Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice

Julian Higgins, Simon Thompson, Jonathan Deeks and Douglas Altman

J Health Serv Res Policy 2002 7: 51
DOI: 10.1258/1355819021927674

Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice

Julian Higgins, Simon Thompson, Jonathan Deeks[1], Douglas Altman[1]

MRC Biostatistics Unit, Cambridge; [1]ICRF/NHS Centre for Statistics in Medicine, Oxford, UK

***Objective*: Heterogeneity between study results can be a problem in any systematic review or meta-analysis of clinical trials. Identifying its presence, investigating its cause and correctly accounting for it in analyses all involve difficult decisions for the researcher. Our objectives were: to collate recommendations on the subject of dealing with heterogeneity in systematic reviews of clinical trials; to investigate current practice in addressing heterogeneity in Cochrane reviews; and to compare current practice with recommendations.**

***Methods*: We review guidelines for those undertaking systematic reviews and examine how heterogeneity is addressed in practice in a sample of systematic reviews, and their protocols, from the Cochrane Database of Systematic Reviews.**

***Results*: Advice to reviewers is on the whole consistent and sensible. However, examination of a sample of Cochrane protocols and reviews demonstrates that the advice is difficult to follow given the small numbers of studies identified in many systematic reviews, the difficulty of pre-specifying important effect modifiers for subgroup analysis or meta-regression and the unresolved debate concerning fixed versus random effects meta-analyses. There was disagreement between protocols and reviews, often either regarding choice of important potential effect modifiers or due to the review identifying too few studies to perform planned analyses.**

***Conclusion*: Guidelines that address practical issues are required to reduce the risk of spurious findings from investigations of heterogeneity. This may involve discouraging statistical investigations such as subgroup analyses and meta-regression, rather than simply adopting a cautious approach to their interpretation, unless a large number of studies is available. The notion of a priori specification of potential effect modifiers for a retrospective review of studies is ill-defined, and the appropriateness of using a statistical test for heterogeneity to decide between analysis strategies is suspect.**

Journal of Health Services Research & Policy Vol 7 No 1, 2002: 51–61

Introduction

Systematic reviews of health care interventions underpin provision of health services and therefore play an important role in both research and health services policy. Any systematic review or meta-analysis of clinical trials will inevitably include studies that are to some extent heterogeneous. Heterogeneity may be of a number of types. Clinical heterogeneity might result from differences between patients, interventions being compared or outcomes collected. Methodological heterogeneity arises through the use of different study designs and different degrees of control over bias. Statistical heterogeneity is a third type of heterogeneity that is a consequence of the first two – that is, variation between trials in the underlying treatment effects being evaluated. It may be detected if variation in the results of the studies is above that compatible with chance alone. This paper largely focuses on statistical heterogeneity, and the word 'heterogeneity' will be used for this purpose. Differences in clinical and methodological aspects of the studies we refer to as 'diversity'. Box 1 provides a glossary of these and other methodological terms.

It is now generally accepted that meta-analysis should attempt to go beyond estimating a single average treatment effect.[1] Reasons for heterogeneity should be explored in order to increase scientific understanding and clinical relevance.[2] But despite the inevitability of heterogeneity, how it should be handled in systematic reviews is far from clear. Questions that face a reviewer include: should we rely on a statistical test for detecting heterogeneity; what should guide one's choice of fixed or random effects meta-analysis techniques; is there an extent of heterogeneity beyond which it is not meaningful to undertake a quantitative synthesis; how many

Julian Higgins PhD, Statistician, **Simon Thompson DSc**, Director, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK. **Jonathan Deeks MSc**, Senior Medical Statistician, **Douglas Altman DSc**, Professor of Statistics in Medicine, ICRF/NHS Centre for Statistics in Medicine, Oxford, UK.

Correspondence to: JH.

Box 1 Glossary of methodological terms

Systematic review

An approach to reviewing literature or a research field that attempts to reduce biases by taking a systematic, replicable approach to the identification, selection and assessment of studies and the presentation of their findings.

Example: A systematic review might seek and summarise the findings of all randomised controlled trials of the drugs amantadine and rimantadine versus placebo for the prevention of influenza (see Box 2).

Diversity

The variation in clinical and methodological characteristics of a set of studies collected together in a systematic review.

Example: Clinical trials will normally differ clinically in terms of the population studied, the exact implementations of the intervention and control and the definition and assessment of outcomes. They may also differ methodologically in terms of study design (for example, cross-over and parallel group trials) and degrees of control over bias (for example, concealment of allocation to interventions, or blinding).

Meta-analysis

Statistical combination of results from separate studies, ideally collated within a systematic review, for the purpose of providing an overall numerical estimate of treatment effect.

Example: A meta-analysis might combine the relative risks of developing influenza while taking amantadine compared with placebo from several clinical trials.

Heterogeneity (statistical heterogeneity)

The situation in which treatment effects being estimated by individual studies in a systematic review are not identical. This will manifest itself in greater variability in the estimates than would be expected by chance (sampling variation) alone.

Example: Relative risks underlying the individual trials in a systematic review may differ because the trials are undertaken in populations with different responses to the drug.

Fixed effect meta-analysis

A meta-analysis that assumes every study is estimating the same unknown treatment effect – i.e. that the underlying treatment effects are identical.

Example: A fixed effect meta-analysis of clinical trials comparing amantadine with placebo might assume that there is a single, unknown, relative risk that all studies are trying to evaluate.

Random effects meta-analysis

A meta-analysis that assumes studies are estimating different, but related, unknown treatment effects, with the differences between these represented by random variation.

Example: A random effects meta-analysis of clinical trials comparing amantadine with placebo might assume that the underlying log relative risks follow a normal distribution across the studies.

Potential effect modifier (covariate)

A study characteristic that might explain differences in results across studies. An ‘effect modifier’ is a characteristic that actually explains different results.

Example: There is evidence from clinical trials in pregnancy and childbirth that lack of concealed allocation to interventions is associated with exaggerated treatment effects.[28] Thus concealment of allocation is a potential effect modifier of treatment effects in other fields.

Subgroup analysis

The separation of different studies in a meta-analysis into subsets of studies, with a meta-analysis being performed on one or more of these subsets. Criteria for subgrouping studies are potential effect modifiers that are categorical in nature.

Example: A meta-analysis of those clinical trials that took great care to conceal allocation to interventions might be performed as a subgroup analysis of all clinical trials.

Meta-regression

An extension to meta-analysis and a generalisation of subgroup analyses. Rather than seeking an overall evaluation of a single unknown treatment effect, meta-regression investigates the relationship between the treatment effect and one or more potential effect modifiers.

Example: It might be posed that the underlying relative risk of developing influenza on amantadine relative to placebo depends on the dose of amantadine administered. Meta-regression may be used to assess the relationship between dose and efficacy (i.e. relative risk) across studies.

Sensitivity analysis

Any investigation that addresses the question ‘Are the results robust to how they were obtained?’. Most research studies and statistical analyses involve decisions about how they should be undertaken. These decisions may or may not be important.

Example: One factor that might be subject to sensitivity analysis in a meta-analysis of clinical trials is a decision to use a fixed effect rather than a random effects meta-analysis.

subgroup analyses should be carried out, and does this depend on the number of studies available; is meta-regression a preferable technique to subgroup analysis? Many of these are difficult questions on which reviewers need advice.

Here we investigate what available guidelines and recommendations say about addressing heterogeneity. In addition, we compare protocols for reviews with the completed review in the Cochrane Database of Systematic Reviews (CDSR).[3] The CDSR is the major source of systematic reviews of clinical trials of health care interventions. Reviews appear in *The Cochrane Library* and are regularly updated to incorporate new or newly identified studies. The Library also contains protocols for reviews in preparation, allowing comparison of pre-specified with subsequently performed analyses. The reviews are produced by problem- or disease-based collaborative review groups. At the time of this study there were 47 registered groups, of which 39 had published completed reviews.

Our objectives were: to collate recommendations on the subject of dealing with heterogeneity in systematic reviews of clinical trials; to investigate current practice in addressing heterogeneity in Cochrane reviews; and to compare current practice with recommendations. In this way, we focus attention on areas in need of further research. We note at the outset that some of the guidelines we discuss are under continual review and subject to change. This paper should be viewed as part of that process of review and change.

Methods

Guidance for addressing heterogeneity

Many papers have addressed the importance of heterogeneity within systematic reviews or meta-analyses,[2,4–6] although none of these present guidelines or recommendations. We examined three principal sources of guidance to people undertaking systematic reviews of clinical trials: the Cochrane Handbook,[7,8] a report from the NHS Centre for Reviews and Dissemination (CRD)[9] and guidelines by the Australian National Health and Medical Research Council (NHMRC).[10] Searches of the Cochrane Review Methodology Database, MEDLINE and the Science Citation Index (to May 2000) for articles mentioning meta-analysis along with a variation of guideline or recommendation yielded two other relevant sets of advice. The first contains widely cited guidelines that arose from a conference on meta-analysis in Potsdam in 1994.[11] The second lists concluding recommendations from an extensive review of methodology for systematic reviews and meta-analysis.[12] This advocates following the Cochrane Handbook, the CRD report and the Potsdam conference guidelines 'for the most part' and offers a series of more specific recommendations largely tailored to reviewing studies in health services research. A final source of which we were aware is the list of conclusions of a panel brought together by the US National Research Council to address combining information in all areas of research, including health care.[13] These constitute recommendations from a statistical perspective, strongly advocating advanced statistical methods such as hierarchical models. These are relevant to the choice of analysis strategy but do not address all heterogeneity-related issues important to systematic reviewers. We thus restrict most of our discussion to the four principal sets of general guidelines,[8–11] which we refer to as 'the Handbook', 'CRD', 'NHMRC' and 'Potsdam', respectively, and draw on the US National Research Council recommendations when appropriate. We note that there also exist recommendations for reporting of meta-analyses,[14] which we discuss later.

Practice when dealing with heterogeneity

To obtain a representative sample of current Cochrane reviews, we looked at the most recently completed systematic review from each collaborative review group in the CDSR (Issue 2, 1999). The original protocol for each identified review was sought from back issues of *The Cochrane Library*. A data extraction form was developed and piloted on five protocol/review pairs not included in the sample. A randomly generated selection of seven reviews had data extracted independently by two of the authors. Disagreements were readily resolved by discussion; the majority were due to different interpretations of the data extraction instructions rather than of the reviews. The remaining reviews had data extracted by the first author only. The text and, where appropriate, data and analyses of both the protocol and the final review were examined and information collected concerning: (1) the objective of the systematic review; (2) identification of heterogeneity (by testing, graphical inspection or other means); (3) whether meta-analyses were performed, and their results for the primary outcome; (4) choices concerning fixed and random effects analyses; (5) subgroup analyses and meta-regression (both pre-specified and post hoc); and (6) action taken as a result of statistical, clinical or methodological heterogeneity. The definition of a 'primary' outcome is given in the footnote to the Table. In reviews with more than one major comparison of treatments, the first presented in a meta-analysis was used. Care was taken to distinguish subgrouping of trials related to possible statistical heterogeneity from subgrouping of trials made for entirely logical grounds. An example of the former may include children and adults; examples of the latter may include comparisons with placebo and comparisons with a standard treatment.

We supplemented our findings from both the guidelines and the systematic reviews with more specific, mainly unpublished, advice prepared by collaborative review groups themselves. To obtain the latter we surveyed the coordinators of the 44 existing groups in May 1999, with the approval of the Collaboration Secretariat, to ask what additional advice they were able to offer reviewers in terms of guidelines, group policies or generic documents. Of the 44 groups we

Table Information extracted from 39 reviews identified in the Cochrane Database of Systematic Reviews (CDSR)

CDSR review number	Topic area	Protocol published	Total number of included trials	Fixed/ random effects[b] in review	Subgroups		Primary outcome[a]			
					Multiple questions[c]	Subgroup fate category[d]	Number of eligible studies	Number of eligible studies with data	Meta-analysis performed[e]	Statistically significant heterogeneity ($P < 0.05$)
0019	Prenatal thyrotropin-releasing hormone for preterm birth	✗	11	r>f	✗	–	11	6	✓	✗
0125	Local opinion leaders	✓	8	–	✓	A	1	1	✗	–
0144	Natural surfactant extract for neonatal respiratory distress syndrome	✗	7	f	✗	–	7	5	✓	✗
0169	Prevention versus treatment for malaria	✗	15	f	✓	–	11	1	✗	–
0209	Vitamin E for tardive dyskinesia	✓	8	f	✗	F	8	4	✓	✗
0235	Pergolide for complications in Parkinson's disease	✓	1	–	✗	A	1	1	✗	–
0251	Managements for people with disorders of sexual preference	✓	3	f	✓	C	1	0	✗	–
0299	Anti-tuberculous therapy for Crohn's disease	✓	7	f	✗	D	7	7	✓	✓
0309	Diabetes routine surveillance	✓	5	f	✗	B	5	4	✓	✗
0419	Piracetam for acute ischaemic stroke	✓	3	f	✗	C	3	3	✓	✗
0473	Treatment for chronic suppurative otitis media	✓	24	r>f	✓	A	2	2	✓	✗
0475	Interventions for adhesions after subfertility surgery	✓	15	f	✓	F	3	0	✗	–
0484	Dieting for hypertension in adults	✓	18	–	✓	C	12	11	✗	✓
0485	Ovarian ablation for early breast cancer	✓	13	f	✗	F	13	12	✓	✗
0992	Compression bandages for venous leg ulcers	✓	22	f	✓	A	6	6	✗	✗
1036	Collection devices for cervical cytology samples	✗	34	f	✓	–	10	8	✓	✓
1048	Therapeutic hypothermia for head injury	✓	8	f	✓	B	7	6	✓	✗
1130	Drugs for dysthymia	✓	15	r	✗	D	15	13	✓	✗
1159	Nerve blocks for hip fractures	✓	6	f	✓	E	4	3	(✓)	✗
1168	Interventions for oral lichen planus	✓	9	f	✓	A	3	2	✓	✗
1169	Amantadine/rimantadine for influenza A	✓	31	r>f	✓	F	18	10	(✓)	✓
1199	(Neo)adjuvant therapy for operable hepatocellular carcinoma	✓	8	–	✗	A	8	0	✗	–
1216	Screening for colorectal cancer using Hemoccult	✗	5	f+r	✗	–	5	5	✓	✗
1287	Mucolytic agents for chronic bronchitis	✓	15	f	✗	D	15	14	✓	✓
1292	Individual behavioural counselling for smoking cessation	✓	11	f	✓	A	10	10	✓	✗
1308	Bladder training for urinary incontinence	✓	5	–	✓	C	3	1	✗	–
1321	Cranberries for urinary tract infections	✗	4	f+r	✗	–	4	0	✗	–
1333	Naltrexone maintenance treatment for opioid dependence	✓	11	–	✓	E	6	5	(✓)	✗
1351	Acupuncture for low back pain	✓	11	–	✓	C	3	1	✗	–
1394	Validation therapy for dementia	✗	2	–	✓	–	1	1	✗	–
1423	Serenoa repens for benign prostatic hyperplasia	✓	18	r	✓	A	13	1	✗	–
1446	Corticosteroids for Guillain–Barré syndrome	✓	6	f	✗	C	6	3	✓	✗
1460	Penicillamine for rheumatoid arthritis	✗	6	f	✗	–	6	3	✓	✗
1487	Graft type for femoropopliteal bypass surgery	✓	9	–	✓	A	2	0	✗	–
1506	Nebulised hypertonic saline for cystic fibrosis	✓	5	f	✓	E	4	1	✗	–
1524	Yoga for epilepsy	✓	1	–	✗	C	1	1	✗	–
1538	Surgery for non-arteritic anterior ischaemic optic neuropathy	✗	1	–	✗	–	1	1	✗	–
1548	Single dose oral ibuprofen/diclofenac for postoperative pain	✗	39	f	✗	–	39	39	(✓)	✓
1560	Primary angioplasty for acute myocardial infarction	✗	10	f	✗	–	10	10	✓	✗

[a] A primary outcome was selected for each review by selecting that which met the first of the following criteria: (1) an outcome described as the primary outcome in the protocol; (2) the first mentioned in the objectives, or failing that the list of outcomes, in the protocol; (3) the first mentioned in the objectives, or failing that the list of outcomes, in the review; (4) mortality; or (5) the outcome of the first meta-analysis presented in the review.
[b] Policy regarding fixed and random effects meta-analysis based on the protocol where possible (f = fixed effect; r = random effects; f + r = both; r> f = fixed effect when studies homogeneous or random effects when not).
[c] Multiple questions were concluded when the review appeared to address more than one question; for example, drug A versus placebo as well as drug A versus drug B. In such instances, the first question addressed was considered in investigations of the 'primary outcome'.
[d] See Figure 1. Only reviews with protocols are classified.
[e] (✓), the reviewers chose not to combine data from all studies in one analysis.

surveyed, 31 replied, yielding essentially four different documents. One of these had been adopted by seven different groups, one was based on the Potsdam guidelines and one has been published as a journal article.[15]

Results

Guidance for addressing heterogeneity

The Handbook, CRD and NHMRC discussed the role of problem formulation and inclusion criteria in defining the scope of a systematic review and hence the likely amount of diversity. The decision on the scope of a review has important consequences for subsequent investigations or explorations of heterogeneity. All three sources stated that issues related to heterogeneity should be addressed in the protocol and clearly recommended that potential subgroup analyses be specified a priori. No clear definitions of 'a priori' were given, a notable omission given the nature of most systematic reviews as summarising historical evidence. CRD suggested 'potential effect modifiers should be identified prior to any data analysis by reviewing the literature on risk factors, and through consultations with experts in the field'. A review with broad inclusion criteria may facilitate the most meaningful investigations of heterogeneity. Alternatively it may, for example, contain several treatment comparisons, or include two or more distinct patient populations, and the line between what is a subgroup analysis and what is a separate analysis may easily become blurred. In practice, subgroup analyses focus on differential effects between different subgroups, rather than on separate effects in different analyses. On the topic of comparing subgroups, NHMRC stressed the need to compare subgroups directly (for example, using a test for interaction) rather than undertake separate analyses in separate subgroups.

It is possible to perform a statistical test for the presence of heterogeneity. A common approach is based on a summary statistic from each study. It has been demonstrated that the test often has low power in practice, so that even a moderate degree of true heterogeneity may not be revealed as statistically significant.[9] All four guidelines mention the possibility of performing such a test. CRD refers to its low power. In contrast, Potsdam referred to the test's excessive power when there are many studies, and therefore suggested the need to 'specify a priori the clinically important degree of heterogeneity'. No source recommended a specific level for statistical significance, though the Handbook did suggest that either a 5% or 10% level might be appropriate, hinting at the low power of the test. NHMRC recommended using the statistic $Q/(k-1)$ as an indication of the amount of heterogeneity, where Q is the χ^2 statistic in a test for heterogeneity and k is the number of studies in the meta-analysis, suggesting that a value above 1 indicates a need to explore heterogeneity even if the test is non-significant.

Guidance was available on what may be done if excessive heterogeneity is identified. Advice was generally consistent, advocating a cautious examination of potential causes of heterogeneity and the use of random effects meta-analyses to account for variation that cannot be explained (either instead of or in addition to fixed effect analyses). Specific guidance on choosing between fixed effect and random effects meta-analyses was not available, except from the Cochrane Eyes and Vision Group, who recommended that for a p-value from the heterogeneity test greater than 0.10 a fixed effect model should be used; for a p-value between 0.05 and 0.10 both models should be used; and for a p-value less than 0.05 no meta-analysis should be performed.

The word caution with respect to handling heterogeneity appeared in three sources – extremely liberally in the Handbook – but this is another term that is insufficiently well defined to be of great use to the reviewer. All sources stressed the likelihood of spurious findings from post hoc subgroup analyses. The Handbook and Potsdam largely followed the advice given by Oxman and Guyatt,[16] highlighting the importance of pre-specification, small numbers of subgroups, support by causal mechanisms, magnitude of effect and statistical significance. The Handbook used the strongest language, stating that subgroup analyses 'are also often misleading ... reviewers need to be cautious about doing subgroup analyses and about interpreting the results of the ones that they feel compelled to do'.

The Handbook and CRD made mention of meta-regression as a tool for investigating heterogeneity. The Potsdam report suggested the need 'to explore the relationship between effect size and study quality, control event rates, and other relevant features' under the heading of 'sensitivity analysis'. The software for preparing Cochrane reviews does not currently offer the possibility of performing meta-regression, so it is understandable that the approach was not emphasised in the Handbook. This document did offer the warning that 'there is a potential for bias when selecting which of many possible factors to include' in a meta-regression.

The principal contributions of the two further sources we examined were regarding statistical methods. Sutton et al[12] recommended the use of 'hierarchical models, including fixed covariates to explain some elements of between study variation, in combination with random effects terms'. The US National Research Council panel[13] 'believes that [meta-analytic] modeling would be improved by the increased use of random effects models in preference to the current default of fixed effect models'.

In summary, generic guidance for addressing heterogeneity in systematic reviews of clinical trials appeared to be fairly consistent, with the key features being a priori specification of potential effect modifiers and planned subgroup analyses, recommendations to examine heterogeneity with regard to both a priori and post hoc specified effect modifiers (though with particularly cautious interpretation of the latter) and cautious use of fixed effect and random effects analyses (or a comparison of the two) when heterogeneity is present but cannot adequately be explained. We noted only a limited amount of specific advice on determining when heterogeneity poses a problem, on whether there is a point at which heterogeneity becomes too excessive for a meaningful meta-analysis, and on how many studies or how many effect modifiers might be suitable for explorations of heterogeneity. There was little discussion of the use of meta-regression as a more flexible alternative to subgroup analyses or on how the choice of effect measure (for example, risk ratio versus risk difference) can affect the extent of heterogeneity.

Practice when dealing with heterogeneity

We present findings so as to illustrate points rather than to summarise all the information we collected, some of which involved subjective interpretation of the protocols and reviews. We supplement this with a case study (Box 2) to illustrate and clarify some of the issues within the context of a specific example.

Summary information from the 39 reviews is presented in the Table. Protocols had been previously published for 28 (69%) of the reviews, two of which were extremely short (one of these referred to an external source of methodology[17] and the other to a generic document for the review group[18]). No review had a primary objective of investigating heterogeneity between studies: all had primary objectives of 'assessing', 'determining', 'evaluating' or 'examining' the effect of the intervention under scrutiny, of testing a null hypothesis or of 'comparing' the effects of two or more interventions. Three reviews stated secondary objectives of determining a 'best' intervention or method of delivery, and three stated secondary objectives of exploring differences (i.e. heterogeneity) between studies. Sixteen protocols (57%) and 28 reviews (72%) were considered to have specifically addressed the topic of heterogeneity in some form.

Pre-specified and post hoc subgroup analyses

Fifteen of the 28 protocols pre-specified potential effect modifiers for use in subgroup analyses or meta-regression. Characteristics that make studies inappropriate to combine in a logical sense (such as different treatment comparisons) are not included as effect modifiers here. The fate of reviews that did and did not pre-specify subgroups (including study characteristics for meta-regression) for the 28 reviews that had protocols is described in Figure 1 (see also the Table, column 7).

Just two reviews performed only subgroup analyses that had been pre-specified in the protocol. Ten reviews did not undertake any of the planned subgroup analyses or meta-regressions, and in four further reviews at least one planned subgroup analysis was not undertaken. Stated reasons for these decisions (some reviews giving more than one reason) were insufficient information to classify studies into subgroups (in five reviews), too few studies (two reviews), planned subgroup analyses more appropriate for dividing patients than studies (two reviews), too much diversity in studies to consider any meta-analysis (two reviews), no studies in a subgroup (one review) and lack of a significant treatment effect across all studies (one review). When no reason for omitting planned subgroup analyses was given, it was clear that having too few studies was a major factor.

Ten reviews undertook subgroup analyses that were not mentioned in the protocol (defined here as post hoc subgroup analyses). Of three that conducted both pre-specified and post hoc subgroup analyses, only one made the distinction clear. Eight of the 11 reviews without protocols undertook subgroup analyses; it was not possible to determine whether these had been pre-specified.

Identifying heterogeneity

Twelve protocols stated plans to perform statistical tests for heterogeneity, only one of which pre-specified the level of significance to be used (as 10%). Five planned to investigate diversity of studies before considering statistical combination of results in meta-analyses. A further six completed reviews specifically considered clinical diversity prior to proceeding with quantitative summaries. In total, 22 reviews (56%) commented on the use of statistical tests for heterogeneity in the text. Two of these stated the significance level as 5%, two as 10%, and one declared a finding of $p = 0.08$ 'borderline'. Five reviews mentioned graphical assessments of heterogeneity, all but one as precursors or supplements to statistical tests.

Fixed and random effects meta-analyses

One way of dealing with statistical heterogeneity is to incorporate a term to account for it in a random effects meta-analysis. Many different policies were described for choosing between fixed effect and random effects meta-analyses. The paths that the reviews followed from protocol to completion are illustrated in Figure 2 (see also the Table, column 5). Ten protocols stated an explicit policy: two to do random effects analyses, four to do fixed effect analyses, two to do both, and two to do random effects analyses in the presence of significant heterogeneity and fixed effect analyses otherwise. No definition of statistical significance was given in either of these last two cases. Plans in the protocol were generally, though not always, followed. Among the 21 reviews that performed (or preferred) only fixed effect meta-analyses, four did this despite having outcomes with statistically significant heterogeneity ($p < 0.05$). One of these had pre-specified the use of fixed effect analyses; the other three had no protocols. Of note is that these four statistically significant findings constitute the majority of the six significant heterogeneity findings we found in total among the main outcomes of the 39 reviews (Table, last column). Only two completed reviews performed (or preferred) random effects meta-analyses alone, whereas three reviews followed a rule of choosing random effects analyses in the presence of statistical heterogeneity. Two of these defined statistical

Box 2 A case study: Amantadine and rimantadine for preventing and treating influenza A in adults[29]

One of the reviews in our selection was of two antiviral drugs, amantadine and rimantadine, for prevention and treatment of influenza A versus either control/placebo or each other (Table, review number 1169). The following analysis of this review contrasts information derived from the protocol and review with explanations from an author of the review (JJD). The two perspectives emphasise difficulties in pre-stating methods for the analysis of heterogeneity in a review and in ascertaining the rationales behind presented analyses.

The protocol for the review specified that a (fixed effect) Mantel–Haenszel approach to meta-analysis would be used, and between-trial heterogeneity would be examined and incorporated in random effects analyses if present. (The subsequent review later defined heterogeneity being present as a test for heterogeneity yielding a p-value less than 0.05.) No formal subgroup analyses were specified a priori, other than it being stated that studies of prevention and treatment would be considered separately as they address different questions. Thus the review is an example of one addressing multiple questions (see Table).

The review identified 17 studies in prevention, involving 18 distinct clinical trials (14 of amantadine versus placebo, one of rimantadine versus placebo and three of amantadine versus rimantadine versus placebo) and ten studies in treatment, involving 13 distinct clinical trials (ten, one and two of the three designs, respectively). The protocol for our investigation dictated that we selected prevention studies with the primary outcome of cases of serologically confirmed clinical influenza. Ten of the 18 prevention trials reported this outcome. The review also addressed adverse effects and, as the major outcome in the treatment studies, fever as a measure of influenza severity. The authors note that this last outcome is not the best way to measure severity of influenza, but it was the only outcome reported by the majority of studies.

One approach to dealing with sources of heterogeneity is to present separate analyses for subgroups of trials. In the published review three such analyses are presented, but only the separation of prevention and treatment trials was pre-specified in the protocol. In addition, placebo-controlled trials of amantadine and rimantadine were analysed separately, and trials that assessed 'clinical influenza' were analysed separately from those that assessed 'serologically confirmed clinical influenza'. The rationale for these last two splits are neither stated in the protocol nor explained explicitly in the review.

The placebo-controlled trials of the two drugs were considered separately as the direct comparison of the two drugs revealed strong evidence of differences in adverse effect profiles. The protocol did not anticipate the existence of direct comparisons between rimantadine and amantadine, but, given that they were located in the search, the scope of the review was expanded to include these trials. Given that these comparisons revealed that the two drugs differ significantly on one set of outcomes there was a strong argument that they should not be combined for other outcomes. This example clearly illustrates the difficulties in pre-specifying all analyses, and the importance of ensuring that the protocol does not become 'a straight-jacket which prevents the review from exploring useful and unexpected issues'.[9]

In contrast, the separation of the two definitions of influenza is an example of a separation so obvious to those working in the field that they did not consider explaining it. In practice, many of the trials contained counts of both outcomes and contributed to both analyses.

The reviewers followed their planned analysis strategy by performing fixed effect meta-analyses of relative risks when there was little evidence of heterogeneity (i.e. when $p > 0.05$) and random effects analyses otherwise (the only protocol/review to follow this path; see Figure 2). A beneficial effect of amantadine, and an effect of similar magnitude for rimantadine (though not achieving statistical significance perhaps due to a rather smaller number of studies), were found by the reviewers. Statistically significant heterogeneity was identified in many meta-analyses, including the primary outcome for the prevention studies ($p < 0.01$). In compiling our Table, we have considered amantadine and rimantadine studies together, pooling both active treatment groups in the trials that investigated both of them.

In the selected meta-analysis, trials were subdivided into those in vaccinated and non-vaccinated populations, a subgroup analysis not mentioned in the protocol (Table, Figure 1) and again without a stated rationale in the review. The reviewers had not anticipated this source of clinical diversity, but thought it important enough to display stratified analyses, even though there appeared to be no relationship between vaccination status and effectiveness. The main motivation was to confirm the wider applicability of the findings.

The reviewers also report that they 'considered carrying out sub-analysis by dose . . ., but decided against this given the small size of the resulting meta-analysis'. They also note among their conclusions that 'investigation of the reasons for the heterogeneity seen between the trials may be of some value'. These analyses and comments were, in fact, included in response to concerns about the residual heterogeneity raised by peer reviewers, whose reports listed a series of additional investigations of heterogeneity. All of these were impossible to undertake as the data required were not stated in the trial reports. The reviewers were also resistant to proposals to undertake additional post hoc investigations.

Discussing the residual heterogeneity, the reviewers report that 'such a pattern has been noted in other reviews in preventive procedures, such as influenza and cholera vaccination, and may reflect differences between the trial populations to natural exposure and immunity to influenza A and similar viruses'. Such a mechanism is supported by the observation that the heterogeneity was less in the meta-analyses of 'serologically confirmed influenza' where the impact of other influenza-like viruses would be reduced. However, such differences should also manifest themselves in variable control group event rates across the trials. Whereas this variability was present, heterogeneity would only be explained if there is a relationship between control group event rates and the relative risk of developing influenza on treatment relative to control. Using methods described by Thompson et al,[30] we have found no clear evidence that this is the case.

significance of the test for heterogeneity, at $p < 0.10$ and $p < 0.05$, respectively. No reviews discussed problems associated with random effects analyses of small numbers of studies (in which case it is difficult to estimate precisely the between-trial variation), though one stated: 'Results of comparable groups of trials were to be pooled using fixed and random effects models. However given the very few studies involved, only the fixed effects model was applied here.'

Numbers of studies

Planned investigations of heterogeneity can be severely limited by the numbers of studies identified. Figure 3 illustrates the numbers of studies addressed by the 39 reviews. Twenty (51%) of the reviews had addressed several questions in a single review by including more than one treatment comparison. The studies not relevant to our primary treatment comparison (or

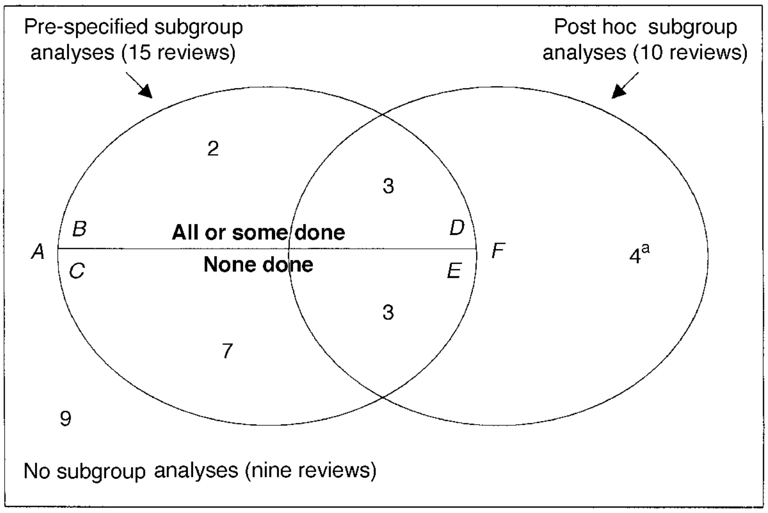
Pre-specified subgroup analyses (15 reviews)
Post hoc subgroup analyses (10 reviews)
2
3
B
All or some done
D
A
F
4a
C
None done
E
3
7
9
No subgroup analyses (nine reviews)

Figure 1 The use of subgroup analyses among 28 Cochrane reviews with published protocols. Pre-specified subgroups (B + C + D + E) refer to subgroup analyses (or meta-regressions) listed in the protocol, which may or may not have been done in the review. When such analyses had been pre-specified, we divide reviews into those that reported at least one of them ('all or some done': B + D) and those that reported none of them ('none done': C + E). Post hoc subgroup analyses (D + E + F) are those performed in the review that were not listed in the protocol. Within each area the number of reviews falling into that category is given; a letter indicates the code used for that category in the Table. [a]Two of these four stated in the review that the subgroup analyses were pre-planned, but they were not mentioned in the protocols we looked at.

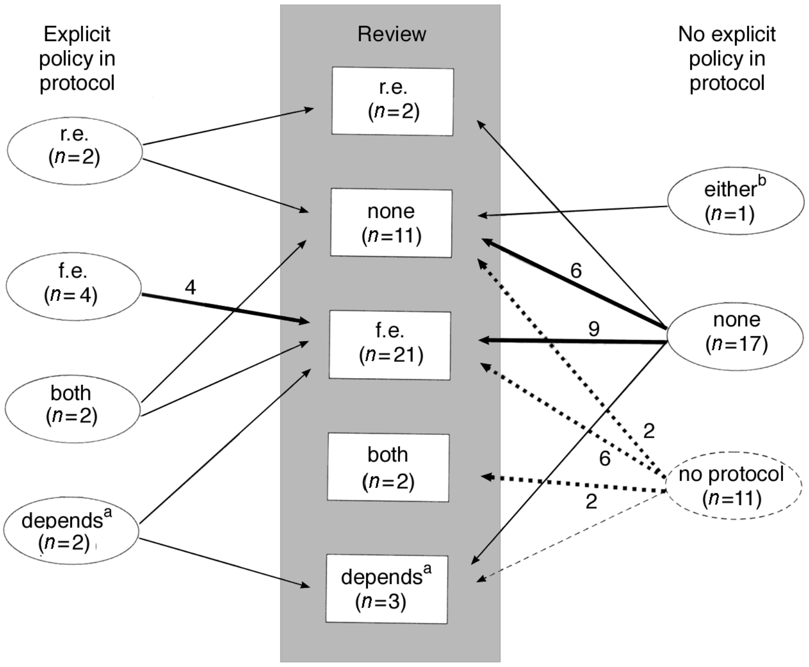

Figure 2 The planning and use of fixed effect (f.e.) and random effects (r.e.) meta-analyses among 39 Cochrane reviews. Analyses planned in the protocols are given in ovals; those actually performed (or stated as preferred) in the reviews appear in rectangles. Narrow lines indicate that a single study followed this route; numbers of studies following routes with bold lines are given. Dotted lines refer to the reviews without protocols. Protocols described as 'none' made no mention of the planned or preferred analysis; reviews described as 'none' either did no meta-analyses or reported no text that indicated a preference. [a]r.e. if significant heterogeneity; f.e. otherwise. [b]This protocol stated 'either fixed or random effects'.

which did not collect data for our primary outcome) are depicted by the upper portion of the bars in Figure 3. For several reviews, studies relevant to the primary comparison were not reported to have collected data for our chosen primary outcome. The middle portions of the bars in Figure 3 illustrate the numbers of studies for which primary outcome data should have been available but were not. Among all 39 reviews, primary outcome data were provided for only 71% of the 285 eligible studies. For only 13 reviews (33%) were data available from every study for the review's primary outcome, and in 13 (33%) fewer than half of the studies provided data. The lower portion of the bar gives actual numbers of studies available for meta-analysis of our chosen outcome for our chosen treatment comparison. Only eight (21%) of the reviews had ten or more such studies and only one had more than 15 studies.

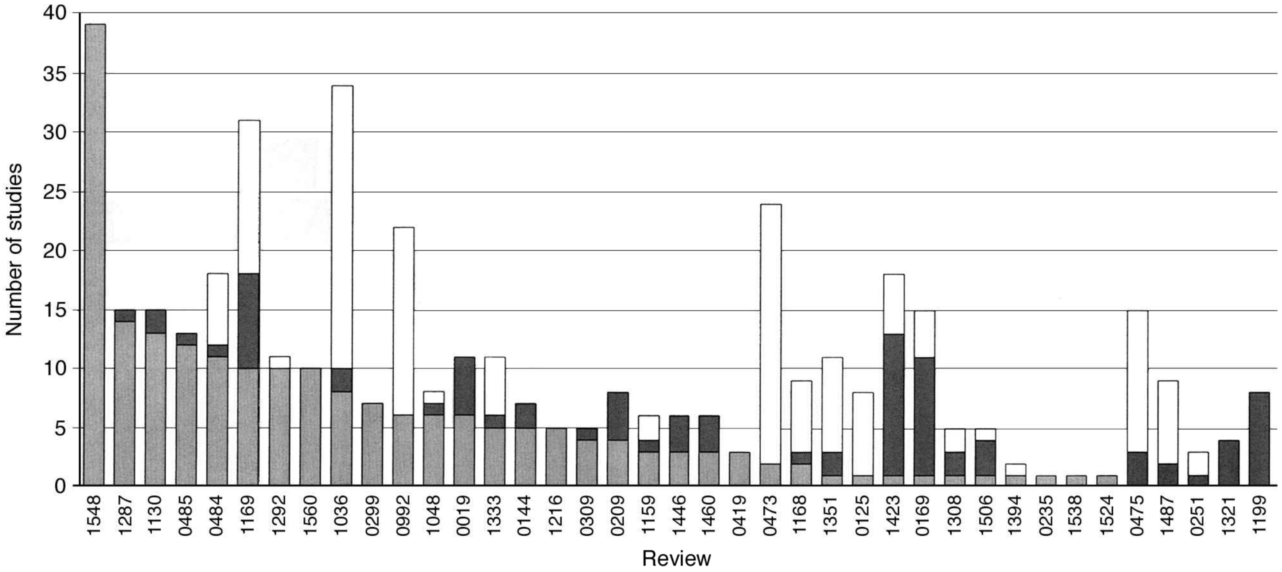

Figure 3 Numbers of studies involved in the 39 Cochrane reviews. The height of the bar gives the number of studies included in the review. The upper portion (no shading) gives the number of studies that either were not relevant to the primary treatment comparison we looked at or did not collect data for the primary outcome we looked at. The middle portion (dark shading) gives the number of studies for which outcome data had been collected but were not presented by the reviewers (so might be considered 'missing'). The lower portion (light shading) gives the number of studies contributing data to the primary meta-analysis.

Discussion

We have adopted a novel approach to empirical research in this paper by comparing published protocols with corresponding completed systematic reviews available in the CDSR.[19] We have thereby found that plans for addressing heterogeneity between studies are seldom followed. This would suggest either that protocols have been inappropriately developed, or that they are poor at predicting sensible options when faced with the studies subsequently identified for review. Published guidelines were generally consistent, and areas of uncertainty or disagreement were those in which knowledge is deficient and further empirical or methodological research is desirable. Here we discuss the practical difficulties in addressing heterogeneity that arose from our investigation of CDSR reviews and relate them to the available guidelines. We remark on areas that would benefit from further research.

It is clear that pre-specification of effect modifiers for use in investigations of heterogeneity is 'easier said than done'. Much emphasis was placed on pre-specification in most of the guidelines, including all the collaborative review groups' specific advice resulting from the survey. Yet few reviews studied only those effect modifiers they had pre-specified, indicating the practical limitations of such a procedure. Of additional concern is the process through which pre-specified effect modifiers are themselves determined, including what, precisely, is meant by 'a priori'? There may in fact be little difference between pre-specifying effect modifiers based on limited knowledge of the studies and selecting post hoc effect modifiers based on direct knowledge of the results. A further problem associated with specifying potential effect modifiers for subgroup analysis is that they may be inappropriate, either because they are suited to investigation at the patient level rather than the study level, or because there is confusion between important prognostic factors and potential effect modifiers.[20] The importance of having a defensible scientific rationale for determining effect modifiers is clear, and indeed a welcome contribution of the recent QUOROM statement[14] (recommendations for reporting meta-analyses of randomised trials) is that 'a rationale for any a priori ... subgroup analyses' be presented. As an extreme example of lack of such a rationale, one review of studies with extreme statistical heterogeneity presented a subgroup with 'less heterogeneity' ($0.05 < p < 0.10$) consisting of studies *not* conducted in a particular country (and a highly heterogeneous subgroup of studies conducted *in* that country).

A degree of caution in conducting post hoc investigations of heterogeneity was advocated widely in the guidelines. Whether introducing post hoc subgroup analyses resulted in spurious findings, and whether interpretations of subgroup differences were correct, can only be speculated upon for the reviews in our sample. Such issues could be properly resolved only by identifying or performing new large-scale randomised trials. It is clear from theoretical arguments that multiple subgroup analyses using a conventional significance level will tend to yield false-positive (spuriously significant) results. However, we are not in a position to judge whether deviations from pre-specified analyses or interpretations of subgroup analyses could be justified on a case-by-base basis. The case study offers some insights for a particular review, in which subgroupings of trials not specified in advance were presented: for reasons obvious to investigators in the field; due to unanticipated differences in adverse effect profiles; and to affirm wider applicability of the findings. The case study also raises the problem of pressure from peer reviewers to incorporate additional subgroup analyses. Different researchers would approach a systematic review with different a priori beliefs and therefore with different lists of pre-specified potential effect modifiers.

Failure either to state effect modifiers in advance or to present a scientific rationale for their investigation does not necessarily imply absence of a sound approach. We acknowledge a limitation of this kind of research in that our interpretations are based solely on what was reported. The case study (Box 2) illustrates the tension between comprehensive reporting of methodology and producing a concise and practically useful summary for the reader. A previous study has resulted in one collaborative review group ensuring that Cochrane protocols include a scientific rationale for specified effect modifiers and that the subgroup analyses in subsequent reviews are consistent with the protocols.[19]

A further problem in conducting subgroup analyses is the number of potential effect modifiers that can be investigated. The likelihood of apparently convincing but spurious findings increases with the number of effect modifiers analysed. None of the guidelines indicated how many studies are required for a reliable investigation of differences between them, and the only comments regarding numbers of effect modifiers stated that the number of subgroup analyses should be 'kept to a minimum'. A few reviews decided, probably wisely, not to investigate heterogeneity on the grounds that there were too few studies. In general, we found that numbers of studies eligible for meta-analysis were small, and that the situation was made more difficult by appropriate data not being available for a large proportion of otherwise suitable studies. This obstacle applies both to outcome measures and to information required to divide studies into subgroups.

Reviewers apparently rely heavily on a statistical test for diagnosing heterogeneity. It is known that the test suffers from low power when there are few studies in a meta-analysis, as in the vast majority of meta-analyses in our study. A consequence of this is that reviewers may fail to detect genuine heterogeneity which, if fully accounted for in the analysis, could impact on the conclusions of the meta-analysis. While it has been suggested that a higher significance level of 10% be used to compensate for this,[21] this was not widely recommended or practised. Indeed, the guideline document that was shared by seven collaborative review groups explicitly states $p = 0.05$ as indicating statistical significance. The choice between fixed and random effects analyses was sometimes determined by the identification of statistically significant heterogeneity. None of the comprehensive guidelines presented an argument that random effects analyses might be the preferred option irrespective of statistical significance of the test, and this was reflected in there being only two reviews to choose this approach. The document used by several collaborative review groups recommended that both fixed and random effects analyses be performed, a strategy that can lead to problems of interpretation. One set of collaborative review group-specific guidelines stated that 'the clinical heterogeneity of the back pain literature suggests that the assumptions underlying the random effects model are better suited'.[15] Most often overall, however, fixed effect meta-analysis was preferred in practice. This is despite the strong and, some would say, untenable assumption of equal underlying treatment effects on which this method is based.[5,22]

Our study highlights areas that would benefit from further research. First, the relationship between the number of studies, the number of potential effect modifiers and the likelihood of spurious findings is unclear, both for performing subgroup analyses and for meta-regression. Second, some qualitative research on the notion of what a priori means in the context of a systematic review may expose limitations to the important process of pre-specifying effect modifiers which has the explicit intention of preventing spurious conclusions. Third, the debate between fixed effect and random effects meta-analysis methods is ongoing; our findings here contribute the information that fixed effect analyses appear to be the most frequent option in Cochrane reviews irrespective of study diversity or heterogeneity. Fourth, the validity of using the statistical test for heterogeneity as a rule for deciding between the two models needs assessing, since in other contexts such a two-step approach to statistical analysis is unwise.[23] Fifth, developing means of improving the availability of existing, but unpublished, data would enable more extensive investigations of heterogeneity. Finally, a cause of heterogeneity that we have not addressed in this paper is the choice of treatment effect measure. Research focusing on a suitable choice of treatment effect for dichotomous outcomes indicates that risk differences are likely to be more heterogeneous than risk ratios or odds ratios.[24] However, it remains unclear whether the observed extent of heterogeneity in a particular review should determine the effect measure chosen.

In this paper, we have considered only reviews in the CDSR. These may differ from systematic reviews published in peer-reviewed journals in terms of topic area, numbers of studies and indeed the handling of heterogeneity.[25] Cochrane reviewers are provided with software for conducting fixed effect and random effects meta-analyses but not for investigating heterogeneity through techniques such as meta-regression. These methods may therefore be under-represented in Cochrane reviews. However, not only are reviews in journals subject to the selective influences of the publications process, but they also rarely have published protocols; so in these it would be impossible to separate intention from practice in terms of handling heterogeneity. Accessible software for undertaking meta-regression has only recently become available, and we might anticipate increasing use of this technique in the future, together with the attendant risks of spurious findings resulting from 'data dredging'. One promising development is the increasing number of systematic reviews based on collating individual patient data. These allow the investigation of whether patient characteristics (rather than simply study characteristics or study-level summaries of patient characteristics) are potential effect modifiers. Analyses based on study-level data, such as meta-regressions and analyses of subgroups of studies, are more prone to bias and confounding[26] and suffer from a loss of power compared with analyses of individual patient data.[27]

In conclusion, we reaffirm that heterogeneity between studies in a systematic review is, and may always be, a difficult issue, philosophically, statistically and practically. In the majority of reviews of randomised trials, such as many of those undertaken by the Cochrane Collaboration, insufficient data will preclude reliable statistical investigations of heterogeneity. Whereas pre-specification of scientifically justified effect modifiers has the theoretical potential for providing meaningful findings, in practice it may not yield characteristics that explain the heterogeneity, and the temptation to view post hoc observations as important is ever-present. We suggest that in many cases it is not clear that sources of heterogeneity should be investigated[2] unless a large number of studies is available.

Acknowledgements

We are grateful to the Cochrane Review Group Co-ordinators for responding to our survey, and to an editor and two reviewers for their comments.

References

1. Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean? BMJ 1997; 315: 1610–1614
2. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994; 309: 1351–1355
3. The Cochrane Collaboration. Cochrane database of systematic reviews. The Cochrane Library (Issue 2). Oxford: Update Software, 2001
4. Bailey KR. Inter-study differences – how should they influence the interpretation and analysis of results. Statistics in Medicine 1987; 6: 351–360
5. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies. American Journal of Epidemiology 1995; 142: 371–382
6. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine 1999; 18: 2693–2708
7. Mulrow CD, Oxman AD, eds. Cochrane collaboration handbook [updated September 1997]. The Cochrane Library. Oxford: Update Software, 1997
8. Clarke M, Oxman AD, eds. Cochrane reviewers' handbook 4.0 [updated July 1999]. The Cochrane Library. Oxford: Update Software, 2000
9. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Report 4. York: Centre for Reviews and Dissemination, 1996
10. National Health and Medical Research Council (Australia). How to review the evidence: systematic identification and review of the scientific literature. Canberra: National Health and Medical Research Council, 2000
11. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. Journal of Clinical Epidemiology 1995; 48: 167–171
12. Sutton AJ, Jones DR, Abrams KR, Sheldon TA, Song F. Systematic reviews and meta-analysis. a structured review of the methodological literature. Journal of Health Services Research & Policy 1999; 4: 49–55
13. National Research Council. Combining information: statistical issues and opportunities for research. Washington, DC: National Academy Press, 1992
14. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised trials: the QUOROM statement. Lancet 1999; 354: 1896–1900
15. Van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders. Spine 1997; 22: 2323–2330
16. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Annals of Internal Medicine 1992; 116: 78–84
17. Chalmers I, Hetherington J, Elbourne D, Keirse MJNC, Enkin M. Materials and methods used in synthesizing evidence to evaluate the effects of care during pregnancy and childbirth. In: Chalmers I, Enkin M, Keirse MJNC, eds. Effective care in pregnancy and childbirth. Oxford: Oxford University Press, 1989: 39–65
18. Simes J, Clarke M, Ganz P, Glasziou P, Henderson IC, Liberati A et al. Cochrane Breast Cancer Group [Review Group Module]. The Cochrane Library (Issue 2). Oxford: Update Software, 1999
19. Hahn S, Garner P, Williamson P. Are systematic reviews taking heterogeneity into account? An analysis from the Infectious Diseases Module of the Cochrane Library. Journal of Evaluation in Clinical Practice 2000; 6: 231–233
20. Deeks JJ. Systematic reviews of published evidence: miracles or minefields? Annals of Oncology 1998; 9: 703–709
21. Fleiss JL. Analysis of data from multiclinic trials. Controlled Clinical Trials 1986; 7: 267–275
22. Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). Annual Review of Public Health 1996; 17: 1–23
23. Senn S. Cross-over trials in clinical research. Chichester: Wiley, 1993
24. Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ, 2001: 313–335
25. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of COCHRANE reviews with articles published in paper-based journals. JAMA 1998; 280: 278–280
26. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine (in press)
27. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of Clinical Epidemiology (in press)
28. 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: 408–412
29. Jefferson TO, DeMicheli V, Deeks JJ, Rivetti D. Amantadine and rimantadine for preventing and treating influenza A in adults (Cochrane Review). The Cochrane Library (Issue 2). Oxford: Update Software, 1999
30. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Statistics in Medicine 1997; 16: 2741–2758