But only included RCTs or 'strong' quasi-experimental studies	No obvious differences	Yes, for one outcome

Review details	Description of intervention	Policy level	Target group	Rationale given by authors for pooling	Obvious differences between RCTs and NRSs (reviewers' judgement)	Sufficient detail for re-analysis? (yes/no)
Legler, 2002[138] Total 38 No of designs unclear	Interventions to promote mammography	Health-care provision/ organisation	Women with historically lower rates of screening	Not given	Not possible to tell, as study design not indicated	No. Unable to identify design of included studies
Monninkhof, 2002[139] 8 RCTs 1 CCT	Self management education for COPD	Health-care provision/ organisation	Patients with COPD	Not given	No obvious differences	No. The CCT is not included in any forest plot (different outcomes)
Thompson O'Brien, 2001[162] 30 RCTs 2 non-equivalent group designs (allocation by a non-random process other than participant choice, data collection contemporaneous, and choice of control site/activity appeared appropriate)	Planned educational activities: meetings, conferences, lectures, workshops, seminars, symposia and courses that occurred off-site from the practice setting	Health professional education provision/ organisation	Qualified health professionals or health professionals in postgraduate training	No explicit statements. However, in the two NEGDs protection against bias was scored as 'moderate' because of adequate blinding and follow-up even though the groups were not randomly assigned. In the RCTs, seven scored as high and 24 as moderate protection against bias	No obvious differences	No
Ziguras, 2000[147] 35 RCTs 9 controlled studies	Mental health case management	Health-care provision/ organisation/ policy for a community	Adults with serious mental illness	Not given	No obvious differences in sample size, model or outcome domain. No further information	No
Szilagyi, 2002[143] 42 RCTs 3 CBAs (latest version)	Interventions aimed at improving immunisation rates	Health-care provision/ organisation	Health-care personnel who deliver immunisations and children or adults who receive immunisations in any setting	Not given. Looks like only RCTs were included in the meta-analyses in the latest version. Exclude?	Differences between all studies in participants and interventions, but no obvious differences between RCTs and NRSs	Yes

Review details	Description of intervention	Policy level	Target group	Rationale given by authors for pooling	Obvious differences between RCTs and NRSs (reviewers' judgement)	Sufficient detail for re-analysis? (yes/no)
Cuijpers, 2002[130] 11 RCTs 1 NRS	School-based drug prevention programmes – studies had to compare peer-led to adult-led programmes	Institution (school) Education provision/ organisation	Schoolchildren	No stated rationale but 'as the examined interventions differed considerably, as well as other characteristics of the studies, the random effects model was chosen'	The one NRS was much smaller than the RCTs and had shorter follow-up	Yes
Prendergast, 2000[142] 12 RCTs 131 NRSs	Drug dependence treatment programmes	Health-care provision/ organisation	Adults with drug dependence in the US or Canada	No stated rationale but 'because of heterogeneity among studies, the data are analysed in terms of type of outcome variable (drug use and crime), type of design (single-group and comparison group) and type of treatment'. Separates controlled and uncontrolled studies, but not RCTs and NRSs	Not possible to tell, data not provided	No
Parker, 2000[140] All randomised or pseudo-randomised (numbers of each not given, 45 in total)	Compares different locations of care for older patients	Institution and community	Older (65+ years) people	No rationale. Studies were all randomised or pseudo-randomised. Authors state that opportunities for meaningful meta-analysis were limited. Where significant heterogeneity was evident, ORs were recalculated using a random effect model	Not possible to tell	No
Allaby, 1999[129] 2 RCTs 7 NRSs (6 with controls, 1 before-after study)	Antenatal anti-D prophylaxis	Institution (hospital)	Rhesus-D negative women who bore rhesus-D positive babies	Before pooling, authors state that because only one of the RCTs used a dosage regimen that is currently considered appropriate, the NRSs have been retained for further consideration. After pooling, authors state that results of the NRSs are very similar to those of the one randomised study that used a currently recommended dose of anti-D, which suggests that the NRSs are unlikely to be seriously biased	Only one of two RCTs used a dosage currently considered appropriate	Yes

Review details	Description of intervention	Policy level	Target group	Rationale given by authors for pooling	Obvious differences between RCTs and NRSs (reviewers' judgement)	Sufficient detail for re-analysis? (yes/no)
Yin, 2002[146] 17 RCTs 9 NRSs	Group and individual interventions to reduce burden on caregivers of the frail elderly	Community	Family caregivers of elderly persons	No clear rationale given. They say that they included both true experimental (random assignment) and quasi-experimental designs with a control group (convenient assignment). There is some investigation of random assignment as a potential moderator in heterogeneity investigation, but only for group interventions. Individual interventions had a non-significant *Q* statistic, indicating homogeneity, therefore no further searching for potential moderators was performed	No obvious differences	No (gives effect size only for each study, no variance)
Yabroff, 2001[145] 54 RCTs 12 NRSs	Patient targeted interventions to increase mammography use both outside and inside the primary care medical setting	Health-care provision organisation Community	Women in the US eligible for mammography	No rationale stated. Only included randomised or 'concurrently controlled' studies. Effect sizes calculated as endpoint differences for RCTs and change score differences for NRSs, but then combined	Not possible to tell, data not provided	No

CBA, controlled before/after study; CCT, controlled clinical trial; COPD, chronic obstructive pulmonary disease; ITS, interrupted time series; NEGD, non-equivalent group design.

Appendix 5

Additional information on variance in our analyses

Here we find algebraic expressions for the standard deviations of nRCTs and RCTs resampled from RCT data. We let y_{1i} and y_{2i} be the log odds in area i in the intervention and control arms respectively; for a quantitative outcome, they would be the means. We let n be the number of areas, so that there are n RCTs and $n(n-1)$ nRCTs. We define

$d_{ij} = (y_{2j} - y_{1i})$, the intervention effect in a resampled study

$\bar{d} = (\bar{y}_2 - \bar{y}_1)$, the average intervention effect

$SS_{RCT} = \sum_{i=1}^{n}\left(d_{ii} - \bar{d}\right)^2$, the sum of squares for the intervention effects in all RCTs

$SS_{ALL} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ii} - \bar{d}\right)^2$, the sum of squares for the intervention effects in all resampled studies

$S_{Bk}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{ki} - \bar{y}_k)^2$, the variance between areas in arm k

$S_{B12} = \frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$, the covariance between arms

$r = s_{B12} / s_{B1}s_{B2}$, the correlation between arms

Note that in this work we define all variances using n rather than $(n-1)$ as denominator. A little algebra then shows that

$$SS_{RCT} = n(s_{B1}^2 + s_{B2}^2 - 2rs_{B1}s_{B2})$$

$$SS_{ALL} = n^2(s_{B1}^2 + s_{B2}^2)$$

and hence the variances V_{RCT} and V_{nRCT} in RCTs and nRCTs are given by

$$V_{RCT} = SS_{RCT} / n = s_{B1}^2 + s_{B2}^2 - 2rs_{B1}s_{B2}$$

$$V_{nRCT} = (SS_{ALL} - SS_{RCT}) / n(n-1) = s_{B1}^2 + s_{B2}^2 + \frac{2}{n-1}rs_{B1}s_{B2}$$

It follows that the variances in RCTs and nRCTs are equal if, and only if, $r = 0$, so that a significance test of $V_{\text{RCT}} = V_{\text{nRCT}}$ is performed by testing $r = 0$.

For the covariate-adjusted analysis, we propose performing an approximate significance test of $V_{\text{RCT}} = V_{\text{nRCT}}$ by a two-stage procedure. In the first stage, we fit the logistic regression model $\text{logodds}_{kil} = a_{ki} + b\,x_{kil}$, where logodds_{kil} is the log odds in the lth individual in the ith area in arm k and x_{kil} is that individual's covariate vector. In the second stage, we compute an adjusted correlation r_{adj} using the a_{ki} in place of the y_{ki} above. The approximate p-value for the test of $V_{\text{RCT}} = V_{\text{nRCT}}$ is the p-value for testing $r_{\text{adj}} = 0$.

We have used the whole of the sampled areas in the resampled studies, but Deeks *et al.*[43] sampled m individuals with replacement from each selected area. This adds an additional term V_2 to the above results:

$$V_{RCT}^* = V_{RCT} + V_2$$
$$V_{nRCT}^* = V_{nRCT} + V_2$$

where $V_2 = \frac{1}{n}\sum_{i=1}^{n}(s_{W1i}^2 + s_{W2i}^2)$

and s_{Wki}^2 = variance of sampled log odds in arm k and i

$$= \frac{1}{mp_{ki}(1-p_{ki})}$$

where p_{ki} = observed proportion in k area i

This does not affect the result that the variances in RCTs and nRCTs are equal if and only if $r = 0$.

Health Technology Assessment reports published to date

Volume 1, 1997

No. 1
Home parenteral nutrition: a systematic review.
By Richards DM, Deeks JJ, Sheldon TA, Shaffer JL.

No. 2
Diagnosis, management and screening of early localised prostate cancer.
A review by Selley S, Donovan J, Faulkner A, Coast J, Gillatt D.

No. 3
The diagnosis, management, treatment and costs of prostate cancer in England and Wales.
A review by Chamberlain J, Melia J, Moss S, Brown J.

No. 4
Screening for fragile X syndrome.
A review by Murray J, Cuckle H, Taylor G, Hewison J.

No. 5
A review of near patient testing in primary care.
By Hobbs FDR, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, *et al*.

No. 6
Systematic review of outpatient services for chronic pain control.
By McQuay HJ, Moore RA, Eccleston C, Morley S, de C Williams AC.

No. 7
Neonatal screening for inborn errors of metabolism: cost, yield and outcome.
A review by Pollitt RJ, Green A, McCabe CJ, Booth A, Cooper NJ, Leonard JV, *et al*.

No. 8
Preschool vision screening.
A review by Snowdon SK, Stewart-Brown SL.

No. 9
Implications of socio-cultural contexts for the ethics of clinical trials.
A review by Ashcroft RE, Chadwick DW, Clark SRL, Edwards RHT, Frith L, Hutton JL.

No. 10
A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment.
By Davis A, Bamford J, Wilson I, Ramkalawan T, Forshaw M, Wright S.

No. 11
Newborn screening for inborn errors of metabolism: a systematic review.
By Seymour CA, Thomason MJ, Chalmers RA, Addison GM, Bain MD, Cockburn F, *et al*.

No. 12
Routine preoperative testing: a systematic review of the evidence.
By Munro J, Booth A, Nicholl J.

No. 13
Systematic review of the effectiveness of laxatives in the elderly.
By Petticrew M, Watt I, Sheldon T.

No. 14
When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies.
A review by Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L.

Volume 2, 1998

No. 1
Antenatal screening for Down's syndrome.
A review by Wald NJ, Kennard A, Hackshaw A, McGuire A.

No. 2
Screening for ovarian cancer: a systematic review.
By Bell R, Petticrew M, Luengo S, Sheldon TA.

No. 3
Consensus development methods, and their use in clinical guideline development.
A review by Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al*.

No. 4
A cost–utility analysis of interferon beta for multiple sclerosis.
By Parkin D, McNamee P, Jacoby A, Miller P, Thomas S, Bates D.

No. 5
Effectiveness and efficiency of methods of dialysis therapy for end-stage renal disease: systematic reviews.
By MacLeod A, Grant A, Donaldson C, Khan I, Campbell M, Daly C, *et al*.

No. 6
Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model.
By Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G.

No. 7
Antimicrobial prophylaxis in colorectal surgery: a systematic review of randomised controlled trials.
By Song F, Glenny AM.

No. 8
Bone marrow and peripheral blood stem cell transplantation for malignancy.
A review by Johnson PWM, Simnett SJ, Sweetenham JW, Morgan GJ, Stewart LA.

No. 9
Screening for speech and language delay: a systematic review of the literature.
By Law J, Boyle J, Harris F, Harkness A, Nye C.

No. 10
Resource allocation for chronic stable angina: a systematic review of effectiveness, costs and cost-effectiveness of alternative interventions.
By Sculpher MJ, Petticrew M, Kelland JL, Elliott RA, Holdright DR, Buxton MJ.

No. 11
Detection, adherence and control of hypertension for the prevention of stroke: a systematic review.
By Ebrahim S.

No. 12
Postoperative analgesia and vomiting, with special reference to day-case surgery: a systematic review.
By McQuay HJ, Moore RA.

No. 13
Choosing between randomised and nonrandomised studies: a systematic review.
By Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C.

No. 14
Evaluating patient-based outcome measures for use in clinical trials.
A review by Fitzpatrick R, Davey C, Buxton MJ, Jones DR.

No. 15
Ethical issues in the design and conduct of randomised controlled trials.
A review by Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J.

No. 16
Qualitative research methods in health technology assessment: a review of the literature.
By Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P.

No. 17
The costs and benefits of paramedic skills in pre-hospital trauma care.
By Nicholl J, Hughes S, Dixon S, Turner J, Yates D.

No. 18
Systematic review of endoscopic ultrasound in gastro-oesophageal cancer.
By Harris KM, Kelly S, Berry E, Hutton J, Roderick P, Cullingworth J, *et al*.

No. 19
Systematic reviews of trials and other studies.
By Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F.

No. 20
Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses.
A review by Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, *et al*.

Volume 3, 1999

No. 1
Informed decision making: an annotated bibliography and systematic review.
By Bekker H, Thornton JG, Airey CM, Connelly JB, Hewison J, Robinson MB, *et al*.

No. 2
Handling uncertainty when performing economic evaluation of healthcare interventions.
A review by Briggs AH, Gray AM.

No. 3
The role of expectancies in the placebo effect and their use in the delivery of health care: a systematic review.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Thomas H.

No. 4
A randomised controlled trial of different approaches to universal antenatal HIV testing: uptake and acceptability. Annex: Antenatal HIV testing – assessment of a routine voluntary approach.
By Simpson WM, Johnstone FD, Boyd FM, Goldberg DJ, Hart GJ, Gormley SM, *et al*.

No. 5
Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.
By Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ.

No. 6
Assessing the costs of healthcare technologies in clinical trials.
A review by Johnston K, Buxton MJ, Jones DR, Fitzpatrick R.

No. 7
Cooperatives and their primary care emergency centres: organisation and impact.
By Hallam L, Henthorne K.

No. 8
Screening for cystic fibrosis.
A review by Murray J, Cuckle H, Taylor G, Littlewood J, Hewison J.

No. 9
A review of the use of health status measures in economic evaluation.
By Brazier J, Deverill M, Green C, Harper R, Booth A.

No. 10
Methods for the analysis of quality-of-life and survival data in health technology assessment.
A review by Billingham LJ, Abrams KR, Jones DR.

No. 11
Antenatal and neonatal haemoglobinopathy screening in the UK: review and economic analysis.
By Zeuner D, Ades AE, Karnon J, Brown J, Dezateux C, Anionwu EN.

No. 12
Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses.
A review by Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, *et al*.

No. 13
'Early warning systems' for identifying new healthcare technologies.
By Robert G, Stevens A, Gabbay J.

No. 14
A systematic review of the role of human papillomavirus testing within a cervical screening programme.
By Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*.

No. 15
Near patient testing in diabetes clinics: appraising the costs and outcomes.
By Grieve R, Beech R, Vincent J, Mazurkiewicz J.

No. 16
Positron emission tomography: establishing priorities for health technology assessment.
A review by Robert G, Milne R.

No. 17 (Pt 1)
The debridement of chronic wounds: a systematic review.
By Bradley M, Cullum N, Sheldon T.

No. 17 (Pt 2)
Systematic reviews of wound care management: (2) Dressings and topical agents used in the healing of chronic wounds.
By Bradley M, Cullum N, Nelson EA, Petticrew M, Sheldon T, Torgerson D.

No. 18
A systematic literature review of spiral and electron beam computed tomography: with particular reference to clinical applications in hepatic lesions, pulmonary embolus and coronary artery disease.
By Berry E, Kelly S, Hutton J, Harris KM, Roderick P, Boyce JC, *et al*.

No. 19
What role for statins? A review and economic model.
By Ebrahim S, Davey Smith G, McCabe C, Payne N, Pickin M, Sheldon TA, *et al*.

No. 20
Factors that limit the quality, number and progress of randomised controlled trials.
A review by Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, *et al*.

No. 21
Antimicrobial prophylaxis in total hip replacement: a systematic review.
By Glenny AM, Song F.

No. 22
Health promoting schools and health promotion in schools: two systematic reviews.
By Lister-Sharp D, Chapman S, Stewart-Brown S, Sowden A.

No. 23
Economic evaluation of a primary care-based education programme for patients with osteoarthritis of the knee.
A review by Lord J, Victor C, Littlejohns P, Ross FM, Axford JS.

Volume 4, 2000

No. 1
The estimation of marginal time preference in a UK-wide sample (TEMPUS) project.
A review by Cairns JA, van der Pol MM.

No. 2
Geriatric rehabilitation following fractures in older people: a systematic review.
By Cameron I, Crotty M, Currie C, Finnegan T, Gillespie L, Gillespie W, *et al*.

No. 3
Screening for sickle cell disease and thalassaemia: a systematic review with supplementary research.
By Davies SC, Cronin E, Gill M, Greengross P, Hickman M, Normand C.

No. 4
Community provision of hearing aids and related audiology services.
A review by Reeves DJ, Alborz A, Hickson FS, Bamford JM.

No. 5
False-negative results in screening programmes: systematic review of impact and implications.
By Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K.

No. 6
Costs and benefits of community postnatal support workers: a randomised controlled trial.
By Morrell CJ, Spiby H, Stewart P, Walters S, Morgan A.

No. 7
Implantable contraceptives (subdermal implants and hormonally impregnated intrauterine systems) versus other forms of reversible contraceptives: two systematic reviews to assess relative effectiveness, acceptability, tolerability and cost-effectiveness.
By French RS, Cowan FM, Mansour DJA, Morris S, Procter T, Hughes D, *et al*.

No. 8
An introduction to statistical methods for health technology assessment.
A review by White SJ, Ashby D, Brown PJ.

No. 9
Disease-modifying drugs for multiple sclerosis: a rapid and systematic review.
By Clegg A, Bryant J, Milne R.

No. 10
Publication and related biases.
A review by Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ.

No. 11
Cost and outcome implications of the organisation of vascular services.
By Michaels J, Brazier J, Palfreyman S, Shackley P, Slack R.

No. 12
Monitoring blood glucose control in diabetes mellitus: a systematic review.
By Coster S, Gulliford MC, Seed PT, Powrie JK, Swaminathan R.

No. 13
The effectiveness of domiciliary health visiting: a systematic review of international studies and a selective review of the British literature.
By Elkan R, Kendrick D, Hewitt M, Robinson JJA, Tolley K, Blair M, *et al*.

No. 14
The determinants of screening uptake and interventions for increasing uptake: a systematic review.
By Jepson R, Clegg A, Forbes C, Lewis R, Sowden A, Kleijnen J.

No. 15
The effectiveness and cost-effectiveness of prophylactic removal of wisdom teeth.
A rapid review by Song F, O'Meara S, Wilson P, Golder S, Kleijnen J.

No. 16
Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views.
By Bricker L, Garcia J, Henderson J, Mugford M, Neilson J, Roberts T, *et al*.

No. 17
A rapid and systematic review of the effectiveness and cost-effectiveness of the taxanes used in the treatment of advanced breast and ovarian cancer.
By Lister-Sharp D, McDonagh MS, Khan KS, Kleijnen J.

No. 18
Liquid-based cytology in cervical screening: a rapid and systematic review.
By Payne N, Chilcott J, McGoogan E.

No. 19
Randomised controlled trial of non-directive counselling, cognitive–behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care.
By King M, Sibbald B, Ward E, Bower P, Lloyd M, Gabbay M, *et al*.

No. 20
Routine referral for radiography of patients presenting with low back pain: is patients' outcome influenced by GPs' referral for plain radiography?
By Kerry S, Hilton S, Patel S, Dundas D, Rink E, Lord J.

No. 21
Systematic reviews of wound care management: (3) antimicrobial agents for chronic wounds; (4) diabetic foot ulceration.
By O'Meara S, Cullum N, Majid M, Sheldon T.

No. 22
Using routine data to complement and enhance the results of randomised controlled trials.
By Lewsey JD, Leyland AH, Murray GD, Boddy FA.

No. 23
Coronary artery stents in the treatment of ischaemic heart disease: a rapid and systematic review.
By Meads C, Cummins C, Jolly K, Stevens A, Burls A, Hyde C.

No. 24
Outcome measures for adult critical care: a systematic review.
By Hayes JA, Black NA, Jenkinson C, Young JD, Rowan KM, Daly K, *et al*.

No. 25
A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding.
By Fairbank L, O'Meara S, Renfrew MJ, Woolridge M, Sowden AJ, Lister-Sharp D.

No. 26
Implantable cardioverter defibrillators: arrhythmias. A rapid and systematic review.
By Parkes J, Bryant J, Milne R.

No. 27
Treatments for fatigue in multiple sclerosis: a rapid and systematic review.
By Brañas P, Jordan R, Fry-Smith A, Burls A, Hyde C.

No. 28
Early asthma prophylaxis, natural history, skeletal development and economy (EASE): a pilot randomised controlled trial.
By Baxter-Jones ADG, Helms PJ, Russell G, Grant A, Ross S, Cairns JA, *et al*.

No. 29
Screening for hypercholesterolaemia versus case finding for familial hypercholesterolaemia: a systematic review and cost-effectiveness analysis.
By Marks D, Wonderling D, Thorogood M, Lambert H, Humphries SE, Neil HAW.

No. 30
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists in the medical management of unstable angina.
By McDonagh MS, Bachmann LM, Golder S, Kleijnen J, ter Riet G.

No. 31
A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma.
By Turner J, Nicholl J, Webber L, Cox H, Dixon S, Yates D.

No. 32
Intrathecal pumps for giving opioids in chronic pain: a systematic review.
By Williams JE, Louw G, Towlerton G.

No. 33
Combination therapy (interferon alfa and ribavirin) in the treatment of chronic hepatitis C: a rapid and systematic review.
By Shepherd J, Waugh N, Hewitson P.

No. 34
A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies.
By MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS.

No. 35
Intravascular ultrasound-guided interventions in coronary artery disease: a systematic literature review, with decision-analytic modelling, of outcomes and cost-effectiveness.
By Berry E, Kelly S, Hutton J, Lindsay HSJ, Blaxill JM, Evans JA, *et al*.

No. 36
A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression.
By Simpson S, Corney R, Fitzgerald P, Beecham J.

No. 37
Systematic review of treatments for atopic eczema.
By Hoare C, Li Wan Po A, Williams H.

No. 38
Bayesian methods in health technology assessment: a review.
By Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR.

No. 39
The management of dyspepsia: a systematic review.
By Delaney B, Moayyedi P, Deeks J, Innes M, Soo S, Barton P, *et al*.

No. 40
A systematic review of treatments for severe psoriasis.
By Griffiths CEM, Clark CM, Chalmers RJG, Li Wan Po A, Williams HC.

Volume 5, 2001

No. 1
Clinical and cost-effectiveness of donepezil, rivastigmine and galantamine for Alzheimer's disease: a rapid and systematic review.
By Clegg A, Bryant J, Nicholson T, McIntyre L, De Broe S, Gerard K, *et al*.

No. 2
The clinical effectiveness and cost-effectiveness of riluzole for motor neurone disease: a rapid and systematic review.
By Stewart A, Sandercock J, Bryan S, Hyde C, Barton PM, Fry-Smith A, *et al*.

No. 3
Equity and the economic evaluation of healthcare.
By Sassi F, Archard L, Le Grand J.

No. 4
Quality-of-life measures in chronic diseases of childhood.
By Eiser C, Morse R.

No. 5
Eliciting public preferences for healthcare: a systematic review of techniques.
By Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, *et al*.

No. 6
General health status measures for people with cognitive impairment: learning disability and acquired brain injury.
By Riemsma RP, Forbes CA, Glanville JM, Eastwood AJ, Kleijnen J.

No. 7
An assessment of screening strategies for fragile X syndrome in the UK.
By Pembrey ME, Barnicoat AJ, Carmichael B, Bobrow M, Turner G.

No. 8
Issues in methodological research: perspectives from researchers and commissioners.
By Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, *et al*.

No. 9
Systematic reviews of wound care management: (5) beds; (6) compression; (7) laser therapy, therapeutic ultrasound, electrotherapy and electromagnetic therapy.
By Cullum N, Nelson EA, Flemming K, Sheldon T.

No. 10
Effects of educational and psychosocial interventions for adolescents with diabetes mellitus: a systematic review.
By Hampson SE, Skinner TC, Hart J, Storey L, Gage H, Foxcroft D, *et al*.

No. 11
Effectiveness of autologous chondrocyte transplantation for hyaline cartilage defects in knees: a rapid and systematic review.
By Jobanputra P, Parry D, Fry-Smith A, Burls A.

No. 12
Statistical assessment of the learning curves of health technologies.
By Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT.

No. 13
The effectiveness and cost-effectiveness of temozolomide for the treatment of recurrent malignant glioma: a rapid and systematic review.
By Dinnes J, Cave C, Huang S, Major K, Milne R.

No. 14
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of debriding agents in treating surgical wounds healing by secondary intention.
By Lewis R, Whiting P, ter Riet G, O'Meara S, Glanville J.

No. 15
Home treatment for mental health problems: a systematic review.
By Burns T, Knapp M, Catty J, Healey A, Henderson J, Watt H, *et al*.

No. 16
How to develop cost-conscious guidelines.
By Eccles M, Mason J.

No. 17
The role of specialist nurses in multiple sclerosis: a rapid and systematic review.
By De Broe S, Christopher F, Waugh N.

No. 18
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of orlistat in the management of obesity.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 19
The clinical effectiveness and cost-effectiveness of pioglitazone for type 2 diabetes mellitus: a rapid and systematic review.
By Chilcott J, Wight J, Lloyd Jones M, Tappenden P.

No. 20
Extended scope of nursing practice: a multicentre randomised controlled trial of appropriately trained nurses and preregistration house officers in preoperative assessment in elective general surgery.
By Kinley H, Czoski-Murray C, George S, McCabe C, Primrose J, Reilly C, *et al*.

No. 21
Systematic reviews of the effectiveness of day care for people with severe mental disorders: (1) Acute day hospital versus admission; (2) Vocational rehabilitation; (3) Day hospital versus outpatient care.
By Marshall M, Crowther R, Almaraz- Serrano A, Creed F, Sledge W, Kluiter H, *et al*.

No. 22
The measurement and monitoring of surgical adverse events.
By Bruce J, Russell EM, Mollison J, Krukowski ZH.

No. 23
Action research: a systematic review and guidance for assessment.
By Waterman H, Tillen D, Dickson R, de Koning K.

No. 24
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of gemcitabine for the treatment of pancreatic cancer.
By Ward S, Morris E, Bansback N, Calvert N, Crellin A, Forman D, *et al*.

No. 25
A rapid and systematic review of the evidence for the clinical effectiveness and cost-effectiveness of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer.
By Lloyd Jones M, Hummel S, Bansback N, Orr B, Seymour M.

No. 26
Comparison of the effectiveness of inhaler devices in asthma and chronic obstructive airways disease: a systematic review of the literature.
By Brocklebank D, Ram F, Wright J, Barry P, Cates C, Davies L, *et al*.

No. 27
The cost-effectiveness of magnetic resonance imaging for investigation of the knee joint.
By Bryan S, Weatherburn G, Bungay H, Hatrick C, Salas C, Parry D, *et al*.

No. 28
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of topotecan for ovarian cancer.
By Forbes C, Shirran L, Bagnall A-M, Duffy S, ter Riet G.

No. 29
Superseded by a report published in a later volume.

No. 30
The role of radiography in primary care patients with low back pain of at least 6 weeks duration: a randomised (unblinded) controlled trial.
By Kendrick D, Fielding K, Bentley E, Miller P, Kerslake R, Pringle M.

No. 31
Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients.
By McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, *et al*.

No. 32
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of paclitaxel, docetaxel, gemcitabine and vinorelbine in non-small-cell lung cancer.
By Clegg A, Scott DA, Sidhu M, Hewitson P, Waugh N.

No. 33
Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives.
By Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G.

No. 34
Depot antipsychotic medication in the treatment of patients with schizophrenia: (1) Meta-review; (2) Patient and nurse attitudes.
By David AS, Adams C.

No. 35
A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression.
By Churchill R, Hunot V, Corney R, Knapp M, McGuire H, Tylee A, *et al*.

No. 36
Cost analysis of child health surveillance.
By Sanderson D, Wright D, Acton C, Duree D.

Volume 6, 2002

No. 1
A study of the methods used to select review criteria for clinical audit.
By Hearnshaw H, Harker R, Cheater F, Baker R, Grimshaw G.

No. 2
Fludarabine as second-line therapy for B cell chronic lymphocytic leukaemia: a technology assessment.
By Hyde C, Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, *et al*.

No. 3
Rituximab as third-line treatment for refractory or recurrent Stage III or IV follicular non-Hodgkin's lymphoma: a systematic review and economic evaluation.
By Wake B, Hyde C, Bryan S, Barton P, Song F, Fry-Smith A, *et al*.

No. 4
A systematic review of discharge arrangements for older people.
By Parker SG, Peet SM, McPherson A, Cannaby AM, Baker R, Wilson A, *et al*.

No. 5
The clinical effectiveness and cost-effectiveness of inhaler devices used in the routine management of chronic asthma in older children: a systematic review and economic evaluation.
By Peters J, Stevenson M, Beverley C, Lim J, Smith S.

No. 6
The clinical effectiveness and cost-effectiveness of sibutramine in the management of obesity: a technology assessment.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 7
The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: a systematic review.
By Berry E, Kelly S, Westwood ME, Davies LM, Gough MJ, Bamford JM, *et al*.

No. 8
Promoting physical activity in South Asian Muslim women through 'exercise on prescription'.
By Carroll B, Ali N, Azam N.

No. 9
Zanamivir for the treatment of influenza in adults: a systematic review and economic evaluation.
By Burls A, Clark W, Stewart T, Preston C, Bryan S, Jefferson T, *et al*.

No. 10
A review of the natural history and epidemiology of multiple sclerosis: implications for resource allocation and health economic models.
By Richards RG, Sampson FC, Beard SM, Tappenden P.

No. 11
Screening for gestational diabetes: a systematic review and economic evaluation.
By Scott DA, Loveman E, McIntyre L, Waugh N.

No. 12
The clinical effectiveness and cost-effectiveness of surgery for people with morbid obesity: a systematic review and economic evaluation.
By Clegg AJ, Colquitt J, Sidhu MK, Royle P, Loveman E, Walker A.

No. 13
The clinical effectiveness of trastuzumab for breast cancer: a systematic review.
By Lewis R, Bagnall A-M, Forbes C, Shirran E, Duffy S, Kleijnen J, *et al*.

No. 14
The clinical effectiveness and cost-effectiveness of vinorelbine for breast cancer: a systematic review and economic evaluation.
By Lewis R, Bagnall A-M, King S, Woolacott N, Forbes C, Shirran L, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of metal-on-metal hip resurfacing arthroplasty for treatment of hip disease.
By Vale L, Wyness L, McCormack K, McKenzie L, Brazzelli M, Stearns SC.

No. 16
The clinical effectiveness and cost-effectiveness of bupropion and nicotine replacement therapy for smoking cessation: a systematic review and economic evaluation.
By Woolacott NF, Jones L, Forbes CA, Mather LC, Sowden AJ, Song FJ, *et al*.

No. 17
A systematic review of effectiveness and economic evaluation of new drug treatments for juvenile idiopathic arthritis: etanercept.
By Cummins C, Connock M, Fry-Smith A, Burls A.

No. 18
Clinical effectiveness and cost-effectiveness of growth hormone in children: a systematic review and economic evaluation.
By Bryant J, Cave C, Mihaylova B, Chase D, McIntyre L, Gerard K, *et al*.

No. 19
Clinical effectiveness and cost-effectiveness of growth hormone in adults in relation to impact on quality of life: a systematic review and economic evaluation.
By Bryant J, Loveman E, Chase D, Mihaylova B, Cave C, Gerard K, *et al*.

No. 20
Clinical medication review by a pharmacist of patients on repeat prescriptions in general practice: a randomised controlled trial.
By Zermansky AG, Petty DR, Raynor DK, Lowe CJ, Freementle N, Vail A.

No. 21
The effectiveness of infliximab and etanercept for the treatment of rheumatoid arthritis: a systematic review and economic evaluation.
By Jobanputra P, Barton P, Bryan S, Burls A.

No. 22
A systematic review and economic evaluation of computerised cognitive behaviour therapy for depression and anxiety.
By Kaltenthaler E, Shackley P, Stevens K, Beverley C, Parry G, Chilcott J.

No. 23
A systematic review and economic evaluation of pegylated liposomal doxorubicin hydrochloride for ovarian cancer.
By Forbes C, Wilby J, Richardson G, Sculpher M, Mather L, Reimsma R.

No. 24
A systematic review of the effectiveness of interventions based on a stages-of-change approach to promote individual behaviour change.
By Riemsma RP, Pattenden J, Bridle C, Sowden AJ, Mather L, Watt IS, *et al*.

No. 25
A systematic review update of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists.
By Robinson M, Ginnelly L, Sculpher M, Jones L, Riemsma R, Palmer S, *et al*.

No. 26
A systematic review of the effectiveness, cost-effectiveness and barriers to implementation of thrombolytic and neuroprotective therapy for acute ischaemic stroke in the NHS.
By Sandercock P, Berge E, Dennis M, Forbes J, Hand P, Kwan J, *et al*.

No. 27
A randomised controlled crossover trial of nurse practitioner versus doctor-led outpatient care in a bronchiectasis clinic.
By Caine N, Sharples LD, Hollingworth W, French J, Keogan M, Exley A, *et al*.

No. 28
Clinical effectiveness and cost – consequences of selective serotonin reuptake inhibitors in the treatment of sex offenders.
By Adi Y, Ashcroft D, Browne K, Beech A, Fry-Smith A, Hyde C.

No. 29
Treatment of established osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Brazier JE, Stevenson M, Calvert NW, Lloyd Jones M.

No. 30
Which anaesthetic agents are cost-effective in day surgery? Literature review, national survey of practice and randomised controlled trial.
By Elliott RA Payne K, Moore JK, Davies LM, Harper NJN, St Leger AS, *et al*.

No. 31
Screening for hepatitis C among injecting drug users and in genitourinary medicine clinics: systematic reviews of effectiveness, modelling study and national survey of current practice.
By Stein K, Dalziel K, Walker A, McIntyre L, Jenkins B, Horne J, *et al*.

No. 32
The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L, *et al*.

No. 33
The effectiveness and cost-effectiveness of imatinib in chronic myeloid leukaemia: a systematic review.
By Garside R, Round A, Dalziel K, Stein K, Royle R.

No. 34
A comparative study of hypertonic saline, daily and alternate-day rhDNase in children with cystic fibrosis.
By Suri R, Wallis C, Bush A, Thompson S, Normand C, Flather M, *et al*.

No. 35
A systematic review of the costs and effectiveness of different models of paediatric home care.
By Parker G, Bhakta P, Lovett CA, Paisley S, Olsen R, Turner D, *et al*.

Volume 7, 2003

No. 1
How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study.
By Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J.

No. 2
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of home versus hospital or satellite unit haemodialysis for people with end-stage renal failure.
By Mowatt G, Vale L, Perez J, Wyness L, Fraser C, MacLeod A, *et al*.

No. 3
Systematic review and economic evaluation of the effectiveness of infliximab for the treatment of Crohn's disease.
By Clark W, Raftery J, Barton P, Song F, Fry-Smith A, Burls A.

No. 4
A review of the clinical effectiveness and cost-effectiveness of routine anti-D prophylaxis for pregnant women who are rhesus negative.
By Chilcott J, Lloyd Jones M, Wight J, Forman K, Wray J, Beverley C, *et al*.

No. 5
Systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma.
By Riley RD, Burchill SA, Abrams KR, Heney D, Lambert PC, Jones DR, *et al*.

No. 6
The cost-effectiveness of screening for *Helicobacter pylori* to reduce mortality and morbidity from gastric cancer and peptic ulcer disease: a discrete-event simulation model.
By Roderick P, Davies R, Raftery J, Crabbe D, Pearce R, Bhandari P, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of routine dental checks: a systematic review and economic evaluation.
By Davenport C, Elley K, Salas C, Taylor-Weetman CL, Fry-Smith A, Bryan S, *et al*.

No. 8
A multicentre randomised controlled trial assessing the costs and benefits of using structured information and analysis of women's preferences in the management of menorrhagia.
By Kennedy ADM, Sculpher MJ, Coulter A, Dwyer N, Rees M, Horsley S, *et al*.

No. 9
Clinical effectiveness and cost–utility of photodynamic therapy for wet age-related macular degeneration: a systematic review and economic evaluation.
By Meads C, Salas C, Roberts T, Moore D, Fry-Smith A, Hyde C.

No. 10
Evaluation of molecular tests for prenatal diagnosis of chromosome abnormalities.
By Grimshaw GM, Szczepura A, Hultén M, MacDonald F, Nevin NC, Sutton F, *et al*.

No. 11
First and second trimester antenatal screening for Down's syndrome: the results of the Serum, Urine and Ultrasound Screening Study (SURUSS).
By Wald NJ, Rodeck C, Hackshaw AK, Walters J, Chitty L, Mackinson AM.

No. 12
The effectiveness and cost-effectiveness of ultrasound locating devices for central venous access: a systematic review and economic evaluation.
By Calvert N, Hind D, McWilliams RG, Thomas SM, Beverley C, Davidson A.

No. 13
A systematic review of atypical antipsychotics in schizophrenia.
By Bagnall A-M, Jones L, Lewis R, Ginnelly L, Glanville J, Torgerson D, *et al.*

No. 14
Prostate Testing for Cancer and Treatment (ProtecT) feasibility study.
By Donovan J, Hamdy F, Neal D, Peters T, Oliver S, Brindle L, *et al*.

No. 15
Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation.
By Boland A, Dundar Y, Bagust A, Haycox A, Hill R, Mujica Mota R, *et al*.

No. 16
Screening for fragile X syndrome: a literature review and modelling.
By Song FJ, Barton P, Sleightholme V, Yao GL, Fry-Smith A.

No. 17
Systematic review of endoscopic sinus surgery for nasal polyps.
By Dalziel K, Stein K, Round A, Garside R, Royle P.

No. 18
Towards efficient guidelines: how to monitor guideline use in primary care.
By Hutchinson A, McIntosh A, Cox S, Gilbert C.

No. 19
Effectiveness and cost-effectiveness of acute hospital-based spinal cord injuries services: systematic review.
By Bagnall A-M, Jones L, Richardson G, Duffy S, Riemsma R.

No. 20
Prioritisation of health technology assessment. The PATHS model: methods and case studies.
By Townsend J, Buxton M, Harper G.

No. 21
Systematic review of the clinical effectiveness and cost-effectiveness of tension-free vaginal tape for treatment of urinary stress incontinence.
By Cody J, Wyness L, Wallace S, Glazener C, Kilonzo M, Stearns S, *et al*.

No. 22
The clinical and cost-effectiveness of patient education models for diabetes: a systematic review and economic evaluation.
By Loveman E, Cave C, Green C, Royle P, Dunn N, Waugh N.

No. 23
The role of modelling in prioritising and planning clinical trials.
By Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P.

No. 24
Cost–benefit evaluation of routine influenza immunisation in people 65–74 years of age.
By Allsup S, Gosney M, Haycox A, Regan M.

No. 25
The clinical and cost-effectiveness of pulsatile machine perfusion versus cold storage of kidneys for transplantation retrieved from heart-beating and non-heart-beating donors.
By Wight J, Chilcott J, Holmes M, Brewer N.

No. 26
Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment.
By Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT.

No. 27
Evaluating non-randomised intervention studies.
By Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*.

No. 28
A randomised controlled trial to assess the impact of a package comprising a patient-orientated, evidence-based self-help guidebook and patient-centred consultations on disease management and satisfaction in inflammatory bowel disease.
By Kennedy A, Nelson E, Reeves D, Richardson G, Roberts C, Robinson A, *et al*.

No. 29
The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review.
By Dinnes J, Loveman E, McIntyre L, Waugh N.

No. 30
The value of digital imaging in diabetic retinopathy.
By Sharp PF, Olson J, Strachan F, Hipwell J, Ludbrook A, O'Donnell M, *et al*.

No. 31
Lowering blood pressure to prevent myocardial infarction and stroke: a new preventive strategy.
By Law M, Wald N, Morris J.

No. 32
Clinical and cost-effectiveness of capecitabine and tegafur with uracil for the treatment of metastatic colorectal cancer: systematic review and economic evaluation.
By Ward S, Kaltenthaler E, Cowan J, Brewer N.

No. 33
Clinical and cost-effectiveness of new and emerging technologies for early localised prostate cancer: a systematic review.
By Hummel S, Paisley S, Morgan A, Currie E, Brewer N.

No. 34
Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system.
By Royle P, Waugh N.

No. 35
Systematic review and economic decision modelling for the prevention and treatment of influenza A and B.
By Turner D, Wailoo A, Nicholson K, Cooper N, Sutton A, Abrams K.

No. 36
A randomised controlled trial to evaluate the clinical and cost-effectiveness of Hickman line insertions in adult cancer patients by nurses.
By Boland A, Haycox A, Bagust A, Fitzsimmons L.

No. 37
Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs.
By MacArthur C, Winter HR, Bick DE, Lilford RJ, Lancashire RJ, Knowles H, *et al*.

No. 38
Estimating implied rates of discount in healthcare decision-making.
By West RR, McNabb R, Thompson AGH, Sheldon TA, Grimley Evans J.

No. 39
Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus*: a review of the literature with epidemiological and economic modelling.
By Cooper BS, Stone SP, Kibbler CC, Cookson BD, Roberts JA, Medley GF, *et al*.

No. 40
Treatments for spasticity and pain in multiple sclerosis: a systematic review.
By Beard S, Hunn A, Wight J.

No. 41
The inclusion of reports of randomised trials published in languages other than English in systematic reviews.
By Moher D, Pham B, Lawson ML, Klassen TP.

No. 42
The impact of screening on future health-promoting behaviours and health beliefs: a systematic review.
By Bankhead CR, Brett J, Bukach C, Webster P, Stewart-Brown S, Munafo M, *et al*.

Volume 8, 2004

No. 1
What is the best imaging strategy for acute stroke?
By Wardlaw JM, Keir SL, Seymour J, Lewis S, Sandercock PAG, Dennis MS, *et al*.

No. 2
Systematic review and modelling of the investigation of acute and chronic chest pain presenting in primary care.
By Mant J, McManus RJ, Oakes RAL, Delaney BC, Barton PM, Deeks JJ, *et al*.

No. 3
The effectiveness and cost-effectiveness of microwave and thermal balloon endometrial ablation for heavy menstrual bleeding: a systematic review and economic modelling.
By Garside R, Stein K, Wyatt K, Round A, Price A.

No. 4
A systematic review of the role of bisphosphonates in metastatic disease.
By Ross JR, Saunders Y, Edmonds PM, Patel S, Wonderling D, Normand C, *et al*.

No. 5
Systematic review of the clinical effectiveness and cost-effectiveness of capecitabine (Xeloda®) for locally advanced and/or metastatic breast cancer.
By Jones L, Hawkins N, Westwood M, Wright K, Richardson G, Riemsma R.

No. 6
Effectiveness and efficiency of guideline dissemination and implementation strategies.
By Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L, *et al*.

No. 7
Clinical effectiveness and costs of the Sugarbaker procedure for the treatment of pseudomyxoma peritonei.
By Bryant J, Clegg AJ, Sidhu MK, Brodin H, Royle P, Davidson P.

No. 8
Psychological treatment for insomnia in the regulation of long-term hypnotic drug use.
By Morgan K, Dixon S, Mathers N, Thompson J, Tomeny M.

No. 9
Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome.
By Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ.

No. 10
A systematic review and economic evaluation of magnetic resonance cholangiopancreatography compared with diagnostic endoscopic retrograde cholangiopancreatography.
By Kaltenthaler E, Bravo Vergel Y, Chilcott J, Thomas S, Blakeborough T, Walters SJ, *et al*.

No. 11
The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis.
By Barton P, Jobanputra P, Wilson J, Bryan S, Burls A.

No. 12
Clinical effectiveness and cost-effectiveness of neonatal screening for inborn errors of metabolism using tandem mass spectrometry: a systematic review.
By Pandor A, Eastham J, Beverley C, Chilcott J, Paisley S.

No. 13
Clinical effectiveness and cost-effectiveness of pioglitazone and rosiglitazone in the treatment of type 2 diabetes: a systematic review and economic evaluation.
By Czoski-Murray C, Warren E, Chilcott J, Beverley C, Psyllaki MA, Cowan J.

No. 14
Routine examination of the newborn: the EMREN study. Evaluation of an extension of the midwife role including a randomised controlled trial of appropriately trained midwives and paediatric senior house officers.
By Townsend J, Wolke D, Hayes J, Davé S, Rogers C, Bloomfield L, *et al*.

No. 15
Involving consumers in research and development agenda setting for the NHS: developing an evidence-based approach.
By Oliver S, Clarke-Jones L, Rees R, Milne R, Buchanan P, Gabbay J, *et al*.

No. 16
A multi-centre randomised controlled trial of minimally invasive direct coronary bypass grafting versus percutaneous transluminal coronary angioplasty with stenting for proximal stenosis of the left anterior descending coronary artery.
By Reeves BC, Angelini GD, Bryan AJ, Taylor FC, Cripps T, Spyt TJ, *et al*.

No. 17
Does early magnetic resonance imaging influence management or improve outcome in patients referred to secondary care with low back pain? A pragmatic randomised controlled trial.
By Gilbert FJ, Grant AM, Gillan MGC, Vale L, Scott NW, Campbell MK, *et al*.

No. 18
The clinical and cost-effectiveness of anakinra for the treatment of rheumatoid arthritis in adults: a systematic review and economic analysis.
By Clark W, Jobanputra P, Barton P, Burls A.

No. 19
A rapid and systematic review and economic evaluation of the clinical and cost-effectiveness of newer drugs for treatment of mania associated with bipolar affective disorder.
By Bridle C, Palmer S, Bagnall A-M, Darba J, Duffy S, Sculpher M, *et al*.

No. 20
Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic analysis.
By Karnon J, Peters J, Platt J, Chilcott J, McGoogan E, Brewer N.

No. 21
Systematic review of the long-term effects and economic consequences of treatments for obesity and implications for health improvement.
By Avenell A, Broom J, Brown TJ, Poobalan A, Aucott L, Stearns SC, *et al*.

No. 22
Autoantibody testing in children with newly diagnosed type 1 diabetes mellitus.
By Dretzke J, Cummins C, Sandercock J, Fry-Smith A, Barrett T, Burls A.

No. 23
Clinical effectiveness and cost-effectiveness of prehospital intravenous fluids in trauma patients.
By Dretzke J, Sandercock J, Bayliss S, Burls A.

No. 24
Newer hypnotic drugs for the short-term management of insomnia: a systematic review and economic evaluation.
By Dündar Y, Boland A, Strobl J, Dodd S, Haycox A, Bagust A, *et al*.

No. 25
Development and validation of methods for assessing the quality of diagnostic accuracy studies.
By Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J.

No. 26
EVALUATE hysterectomy trial: a multicentre randomised trial comparing abdominal, vaginal and laparoscopic methods of hysterectomy.
By Garry R, Fountain J, Brown J, Manca A, Mason S, Sculpher M, *et al*.

No. 27
Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-β and glatiramer acetate for multiple sclerosis.
By Tappenden P, Chilcott JB, Eggington S, Oakley J, McCabe C.

No. 28
Effectiveness and cost-effectiveness of imatinib for first-line treatment of chronic myeloid leukaemia in chronic phase: a systematic review and economic analysis.
By Dalziel K, Round A, Stein K, Garside R, Price A.

No. 29
VenUS I: a randomised controlled trial of two types of bandage for treating venous leg ulcers.
By Iglesias C, Nelson EA, Cullum NA, Torgerson DJ, on behalf of the VenUS Team.

No. 30
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of myocardial perfusion scintigraphy for the diagnosis and management of angina and myocardial infarction.
By Mowatt G, Vale L, Brazzelli M, Hernandez R, Murray A, Scott N, *et al*.

No. 31
A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme.
By Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S.

No. 32
The Social Support and Family Health Study: a randomised controlled trial and economic evaluation of two alternative forms of postnatal support for mothers living in disadvantaged inner-city areas.
By Wiggins M, Oakley A, Roberts I, Turner H, Rajan L, Austerberry H, *et al*.

No. 33
Psychosocial aspects of genetic screening of pregnant women and newborns: a systematic review.
By Green JM, Hewison J, Bekker HL, Bryant, Cuckle HS.

No. 34
Evaluation of abnormal uterine bleeding: comparison of three outpatient procedures within cohorts defined by age and menopausal status.
By Critchley HOD, Warner P, Lee AJ, Brechin S, Guise J, Graham B.

No. 35
Coronary artery stents: a rapid systematic review and economic evaluation.
By Hill R, Bagust A, Bakhai A, Dickson R, Dündar Y, Haycox A, *et al*.

No. 36
Review of guidelines for good practice in decision-analytic modelling in health technology assessment.
By Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, *et al*.

No. 37
Rituximab (MabThera®) for aggressive non-Hodgkin's lymphoma: systematic review and economic evaluation.
By Knight C, Hind D, Brewer N, Abbott V.

No. 38
Clinical effectiveness and cost-effectiveness of clopidogrel and modified-release dipyridamole in the secondary prevention of occlusive vascular events: a systematic review and economic evaluation.
By Jones L, Griffin S, Palmer S, Main C, Orton V, Sculpher M, *et al.*

No. 39
Pegylated interferon α-2a and -2b in combination with ribavirin in the treatment of chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Brodin H, Cave C, Waugh N, Price A, Gabbay J.

No. 40
Clopidogrel used in combination with aspirin compared with aspirin alone in the treatment of non-ST-segment-elevation acute coronary syndromes: a systematic review and economic evaluation.
By Main C, Palmer S, Griffin S, Jones L, Orton V, Sculpher M, *et al.*

No. 41
Provision, uptake and cost of cardiac rehabilitation programmes: improving services to under-represented groups.
By Beswick AD, Rees K, Griebsch I, Taylor FC, Burke M, West RR, *et al.*

No. 42
Involving South Asian patients in clinical trials.
By Hussain-Gambles M, Leese B, Atkin K, Brown J, Mason S, Tovey P.

No. 43
Clinical and cost-effectiveness of continuous subcutaneous insulin infusion for diabetes.
By Colquitt JL, Green C, Sidhu MK, Hartwell D, Waugh N.

No. 44
Identification and assessment of ongoing trials in health technology assessment reviews.
By Song FJ, Fry-Smith A, Davenport C, Bayliss S, Adi Y, Wilson JS, *et al*.

No. 45
Systematic review and economic evaluation of a long-acting insulin analogue, insulin glargine
By Warren E, Weatherley-Jones E, Chilcott J, Beverley C.

No. 46
Supplementation of a home-based exercise programme with a class-based programme for people with osteoarthritis of the knees: a randomised controlled trial and health economic analysis.
By McCarthy CJ, Mills PM, Pullen R, Richardson G, Hawkins N, Roberts CR, *et al*.

No. 47
Clinical and cost-effectiveness of once-daily versus more frequent use of same potency topical corticosteroids for atopic eczema: a systematic review and economic evaluation.
By Green C, Colquitt JL, Kirby J, Davidson P, Payne E.

No. 48
Acupuncture of chronic headache disorders in primary care: randomised controlled trial and economic analysis.
By Vickers AJ, Rees RW, Zollman CE, McCarney R, Smith CM, Ellis N, *et al*.

No. 49
Generalisability in economic evaluation studies in healthcare: a review and case studies.
By Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, *et al*.

No. 50
Virtual outreach: a randomised controlled trial and economic evaluation of joint teleconferenced medical consultations.
By Wallace P, Barber J, Clayton W, Currell R, Fleming K, Garner P, *et al*.

Volume 9, 2005

No. 1
Randomised controlled multiple treatment comparison to provide a cost-effectiveness rationale for the selection of antimicrobial therapy in acne.
By Ozolins M, Eady EA, Avery A, Cunliffe WJ, O'Neill C, Simpson NB, *et al*.

No. 2
Do the findings of case series studies vary significantly according to methodological characteristics?
By Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L.

No. 3
Improving the referral process for familial breast cancer genetic counselling: findings of three randomised controlled trials of two interventions.
By Wilson BJ, Torrance N, Mollison J, Wordsworth S, Gray JR, Haites NE, *et al*.

No. 4
Randomised evaluation of alternative electrosurgical modalities to treat bladder outflow obstruction in men with benign prostatic hyperplasia.
By Fowler C, McAllister W, Plail R, Karim O, Yang Q.

No. 5
A pragmatic randomised controlled trial of the cost-effectiveness of palliative therapies for patients with inoperable oesophageal cancer.
By Shenfine J, McNamee P, Steen N, Bond J, Griffin SM.

No. 6
Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography.
By Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H.

No. 7
Issues in data monitoring and interim analysis of trials.
By Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, *et al*.

No. 8
Lay public's understanding of equipoise and randomisation in randomised controlled trials.
By Robinson EJ, Kerr CEP, Stevens AJ, Lilford RJ, Braunholtz DA, Edwards SJ, *et al*.

No. 9
Clinical and cost-effectiveness of electroconvulsive therapy for depressive illness, schizophrenia, catatonia and mania: systematic reviews and economic modelling studies.
By Greenhalgh J, Knight C, Hind D, Beverley C, Walters S.

No. 10
Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology.
By Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, *et al*.

No. 11
Clinical effectiveness and cost-effectiveness of drotrecogin alfa (activated) (Xigris®) for the treatment of severe sepsis in adults: a systematic review and economic evaluation.
By Green C, Dinnes J, Takeda A, Shepherd J, Hartwell D, Cave C, *et al*.

No. 12
A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy.
By Dinnes J, Deeks J, Kirby J, Roderick P.

No. 13
Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK.
By Willis BH, Barton P, Pearmain P, Bryan S, Hyde C.

No. 14
Laparoscopic surgery for inguinal hernia repair: systematic review of effectiveness and economic evaluation.
By McCormack K, Wake B, Perez J, Fraser C, Cook J, McIntosh E, *et al*.

No. 15
Clinical effectiveness, tolerability and cost-effectiveness of newer drugs for epilepsy in adults: a systematic review and economic evaluation.
By Wilby J, Kainth A, Hawkins N, Epstein D, McIntosh H, McDaid C, *et al*.

No. 16
A randomised controlled trial to compare the cost-effectiveness of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine.
By Peveler R, Kendrick T, Buxton M, Longworth L, Baldwin D, Moore M, *et al*.

No. 17
Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation.
By Hartwell D, Colquitt J, Loveman E, Clegg AJ, Brodin H, Waugh N, *et al*.

No. 18
A randomised controlled comparison of alternative strategies in stroke care.
By Kalra L, Evans A, Perez I, Knapp M, Swift C, Donaldson N.

No. 19
The investigation and analysis of critical incidents and adverse events in healthcare.
By Woloshynowych M, Rogers S, Taylor-Adams S, Vincent C.

No. 20
Potential use of routine databases in health technology assessment.
By Raftery J, Roderick P, Stevens A.

No. 21
Clinical and cost-effectiveness of newer immunosuppressive regimens in renal transplantation: a systematic review and modelling study.
By Woodroffe R, Yao GL, Meads C, Bayliss S, Ready A, Raftery J, *et al*.

No. 22
A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis.
By Stevenson M, Lloyd Jones M, De Nigris E, Brewer N, Davis S, Oakley J.

No. 23
A systematic review to examine the impact of psycho-educational interventions on health outcomes and costs in adults and children with difficult asthma.
By Smith JR, Mugford M, Holland R, Candy B, Noble MJ, Harrison BDW, *et al*.

No. 24
An evaluation of the costs, effectiveness and quality of renal replacement therapy provision in renal satellite units in England and Wales.
By Roderick P, Nicholson T, Armitage A, Mehta R, Mullee M, Gerard K, *et al*.

No. 25
Imatinib for the treatment of patients with unresectable and/or metastatic gastrointestinal stromal tumours: systematic review and economic evaluation.
By Wilson J, Connock M, Song F, Yao G, Fry-Smith A, Raftery J, *et al*.

No. 26
Indirect comparisons of competing interventions.
By Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, *et al*.

No. 27
Cost-effectiveness of alternative strategies for the initial medical management of non-ST elevation acute coronary syndrome: systematic review and decision-analytical modelling.
By Robinson M, Palmer S, Sculpher M, Philips Z, Ginnelly L, Bowens A, *et al*.

No. 28
Outcomes of electrically stimulated gracilis neosphincter surgery.
By Tillin T, Chambers M, Feldman R.

No. 29
The effectiveness and cost-effectiveness of pimecrolimus and tacrolimus for atopic eczema: a systematic review and economic evaluation.
By Garside R, Stein K, Castelnuovo E, Pitt M, Ashcroft D, Dimmock P, *et al*.

No. 30
Systematic review on urine albumin testing for early detection of diabetic complications.
By Newman DJ, Mattock MB, Dawnay ABS, Kerry S, McGuire A, Yaqoob M, *et al*.

No. 31
Randomised controlled trial of the cost-effectiveness of water-based therapy for lower limb osteoarthritis.
By Cochrane T, Davey RC, Matthes Edwards SM.

No. 32
Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain.
By Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, *et al*.

No. 33
Cost-effectiveness and safety of epidural steroids in the management of sciatica.
By Price C, Arden N, Coglan L, Rogers P.

No. 34
The British Rheumatoid Outcome Study Group (BROSG) randomised controlled trial to compare the effectiveness and cost-effectiveness of aggressive versus symptomatic therapy in established rheumatoid arthritis.
By Symmons D, Tricker K, Roberts C, Davies L, Dawes P, Scott DL.

No. 35
Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials.
By King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, *et al*.

No. 36
The clinical and cost-effectiveness of implantable cardioverter defibrillators: a systematic review.
By Bryant J, Brodin H, Loveman E, Payne E, Clegg A.

No. 37
A trial of problem-solving by community mental health nurses for anxiety, depression and life difficulties among general practice patients. The CPN-GP study.
By Kendrick T, Simons L, Mynors-Wallis L, Gray A, Lathlean J, Pickering R, *et al*.

No. 38
The causes and effects of socio-demographic exclusions from clinical trials.
By Bartlett C, Doyal L, Ebrahim S, Davey P, Bachmann M, Egger M, *et al*.

No. 39
Is hydrotherapy cost-effective? A randomised controlled trial of combined hydrotherapy programmes compared with physiotherapy land techniques in children with juvenile idiopathic arthritis.
By Epps H, Ginnelly L, Utley M, Southwood T, Gallivan S, Sculpher M, *et al*.

No. 40
A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study.
By Hobbs FDR, Fitzmaurice DA, Mant J, Murray E, Jowett S, Bryan S, *et al*.

No. 41
Displaced intracapsular hip fractures in fit, older people: a randomised comparison of reduction and fixation, bipolar hemiarthroplasty and total hip arthroplasty.
By Keating JF, Grant A, Masson M, Scott NW, Forbes JF.

No. 42
Long-term outcome of cognitive behaviour therapy clinical trials in central Scotland.
By Durham RC, Chambers JA, Power KG, Sharp DM, Macdonald RR, Major KA, *et al*.

No. 43
The effectiveness and cost-effectiveness of dual-chamber pacemakers compared with single-chamber pacemakers for bradycardia due to atrioventricular block or sick sinus syndrome: systematic review and economic evaluation.
By Castelnuovo E, Stein K, Pitt M, Garside R, Payne E.

No. 44
Newborn screening for congenital heart defects: a systematic review and cost-effectiveness analysis.
By Knowles R, Griebsch I, Dezateux C, Brown J, Bull C, Wren C.

No. 45
The clinical and cost-effectiveness of left ventricular assist devices for end-stage heart failure: a systematic review and economic evaluation.
By Clegg AJ, Scott DA, Loveman E, Colquitt J, Hutchinson J, Royle P, *et al*.

No. 46
The effectiveness of the Heidelberg Retina Tomograph and laser diagnostic glaucoma scanning system (GDx) in detecting and monitoring glaucoma.
By Kwartz AJ, Henson DB, Harper RA, Spencer AF, McLeod D.

No. 47
Clinical and cost-effectiveness of autologous chondrocyte implantation for cartilage defects in knee joints: systematic review and economic evaluation.
By Clar C, Cummins E, McIntyre L, Thomas S, Lamb J, Bain L, *et al*.

No. 48
Systematic review of effectiveness of different treatments for childhood retinoblastoma.
By McDaid C, Hartley S, Bagnall A-M, Ritchie G, Light K, Riemsma R.

No. 49
Towards evidence-based guidelines for the prevention of venous thromboembolism: systematic reviews of mechanical methods, oral anticoagulation, dextran and regional anaesthesia as thromboprophylaxis.
By Roderick P, Ferris G, Wilson K, Halls H, Jackson D, Collins R, *et al*.

No. 50
The effectiveness and cost-effectiveness of parent training/education programmes for the treatment of conduct disorder, including oppositional defiant disorder, in children.
By Dretzke J, Frew E, Davenport C, Barlow J, Stewart-Brown S, Sandercock J, *et al*.

Volume 10, 2006

No. 1
The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease.
By Loveman E, Green C, Kirby J, Takeda A, Picot J, Payne E, *et al*.

No. 2
FOOD: a multicentre randomised trial evaluating feeding policies in patients admitted to hospital with a recent stroke.
By Dennis M, Lewis S, Cranswick G, Forbes J.

No. 3
The clinical effectiveness and cost-effectiveness of computed tomography screening for lung cancer: systematic reviews.
By Black C, Bagust A, Boland A, Walker S, McLeod C, De Verteuil R, *et al*.

No. 4
A systematic review of the effectiveness and cost-effectiveness of neuroimaging assessments used to visualise the seizure focus in people with refractory epilepsy being considered for surgery.
By Whiting P, Gupta R, Burch J, Mujica Mota RE, Wright K, Marson A, *et al*.

No. 5
Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies.
By Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR.

No. 6
Systematic review and evaluation of methods of assessing urinary incontinence.
By Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of newer drugs for children with epilepsy. A systematic review.
By Connock M, Frew E, Evans B-W, Bryan S, Cummins C, Fry-Smith A, *et al*.

No. 8
Surveillance of Barrett's oesophagus: exploring the uncertainty through systematic review, expert workshop and economic modelling.
By Garside R, Pitt M, Somerville M, Stein K, Price A, Gilbert N.

No. 9
Topotecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for second-line or subsequent treatment of advanced ovarian cancer: a systematic review and economic evaluation.
By Main C, Bojke L, Griffin S, Norman G, Barbieri M, Mather L, *et al*.

No. 10
Evaluation of molecular techniques in prediction and diagnosis of cytomegalovirus disease in immunocompromised patients.
By Szczepura A, Westmoreland D, Vinogradova Y, Fox J, Clark M.

No. 11
Screening for thrombophilia in high-risk situations: systematic review and cost-effectiveness analysis. The Thrombosis: Risk and Economic Assessment of Thrombophilia Screening (TREATS) study.
By Wu O, Robertson L, Twaddle S, Lowe GDO, Clark P, Greaves M, *et al*.

No. 12
A series of systematic reviews to inform a decision analysis for sampling and treating infected diabetic foot ulcers.
By Nelson EA, O'Meara S, Craig D, Iglesias C, Golder S, Dalton J, *et al*.

No. 13
Randomised clinical trial, observational study and assessment of cost-effectiveness of the treatment of varicose veins (REACTIV trial).
By Michaels JA, Campbell WB, Brazier JE, MacIntyre JB, Palfreyman SJ, Ratcliffe J, *et al*.

No. 14
The cost-effectiveness of screening for oral cancer in primary care.
By Speight PM, Palmer S, Moles DR, Downer MC, Smith DH, Henriksson M, *et al*.

No. 15
Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis.
By Goodacre S, Sampson F, Stevenson M, Wailoo A, Sutton A, Thomas S, *et al*.

No. 16
Systematic review of the effectiveness and cost-effectiveness of HealOzone® for the treatment of occlusal pit/fissure caries and root caries.
By Brazzelli M, McKenzie L, Fielding S, Fraser C, Clarkson J, Kilonzo M, *et al*.

No. 17
Randomised controlled trials of conventional antipsychotic versus new atypical drugs, and new atypical drugs versus clozapine, in people with schizophrenia responding poorly to, or intolerant of, current drug treatment.
By Lewis SW, Davies L, Jones PB, Barnes TRE, Murray RM, Kerwin R, *et al*.

No. 18
Diagnostic tests and algorithms used in the investigation of haematuria: systematic reviews and economic evaluation.
By Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D, *et al*.

No. 19
Cognitive behavioural therapy in addition to antispasmodic therapy for irritable bowel syndrome in primary care: randomised controlled trial.
By Kennedy TM, Chalder T, McCrone P, Darnley S, Knapp M, Jones RH, *et al*.

No. 20
A systematic review of the clinical effectiveness and cost-effectiveness of enzyme replacement therapies for Fabry's disease and mucopolysaccharidosis type 1.
By Connock M, Juarez-Garcia A, Frew E, Mans A, Dretzke J, Fry-Smith A, *et al*.

No. 21
Health benefits of antiviral therapy for mild chronic hepatitis C: randomised controlled trial and economic evaluation.
By Wright M, Grieve R, Roberts J, Main J, Thomas HC, on behalf of the UK Mild Hepatitis C Trial Investigators.

No. 22
Pressure relieving support surfaces: a randomised evaluation.
By Nixon J, Nelson EA, Cranny G, Iglesias CP, Hawkins K, Cullum NA, *et al*.

No. 23
A systematic review and economic model of the effectiveness and cost-effectiveness of methylphenidate, dexamfetamine and atomoxetine for the treatment of attention deficit hyperactivity disorder in children and adolescents.
By King S, Griffin S, Hodges Z, Weatherly H, Asseburg C, Richardson G, *et al.*

No. 24
The clinical effectiveness and cost-effectiveness of enzyme replacement therapy for Gaucher's disease: a systematic review.
By Connock M, Burls A, Frew E, Fry-Smith A, Juarez-Garcia A, McCabe C, *et al*.

No. 25
Effectiveness and cost-effectiveness of salicylic acid and cryotherapy for cutaneous warts. An economic decision model.
By Thomas KS, Keogh-Brown MR, Chalmers JR, Fordham RJ, Holland RC, Armstrong SJ, *et al*.

No. 26
A systematic literature review of the effectiveness of non-pharmacological interventions to prevent wandering in dementia and evaluation of the ethical implications and acceptability of their use.
By Robinson L, Hutchings D, Corner L, Beyer F, Dickinson H, Vanoli A, *et al*.

No. 27
A review of the evidence on the effects and costs of implantable cardioverter defibrillator therapy in different patient groups, and modelling of cost-effectiveness and cost–utility for these groups in a UK context.
By Buxton M, Caine N, Chase D, Connelly D, Grace A, Jackson C, *et al*.

No. 28
Adefovir dipivoxil and pegylated interferon alfa-2a for the treatment of chronic hepatitis B: a systematic review and economic evaluation.
By Shepherd J, Jones J, Takeda A, Davidson P, Price A.

No. 29
An evaluation of the clinical and cost-effectiveness of pulmonary artery catheters in patient management in intensive care: a systematic review and a randomised controlled trial.
By Harvey S, Stevens K, Harrison D, Young D, Brampton W, McCabe C, *et al*.

No. 30
Accurate, practical and cost-effective assessment of carotid stenosis in the UK.
By Wardlaw JM, Chappell FM, Stevenson M, De Nigris E, Thomas S, Gillard J, *et al*.

No. 31
Etanercept and infliximab for the treatment of psoriatic arthritis: a systematic review and economic evaluation.
By Woolacott N, Bravo Vergel Y, Hawkins N, Kainth A, Khadjesari Z, Misso K, *et al*.

No. 32
The cost-effectiveness of testing for hepatitis C in former injecting drug users.
By Castelnuovo E, Thompson-Coon J, Pitt M, Cramp M, Siebert U, Price A, *et al*.

No. 33
Computerised cognitive behaviour therapy for depression and anxiety update: a systematic review and economic evaluation.
By Kaltenthaler E, Brazier J, De Nigris E, Tumur I, Ferriter M, Beverley C, *et al*.

No. 34
Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy.
By Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, *et al*.

No. 35
Psychological therapies including dialectical behaviour therapy for borderline personality disorder: a systematic review and preliminary economic evaluation.
By Brazier J, Tumur I, Holmes M, Ferriter M, Parry G, Dent-Brown K, *et al*.

No. 36
Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model.
By Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J, *et al*.

No. 37
Cognitive behavioural therapy in chronic fatigue syndrome: a randomised controlled trial of an outpatient group programme.
By O'Dowd H, Gladwell P, Rogers CA, Hollinghurst S, Gregory A.

No. 38
A comparison of the cost-effectiveness of five strategies for the prevention of nonsteroidal anti-inflammatory drug-induced gastrointestinal toxicity: a systematic review with economic modelling.
By Brown TJ, Hooper L, Elliott RA, Payne K, Webb R, Roberts C, et al.

No. 39
The effectiveness and cost-effectiveness of computed tomography screening for coronary artery disease: systematic review.
By Waugh N, Black C, Walker S, McIntyre L, Cummins E, Hillis G.

No. 40
What are the clinical outcome and cost-effectiveness of endoscopy undertaken by nurses when compared with doctors? A Multi-Institution Nurse Endoscopy Trial (MINuET).
By Williams J, Russell I, Durai D, Cheung W-Y, Farrin A, Bloor K, *et al*.

No. 41
The clinical and cost-effectiveness of oxaliplatin and capecitabine for the adjuvant treatment of colon cancer: systematic review and economic evaluation.
By Pandor A, Eggington S, Paisley S, Tappenden P, Sutcliffe P.

No. 42
A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness.
By Chen Y-F, Jobanputra P, Barton P, Jowett S, Bryan S, Clark W, *et al*.

No. 43
Telemedicine in dermatology: a randomised controlled trial.
By Bowns IR, Collins K, Walters SJ, McDonagh AJG.

No. 44
Cost-effectiveness of cell salvage and alternative methods of minimising perioperative allogeneic blood transfusion: a systematic review and economic model.
By Davies L, Brown TJ, Haynes S, Payne K, Elliott RA, McCollum C.

No. 45
Clinical effectiveness and cost-effectiveness of laparoscopic surgery for colorectal cancer: systematic reviews and economic evaluation.
By Murray A, Lourenco T, de Verteuil R, Hernandez R, Fraser C, McKinley A, *et al*.

No. 46
Etanercept and efalizumab for the treatment of psoriasis: a systematic review.
By Woolacott N, Hawkins N, Mason A, Kainth A, Khadjesari Z, Bravo Vergel Y, *et al*.

No. 47
Systematic reviews of clinical decision tools for acute abdominal pain.
By Liu JLY, Wyatt JC, Deeks JJ, Clamp S, Keen J, Verde P, *et al*.

No. 48
Evaluation of the ventricular assist device programme in the UK.
By Sharples L, Buxton M, Caine N, Cafferty F, Demiris N, Dyer M, *et al*.

No. 49
A systematic review and economic model of the clinical and cost-effectiveness of immunosuppressive therapy for renal transplantation in children.
By Yao G, Albon E, Adi Y, Milford D, Bayliss S, Ready A, *et al*.

No. 50
Amniocentesis results: investigation of anxiety. The ARIA trial.
By Hewison J, Nixon J, Fountain J, Cocks K, Jones C, Mason G, *et al*.

Volume 11, 2007

No. 1
Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation.
By Dundar Y, Bagust A, Dickson R, Dodd S, Green J, Haycox A, *et al*.

No. 2
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer.
By Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K, *et al*.

No. 3
A systematic review of rapid diagnostic tests for the detection of tuberculosis infection.
By Dinnes J, Deeks J, Kunst H, Gibson A, Cummins E, Waugh N, *et al*.

No. 4
The clinical effectiveness and cost-effectiveness of strontium ranelate for the prevention of osteoporotic fragility fractures in postmenopausal women.
By Stevenson M, Davis S, Lloyd-Jones M, Beverley C.

No. 5
A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines.
By Raynor DK, Blenkinsopp A, Knapp P, Grime J, Nicolson DJ, Pollock K, *et al*.

No. 6
Oral naltrexone as a treatment for relapse prevention in formerly opioid-dependent drug users: a systematic review and economic evaluation.
By Adi Y, Juarez-Garcia A, Wang D, Jowett S, Frew E, Day E, *et al*.

No. 7
Glucocorticoid-induced osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Stevenson M, McCloskey EV, Davis S, Lloyd-Jones M.

No. 8
Epidemiological, social, diagnostic and economic evaluation of population screening for genital chlamydial infection.
By Low N, McCarthy A, Macleod J, Salisbury C, Campbell R, Roberts TE, *et al*.

No. 9
Methadone and buprenorphine for the management of opioid dependence: a systematic review and economic evaluation.
By Connock M, Juarez-Garcia A, Jowett S, Frew E, Liu Z, Taylor RJ, *et al*.

No. 10
Exercise Evaluation Randomised Trial (EXERT): a randomised trial comparing GP referral for leisure centre-based exercise, community-based walking and advice only.
By Isaacs AJ, Critchley JA, See Tai S, Buckingham K, Westley D, Harridge SDR, *et al*.

No. 11
Interferon alfa (pegylated and non-pegylated) and ribavirin for the treatment of mild chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Jones J, Hartwell D, Davidson P, Price A, Waugh N.

No. 12
Systematic review and economic evaluation of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer.
By Tappenden P, Jones R, Paisley S, Carroll C.

No. 13
A systematic review and economic evaluation of epoetin alfa, epoetin beta and darbepoetin alfa in anaemia associated with cancer, especially that attributable to cancer treatment.
By Wilson J, Yao GL, Raftery J, Bohlius J, Brunskill S, Sandercock J, *et al*.

No. 14
A systematic review and economic evaluation of statins for the prevention of coronary events.
By Ward S, Lloyd Jones M, Pandor A, Holmes M, Ara R, Ryan A, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of different models of community-based respite care for frail older people and their carers.
By Mason A, Weatherly H, Spilsbury K, Arksey H, Golder S, Adamson J, *et al*.

No. 16
Additional therapy for young children with spastic cerebral palsy: a randomised controlled trial.
By Weindling AM, Cunningham CC, Glenn SM, Edwards RT, Reeves DJ.

No. 17
Screening for type 2 diabetes: literature review and economic modelling.
By Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, *et al*.

No. 18
The effectiveness and cost-effectiveness of cinacalcet for secondary hyperparathyroidism in end-stage renal disease patients on dialysis: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Mealing S, Roome C, Snaith A, *et al*.

No. 19
The clinical effectiveness and cost-effectiveness of gemcitabine for metastatic breast cancer: a systematic review and economic evaluation.
By Takeda AL, Jones J, Loveman E, Tan SC, Clegg AJ.

No. 20
A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease.
By Collins R, Cranny G, Burch J, Aguiar-Ibáñez R, Craig D, Wright K, *et al*.

No. 21
The clinical effectiveness and cost-effectiveness of treatments for children with idiopathic steroid-resistant nephrotic syndrome: a systematic review.
By Colquitt JL, Kirby J, Green C, Cooper K, Trompeter RS.

No. 22
A systematic review of the routine monitoring of growth in children of primary school age to identify growth-related conditions.
By Fayter D, Nixon J, Hartley S, Rithalia A, Butler G, Rudolf M, *et al*.

No. 23
Systematic review of the effectiveness of preventing and treating *Staphylococcus aureus* carriage in reducing peritoneal catheter-related infections.
By McCormack K, Rabindranath K, Kilonzo M, Vale L, Fraser C, McIntyre L, *et al*.

No. 24
The clinical effectiveness and cost of repetitive transcranial magnetic stimulation versus electroconvulsive therapy in severe depression: a multicentre pragmatic randomised controlled trial and economic analysis.
By McLoughlin DM, Mogg A, Eranti S, Pluck G, Purvis R, Edwards D, *et al*.

No. 25
A randomised controlled trial and economic evaluation of direct versus indirect and individual versus group modes of speech and language therapy for children with primary language impairment.
By Boyle J, McCartney E, Forbes J, O'Hare A.

No. 26
Hormonal therapies for early breast cancer: systematic review and economic evaluation.
By Hind D, Ward S, De Nigris E, Simpson E, Carroll C, Wyld L.

No. 27
Cardioprotection against the toxic effects of anthracyclines given to children with cancer: a systematic review.
By Bryant J, Picot J, Levitt G, Sullivan I, Baxter L, Clegg A.

No. 28
Adalimumab, etanercept and infliximab for the treatment of ankylosing spondylitis: a systematic review and economic evaluation.
By McLeod C, Bagust A, Boland A, Dagenais P, Dickson R, Dundar Y, *et al*.

No. 29
Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses.
By Colbourn T, Asseburg C, Bojke L, Philips Z, Claxton K, Ades AE, *et al*.

No. 30
Clinical effectiveness and cost-effectiveness of bone morphogenetic proteins in the non-healing of fractures and spinal fusion: a systematic review.
By Garrison KR, Donell S, Ryder J, Shemilt I, Mugford M, Harvey I, *et al*.

No. 31
A randomised controlled trial of postoperative radiotherapy following breast-conserving surgery in a minimum-risk older population. The PRIME trial.
By Prescott RJ, Kunkler IH, Williams LJ, King CC, Jack W, van der Pol M, *et al*.

No. 32
Current practice, accuracy, effectiveness and cost-effectiveness of the school entry hearing screen.
By Bamford J, Fortnum H, Bristow K, Smith J, Vamvakas G, Davies L, *et al*.

No. 33
The clinical effectiveness and cost-effectiveness of inhaled insulin in diabetes mellitus: a systematic review and economic evaluation.
By Black C, Cummins E, Royle P, Philip S, Waugh N.

No. 34
Surveillance of cirrhosis for hepatocellular carcinoma: systematic review and economic analysis.
By Thompson Coon J, Rogers G, Hewson P, Wright D, Anderson R, Cramp M, *et al*.

No. 35
The Birmingham Rehabilitation Uptake Maximisation Study (BRUM). Homebased compared with hospital-based cardiac rehabilitation in a multi-ethnic population: cost-effectiveness and patient adherence.
By Jolly K, Taylor R, Lip GYH, Greenfield S, Raftery J, Mant J, *et al*.

No. 36
A systematic review of the clinical, public health and cost-effectiveness of rapid diagnostic tests for the detection and identification of bacterial intestinal pathogens in faeces and food.
By Abubakar I, Irvine L, Aldus CF, Wyatt GM, Fordham R, Schelenz S, *et al*.

No. 37
A randomised controlled trial examining the longer-term outcomes of standard versus new antiepileptic drugs. The SANAD trial.
By Marson AG, Appleton R, Baker GA, Chadwick DW, Doughty J, Eaton B, *et al*.

No. 38
Clinical effectiveness and cost-effectiveness of different models of managing long-term oral anti-coagulation therapy: a systematic review and economic modelling.
By Connock M, Stevens C, Fry-Smith A, Jowett S, Fitzmaurice D, Moore D, *et al*.

No. 39
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of interventions for preventing relapse in people with bipolar disorder.
By Soares-Weiser K, Bravo Vergel Y, Beynon S, Dunn G, Barbieri M, Duffy S, *et al*.

No. 40
Taxanes for the adjuvant treatment of early breast cancer: systematic review and economic evaluation.
By Ward S, Simpson E, Davis S, Hind D, Rees A, Wilkinson A.

No. 41
The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation.
By Burr JM, Mowatt G, Hernández R, Siddiqui MAR, Cook J, Lourenco T, *et al*.

No. 42
Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models.
By Davis A, Smith P, Ferguson M, Stephens D, Gianopoulos I.

No. 43
Contamination in trials of educational interventions.
By Keogh-Brown MR, Bachmann MO, Shepstone L, Hewitt C, Howe A, Ramsay CR, *et al*.

No. 44
Overview of the clinical effectiveness of positron emission tomography imaging in selected cancers.
By Facey K, Bradbury I, Laking G, Payne E.

No. 45
The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Rogers G, Dyer M, Mealing S, *et al*.

No. 46
Drug-eluting stents: a systematic review and economic evaluation.
By Hill RA, Boland A, Dickson R, Dündar Y, Haycox A, McLeod C, *et al*.

No. 47
The clinical effectiveness and cost-effectiveness of cardiac resynchronisation (biventricular pacing) for heart failure: systematic review and economic model.
By Fox M, Mealing S, Anderson R, Dean J, Stein K, Price A, *et al*.

No. 48
Recruitment to randomised trials: strategies for trial enrolment and participation study. The STEPS study.
By Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, *et al*.

No. 49
Cost-effectiveness of functional cardiac testing in the diagnosis and management of coronary artery disease: a randomised controlled trial. The CECaT trial.
By Sharples L, Hughes V, Crean A, Dyer M, Buxton M, Goldsmith K, *et al*.

No. 50
Evaluation of diagnostic tests when there is no gold standard. A review of methods.
By Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM.

No. 51
Systematic reviews of the clinical effectiveness and cost-effectiveness of proton pump inhibitors in acute upper gastrointestinal bleeding.
By Leontiadis GI, Sreedharan A, Dorward S, Barton P, Delaney B, Howden CW, *et al*.

No. 52
A review and critique of modelling in prioritising and designing screening programmes.
By Karnon J, Goyder E, Tappenden P, McPhie S, Towers I, Brazier J, *et al*.

No. 53
An assessment of the impact of the NHS Health Technology Assessment Programme.
By Hanney S, Buxton M, Green C, Coulson D, Raftery J.

Volume 12, 2008

No. 1
A systematic review and economic model of switching from nonglycopeptide to glycopeptide antibiotic prophylaxis for surgery.
By Cranny G, Elliott R, Weatherly H, Chambers D, Hawkins N, Myers L, *et al*.

No. 2
'Cut down to quit' with nicotine replacement therapies in smoking cessation: a systematic review of effectiveness and economic analysis.
By Wang D, Connock M, Barton P, Fry-Smith A, Aveyard P, Moore D.

No. 3
A systematic review of the effectiveness of strategies for reducing fracture risk in children with juvenile idiopathic arthritis with additional data on long-term risk of fracture and cost of disease management.
By Thornton J, Ashcroft D, O'Neill T, Elliott R, Adams J, Roberts C, *et al*.

No. 4
Does befriending by trained lay workers improve psychological well-being and quality of life for carers of people with dementia, and at what cost? A randomised controlled trial.
By Charlesworth G, Shepstone L, Wilson E, Thalanany M, Mugford M, Poland F.

No. 5
A multi-centre retrospective cohort study comparing the efficacy, safety and cost-effectiveness of hysterectomy and uterine artery embolisation for the treatment of symptomatic uterine fibroids. The HOPEFUL study.
By Hirst A, Dutton S, Wu O, Briggs A, Edwards C, Waldenmaier L, *et al*.

No. 6
Methods of prediction and prevention of pre-eclampsia: systematic reviews of accuracy and effectiveness literature with economic modelling.
By Meads CA, Cnossen JS, Meher S, Juarez-Garcia A, ter Riet G, Duley L, *et al*.

No. 7
The use of economic evaluations in NHS decision-making: a review and empirical investigation.
By Williams I, McIver S, Moore D, Bryan S.

No. 8
Stapled haemorrhoidectomy (haemorrhoidopexy) for the treatment of haemorrhoids: a systematic review and economic evaluation.
By Burch J, Epstein D, Baba-Akbari A, Weatherly H, Fox D, Golder S, *et al*.

No. 9
The clinical effectiveness of diabetes education models for Type 2 diabetes: a systematic review.
By Loveman E, Frampton GK, Clegg AJ.

No. 10
Payment to healthcare professionals for patient recruitment to trials: systematic review and qualitative study.
By Raftery J, Bryant J, Powell J, Kerr C, Hawker S.

No. 11
Cyclooxygenase-2 selective non-steroidal anti-inflammatory drugs (etodolac, meloxicam, celecoxib, rofecoxib, etoricoxib, valdecoxib and lumiracoxib) for osteoarthritis and rheumatoid arthritis: a systematic review and economic evaluation.
By Chen Y-F, Jobanputra P, Barton P, Bryan S, Fry-Smith A, Harris G, *et al*.

No. 12
The clinical effectiveness and cost-effectiveness of central venous catheters treated with anti-infective agents in preventing bloodstream infections: a systematic review and economic evaluation.
By Hockenhull JC, Dwan K, Boland A, Smith G, Bagust A, Dundar Y, *et al*.

No. 13
Stepped treatment of older adults on laxatives. The STOOL trial.
By Mihaylov S, Stark C, McColl E, Steen N, Vanoli A, Rubin G, *et al*.

No. 14
A randomised controlled trial of cognitive behaviour therapy in adolescents with major depression treated by selective serotonin reuptake inhibitors. The ADAPT trial.
By Goodyer IM, Dubicka B, Wilkinson P, Kelvin R, Roberts C, Byford S, *et al*.

No. 15
The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation.
By Hind D, Tappenden P, Tumur I, Eggington E, Sutcliffe P, Ryan A.

No. 16
Ranibizumab and pegaptanib for the treatment of age-related macular degeneration: a systematic review and economic evaluation.
By Colquitt JL, Jones J, Tan SC, Takeda A, Clegg AJ, Price A.

No. 17
Systematic review of the clinical effectiveness and cost-effectiveness of 64-slice or higher computed tomography angiography as an alternative to invasive coronary angiography in the investigation of coronary artery disease.
By Mowatt G, Cummins E, Waugh N, Walker S, Cook J, Jia X, *et al*.

No. 18
Structural neuroimaging in psychosis: a systematic review and economic evaluation.
By Albon E, Tsourapas A, Frew E, Davenport C, Oyebode F, Bayliss S, *et al*.

No. 19
Systematic review and economic analysis of the comparative effectiveness of different inhaled corticosteroids and their usage with long-acting beta$_2$ agonists for the treatment of chronic asthma in adults and children aged 12 years and over.
By Shepherd J, Rogers G, Anderson R, Main C, Thompson-Coon J, Hartwell D, *et al*.

No. 20
Systematic review and economic analysis of the comparative effectiveness of different inhaled corticosteroids and their usage with long-acting beta$_2$ agonists for the treatment of chronic asthma in children under the age of 12 years.
By Main C, Shepherd J, Anderson R, Rogers G, Thompson-Coon J, Liu Z, *et al*.

No. 21
Ezetimibe for the treatment of hypercholesterolaemia: a systematic review and economic evaluation.
By Ara R, Tumur I, Pandor A, Duenas A, Williams R, Wilkinson A, *et al*.

No. 22
Topical or oral ibuprofen for chronic knee pain in older people. The TOIB study.
By Underwood M, Ashby D, Carnes D, Castelnuovo E, Cross P, Harding G, *et al*.

No. 23
A prospective randomised comparison of minor surgery in primary and secondary care. The MiSTIC trial.
By George S, Pockney P, Primrose J, Smith H, Little P, Kinley H, *et al*.

No. 24
A review and critical appraisal of measures of therapist–patient interactions in mental health settings.
By Cahill J, Barkham M, Hardy G, Gilbody S, Richards D, Bower P, *et al*.

No. 25
The clinical effectiveness and cost-effectiveness of screening programmes for amblyopia and strabismus in children up to the age of 4–5 years: a systematic review and economic evaluation.
By Carlton J, Karnon J, Czoski-Murray C, Smith KJ, Marr J.

No. 26
A systematic review of the clinical effectiveness and cost-effectiveness and economic modelling of minimal incision total hip replacement approaches in the management of arthritic disease of the hip.
By de Verteuil R, Imamura M, Zhu S, Glazener C, Fraser C, Munro N, *et al*.

No. 27
A preliminary model-based assessment of the cost–utility of a screening programme for early age-related macular degeneration.
By Karnon J, Czoski-Murray C, Smith K, Brand C, Chakravarthy U, Davis S, *et al*.

No. 28
Intravenous magnesium sulphate and sotalol for prevention of atrial fibrillation after coronary artery bypass surgery: a systematic review and economic evaluation.
By Shepherd J, Jones J, Frampton GK, Tanajewski L, Turner D, Price A.

No. 29
Absorbent products for urinary/faecal incontinence: a comparative evaluation of key product categories.
By Fader M, Cottenden A, Getliffe K, Gage H, Clarke-O'Neill S, Jamieson K, *et al*.

No. 30
A systematic review of repetitive functional task practice with modelling of resource use, costs and effectiveness.
By French B, Leathley M, Sutton C, McAdam J, Thomas L, Forster A, *et al*.

No. 31
The effectiveness and cost-effectivness of minimal access surgery amongst people with gastro-oesophageal reflux disease – a UK collaborative study. The REFLUX trial.
By Grant A, Wileman S, Ramsay C, Bojke L, Epstein D, Sculpher M, *et al.*

No. 32
Time to full publication of studies of anti-cancer medicines for breast cancer and the potential for publication bias: a short systematic review.
By Takeda A, Loveman E, Harris P, Hartwell D, Welch K.

No. 33
Performance of screening tests for child physical abuse in accident and emergency departments.
By Woodman J, Pitt M, Wentz R, Taylor B, Hodes D, Gilbert RE.

No. 34
Curative catheter ablation in atrial fibrillation and typical atrial flutter: systematic review and economic evaluation.
By Rodgers M, McKenna C, Palmer S, Chambers D, Van Hout S, Golder S, *et al.*

No. 35
Systematic review and economic modelling of effectiveness and cost utility of surgical treatments for men with benign prostatic enlargement.
By Lourenco T, Armstrong N, N'Dow J, Nabi G, Deverill M, Pickard R, *et al.*

No. 36
Immunoprophylaxis against respiratory syncytial virus (RSV) with palivizumab in children: a systematic review and economic evaluation.
By Wang D, Cummins C, Bayliss S, Sandercock J, Burls A.

Volume 13, 2009

No. 1
Deferasirox for the treatment of iron overload associated with regular blood transfusions (transfusional haemosiderosis) in patients suffering with chronic anaemia: a systematic review and economic evaluation.
By McLeod C, Fleeman N, Kirkham J, Bagust A, Boland A, Chu P, *et al.*

No. 2
Thrombophilia testing in people with venous thromboembolism: systematic review and cost-effectiveness analysis.
By Simpson EL, Stevenson MD, Rawdin A, Papaioannou D.

No. 3
Surgical procedures and non-surgical devices for the management of non-apnoeic snoring: a systematic review of clinical effects and associated treatment costs.
By Main C, Liu Z, Welch K, Weiner G, Quentin Jones S, Stein K.

No. 4
Continuous positive airway pressure devices for the treatment of obstructive sleep apnoea–hypopnoea syndrome: a systematic review and economic analysis.
By McDaid C, Griffin S, Weatherly H, Durée K, van der Burgt M, van Hout S, Akers J, *et al.*

No. 5
Use of classical and novel biomarkers as prognostic risk factors for localised prostate cancer: a systematic review.
By Sutcliffe P, Hummel S, Simpson E, Young T, Rees A, Wilkinson A, *et al.*

No. 6
The harmful health effects of recreational ecstasy: a systematic review of observational evidence.
By Rogers G, Elston J, Garside R, Roome C, Taylor R, Younger P, *et al.*

No. 7
Systematic review of the clinical effectiveness and cost-effectiveness of oesophageal Doppler monitoring in critically ill and high-risk surgical patients.
By Mowatt G, Houston G, Hernández R, de Verteuil R, Fraser C, Cuthbertson B, *et al.*

No. 8
The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports.
By Taylor RS, Elston J.

No. 9
Controlling Hypertension and Hypotension Immediately Post Stroke (CHHIPS) – a randomised controlled trial.
By Potter J, Mistri A, Brodie F, Chernova J, Wilson E, Jagger C, *et al.*

No. 10
Routine antenatal anti-D prophylaxis for RhD-negative women: a systematic review and economic evaluation.
By Pilgrim H, Lloyd-Jones M, Rees A.

No. 11
Amantadine, oseltamivir and zanamivir for the prophylaxis of influenza (including a review of existing guidance no. 67): a systematic review and economic evaluation.
By Tappenden P, Jackson R, Cooper K, Rees A, Simpson E, Read R, *et al.*

No. 12
Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods.
By Hobart J, Cano S.

No. 13
Treatment of severe ankle sprain: a pragmatic randomised controlled trial comparing the clinical effectiveness and cost-effectiveness of three types of mechanical ankle support with tubular bandage. The CAST trial.
By Cooke MW, Marsh JL, Clark M, Nakash R, Jarvis RM, Hutton JL, *et al.*, on behalf of the CAST trial group.

No. 14
Non-occupational postexposure prophylaxis for HIV: a systematic review.
By Bryant J, Baxter L, Hird S.

No. 15
Blood glucose self-monitoring in type 2 diabetes: a randomised controlled trial.
By Farmer AJ, Wade AN, French DP, Simon J, Yudkin P, Gray A, *et al.*

No. 16
How far does screening women for domestic (partner) violence in different health-care settings meet criteria for a screening programme? Systematic reviews of nine UK National Screening Committee criteria.
By Feder G, Ramsay J, Dunne D, Rose M, Arsene C, Norman R, *et al.*

No. 17
Spinal cord stimulation for chronic pain of neuropathic or ischaemic origin: systematic review and economic evaluation.
By Simpson, EL, Duenas A, Holmes MW, Papaioannou D, Chilcott J.

No. 18
The role of magnetic resonance imaging in the identification of suspected acoustic neuroma: a systematic review of clinical and cost-effectiveness and natural history.
By Fortnum H, O'Neill C, Taylor R, Lenthall R, Nikolopoulos T, Lightfoot G, *et al.*

No. 19
Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort and qualitative study.
By Little P, Turner S, Rumsby K, Warner G, Moore M, Lowes JA, *et al.*

No. 20
Systematic review of respite care in the frail elderly.
By Shaw C, McNamara R, Abrams K, Cannings-John R, Hood K, Longo M, *et al.*

No. 21
Neuroleptics in the treatment of aggressive challenging behaviour for people with intellectual disabilities: a randomised controlled trial (NACHBID).
By Tyrer P, Oliver-Africano P, Romeo R, Knapp M, Dickens S, Bouras N, *et al.*

No. 22
Randomised controlled trial to determine the clinical effectiveness and cost-effectiveness of selective serotonin reuptake inhibitors plus supportive care, versus supportive care alone, for mild to moderate depression with somatic symptoms in primary care: the THREAD (THREshold for AntiDepressant response) study.
By Kendrick T, Chatwin J, Dowrick C, Tylee A, Morriss R, Peveler R, *et al.*

No. 23
Diagnostic strategies using DNA testing for hereditary haemochromatosis in at-risk populations: a systematic review and economic evaluation.
By Bryant J, Cooper K, Picot J, Clegg A, Roderick P, Rosenberg W, *et al.*

No. 24
Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis.
By McKenna C, McDaid C, Suekarran S, Hawkins N, Claxton K, Light K, *et al.*

No. 25
Development of a decision support tool for primary care management of patients with abnormal liver function tests without clinically apparent liver disease: a record-linkage population cohort study and decision analysis (ALFIE).
By Donnan PT, McLernon D, Dillon JF, Ryder S, Roderick P, Sullivan F, *et al.*

No. 26
A systematic review of presumed consent systems for deceased organ donation.
By Rithalia A, McDaid C, Suekarran S, Norman G, Myers L, Sowden A.

No. 27
Paracetamol and ibuprofen for the treatment of fever in children: the PITCH randomised controlled trial.
By Hay AD, Redmond NM, Costelloe C, Montgomery AA, Fletcher M, Hollinghurst S, *et al.*

No. 28
A randomised controlled trial to compare minimally invasive glucose monitoring devices with conventional monitoring in the management of insulin-treated diabetes mellitus (MITRE).
By Newman SP, Cooke D, Casbard A, Walker S, Meredith S, Nunn A, *et al.*

No. 29
Sensitivity analysis in economic evaluation: an audit of NICE current practice and a review of its use and value in decision-making.
By Andronis L, Barton P, Bryan S.

Suppl. 1
Trastuzumab for the treatment of primary breast cancer in HER2-positive women: a single technology appraisal.
By Ward S, Pilgrim H, Hind D.

Docetaxel for the adjuvant treatment of early node-positive breast cancer: a single technology appraisal.
By Chilcott J, Lloyd Jones M, Wilkinson A.

The use of paclitaxel in the management of early stage breast cancer.
By Griffin S, Dunn G, Palmer S, Macfarlane K, Brent S, Dyker A, *et al.*

Rituximab for the first-line treatment of stage III/IV follicular non-Hodgkin's lymphoma.
By Dundar Y, Bagust A, Hounsome J, McLeod C, Boland A, Davis H, *et al.*

Bortezomib for the treatment of multiple myeloma patients.
By Green C, Bryant J, Takeda A, Cooper K, Clegg A, Smith A, *et al.*

Fludarabine phosphate for the first-line treatment of chronic lymphocytic leukaemia.
By Walker S, Palmer S, Erhorn S, Brent S, Dyker A, Ferrie L, *et al.*

Erlotinib for the treatment of relapsed non-small cell lung cancer.
By McLeod C, Bagust A, Boland A, Hockenhull J, Dundar Y, Proudlove C, *et al.*

Cetuximab plus radiotherapy for the treatment of locally advanced squamous cell carcinoma of the head and neck.
By Griffin S, Walker S, Sculpher M, White S, Erhorn S, Brent S, *et al.*

Infliximab for the treatment of adults with psoriasis.
By Loveman E, Turner D, Hartwell D, Cooper K, Clegg A.

No. 30
Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation. The PoNDER trial.
By Morrell CJ, Warner R, Slade P, Dixon S, Walters S, Paley G, *et al.*

No. 31
The effect of different treatment durations of clopidogrel in patients with non-ST-segment elevation acute coronary syndromes: a systematic review and value of information analysis.
By Rogowski R, Burch J, Palmer S, Craigs C, Golder S, Woolacott N.

No. 32
Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care.
By Mant J, Doust J, Roalfe A, Barton P, Cowie MR, Glasziou P, *et al.*

No. 33
A multicentre randomised controlled trial of the use of continuous positive airway pressure and non-invasive positive pressure ventilation in the early treatment of patients presenting to the emergency department with severe acute cardiogenic pulmonary oedema: the 3CPO trial.
By Gray AJ, Goodacre S, Newby DE, Masson MA, Sampson F, Dixon S, *et al.*, on behalf of the 3CPO study investigators.

No. 34
Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation.
By Ara R, Pandor A, Stevens J, Rees A, Rafia R.

No. 35
Adefovir dipivoxil and pegylated interferon alpha for the treatment of chronic hepatitis B: an updated systematic review and economic evaluation.
By Jones J, Shepherd J, Baxter L, Gospodarevskaya E, Hartwell D, Harris P, *et al.*

No. 36
Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis.
By Hewitt CE, Gilbody SM, Brealey S, Paulden M, Palmer S, Mann R, *et al.*

No. 37
A double-blind randomised placebo-controlled trial of topical intranasal corticosteroids in 4- to 11-year-old children with persistent bilateral otitis media with effusion in primary care.
By Williamson I, Benge S, Barton S, Petrou S, Letley L, Fasey N, *et al.*

No. 38
The effectiveness and cost-effectiveness of methods of storing donated kidneys from deceased donors: a systematic review and economic model.
By Bond M, Pitt M, Akoh J, Moxham T, Hoyle M, Anderson R.

No. 39
Rehabilitation of older patients: day hospital compared with rehabilitation at home. A randomised controlled trial.
By Parker SG, Oliver P, Pennington M, Bond J, Jagger C, Enderby PM, *et al.*

No. 40
Breastfeeding promotion for infants in neonatal units: a systematic review and economic analysis
By Renfrew MJ, Craig D, Dyson L, McCormick F, Rice S, King SE, *et al.*

No. 41
The clinical effectiveness and cost-effectiveness of bariatric (weight loss) surgery for obesity: a systematic review and economic evaluation.
By Picot J, Jones J, Colquitt JL, Gospodarevskaya E, Loveman E, Baxter L, *et al.*

No. 42
Rapid testing for group B streptococcus during labour: a test accuracy study with evaluation of acceptability and cost-effectiveness.
By Daniels J, Gray J, Pattison H, Roberts T, Edwards E, Milner P, *et al.*

No. 43
Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling.
By Honest H, Forbes CA, Durée KH, Norman G, Duffy SB, Tsourapas A, *et al.*

No. 44
The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model.
By Bond M, Mealing S, Anderson R, Elston J, Weiner G, Taylor RS, *et al.*

Suppl. 2
Gemcitabine for the treatment of metastatic breast cancer.
By Jones J, Takeda A, Tan SC, Cooper K, Loveman E, Clegg A.

Varenicline in the management of smoking cessation: a single technology appraisal.
By Hind D, Tappenden P, Peters J, Kenjegalieva K.

Alteplase for the treatment of acute ischaemic stroke: a single technology appraisal.
By Lloyd Jones M, Holmes M.

Rituximab for the treatment of rheumatoid arthritis.
By Bagust A, Boland A, Hockenhull J, Fleeman N, Greenhalgh J, Dundar Y, *et al.*

Omalizumab for the treatment of severe persistent allergic asthma.
By Jones J, Shepherd J, Hartwell D, Harris P, Cooper K, Takeda A, *et al.*

Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-Hodgkin's lymphoma.
By Boland A, Bagust A, Hockenhull J, Davis H, Chu P, Dickson R.

Adalimumab for the treatment of psoriasis.
By Turner D, Picot J, Cooper K, Loveman E.

Dabigatran etexilate for the prevention of venous thromboembolism in patients undergoing elective hip and knee surgery: a single technology appraisal.
By Holmes M, C Carroll C, Papaioannou D.

Romiplostim for the treatment of chronic immune or idiopathic thrombocytopenic purpura: a single technology appraisal.
By Mowatt G, Boachie C, Crowther M, Fraser C, Hernández R, Jia X, *et al.*

Sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer.
By Bond M, Hoyle M, Moxham T, Napier M, Anderson R.

No. 45
Vitamin K to prevent fractures in older women: systematic review and economic evaluation.
By Stevenson M, Lloyd-Jones M, Papaioannou D.

No. 46
The effects of biofeedback for the treatment of essential hypertension: a systematic review.
By Greenhalgh J, Dickson R, Dundar Y.

No. 47
A randomised controlled trial of the use of aciclovir and/or prednisolone for the early treatment of Bell's palsy: the BELLS study.
By Sullivan FM, Swan IRC, Donnan PT, Morrison JM, Smith BH, McKinstry B, *et al.*

Suppl. 3
Lapatinib for the treatment of HER2-overexpressing breast cancer.
By Jones J, Takeda A, Picot J, von Keyserlingk C, Clegg A.

Infliximab for the treatment of ulcerative colitis.
By Hyde C, Bryan S, Juarez-Garcia A, Andronis L, Fry-Smith A.

Rimonabant for the treatment of overweight and obese people.
By Burch J, McKenna C, Palmer S, Norman G, Glanville J, Sculpher M, *et al.*

Telbivudine for the treatment of chronic hepatitis B infection.
By Hartwell D, Jones J, Harris P, Cooper K.

Entecavir for the treatment of chronic hepatitis B infection.
By Shepherd J, Gospodarevskaya E, Frampton G, Cooper, K.

Febuxostat for the treatment of hyperuricaemia in people with gout: a single technology appraisal.
By Stevenson M, Pandor A.

Rivaroxaban for the prevention of venous thromboembolism: a single technology appraisal.
By Stevenson M, Scope A, Holmes M, Rees A, Kaltenthaler E.

Cetuximab for the treatment of recurrent and/or metastatic squamous cell carcinoma of the head and neck.
By Greenhalgh J, Bagust A, Boland A, Fleeman N, McLeod C, Dundar Y, *et al.*

Mifamurtide for the treatment of osteosarcoma: a single technology appraisal.
By Pandor A, Fitzgerald P, Stevenson M, Papaioannou D.

Ustekinumab for the treatment of moderate to severe psoriasis.
By Gospodarevskaya E, Picot J, Cooper K, Loveman E, Takeda A.

No. 48
Endovascular stents for abdominal aortic aneurysms: a systematic review and economic model.
By Chambers D, Epstein D, Walker S, Fayter D, Paton F, Wright K, *et al.*

No. 49
Clinical and cost-effectiveness of epoprostenol, iloprost, bosentan, sitaxentan and sildenafil for pulmonary arterial hypertension within their licensed indications: a systematic review and economic evaluation.
By Chen Y-F, Jowett S, Barton P, Malottki K, Hyde C, Gibbs JSR, *et al.*

No. 50
Cessation of attention deficit hyperactivity disorder drugs in the young (CADDY) – a pharmacoepidemiological and qualitative study.
By Wong ICK, Asherson P, Bilbow A, Clifford S, Coghill D, R DeSoysa R, *et al.*

No. 51
ARTISTIC: a randomised trial of human papillomavirus (HPV) testing in primary cervical screening.
By Kitchener HC, Almonte M, Gilham C, Dowie R, Stoykova B, Sargent A, *et al.*

No. 52
The clinical effectiveness of glucosamine and chondroitin supplements in slowing or arresting progression of osteoarthritis of the knee: a systematic review and economic evaluation.
By Black C, Clar C, Henderson R, MacEachern C, McNamee P, Quayyum Z, *et al.*

No. 53
Randomised preference trial of medical versus surgical termination of pregnancy less than 14 weeks' gestation (TOPS).
By Robson SC, Kelly T, Howel D, Deverill M, Hewison J, Lie MLS, *et al.*

No. 54
Randomised controlled trial of the use of three dressing preparations in the management of chronic ulceration of the foot in diabetes.
By Jeffcoate WJ, Price PE, Phillips CJ, Game FL, Mudge E, Davies S, *et al.*

No. 55
VenUS II: a randomised controlled trial of larval therapy in the management of leg ulcers.
By Dumville JC, Worthy G, Soares MO, Bland JM, Cullum N, Dowson C, *et al.*

No. 56
A prospective randomised controlled trial and economic modelling of antimicrobial silver dressings versus non-adherent control dressings for venous leg ulcers: the VULCAN trial
By Michaels JA, Campbell WB, King BM, MacIntyre J, Palfreyman SJ, Shackley P, *et al.*

No. 57
Communication of carrier status information following universal newborn screening for sickle cell disorders and cystic fibrosis: qualitative study of experience and practice.
By Kai J, Ulph F, Cullinan T, Qureshi N.

No. 58
Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation.
By Burch J, Paulden M, Conti S, Stock C, Corbett M, Welton NJ, *et al.*

No. 59
Development of a toolkit and glossary to aid in the adaptation of health technology assessment (HTA) reports for use in different contexts.
By Chase D, Rosten C, Turner S, Hicks N, Milne R.

No. 60
Colour vision testing for diabetic retinopathy: a systematic review of diagnostic accuracy and economic evaluation.
By Rodgers M, Hodges R, Hawkins J, Hollingworth W, Duffy S, McKibbin M, *et al.*

No. 61
Systematic review of the effectiveness and cost-effectiveness of weight management schemes for the under fives: a short report.
By Bond M, Wyatt K, Lloyd J, Welch K, Taylor R.

No. 62
Are adverse effects incorporated in economic models? An initial review of current practice.
By Craig D, McDaid C, Fonseca T, Stock C, Duffy S, Woolacott N.

Volume 14, 2010

No. 1
Multicentre randomised controlled trial examining the cost-effectiveness of contrast-enhanced high field magnetic resonance imaging in women with primary breast cancer scheduled for wide local excision (COMICE).
By Turnbull LW, Brown SR, Olivier C, Harvey I, Brown J, Drew P, *et al.*

No. 2
Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: a systematic review and economic evaluation.
By Thompson Coon J, Hoyle M, Green C, Liu Z, Welch K, Moxham T, *et al.*

No. 3
The clinical effectiveness and cost-effectiveness of testing for cytochrome P450 polymorphisms in patients with schizophrenia treated with antipsychotics: a systematic review and economic evaluation.
By Fleeman N, McLeod C, Bagust A, Beale S, Boland A, Dundar Y, *et al.*

No. 4
Systematic review of the clinical effectiveness and cost-effectiveness of photodynamic diagnosis and urine biomarkers (FISH, ImmunoCyt, NMP22) and cytology for the detection and follow-up of bladder cancer.
By Mowatt G, Zhu S, Kilonzo M, Boachie C, Fraser C, Griffiths TRL, *et al.*

No. 5
Effectiveness and cost-effectiveness of arthroscopic lavage in the treatment of osteoarthritis of the knee: a mixed methods study of the feasibility of conducting a surgical placebo-controlled trial (the KORAL study).
By Campbell MK, Skea ZC, Sutherland AG, Cuthbertson BH, Entwistle VA, McDonald AM, *et al.*

No. 6
A randomised 2 × 2 trial of community versus hospital pulmonary rehabilitation for chronic obstructive pulmonary disease followed by telephone or conventional follow-up.
By Waterhouse JC, Walters SJ, Oluboyede Y, Lawson RA.

No. 7
The effectiveness and cost-effectiveness of behavioural interventions for the prevention of sexually transmitted infections in young people aged 13–19: a systematic review and economic evaluation.
By Shepherd J, Kavanagh J, Picot J, Cooper K, Harden A, Barnett-Page E, *et al.*

No. 8
Dissemination and publication of research findings: an updated review of related biases.
By Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, *et al.*

No. 9
The effectiveness and cost-effectiveness of biomarkers for the prioritisation of patients awaiting coronary revascularisation: a systematic review and decision model.
By Hemingway H, Henriksson M, Chen R, Damant J, Fitzpatrick N, Abrams K, *et al.*

No. 10
Comparison of case note review methods for evaluating quality and safety in health care.
By Hutchinson A, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, *et al.*

No. 11
Clinical effectiveness and cost-effectiveness of continuous subcutaneous insulin infusion for diabetes: systematic review and economic evaluation.
By Cummins E, Royle P, Snaith A, Greene A, Robertson L, McIntyre L, *et al.*

No. 12
Self-monitoring of blood glucose in type 2 diabetes: systematic review.
By Clar C, Barnard K, Cummins E, Royle P, Waugh N.

No. 13
North of England and Scotland Study of Tonsillectomy and Adeno-tonsillectomy in Children (NESSTAC): a pragmatic randomised controlled trial with a parallel non-randomised preference study.
By Lock C, Wilson J, Steen N, Eccles M, Mason H, Carrie S, *et al.*

No. 14
Multicentre randomised controlled trial of the clinical and cost-effectiveness of a bypass-surgery-first versus a balloon-angioplasty-first revascularisation strategy for severe limb ischaemia due to infrainguinal disease. The Bypass versus Angioplasty in Severe Ischaemia of the Leg (BASIL) trial.
By Bradbury AW, Adam DJ, Bell J, Forbes JF, Fowkes FGR, Gillespie I, *et al.*

No. 15
A randomised controlled multicentre trial of treatments for adolescent anorexia nervosa including assessment of cost-effectiveness and patient acceptability – the TOuCAN trial.
By Gowers SG, Clark AF, Roberts C, Byford S, Barrett B, Griffiths A, *et al.*

Health Technology Assessment programme

Prioritisation Strategy Group

Members

HTA Commissioning Board

Members

Observers

Diagnostic Technologies & Screening Panel

Members

Chair,
Professor Paul Glasziou,
Professor of Evidence-Based Medicine, University of Oxford

Deputy Chair,
Dr David Elliman,
Consultant Paediatrician and Honorary Senior Lecturer, Great Ormond Street Hospital, London

Professor Judith E Adams, Consultant Radiologist, Manchester Royal Infirmary, Central Manchester & Manchester Children's University Hospitals NHS Trust, and Professor of Diagnostic Radiology, Imaging Science and Biomedical Engineering, Cancer & Imaging Sciences, University of Manchester

Ms Jane Bates, Consultant Ultrasound Practitioner, Ultrasound Department, Leeds Teaching Hospital NHS Trust

Dr Stephanie Dancer, Consultant Microbiologist, Hairmyres Hospital, East Kilbride

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales

Dr Ron Gray, Consultant Clinical Epidemiologist, Department of Public Health, University of Oxford

Professor Paul D Griffiths, Professor of Radiology, University of Sheffield

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medicines & Healthcare Products Regulatory Agency, London

Dr Anne Mackie, Director of Programmes, UK National Screening Committee

Dr Michael Millar, Consultant Senior Lecturer in Microbiology, Barts and The London NHS Trust, Royal London Hospital

Mr Stephen Pilling, Director, Centre for Outcomes, Research & Effectiveness, Joint Director, National Collaborating Centre for Mental Health, University College London

Mrs Una Rennard, Service User Representative

Dr Phil Shackley, Senior Lecturer in Health Economics, School of Population and Health Sciences, University of Newcastle upon Tyne

Dr W Stuart A Smellie, Consultant in Chemical Pathology, Bishop Auckland General Hospital

Dr Nicholas Summerton, Consultant Clinical and Public Health Advisor, NICE

Ms Dawn Talbot, Service User Representative

Dr Graham Taylor, Scientific Advisor, Regional DNA Laboratory, St James's University Hospital, Leeds

Professor Lindsay Wilson Turnbull, Scientific Director of the Centre for Magnetic Resonance Investigations and YCR Professor of Radiology, Hull Royal Infirmary

Observers

Dr Tim Elliott, Team Leader, Cancer Screening, Department of Health

Dr Catherine Moody, Programme Manager, Neuroscience and Mental Health Board

Dr Ursula Wells, Principal Research Officer, Department of Health

Pharmaceuticals Panel

Members

Chair,
Professor Robin Ferner,
Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Deputy Chair,
Professor Imti Choonara,
Professor in Child Health, University of Nottingham

Mrs Nicola Carey, Senior Research Fellow, School of Health and Social Care, The University of Reading

Mr John Chapman, Service User Representative

Dr Peter Elton, Director of Public Health, Bury Primary Care Trust

Dr Ben Goldacre, Research Fellow, Division of Psychological Medicine and Psychiatry, King's College London

Mrs Barbara Greggains, Service User Representative

Dr Bill Gutteridge, Medical Adviser, London Strategic Health Authority

Dr Dyfrig Hughes, Reader in Pharmacoeconomics and Deputy Director, Centre for Economics and Policy in Health, IMSCaR, Bangor University

Professor Jonathan Ledermann, Professor of Medical Oncology and Director of the Cancer Research UK and University College London Cancer Trials Centre

Dr Yoon K Loke, Senior Lecturer in Clinical Pharmacology, University of East Anglia

Professor Femi Oyebode, Consultant Psychiatrist and Head of Department, University of Birmingham

Dr Andrew Prentice, Senior Lecturer and Consultant Obstetrician and Gynaecologist, The Rosie Hospital, University of Cambridge

Dr Martin Shelly, General Practitioner, Leeds, and Associate Director, NHS Clinical Governance Support Team, Leicester

Dr Gillian Shepherd, Director, Health and Clinical Excellence, Merck Serono Ltd

Mrs Katrina Simister, Assistant Director New Medicines, National Prescribing Centre, Liverpool

Mr David Symes, Service User Representative

Dr Lesley Wise, Unit Manager, Pharmacoepidemiology Research Unit, VRMM, Medicines & Healthcare Products Regulatory Agency

Observers

Ms Kay Pattison, Section Head, NHS R&D Programme, Department of Health

Mr Simon Reeve, Head of Clinical and Cost-Effectiveness, Medicines, Pharmacy and Industry Group, Department of Health

Dr Heike Weber, Programme Manager, Medical Research Council

Dr Ursula Wells, Principal Research Officer, Department of Health

Current and past membership details of all HTA programme 'committees' are available from the HTA website (www.hta.ac.uk)

Therapeutic Procedures Panel

Members

Chair,
Dr John C Pounsford,
Consultant Physician, North Bristol NHS Trust

Deputy Chair,
Professor Scott Weich,
Professor of Psychiatry, Division of Health in the Community, University of Warwick, Coventry

Professor Jane Barlow,
Professor of Public Health in the Early Years, Health Sciences Research Institute, Warwick Medical School, Coventry

Ms Maree Barnett,
Acting Branch Head of Vascular Programme, Department of Health

Mrs Val Carlill,
Service User Representative

Mrs Anthea De Barton-Watson,
Service User Representative

Mr Mark Emberton,
Senior Lecturer in Oncological Urology, Institute of Urology, University College Hospital, London

Professor Steve Goodacre,
Professor of Emergency Medicine, University of Sheffield

Professor Christopher Griffiths,
Professor of Primary Care, Barts and The London School of Medicine and Dentistry

Mr Paul Hilton,
Consultant Gynaecologist and Urogynaecologist, Royal Victoria Infirmary, Newcastle upon Tyne

Professor Nicholas James,
Professor of Clinical Oncology, University of Birmingham, and Consultant in Clinical Oncology, Queen Elizabeth Hospital

Dr Peter Martin,
Consultant Neurologist, Addenbrooke's Hospital, Cambridge

Dr Kate Radford,
Senior Lecturer (Research), Clinical Practice Research Unit, University of Central Lancashire, Preston

Mr Jim Reece
Service User Representative

Dr Karen Roberts,
Nurse Consultant, Dunston Hill Hospital Cottages

Observers

Dr Phillip Leech,
Principal Medical Officer for Primary Care, Department of Health

Ms Kay Pattison,
Section Head, NHS R&D Programme, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Medical Research Council

Professor Tom Walley,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Department of Health

Disease Prevention Panel

Members

Chair,
Dr Edmund Jessop,
Medical Adviser, National Specialist, National Commissioning Group (NCG), London

Deputy Chair,
Dr David Pencheon,
Director, NHS Sustainable Development Unit, Cambridge

Dr Elizabeth Fellow-Smith,
Medical Director, West London Mental Health Trust, Middlesex

Dr John Jackson,
General Practitioner, Parkway Medical Centre, Newcastle upon Tyne

Professor Mike Kelly,
Director, Centre for Public Health Excellence, NICE, London

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Corfe Mullen, Dorset

Ms Jeanett Martin,
Director of Nursing, BarnDoc Limited, Lewisham Primary Care Trust

Dr Julie Mytton,
Locum Consultant in Public Health Medicine, Bristol Primary Care Trust

Miss Nicky Mullany,
Service User Representative

Professor Ian Roberts,
Professor of Epidemiology and Public Health, London School of Hygiene & Tropical Medicine

Professor Ken Stein,
Senior Clinical Lecturer in Public Health, University of Exeter

Dr Kieran Sweeney,
Honorary Clinical Senior Lecturer, Peninsula College of Medicine and Dentistry, Universities of Exeter and Plymouth

Professor Carol Tannahill,
Glasgow Centre for Population Health

Professor Margaret Thorogood,
Professor of Epidemiology, University of Warwick Medical School, Coventry

Observers

Ms Christine McGuire,
Research & Development, Department of Health

Dr Caroline Stone,
Programme Manager, Medical Research Council

Expert Advisory Network

Members

Professor Douglas Altman, Professor of Statistics in Medicine, Centre for Statistics in Medicine, University of Oxford

Professor John Bond, Professor of Social Gerontology & Health Services Research, University of Newcastle upon Tyne

Professor Andrew Bradbury, Professor of Vascular Surgery, Solihull Hospital, Birmingham

Mr Shaun Brogan, Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE, Chief Executive, Regulation and Improvement Authority, Belfast

Ms Tracy Bury, Project Manager, World Confederation for Physical Therapy, London

Professor Iain T Cameron, Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Dr Christine Clark, Medical Writer and Consultant Pharmacist, Rossendale

Professor Collette Clifford, Professor of Nursing and Head of Research, The Medical School, University of Birmingham

Professor Barry Cookson, Director, Laboratory of Hospital Infection, Public Health Laboratory Service, London

Dr Carl Counsell, Clinical Senior Lecturer in Neurology, University of Aberdeen

Professor Howard Cuckle, Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Dr Katherine Darton, Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux, Professor of Paediatric Epidemiology, Institute of Child Health, London

Mr John Dunning, Consultant Cardiothoracic Surgeon, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw, Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles, Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby, Dean of Faculty of Medicine, Institute of General Practice and Primary Care, University of Sheffield

Professor Gene Feder, Professor of Primary Care Research & Development, Centre for Health Sciences, Barts and The London School of Medicine and Dentistry

Mr Leonard R Fenwick, Chief Executive, Freeman Hospital, Newcastle upon Tyne

Mrs Gillian Fletcher, Antenatal Teacher and Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn, Professor of Medicine, University of Birmingham

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Professor Fiona Gilbert, Consultant Radiologist and NCRN Member, University of Aberdeen

Professor Paul Gregg, Professor of Orthopaedic Surgical Science, South Tees Hospital NHS Trust

Bec Hanley, Co-director, TwoCan Associates, West Sussex

Dr Maryann L Hardy, Senior Lecturer, University of Bradford

Mrs Sharon Hart, Healthcare Management Consultant, Reading

Professor Robert E Hawkins, CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor Richard Hobbs, Head of Department of Primary Care & General Practice, University of Birmingham

Professor Alan Horwich, Dean and Section Chairman, The Institute of Cancer Research, London

Professor Allen Hutchinson, Director of Public Health and Deputy Dean of ScHARR, University of Sheffield

Professor Peter Jones, Professor of Psychiatry, University of Cambridge, Cambridge

Professor Stan Kaye, Cancer Research UK Professor of Medical Oncology, Royal Marsden Hospital and Institute of Cancer Research, Surrey

Dr Duncan Keeley, General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping, Research Degrees Programme Director and Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Mr George Levvy, Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay, Professor of Psychiatry for the Elderly, University of Leicester

Professor Julian Little, Professor of Human Genome Epidemiology, University of Ottawa

Professor Alistaire McGuire, Professor of Health Economics, London School of Economics

Professor Rajan Madhok, Medical Director and Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor Alexander Markham, Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Peter Moore, Freelance Science Writer, Ashtead

Dr Andrew Mortimore, Public Health Director, Southampton City Primary Care Trust

Dr Sue Moss, Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Professor Miranda Mugford, Professor of Health Economics and Group Co-ordinator, University of East Anglia

Professor Jim Neilson, Head of School of Reproductive & Developmental Medicine and Professor of Obstetrics and Gynaecology, University of Liverpool

Mrs Julietta Patnick, National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler, Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Professor Chris Price, Director of Clinical Research, Bayer Diagnostics Europe, Stoke Poges

Professor William Rosenberg, Professor of Hepatology and Consultant Physician, University of Southampton

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Dr Susan Schonfield, Consultant in Public Health, Hillingdon Primary Care Trust, Middlesex

Dr Eamonn Sheridan, Consultant in Clinical Genetics, St James's University Hospital, Leeds

Dr Margaret Somerville, Director of Public Health Learning, Peninsula Medical School, University of Plymouth

Professor Sarah Stewart-Brown, Professor of Public Health, Division of Health in the Community, University of Warwick, Coventry

Professor Ala Szczepura, Professor of Health Service Research, Centre for Health Services Studies, University of Warwick, Coventry

Mrs Joan Webster, Consumer Member, Southern Derbyshire Community Health Council

Professor Martin Whittle, Clinical Co-director, National Co-ordinating Centre for Women's and Children's Health, Lymington

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (www.hta.ac.uk) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

NETSCC, Health Technology Assessment
Alpha House
University of Southampton Science Park
Southampton SO16 7NS, UK
Email: hta@hta.ac.uk
www.hta.ac.uk

ISSN 1366-5278