

Randomisation to protect against selection bias in healthcare trials (Review)

Kunz R, Vist GE, Oxman AD



**THE COCHRANE
COLLABORATION®**

This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2008, Issue 2

<http://www.thecochranelibrary.com>



TABLE OF CONTENTS

HEADER	1
ABSTRACT	1
PLAIN LANGUAGE SUMMARY	2
BACKGROUND	2
OBJECTIVES	3
METHODS	3
RESULTS	4
DISCUSSION	9
AUTHORS' CONCLUSIONS	10
ACKNOWLEDGEMENTS	10
REFERENCES	11
CHARACTERISTICS OF STUDIES	15
DATA AND ANALYSES	17
WHAT'S NEW	30
HISTORY	30
CONTRIBUTIONS OF AUTHORS	30
DECLARATIONS OF INTEREST	30
SOURCES OF SUPPORT	31
INDEX TERMS	31

[Methodology Review]

Randomisation to protect against selection bias in healthcare trials

Regina Kunz¹, Gunn Elisabeth Vist², Andrew D Oxman²

¹Basler Institute for Clinical Epidemiology , Siegburg, Germany. ²Department of Evidence-Based Health Services, Norwegian Knowledge Centre for Health Services, Oslo, Norway

Contact address: Regina Kunz, Basler Institute for Clinical Epidemiology , Gemeinsamer Bundesausschuss, Auf dem Seidenberg 3A, Siegburg, 53707, Germany. RKunz@uhbs.ch .

Editorial group: Cochrane Methodology Review Group.

Publication status and date: Edited (no change to conclusions), published in Issue 2, 2008.

Review content assessed as up-to-date: 19 February 2007.

Citation: Kunz R, Vist GE, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art. No.: MR000012. DOI: 10.1002/14651858.MR000012.pub2.

Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

ABSTRACT

Background

Randomised trials use the play of chance to assign participants to comparison groups. The unpredictability of the process, if not subverted, should prevent systematic differences between comparison groups (selection bias), provided that a sufficient number of people are randomised.

Objectives

To assess the effects of randomisation and concealment of allocation on the results of healthcare trials.

Search strategy

We searched the Cochrane Methodology Register, MEDLINE, SciSearch, reference lists up to August 2000 and used personal communication.

Selection criteria

Cohorts of trials, systematic reviews or meta-analyses of healthcare interventions that compared outcomes or prognostic factors for one of the following comparisons: randomised versus non-randomised trials, randomised trials with adequately versus inadequately concealed allocation, or high versus low quality trials where selection bias could not be separated from other sources of bias.

Data collection and analysis

One of us went through all of the citations in the Cochrane Methodology Register and accumulated reference lists. Studies that appeared to meet the inclusion criteria were retrieved and assessed independently by two of the reviewers. The methodological quality of included studies was appraised and information extracted by one of us and checked by a second. Tabular summaries of the results were prepared for each comparison and the results across studies were assessed qualitatively to identify common trends or discrepancies.

Main results

We identified 32 studies including over 3000 trials. Twenty-two studies compared randomised versus non-randomised trials, three compared adequately versus inadequately concealed allocation, and nine compared high versus low quality trials (some studies included more than one comparison). Five studies were of high methodological quality.

In 15 of the 22 studies that compared randomised and non-randomised trials of the same intervention, important differences were found in the estimates of effect. Some of these differences were due to a poorer prognosis in the control groups in the non-randomised trials. The results of the other seven studies that compared randomised and non-randomised trials across different interventions are less clear.

Comparisons of adequately and inadequately concealed allocation in randomised trials of the same intervention provided high quality evidence that concealment can be crucial in achieving similar treatment groups and, therefore, unbiased estimates of treatment effects. Studies with inadequate concealment tended to overestimate treatment effects.

Comparisons of high and low quality trials of the same intervention have found important differences in estimates of effect, but it is not possible to determine the extent to which these differences can be attributed to randomisation or concealment of allocation.

Omitting comparisons between randomised trials and non-randomised trials using historical controls did not substantially alter the results or conclusions of our review.

Authors' conclusions

On average, non-randomised trials and randomised trials with inadequate concealment of allocation tend to result in larger estimates of effect than randomised trials with adequately concealed allocation. However, it is not generally possible to predict the magnitude, or even the direction, of possible selection biases and consequent distortions of treatment effects.

PLAIN LANGUAGE SUMMARY

Randomised controlled trials as safeguard against over- or underestimation of treatment effects

Randomised controlled trials (RCTs) use the play of chance to assign participants to comparison groups to prevent selection bias. Other means of treatment allocation are more prone to bias because decisions can be related to prognosis and responsiveness to treatment. This review compares randomised versus non-randomised trials and adequately versus inadequately concealed allocation. On average not using concealed random allocation results in larger estimates of effect, but it can also result in smaller estimates or even reversal of the direction of effect (from harmful to beneficial or vice versa). It is a paradox that the unpredictability of random allocation is the best protection against the unpredictability of the extent to which non-randomised trials may be biased.

BACKGROUND

Discoveries of dramatically effective health care interventions, like epinephrine for anaphylaxis, are not common. The majority of health care interventions are at best moderately superior to conventional care or a placebo. Some interventions that are believed to be beneficial are, in fact, no more effective than a placebo and some are even harmful. Well-intentioned clinicians have, for example, treated stroke by applying leeches to the anus (Gubler 1971), treated neurosyphilis by injecting malarial parasites (Austin 1992), treated angina with internal mammary artery ligation (Valenstein 1998), treated symptomatic atherosclerotic disease of the internal carotid artery with extracranial-intracranial bypass surgery (EC/IC Bypass 1985), and treated asymptomatic ventricular arrhythmia after myocardial infarction with antiarrhythmic drugs (

Echt 1991). It is estimated that tens of thousands of patients died prematurely from widespread use of class I antiarrhythmic drugs alone (Moore 1995), which caused one death for every 20 patients who were treated (Teo 1993). Failure to adequately evaluate interventions has also delayed the use of effective interventions, such as magnesium sulphate instead of diazepam or phenytoin for the treatment of eclampsia (Eclampisia 1995).

As stated by Archie Cochrane: "Observational evidence is clearly better than opinion, but it is thoroughly unsatisfactory. All research on the effectiveness of therapy was in this unfortunate state until the early 1950s. The only exceptions were the drugs whose effect on immediate mortality were so obvious that no randomised trials were necessary, such as insulin, sulphonamide, and penicillin." (

[Cochrane 1972](#)) Cochrane, along with many others, credits Austin Bradford Hill with bringing an experimental approach into clinical medicine. The 1948 report of the randomised trial of streptomycin for pulmonary tuberculosis by Hill and his colleagues is widely recognised as a landmark study in this regard ([MRC 1948](#)).

“The basic idea, like most good things, is very simple.” ([Cochrane 1972](#)) The primary reason for random assignment is to remove the potential of bias in the assignment of people to one intervention or another; i.e. to protect against any possible systematic connection between the treatment that people receive and their prognosis. In addition to producing comparable groups of treatment and control patients, which other means of allocation such as alternation can also do, randomisation introduces unpredictability. When alternation or any other pre-set plan (such as time of admission) is used, a clinician that wants a patient to receive a particular treatment can learn the plan and can then arrange to enter a patient into the study at an opportune moment. With randomisation, however, each patient’s treatment is assigned according to the play of chance. This unpredictability, unless subverted by clinicians who peek at an inadequately concealed randomisation schedule, should prevent systematic differences in the prognosis of the groups of patients that are being compared, provided that a sufficient number of patients are randomised. Although it is possible to control for differences between comparison groups in other ways, such as statistical adjustment, this is only possible for factors that are known and measured. Randomisation is the only means of controlling for unknown and unmeasured differences as well as those that are known and measured.

Despite this simple logic and many anecdotal examples of harm being done because of delays in conducting randomised trials, there are limitations to the use of randomised trials, both real and imagined, and scepticism about the importance of randomisation ([US Office HTA 1994](#); [Black 1996](#); [Weiss 1998](#); [Pocock 2000](#)). We believe this scepticism is healthy. It is important to question assumptions about research methods, and to test these assumptions empirically, just as it is important to test assumptions about the effects of health care. Methodological hubris can be just as dangerous as medical hubris. Empirical comparisons of randomised versus non-randomised evaluations of the effects of healthcare represent important steps away from hubris. This review of such comparisons has been updated from a previously published review ([Kunz 1998](#)) (see What’s new). It differs from other similar reviews ([Reeves 1998](#); [McKee 1999](#)) in the questions that are addressed and the methods that were used, but there is not a major disagreement in the conclusions of these reviews ([Kunz 1999](#); [Britton 1999](#)).

OBJECTIVES

To assess the effects of randomisation and allocation concealment on the results of healthcare trials.

METHODS

Criteria for considering studies for this review

Types of studies

Cohorts of trials, systematic reviews or meta-analyses that compared random allocation or concealment of allocation (the process used to prevent foreknowledge of group assignment in a randomised trial) to non-random allocation or non-concealed allocation. Single case studies or simulation studies were excluded.

Types of data

Our “study population” was healthcare trials, including trials of clinical interventions (“clinical trials”) and non-clinical interventions where the effects of the intervention on one or more health outcomes were measured. This includes randomised trials (“randomised controlled trials” or “RCTs”), non-randomised trials with concurrent controls (in which a non-random method of allocation, such as alternation, was used to assign participants to the comparison groups; frequently called “quasi-randomised trials”, “concurrently controlled trials” or “CCTs”), and non-randomised trials using historical controls (people treated earlier than those who received the intervention that is being evaluated, frequently called “historically controlled trials” or “HCTs”). Classical observational studies, including cohort studies, case-control studies and “outcomes studies” (evaluations using large administrative or clinical databases) were not included, unless they were incorporated in studies that also included a comparison between randomised and non-randomised trials.

Types of methods

1) Randomised versus non-randomised trials of the same intervention, 2) randomised versus non-randomised trials across different interventions, 3) randomised trials with adequately versus inadequately concealed allocation, and 4) high versus low quality trials where the effect of randomisation or allocation concealment could not be separated from the effects of other methodological manoeuvres. An explicit definition for high and low quality by the authors was required and random allocation or allocation concealment had to be included as a criterion.

Types of outcome measures

The magnitude and direction of estimates of effect (e.g. relative risk reductions, odds ratios, standardised effect sizes) and imbalances in prognostic factors.

Search methods for identification of studies

Studies were identified using the Cochrane Methodology Register, bibliographies, MEDLINE, SciSearch, hand-searching, personal communication with methodologists and the reference lists of relevant articles up to August 2000. Exploratory hand-searching of methodological journals (Controlled Clinical Trials, Statistics in Medicine, the Journal of Clinical Epidemiology) for four volumes (1970, 1980, 1990 and 1995) was not productive. Repeated efforts have been undertaken to develop an efficient electronic search strategy using MEDLINE since 1994. Early attempts were not efficient due to poor indexing of methodological studies. Since 1999 MEDLINE searches have been more successful, particularly by searching for "Related Articles" in PubMed using seven key articles ([Chalmers 1977](#); [Colditz 1989](#); [Emerson 1990](#); [Kunz 1998](#), [Ottenbacher 1992](#); [Sacks 1982](#); [Schulz 1995](#)). This was supplemented with a search strategy using the following combinations of MeSH-terms:

- [Random Allocation OR Randomised Controlled Trial (exp)] AND Bias (epidemiology)
- [Random Allocation OR Randomised Controlled Trial (exp)] AND research /cl,mt,sn,st,rd
 - [Random Allocation OR Randomised Controlled Trial (exp)] /cl,mt,sn,st,td,ut AND Double Blind Method
 - [Random Allocation OR Randomised Controlled Trial (exp)] /cl,mt,sn,st,td,ut AND Clinical Trials /cl,mt,sn,st,td,ut
 - Randomised Controlled Trial (exp) AND Selection Bias
 - Randomised Controlled Trial (exp) AND Follow-Up Studies
 - Randomised Controlled Trial /mt,sn,st,ut AND Follow-Up Studies

SciSearch was searched for articles that cited the following articles: [Chalmers 1977](#); [Colditz 1989](#); [Emerson 1990](#); [Kunz 1998](#); [Miller 1989](#); [Ottenbacher 1992](#); [Sacks 1982](#); [Schulz 1995](#).

A large proportion of studies were assembled through personal contacts with methodologists and from bibliographies and reference lists.

Data collection and analysis

Potentially relevant articles were retrieved and assessed for inclusion independently by RK and either GEV or ADO. Disagreements were resolved by discussion.

The following criteria were used to appraise the methodological quality of included studies:

- Were explicit criteria used to select the trials?
- Did two or more investigators agree regarding the selection of trials?
- Was there a consecutive or complete sample of trials?
- Did the study control for other methodological differences such as double blinding and complete follow-up?

- Did the study control for clinical differences in the participants and interventions in the included trials?
- Were similar outcome measures used in the included trials?

The overall quality of each study was summarised as: no important flaws, possibly important flaws, or major flaws.

For each study one of us (RK) extracted information about the sample of trials, the comparisons that were made, the type of analysis and the results. One of the other reviewers checked the extracted data against the published article. The reported relationship between randomisation and estimates of effect was recorded and, if possible, converted to the relative over- or underestimation of the relative risk reduction using the results of randomised trials, randomised trials with concealed allocation and randomised high quality trials as the reference. Tables were prepared for each type of comparison to facilitate a qualitative analysis of the extent to which the included studies yielded similar results. Heterogeneity in the included studies was explored both within and across comparisons.

Two sensitivity analyses were performed in this update: one about the contribution of historical controls on effect size in non-randomised trials. We thereby performed the analysis excluding historical controls from the non-randomised trials. In a second sensitivity analysis we have excluded studies with major methodological flaws.

RESULTS

Description of studies

See: [Characteristics of excluded studies](#).

See Results and Tables 01 - 04 under Comparisons and Data. We found 32 studies that met our inclusion criteria with a total of more than 3000 healthcare trials. There were 15 studies comparing randomised and non-randomised trials of the same intervention, seven studies of randomised versus non-randomised trials across different interventions, three studies of randomised trials with adequately compared with inadequately concealed allocation, and nine studies of high compared with low quality trials. Some studies included more than one comparison; therefore the total number of comparisons exceeds the number of studies. ([Aronson 1996](#); [Benson 2000](#); [Bhansali 1996](#); [Carroll 1996](#); [Chalmers 1977](#); [Chalmers 1983](#); [Colditz 1989](#); [Diehl 1986](#); [Emerson 1990](#); [Forgie 1998](#); [Guyatt 2000](#); [Imperiale 1990](#); [Khan 1996](#); [Lipsey 1993](#); [Miller 1989](#); [Moher 1998](#); [Mullen 1997](#); [Nurmohamed 1992](#); [Ortiz 1998](#); [Ottenbacher 1991](#); [Ottenbacher 1992](#); [Ottenbacher 1993](#); [Potter 1998](#); [Pyorala 1995](#); [Reimold 1992](#); [RMIT Group 1994](#); [Sacks 1982](#); [Schulz 1995](#); [Shadish 1996](#); [Stanton 1997](#); [Watson 1994](#); [Wortman 1983](#)).

Risk of bias in included studies

See Additional Tables 01. The methodological quality of the included studies varied. Five studies met all of our criteria (Imperiale 1990; Khan 1996; Nurmohamed 1992; Schulz 1995; Shadish 1996). Two of these investigated the association between different methodological features and estimates of effect (Schulz 1995; Shadish 1996). The other three were meta-analyses that assessed the association between methodological quality and the magnitude of effect for a specific healthcare intervention (Imperiale 1990; Khan 1996; Nurmohamed 1992). The other 27 studies had one or more methodological flaws, including not controlling for other methodological manoeuvres (Aronson 1996; Bhansali 1996; Chalmers 1977; Chalmers 1983; Diehl 1986; Forgie 1998; Guyatt 2000; Lipsey 1993; Mullen 1997; Ottenbacher 1991; Ottenbacher 1992; Ottenbacher 1993; Pyorala 1995; Sacks 1982; Wortman 1983) or clinical differences (Aronson 1996; Carroll 1996; Chalmers 1977; Colditz 1989; Emerson 1990; Forgie 1998; Lipsey 1993; Miller 1989; Ortiz 1998; Ottenbacher 1991; Ottenbacher 1992; Ottenbacher 1993; Potter 1998; Pyorala 1995; Stanton 1997; Wortman 1983).

Effect of methods

Randomised versus non-randomised trials of the same intervention

Fifteen studies including 35 comparisons of randomised and non-randomised trials of the same intervention are summarised in Results Table 1. In 22 comparisons, estimates of effect were larger in non-randomised trials. Outcomes in the randomised and non-randomised treatment groups were frequently similar, but worse outcomes among historical controls spuriously increased the estimated treatment effects. Eight comparisons found similar results for randomised and non-randomised trials, and four comparisons found smaller treatment effects in non-randomised trials. In one comparison the smaller estimate of effect was explained by a poorer prognosis for patients in the non-randomised treatment groups. The deviation of the effect estimates for non-randomised compared with randomised trials ranged from a 76% smaller to a 400% larger effect. One comparison found reverse effects: randomised trials indicated a harmful effect while non-randomised trials using historical controls suggested a beneficial effect.

Table 1. Critical appraisal of the methodological quality

Study	Sample	Confounding control?	Reproducibility	Outcome measure	Overall judgement
	Consecutive or complete sample?	1. Other methodological criteria: 2. Participants + clinical interventions	1. Explicit description 2. Assessment of agreement		
Aronson 1996	Yes	1 No 2 No	1 Yes 2 No	Same outcome measure	Possibly important - major flaw
Benson 2000	Yes	1 Partly 2 Partly	1 No 2 No	Same outcome measure	Possibly important flaw
Bhansali 1996	Yes	1 No 2 Partly	1 Yes 2 No	Same outcome measure	Possibly important flaw
Carroll 1996	Yes	1 Partly 2 No	1 partly 2 No	p-value or vote counting	Possibly important flaw
Chalmers 1977	Yes	1 No 2 No	1 Yes 2 No	Same outcome measure	Major flaw => overestimation
Chalmers 1983	Yes	1 No 2 Partly	1 Yes 2 No	Same outcome measure	Possibly important flaw => overestimation

Table 1. Critical appraisal of the methodological quality (Continued)

Colditz 1989	Yes	1 Partly 2 No	1 Yes 2 partly	Standardized outcome measure	Possibly important - major flaw =>underestimation
Diehl 1986	No	1 No 2 Yes	1 Yes 2 No	Same outcome measure	Possibly important flaw -- direction of bias unclear
Emerson 1990	No	1 Yes 2 No	1 Yes 2 partly	Same outcome measure	Possibly important flaw => underestimation
Forgie 1998	Yes	1 No 2 No	1 Yes 2 Yes	Same outcome measure	Possibly important flaw
Guyatt 2000	Yes	1 No 2 Partly	1 Yes 2 No	Same outcome measure	Possibly important flaw
Imperiale 1990	Yes	1 Yes 2 Yes	1 Yes 2 Yes	Same outcome measure	No important flaw
Khan 1996	Yes	1 Yes 2 Yes	1 Yes 2 partly	Same outcome measure	No important flaw
Lipsey 1993	Yes	1 No 2 No	1 No 2 No	Standardized outcome measure	Major flaws
Miller 1989	Yes	1 Partly 2 No	1 Yes 2 partly	Standardized outcome measure	Possibly important - major flaw =>underestimation
Moher 1998	No	1 Yes 2 No	1 Yes 2 No	Standardized outcome measure	No - Possibly important flaw
Mullen 1997	Yes	1 No 2 No	1 Yes 2 Yes	Standardized outcome measure	Possibly important flaw
Nurmohamed 1992	Yes	1 Yes 2 Yes	1 Yes 2 partly	Same outcome measure	No important flaw
Ortiz 1998	Yes	1 Yes 2 No	1 Yes 2 Yes	Same outcome measure	No - possibly important flaw
Ottenbacher 1991	No	1 No 2 No	1 Yes 2 Yes	Standardized outcome measure	Major flaw
Ottenbacher 1992	No	1 No 2 No	1 Yes 2 Yes	Standardized outcome measure	Possibly important - major flaw =>any direction

Table 1. Critical appraisal of the methodological quality (Continued)

Ottenbacher 1993	Yes	1 Partly 2 No	1 Yes 2 Yes	Standardized outcome measure	Possibly important flaw
Potter 1998	Yes	1 Partly 2 No	1 Yes 2 No	Same outcome measure	Possibly important - major flaw
Pyorala 1995	Yes	1 No 2 No	1 partly 2 No	Same outcome measure	Possibly important - major flaw
Reimold 1992	Yes	1 Partly 2 Partly	1 Yes 2 No	Same outcome measure	Possibly important flaw
RMIT Group 1994	Yes	1 partly 2 Yes	1 Yes 2 partly	Same outcome measure	Possibly important flaw
Sacks 1982	No	1 No 2 Partly	1 Yes 2 No	p-value or vote counting	Major flaw => over-estimation
Schulz 1995	Yes	1 Yes 2 Yes	1 Yes 2 partly	Same outcome measure	No important flaw
Shadish 1996	Yes	1 Yes 2 Partly	1 Yes 2 Yes	Standardized outcome measure	No important flaw
Stanton 1997	Yes	1 Partly 2 No	1 No 2 No	Standardized outcome measure	No - possibly important flaw
Watson 1994	Yes	1 Partly 2 Partly	1 Yes 2 No	Same outcome measure	Possibly important flaw
Wortman 1983	Yes	1 No 2 No	1 No 2 No	Same + standardized outcome measure	Major flaws

Randomised versus non-randomised trials across different interventions

Seven studies compared randomised and non-randomised trials across different interventions (Results Table 02), four of which included a mixture of study designs such as crossover trials and classical observational studies, in addition to randomised and non-randomised trials. The results of these studies are less clear than the results of comparisons of trials of the same intervention. All seven studies made comparisons across different clinical conditions using standardised effect sizes. Three studies were inconclusive, and no

consistent relationship could be detected between study design or quality and effect size in the other four studies. The presence of substantial clinical heterogeneity and other factors could have distorted or masked a possible association between randomisation and estimates of effect in all of these comparisons.

Adequately versus inadequately concealed allocation within randomised trials

Three studies included comparisons between adequately and inadequately concealed allocation within randomised trials (Results

Table 03). All three detected larger effect sizes with inadequate concealment: Chalmers and colleagues ([Chalmers 1983](#)) found that failure to adequately conceal allocation was associated with larger imbalances in prognostic factors and larger treatment effects. They found treatment effects that were more than seven-fold larger in randomised trials with inadequate compared to adequate allocation concealment. However, lack of control for other methodological factors limits the validity of their descriptive analysis. Schulz and colleagues ([Schulz 1995](#)) conducted a multivariate analysis controlling for blinding and completeness of follow-up. They found a similar, but less dramatic association between inadequately concealed allocation and the magnitude of effect observed in randomised trials of the same intervention. In this study odds ratios were, on average, 40% larger in randomised trials with unclear/inadequately compared with adequately concealed allocation. Moher and colleagues ([Moher 1998](#)) found the same association in a similar study using a different sample of randomised trials. Inadequately concealed allocation resulted in 35% larger estimates of treatment effect in this study.

High versus low quality trials

Nine studies matched the criteria for this category (Results Table 04). All studies but one ([Lipsey 1993](#)) were restricted to randomised controlled trials. Lipsey had included various randomised and non-randomised comparative designs, not all of which had been explicitly described. Three methodological studies ([Emerson 1990](#); [Lipsey 1993](#); [Moher 1998](#)) including seven, 137 and 11 meta-analyses respectively compared estimates of effect between low and high quality trials. Two of these ([Emerson 1990](#); [Lipsey 1993](#)) did not find a consistent difference. The third ([Moher 1998](#)) found that lower quality trials had, on average, estimates of effect that were 34% larger than those from high quality trials. Six other meta-analyses, two of which included comparisons for two different outcomes, are included in this review. In four comparisons ([Nurmohamed 1992](#); [Stanton 1997](#); [Potter 1998](#)) the estimates of effect were larger in low quality trials (the study by Nurmohamed had two comparisons, surgical and orthopedic patients); in another ([Ortiz 1998](#)), lower quality trials reported a lower estimate of effect for side effects; and in two comparisons ([Khan 1996](#); [Potter 1998](#)) there was a reversal of effects with the low quality trials indicating a beneficial effect and the high quality trials indicating a harmful effect. In one meta-analysis ([Imperiale 1990](#)) the estimate of effect was smaller in low quality trials. Overall, in these analyses lower quality trials had estimates that were anywhere from 55% smaller to 350% larger than estimates of effect from high quality trials of the same intervention.

Sensitivity analyses

Exclusion of non-randomised trials using historical controls

Non-randomised trials using historical controls were only incorporated in comparison 1 and 2 and were consequently removed from the analysis to investigate their impact on the results. For the first comparison (randomised versus non-randomised trials of the same intervention) we excluded six studies: Four studies compared non-randomised trials using historical controls to randomised trials ([Sacks 1982](#); [Diehl 1986](#); [Aronson 1996](#); [Bhansali 1996](#)), while in two studies, results for comparisons with non-randomised concurrent trials could not be separated from comparisons with non-randomised trials using historical controls ([Carroll 1996](#); [Watson 1994](#)). In the remaining nine studies that compared non-randomised with randomised trials, larger effects for non-randomised trials were detected in 12 comparisons (52% versus 63% in all studies), similar results in seven comparisons (30% versus 23%), and smaller treatment effects for non-randomised trials in four comparisons (17% versus 11%). This difference was not significant ($p= 0.7$), and the results of these comparisons were quite similar with and without the inclusion of non-randomised trials using historical controls. Non-randomised trials using concurrent controls had results that ranged between 12% to 400% (median: 49%) larger and 20% to 76% smaller than in RCTs.

In comparison 2 (randomised versus non-randomised trials across different interventions), only the study by Miller ([Miller 1989](#)) incorporated non-randomised trials using historical controls, which was one of five non-randomised trial designs. Excluding trials using historical controls did not change the results of this study, which remained inconclusive.

Exclusion of studies with major methodological flaws

According to our explicit methodological assessment, eight of the included studies had no major flaws or were unlikely to have a major flaw ([Imperiale 1990](#); [Nurmohamed 1992](#); [Schulz 1995](#); [Khan 1996](#); [Shadish 1996](#); [Stanton 1997](#); [Moher 1998](#); [Ortiz 1998](#)).

Comparison 1 and 2 (Randomised versus non-randomised trials across the same or different interventions): Only the study by Shadish ([Shadish 1996](#)) which incorporated trials across different health care interventions achieved high methodological marks among the 22 studies. It found that, on average, non-randomised trials had effect sizes that were smaller than those from randomised trials.

Comparison 3 (Adequately versus inadequately concealed allocation within randomised trials: Two ([Schulz 1995](#); [Moher 1998](#)) of the three studies were of high quality and had consistent findings indicating that, on average, randomised trials with unclear / inadequately concealed allocation have larger estimates of effect than randomised trials with adequately concealed allocation.

Comparison 4 (High versus low quality trials): Six ([Imperiale 1990](#); [Nurmohamed 1992](#); [Khan 1996](#); [Stanton 1997](#); [Moher 1998](#); [Ortiz 1998](#)) of the nine studies were of high quality, all of which were restricted to randomised trials. One study ([Moher](#)

1998) that included 11 meta-analyses found, on average, larger treatment effects in trials of lower quality. The other five meta-analyses reported larger effects in low quality trials in three comparisons, smaller effects for side effects in one comparison, a reversal of effects (beneficial in low and harmful in high quality trials) in one comparison, and smaller effects in low quality trials in one comparison. These results are consistent with the overall results for all nine of the studies included in this comparison.

DISCUSSION

It has been difficult to develop efficient search strategies for locating empirical methodological studies such as the ones included in this review. We did it in parallel to the development of the Cochrane Database of Methodological Studies which was a cumbersome and fuzzy process and took place over several years. In the end it was impossible to reproduce the various search strategies and their results. However, we believe it is unlikely that there are many published methodological studies like those included in the first two comparisons, such as [Sacks 1982](#) and [Shadish 1996](#), those included in the third comparison (adequately versus inadequately concealed allocation) by [Chalmers 1983](#), [Schulz 1995](#), [Moher 1998](#), or methodological studies such as [Emerson 1990](#) included in the fourth comparison (high versus low quality trials). It is, however, possible that there are unpublished or ongoing studies that we have not identified. We are aware of at least two ongoing studies that will include an assessment of the relationship between concealment of allocation (and other methodological manoeuvres) and estimates of effect ([Als-Nielsen 2002](#), personal communication; [Devereaux 2002](#), personal communication) and a third study that will compare RCTs and non-RCTs of surgical interventions (<http://www.cochrane.dk/nrsmg/surgery/>, Accessed 31 May 2002).

It is possible that there is a publication bias or that we have identified a non-representative sample of published studies, given the inefficiency of the search strategies that we have used and a possible bias amongst the people we have contacted. However, two other systematic reviews using different search strategies and methods ([Reeves 1998](#); [McKee 1999](#)) did not identify any studies that we have not included and additional studies have not been identified in either published or personal correspondence following publication of an earlier version of this review.

The Cochrane Library (2002, Issue 1) contains 1297 completed Cochrane reviews, 1013 protocols, and 3299 entries in the Database of Abstracts of Reviews of Effectiveness. We have not systematically gone through all of these systematic reviews. Many of these are likely to include comparisons between high and low quality trials and between adequately and inadequately concealed allocation, including those listed under "References to studies awaiting assessment". Some may also contain comparisons of randomised

versus non-randomised trials. A more recent analysis ([Juni 1999](#)) demonstrating the limited ability of scores to distinguish reliably between high and low quality studies cautions against the use of scores. This finding has to be taken into account and should lead to a careful interpretation of the comparison "high versus low quality studies". Given the results of this analysis and of our review, it appears that further comparisons of high versus low quality studies are not likely to be informative and this comparison will not be included in future updates of this review for this reason.

We have not included comparisons between randomised trials and cohort studies ([Concato 2000](#)), case-control studies ([Colditz 1994](#); [Stieb 1990](#)) or "outcomes studies" (evaluations of effectiveness using large administrative or clinical databases) ([US Office HTA 1994](#)), although some of the studies in this review included observational studies as well as non-randomised trials. Observational studies often provide valuable information that is complementary to the results of trials. For example, case-control studies may be the best available study design for evaluating rare adverse effects, and large database studies may provide important information about the extent to which effects that are expected based on randomised trials are achieved in routine practice. However, as noted in the background section, it is only possible to control for confounders that are known and measured in observational studies. We should be wary of hubris and its consequences in assuming that we know all there is to know about any disease.

As with any review, the quality of the data is limited by the quality of the studies that we have reviewed. Most of the studies included in the review have one or more methodological flaws. Five of the studies met all of our criteria for assessing methodological quality ([Imperiale 1990](#); [Khan 1996](#); [Nurmohamed 1992](#); [Schulz 1995](#); [Shadish 1996](#)). However, three of those ([Imperiale 1990](#); [Khan 1996](#); [Nurmohamed 1992](#)) compared high versus low quality studies and their results have to be interpreted with caution. One study, in particular, provides strong support for the conclusion that clinical trials that lack adequately concealed allocation produce estimates of effect that are, on average, 40% larger than trials with adequately concealed allocation ([Schulz 1995](#)). Moher ([Moher 1998](#)) has replicated those findings in a study using similar methods. The studies by Schulz ([Schulz 1995](#)) and Moher ([Moher 1998](#)) also demonstrate the potential contribution that systematic reviews, and particularly the Cochrane Database of Systematic Reviews, can make towards developing an empirical basis for methodological decisions in evaluations of healthcare. Currently this empirical basis is lacking for many methodological decisions, and many methodological debates rely more on logic or rhetoric than evidence. Analyses such as the one undertaken by [Schulz 1995](#) and [Moher 1998](#), in which methodological comparisons are made among trials of the same intervention are likely to yield more reliable results than comparisons that are made across different interventions ([Colditz 1989](#), [Miller 1989](#), [Ottenbacher 1992](#)) which have a great deal of clinical and methodological heterogeneity and

thus not surprisingly, tend to have inconclusive results.

We have used randomised trials, randomised trials with adequately controlled concealment of allocation and high quality randomised trials as the reference in the comparisons we have made. Implicit in this is an assumption that differences in results are best explained by bias and the reference randomised trials are less likely to be biased. This assumption is, to some extent, supported by findings of larger imbalances in prognostic factors in non-randomised trials. However, it is possible that randomised trials sometimes underestimate the effects of an intervention in routine practice by forcing health care professionals and patients to acknowledge their uncertainty and, thereby, reduce the strength of placebo effects ([Black 1996](#); [Chalmers 1997](#); [Kleijnen 1997](#)). It is also possible that publication bias can partly explain some of the differences in results observed in studies such as the one by Sacks and colleagues ([Sacks 1982](#)). This would be the case if randomised trials were more likely to be published regardless of the effect size than non-randomised trials with historical controls. We are not aware of any evidence that supports this hypothesis and the available evidence shows consistently that randomised trials, like other research, are also more likely to be published if they have “significant” results ([Dickersin 1993](#); [Dickersin 1997](#); [Stern 1997](#); [Ioannidis 1998](#)).

There are a number of other possible explanations for discrepancies between estimates of effect derived from randomised and non-randomised trials. For example, it can be argued that estimates of effect might be larger in randomised trials if the care provided in the context of randomised trials is better than that in routine practice, assuming this is the case for the treatment group and not the control group. Similarly, strict eligibility criteria might select people with a higher capacity to benefit from a treatment, resulting in larger estimates of effect in randomised than non-randomised trials with less strict eligibility criteria. If patients with a poorer prognosis were more likely to be allocated to the treatment group in non-randomised trials for some reason, this would also result in larger estimates of effect in randomised trials. Conversely, if patients with a poorer prognosis are more likely to be allocated to the control group in non-randomised trials, as often appears to be the case based on the results of this review, this would result in larger estimates of effect in non-randomised trials.

A U T H O R S ’ C O N C L U S I O N S

Implication for systematic reviews and evaluations of healthcare

Overall, this review supports the logical arguments for using random allocation and ensuring that randomisation schedules are concealed in healthcare trials. The effect of not using random allocation with adequate concealment can be as large or larger than the expected effects of interventions. On average non-randomised trials and randomised trials with inadequately concealed allocation result in overestimates of effect. However, this bias can go in

either direction, can reverse the direction of effect, or can mask an effect. It is a paradox that the introduction of unpredictability by using concealed random allocation in clinical trials is the best protection we have against the unpredictability of the extent to which the results of non-randomised trials may be biased.

For those undertaking trials this review provides support for using randomisation to assemble comparison groups ([Chalmers 1997](#)). For those undertaking systematic reviews of trials, the review provides support for considering sensitivity analyses based on the adequacy of allocation concealment in addition to or instead of based on overall quality scores, which may be misleading ([Juni 1999](#); [Juni 2001](#)).

As Cochrane stated: “The RCT is a very beautiful technique, of wide applicability, but as with everything else there are snags.” ([Cochrane 1972](#)) Those making decisions on the basis of randomised trials need to be cautious of small trials, even when they are properly randomised, and systematic reviews of small randomised trials, both because of chance effects and the risk of biased reporting ([Counsell 1994](#); [Egger 1997](#)). It is also, of course, possible to introduce bias into a randomised trial despite adequate allocation concealment ([Guyatt 2002](#); [Schulz 1995](#)). Finally, even when the risk of error due to either bias or chance is small, judgements must be made about the applicability of the results to individual patients ([Dans 2002](#)) and about the relative value of the probable benefits, harms and costs.

I M P L I C A T I O N F O R M E T H O D O L O G I C A L R E S E A R C H

Additional well designed studies comparing randomised and non-randomised trials, in particular, and adequately and inadequately concealed allocation in randomised trials of the same intervention could help strengthen inferences about the importance of randomisation and allocation concealment or potentially modify the above inferences. Further comparisons across different interventions and comparisons of trials based on quality scales are of questionable value ([Juni 1999](#); [Juni 2001](#)). A methodology review of comparisons between randomised trials and observational studies, including cohort studies, case-control studies and “outcomes research” (evaluations using large administrative or clinical databases) is needed. To adequately investigate the role of varying baseline risk, heterogeneity or study quality, individual trial analysis might be required.

A C K N O W L E D G E M E N T S

We are grateful to Alex Jadad, Steve Halpern and David Cowan for help locating studies; to Dave Sackett and Iain Chalmers for encouragement and advice; to Mike Clarke for reviewing the manuscript; to Kay Dickersin, Annie Britton and other colleagues who have generously provided us with their bibliographies on research methodology; and to the investigators who have conducted

the studies we have reviewed. We are also grateful to the referees and editors who have commented on earlier drafts of this review and helped to improve it. However, we cannot hold them responsible for our errors. We thank the Norwegian Directorate for Health and Social Welfare and the Humboldt-University / Land Berlin, Germany for support.

REFERENCES

References to studies included in this review

Aronson 1996 {published data only}

Aronson R, Offman HJ, Joffe RT, Naylor CD. Triiodothyronine augmentation in the treatment of refractory depression. A meta-analysis. *Arch Gen Psychiatry* 1996;53(9):842–8. [MEDLINE: 5]

Benson 2000 {published data only}

Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–86.

Bhansali 1996 {published data only}

Bhansali MS, Vaidya JS, Bhatt RG, Patil PK, Badwe RA, Desai PB. Chemotherapy for carcinoma of the esophagus: a comparison of evidence from meta-analyses of randomized trials and of historical control studies. *Ann Oncol* 1996;7(4):355–9. [MEDLINE: 16]

Carroll 1996 {published data only}

Carroll D, Tramer M, McQuay H, Nye B, Moore A. Randomization is important in studies with pain outcomes: systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. *Br J Anaesth*. *Br J Anaesth* 1996;77: 798–803.

Chalmers 1977 {published data only}

Chalmers TC, Matta RJ, Smith H Jr, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1997;297:1091–6.

Chalmers 1983 {published data only}

Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med*. *N Engl J Med* 1983;309:1358–61.

Colditz 1989 {published data only}

Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med* 1989;8: 441–54.

Diehl 1986 {published data only}

Diehl LF, Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid?. *J Clin Oncol* 1986;4:1114–20.

Emerson 1990 {published data only}

Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials* 1990;11:339–52.

Forgie 1998 {published data only}

* Forgie MA, Wells PS, Laupacis A, Ferguson D. Preoperative autologous donation decreases allogeneic transfusion but increases

exposure to all red blood cell transfusion: results of a meta-analysis. International Study of Perioperative Transfusion (ISPOT) Investigators. *Arch Intern Med* 1998;158(6):610–6. [MEDLINE: 322]

Guyatt 2000 {published data only}

Guyatt GH, DiCenso A, Farewell V, Willan A, Griffith L. Randomized trials versus observational studies in adolescent pregnancy prevention. *J Clin Epidemiol* 2000;53(2):167–74. [MEDLINE: 323]

Imperiale 1990 {published data only}

Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta analysis of the randomized trials. *Ann Intern Med*. *Ann Intern Med* 1990;113:299–307.

Khan 1996 {published data only}

Khan KS, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med*. *Arch Intern Med* 1996;156:661–6.

Lipsey 1993 {published data only}

Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *Am Psychol* 1993;48(12):1181–209. [MEDLINE: 312]

Miller 1989 {published data only}

Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II. *Surgical Stat Med* 1989;8: 455–66.

Moher 1998 {published data only}

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? [see comments]. *Lancet* 1998;352(9128):609–13. [MEDLINE: 92]

Mullen 1997 {published data only}

Mullen PD, Simons-Morton DG, Ramirez G, Frankowski RF, Green LW, Mains DA. A meta-analysis of trials evaluating patient education and counseling for three groups of preventive health behaviors. *Patient Educ Couns* 1997;32(3):157–73. [MEDLINE: 313]

Nurmohamed 1992 {published data only}

Nurmohamed MT, Rosendaal FR, Buller HR, Dekker E, Hommes DW, Vandebroucke JP, Briet E. Low molecular weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *Lancet* 1992;340:152–6.

Ortiz 1998 {published data only}

Ortiz Z, Shea B, Suarez Almazor ME, Moher D, Wells GA, Tugwell P. The efficacy of folic acid and folinic acid in reducing

- methotrexate gastrointestinal toxicity in rheumatoid arthritis. A meta-analysis of randomized controlled trials. *J Rheumatol* 1998;25:36–43.
- Ottenbacher 1991 {published data only}**
Ottenbacher KJ. Epistemology and experimentation: an examination of quality factors in research design. *Am J Occup Ther* 1991;45(10):917–23. [MEDLINE: 96]
- Ottenbacher 1992 {published data only}**
Ottenbacher K. Impact of random assignment on study outcome: an empirical examination. *Control Clin Trials* 1992;13:50–61.
- Ottenbacher 1993 {published data only}**
Ottenbacher KJ, Jannell S. The results of clinical trials in stroke rehabilitation research [see comments]. *Arch Neurol* 1993;50(1):37–44. [MEDLINE: 97]
- Potter 1998 {published data only}**
Potter J, Langhorne P, Roberts M. Routine protein energy supplementation in adults: systematic review. *BMJ* 1998;317:495–501.
- Pyorala 1995 {published data only}**
Pyorala S, Huttunen NP, Uhari M. A review and meta-analysis of hormonal treatment of cryptorchidism. *J Clin Endocrinol Metab* 1995;80:2795–9.
- Reimold 1992 {published data only}**
Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug related mortality. *Am Heart J* 1992;124:924–32.
- RMIT Group 1994 {published data only}**
Recurrent Miscarriage Immunotherapy Trialists Group. Worldwide collaborative observational study and meta analysis on allogenic leukocyte immunotherapy for recurrent spontaneous abortion. *Am J Reprod Immunol* 1994;32:55–72.
- Sacks 1982 {published data only}**
Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233–40.
- Schulz 1995 {published data only}**
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.
- Shadish 1996 {published data only}**
Shadish WR, Ragsdale K. Random versus nonrandom assignment in controlled experiments: do you get the same answer? *J Consult Clin Psychol* 1996;64(6):1290–305. [MEDLINE: 319]
- Stanton 1997 {published data only}**
Stanton MD, Shadish WR. Outcome, attrition and family-couples treatment for drug abuse. *Psychological Bulletin* 1997;122:170–91.
- Watson 1994 {published data only}**
Watson A, Vandekerckhove P, Lilford R, Vail A, Brosens I, Hughes E. A meta-analysis of the therapeutic role of oil soluble contrast media at hysterosalpingography: a surprising result? *Fertil Steril* 1994;61:470–7.
- Wortman 1983 {published data only}**
Wortman P, Yeaton WH. Synthesis of results in trials of coronary artery bypass graft surgery. In: Light R editor(s). *Evaluation Studies Review Annual*. Vol. 8, Beverley Hills, CA: Sage Publications, 1983:536–57.
- ### References to studies excluded from this review
- Colditz 1994 {published data only}**
Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, Mosteller F. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta analysis of the published literature. *JAMA* 1994;271:698–702.
- Concato 2000 {published data only}**
Concato J, Shah N, Horwitz RI. Randomized Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *N Engl J Med* 2000;342:1887–92.
- Gilbert 1977 {published data only}**
Gilbert JP, McPeek B, Mosteller F. Statistics and ethics in surgery and anesthesia. *Science* 1977;198(4318):684–689.
- Hovell 1982 {published data only}**
Hovell MF. The experimental evidence for weight-loss treatment of essential hypertension: a critical review. *Am J Public Health* 1982;72(4):359–368.
- Hutchinson 1999 {published data only}**
Hutchinson BG, Oxman AD, Shannon HS, Lloyd S, Altmayer CA, Thomas K. Clinical effectiveness of pneumococcal vaccine. Meta-analysis. *Can Fam Physician* 1999;45:2381–93.
- Kerlikowske 1995 {published data only}**
Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernstner VL. Efficacy of screening mammography. A meta-analysis. *JAMA* 1995;273(2):149–154.
- Koes 1994 {published data only}**
Koes BW, van Tulder MW, van der Windt WM, Bouter LM. The efficacy of back schools: a review of randomized clinical trials. *J Clin Epidemiol* 1994;47(8):851–862.
- Kownacki 1996 {published data only}**
Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;78(4):497–510.
- MacArthur 1995 {published data only}**
MacArthur C, Foran PJ, Bailar JC 3rd. Qualitative assessment of studies included in a meta-analysis: DES and the risk of pregnancy loss. *J Clin Epidemiol* 1995;48(6):739–747.
- Mehta 1999 {published data only}**
Mehta RH, Bates ER. Coronary stent implantation in acute myocardial infarction. *Am Heart J* 1999;137:603–611.
- Rozenberg 1999 {published data only}**
Rozenberg S, Dubourg G, Khalifa P, Paolozzi L, Maheu E, Ravaud P. Efficacy of epidural steroids in low back pain and sciatica. A critical appraisal by a French Task Force of randomized trials. Critical Analysis Group of the French Society for Rheumatology. *Rev Rhum Engl Ed* 1999;66(2):79–85.
- Stieb 1990 {published data only}**
Stieb D, Frayha HH, Oxman AD, Shannon HS, Hutchinson BG, Crombie F. The effectiveness and usefulness of Haemophilus influenzae Type b Vaccines: a systematic overview (meta-analysis). *Can Med Assoc J* 1990;142:719–32.

Weisburd 2001 {published data only}

Weisburd D, Lum CM, Petrosino A. Does research design affect study outcomes in criminal justice. *Annals of the American Academy of Political and Social Science* 2001;578:50–70.

References to studies awaiting assessment**Abrutyn 1991 {published data only}**

Abrutyn E, Berlin JA. Intrathecal therapy in tetanus. A meta-analysis. *JAMA* 1991;266:2262–7.

Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;78:497–510.

Black 1996 (P) {published data only}

Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;78:497–510.

Brosseau 2002 {published data only}

Brosseau L, Welch V, Wells G, deBie R, Gam A, Harman K, Morin M, Shea B, Tugwell P. Low level laser therapy (Classes I, II and III) for treating rheumatoid arthritis. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD002049.pub2]

Callahan 1991 {published data only}

Callahan CM, Dittus RS, Katz BP. Oral corticosteroid therapy for patients with stable chronic obstructive pulmonary disease. A meta-analysis. *Ann Intern Med* 1991;114:216–23.

Camma 1996 {published data only}

Camma C, Almasio P, Craxi A. Interferon as treatment for acute hepatitis C. A meta-analysis. *Dig Dis Sci* 1996;41:1248–55.

Chinoy 1999 {published data only}

Chinoy MA, Parker MJ. Fixed nail plates versus sliding hip systems for the treatment of trochanteric femoral fractures: a meta-analysis of 14 studies. *Injury* 1999;30:157–63.

Cook 1998 {published data only}

Cook RL, Rosenberg MJ. Do spermicides containing nonoxynol-9 prevent sexually transmitted infections? A meta-analysis. *Sex Transm Dis* 1998;25:144–50.

Flor 1992 {published data only}

Flor H, Fydrich T, Turk DC. Efficacy of multidisciplinary pain treatment centers: a meta-analytic review. *Pain* 1992;49:221–30.

Fouque 1992 {published data only}

Fouque D, Laville M, Boissel JP, Chifflet R, Labeeuw M, Zech PY. Controlled low protein diets in chronic renal insufficiency: meta-analysis. *BMJ* 1992;304:216–20.

Gifford 1995 {published data only}

Gifford DS, Morton SC, Fiske M, Kahn K. A meta-analysis of infant outcomes after breech delivery. *Obstet Gynecol* 1995;85:1047–54.

Glasziou 1993 {published data only}

Glasziou PP, Mackerras DE. Vitamin A supplementation in infectious diseases: a meta-analysis. *BMJ* 1993;306:366–70.

Grullon 1997 {published data only}

Grullon KE, Grimes DA. The safety of early postpartum discharge: a review and critique. *Obstet. Gynecol* 1997;90:860–5.

Hart 2000 {published data only}

Hart RG, Halperin JL, McBride R, Benavente O, Man-Son-Hing M, Kronmal RA. Aspirin for the primary prevention of stroke and other major vascular events: meta-analysis and hypotheses. *Arch Neurol* 2000;57:326–32.

Heinsman 1996 {published data only}

Heinsman DT, Shadish WR. Assignment methods in experimentation: When do non-randomised experiments approximate answers from randomised experiments?. *Psychological Methods* 1996;1:154–69.

Hommes 1992 {published data only}

Hommes DW, Bura A, Mazzolai L, Buller HR, ten Cate JW. Subcutaneous heparin compared with continuous intravenous heparin administration in the initial treatment of deep vein thrombosis. A meta-analysis. *Ann Intern Med* 1992;116:279–84.

Imperiale 1991 {published data only}

Imperiale TF, Petrulis AS. A meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease. *JAMA* 1991;266:260–4.

Insua 1994 {published data only}

Insua JT, Sacks HS, Lau TS, Lau J, Reitman D, Pagano D, et al. Drug treatment of hypertension in the elderly: a meta-analysis. *Ann Intern Med* 1994;121:355–62.

Ioannidis 2001 {published data only}

Ioannidis JP, Salem D, Chew PW, Lau J. Accuracy and clinical effect of out-of-hospital electrocardiography in the diagnosis of acute cardiac ischemia: a meta-analysis. *Ann Emerg Med* 2001;37:461–70.

Kasiske 1993 {published data only}

Kasiske BL, Heim-Duthoy K, Ma JZ. Elective cyclosporine withdrawal after renal transplantation. A meta-analysis. *JAMA* 1993;269:395–400.

Kasiske 1998 {published data only}

Kasiske BL, Lakatua JD, Ma JZ, Louis TA. A meta-analysis of the effects of dietary protein restriction on the rate of decline in renal function. *Am J Kidney Dis* 1998;31:954–61.

McAlister 1998 {published data only}

McAlister FA, Clark HD, Wells PS, Laupacis A. Perioperative allogeneic blood transfusion does not cause adverse sequelae in patients with cancer: a meta-analysis of unconfounded studies. *Br Surg* 1998;85:171–8.

Patel 2002 {published data only}

Patel M, Lee CK. Polysaccharide vaccines for preventing serogroup A meningococcal meningitis. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD001093.pub2]

Psaty 1997 {published data only}

Psaty BM, Smith NL, Siscovick DS, Koepsell TD, Weiss NS, Heckbert SR, et al. Health outcomes associated with antihypertensive therapies used as first-line agents. A systematic review and meta-analysis. *JAMA* 1997;277:739–45.

Ramsey 2000 {published data only}

Ramsey MJ, DerSimonian R, Holtel MR, Burgess LP. Corticosteroid treatment for idiopathic facial nerve palsy: a meta-analysis. *Laryngoscope* 2000;110:335–41.

Shadish 1992 {published data only}

Shadish WR. Do family and marital psychotherapies change what people do? A meta-analysis of behavioural outcomes. In: Cook TD, Cooper H, Cordray DS, Hartmann H, Hedges LV, Light RJ, et al. editor(s). *Meta-analysis for Exploration: A Casebook*. New York: Russel Sage Foundation, 1992:129–208.

Shadish 1997 {published data only}

Shadish WR, Heinsman DT. Experiments versus quasi-experiments: do they yield the same answer?. In: Bukoski WJ editor(s). *Meta-analysis of Drug Abuse Prevention Programs*. Washington DC, 1997:147–64.

Shadish 2001 {published data only}

Shadish WR, Glaser RR. Differences between randomised and nonrandomised experiments using meta-analysis: a methodological note. 2001.

Srisurapanont 2002 {published data only}

Srisurapanont M, Jarusuraisin N. Opioid antagonists for alcohol dependence. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD001867.pub2]

Tangkanakul 2002 {published data only}

Tangkanakul C, Counsell C, Warlow C. Local versus general anaesthesia for carotid endarterectomy. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD000126.pub2]

van Nieuwenhoven 2001 {published data only}

van Nieuwenhoven CA, Buskens E, van Tiel FH, Bonten MJ. Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. *JAMA* 2001;286:335–40.

Vandekerckhove 2002 {published data only}

Vandekerckhove P, Watson A, Lilford R, Harada T, Hughes E. Oil-soluble versus water-soluble media for assessing tubal patency with hysterosalpingography or laparoscopy in subfertile women. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD000092]

Vandenbroucke 1991 {published data only}

Vandenbroucke-Grauls CM, Vandenbroucke JP. Effect of selective decontamination of the digestive tract on respiratory tract infections and mortality in the intensive care unit. *Lancet* 1991;338:859–62.

Watson 2002 {published data only}

Watson A, Vandekerckhove P, Lilford R. Techniques for pelvic surgery in subfertility. *Cochrane Database of Systematic Reviews* 2002, Issue 1. [DOI: 10.1002/14651858.CD000221.pub3]

Wells 1994 {published data only}

Wells PS, Lensing AW, Hirsh J. Graduated compression stockings in the prevention of postoperative venous thromboembolism. A meta-analysis. *Arch Intern Med* 1994;154:67–72.

References to ongoing studies

Als-Nielsen 2002 {published data only (unpublished sought but not used)}

Ongoing study Starting date of trial not provided. Contact author for more information.

Devereaux 2002 {published data only (unpublished sought but not used)}

Ongoing study Starting date of trial not provided. Contact author for more information.

Additional references

Austin 1992

Austin SC, Stolley PD, Lasky T. The history of malariotherapy for neurosyphilis. *JAMA* 1992;268:516–9.

Black 1996

Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215–8.

Britton 1999

Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Three systematic reviews - not so different answers. *BMJ* Electronic letters 29th September 1999.

Chalmers 1997

Chalmers I. Assembling comparison groups to assess the effects of health care. *J R Soc Med* 1997;90:379–86.

Cochrane 1972

Cochrane AL. *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust, 1972: 20–5.

Counsell 1994

Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?. *BMJ* 1994;309:1677–81.

Dans 2002

Dans A, McAlister F, Dans L, Richardson WS, Straus S, Guyatt G. Applying the results to individual patients.. In: Guyatt G, Rennie D editor(s). *Users' Guides to the Medical Literature*. Chicago: AMA Press, 2002:369–84.

Dickersin 1993

Dickersin K, Min YI. NIH clinical trials and publication bias. *Online J Curr Clin Trials* 1993;Doc No 50:4967. [MEDLINE: 332]

Dickersin 1997

Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997;9(1 Suppl):15–21. [MEDLINE: 333]

EC/IC Bypass 1985

The EC/IC Bypass Study Group. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an international randomized trial. *N Engl J Med* 1985;313:1191–200.

Echt 1991

Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781–8.

Eclampsia 1995

Eclampsia Trial Collaborative Group. Which anticonvulsant for women with eclampsia?. *Lancet* 1995;345:1455–63.

Egger 1997

Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34.

Gubler 1971

Gubler A. Alternating hemiplegia, a sign of pontine lesion, and documentation of the proof of the facial decussation. In: Wolf JK editor(s). *The Classical Brainstem Syndromes*. Springfield: Charles C Thomas, 1971:9–13.

Guyatt 2002

Guyatt GH, Cook D, Devereaux PJ, Meade M, Straus S. Therapy. *Users' guides to the medical literature*. Chicago: AMA Press, 2002: 55–79.

Ioannidis 1998

Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998;279(4):281–6. [MEDLINE: 59]

Juniper 1999

Juniper P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis [see comments]. *JAMA* 1999;282(11):1054–60. [MEDLINE: 1064]

Juniper 2001

Juniper P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Smith GD, Altman D editor(s). *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Books, 2001.

Kleijnen 1997

Kleijnen J, Gøtzsche P, Kunz RH, Oxman AD, Chalmers I. So what's so special about randomisation?. In: Maynard A, Chalmers I editor(s). *Non-random Reflections on Health Services Research: on the 25th Anniversary of Archie Cochrane's Effectiveness and Efficiency*. London: BMJ Books, 1997:93–106.

Kownacki 1999

Kownacki RJ, Shadish WR. Does Alcoholics Anonymous work? The results from a meta-analysis of controlled experiments. *Subst Use Misuse* 1999;34:1897–916.

Kunz 1999

Kunz R, Oxman AD. Two systematic reviews - two different answers?. *BMJ* Electronic letters 28th August 1999.

McKee 1999

McKee M, Britton A, Black N, Mcpherson K, SandersonC, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999;319:312–5.

Moore 1995

Moore TJ. *Deadly Medicine*. New York: Simon and Schuster, 1995.

MRC 1948

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948;ii:769–82.

Pocock 2000

Pocock SJ, Elbourne DR. Randomised trials or observational tribulations?. *N Engl J Med* 2000;324:1907–9.

Reeves 1998

Reeves BC, MacLennan RR Harvey IM, Sheldon TA, Russell IT, Black AMS. Comparison of effect size estimates derived from randomised and non-randomised studies. In: Black N, Brazier J, Fitzpatrick R, Reeves B editor(s). *Health Services Research Methods: A Guide to Best Practice*. London: BMJ Publishing Group, 1998: 73–85.

Stern 1997

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;315(7109):640–5. [MEDLINE: 334]

Teo 1993

Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. An overview of results from randomized trials. *JAMA* 1993;270: 1589–95.

US Office HTA 1994

US Congress, Office of Technology Assessment. Identifying Health technologies that Work: Searching for Evidence. *OTA-H-608*. Washington DC: US Government Printing Office, 1994:41–51.

Valenstein 1998

Valenstein ES. *Great and Desperate Cures: The Rise and Decline of Psychosurgery and Other Radical Treatments for Mental Illness*. Vol. 338, New York: Basic Books, 1986.

Weiss 1998

Weiss CH. *Evaluation. Methods for Studying Programs and Policies*. 2nd Edition. Upper Saddle River: Prentice Hall, 1998:229–33.

References to other published versions of this review**Kunz 1998**

Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317(7167):1185–90. [MEDLINE: 72]

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of excluded studies *[ordered by study ID]*

Colditz 1994	Comparison of randomised trials versus case-control studies
Concato 2000	Comparison of randomised trials versus cohort studies
Gilbert 1977	Narrative assessment only
Hovell 1982	No systematic review: no search strategy, no explicit inclusion or exclusion criteria; no explicit research question; narrative assessment
Hutchinson 1999	Comparison of randomised trials versus case-control studies
Kerlikowske 1995	Comparison of randomised trials versus case-control studies
Koes 1994	Descriptive assessment of individual studies, no summary comparison provided
Kownacki 1996	Comparison of RCT vs. non-RCTs on the effect of alcoholics anonymous. Coerced participation in group sessions in RCTs and voluntary participation in non-RCTs together with contradicting results raise high suspicion of confounder , as acknowledged by the authors.
MacArthur 1995	Narrative assessment only
Mehta 1999	Descriptive assessment of individual studies, no summary comparison provided
Rozenberg 1999	Narrative assessment only
Stieb 1990	Comparison of randomised trials versus case-control studies
Weisbord 2001	Non-healthcare interventions

DATA AND ANALYSES

Comparison 1. Studies of randomised compared with non-randomised trials of the same intervention.

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Undefined		Other data		No numeric data

Comparison 2. Studies of randomised compared with non-randomised trials across different interventions.

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Undefined		Other data		No numeric data

Comparison 3. Studies of trials with adequately compared with inadequately concealed allocation.

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Undefined		Other data		No numeric data

Comparison 4. Studies of high compared with low quality trials.

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1 Undefined		Other data		No numeric data

Analysis I.1. Comparison I Studies of randomised compared with non-randomised trials of the same intervention., Outcome I Undefined.

Undefined

Aronson 1996	4 RCTs and 4 HCTs of T3- therapy in euthyroid depressed patients refractory to tricyclic antidepressant therapy identified by systematic search	Comparison of RCTs vs. HCTs on response rate to tricyclic antidepressant drugs with or without triiodothyronine augmentation in the treatment of refractory depression	The response rate on T3-supplementation observed in HCTs was exaggerated by 76% compared to the response rate observed in RCTs.	Overestimation of the effect	
Benson 2000	Cohorts of studies comparing RCTs vs. non-RCTs on 3 issues: 1) Bone density after hormone replacement therapy (10 RCTs / 5 non-RCTs) 2) Ca-channel blocker after kidney transplantation (9 RCTs / 5 non-RCTs) 3) Wound infection after laparoscopic vs. open appendectomy: (16 RCTs / 8 non-RCTs), identified by a systematic search	Comparison of RCTs and non-RCTs on 1) %-change in bone density 2) graft survival 3) infection rates	Only graphical display of results was provided. There, similar effects in RCTs and non-RCTs were detected in all three comparisons.	Similar effects	
Bhansali 1996	12 RCTs and 8 HCTs on chemotherapy of oesophageal cancer identified by a systematic search	Comparison of RCTs and HCTs on patient survival at 1, 2 resp. 3 years	While HCTs detected a highly significant benefit of chemotherapy on survival by reducing the OR of death by 68%, no effect was observed by RCTs. This discrepancy was partly explained by better outcome in the treatment arm of HCTs compared to RCTs ($p<0.0001$)	Overestimation of the effect	

Undefined (*Continued*)

			and worse outcome in the control arm of HCTs compared to RCTs ($p<0.0001$)		
Carroll 1996	17 RCTs and 19 non-RCTs (including HCTs or trials with inadequate randomisation procedures) on transcutaneous electric nerve stimulation (TENS), identified by a systematic search	Comparison of RCTs and non-RCTs of postoperative control pain.	TENS was judged ineffective to improve postoperative pain in 85% of RCTs, while 89% of non-RCTs concluded that TENS improved postoperative pain.	Overestimation of the effect	
Chalmers 1977	32 controlled studies of anticoagulation in acute myocardial infarction identified by a systematic search	Comparison of RCTs with CCTs and HCTs on case-fatality rate, rate of thromboembolism and haemorrhages	The relative risk reduction for mortality was overestimated by 35% in HCTs and 6% in CCTs compared with RCTs. The case-fatality rate was highest in historical controls (38.3), compared with randomised controls (19.6) and concurrent controls (29.2). A similar pattern was found for thromboembolism.	Overestimation of the effect	
Diehl 1986	19 RCTs and 17 HCTs for 6 types of cancer (breast, colon, stomach, lung cancer, melanoma, soft tissue sarcoma), identified from the reference lists of two textbooks	Matching of randomised and historical controls for disease, stage, and follow up and comparison on survival and relapse free survival	18 of the 43 matched control groups (42%) varied by more than 10% (absolute difference in either outcome), 9 (21%) by more than 20%, and 2 (5%) by more than 30%. Survival or relapse free survival was better in randomised controls compared with	Overestimation of the effect	

Undefined (*Continued*)

			historical controls in 17/18 matches.		
Forgie 1998	6 RCTs and 9 non-RCTs including cohort studies identified by a systematic search	Com- parison of RCTs and non-RCTs on pre-operative autologous blood donation on exposure to allogeneic and reception of any blood transfusion	There was no difference in treatment effect in RCTs and non-RCTs for the outcome “exposure to allogeneic blood transfusion”, however, there was a fourfold increase in the outcome “reception of any blood transfusion” in non-RCTs.	Overestimation of the effect	
Guyatt 2000	13 RCTs and 17 non-RCTs, including cohort studies, on different interventions to prevent adolescent pregnancy identified by a systematic search	Com- parison of RCTs and non-RCTs / cohort studies on 1) initiation of intercourse; 2) pregnancy; 3) responsible sexual behaviour; 4) birth control use, using separate analyses for male and female teenagers	Among females, non-RCTs / cohort studies demonstrated a significant benefit for all 4 endpoints, ranging between 25% to 41% in change of odds ratio, which was not detected in RCTs. Among male teenagers, non-RCTs / cohort studies demonstrated a significant benefit for 2 / 4 endpoints (12%, resp. 29% in change of odds ratio), which was not detected in RCTs.	Overestimation of the effect	
Mullen 1997	52 RCTs and 22 non-RCTs on educational interventions for preventive health behaviour, identified by a systematic search	Comparison of the effect size of RCTs and non-RCTs in three different health areas: 1) smoking / alcohol; 2) nutrition / weight; 3) other health behaviour	Compared to RCTs, non-RCTs on “smoking / alcohol” or “other interventions” detected a nonsignificant trend of 25% resp. 20% reduction in benefit, but a significant 3,2-fold increase for “nutrition / weight”.	Trend for underestimation of effect in comparison 1 and 3, overestimation of the effect effect in comparison 2.	

Undefined (*Continued*)

Pyorala 1995	11 RCTs and 22 (not further specified) non-RCTs on hormonal therapy in cryptorchidism, identified by a systematic search	Comparison of RCTs and non-RCTs on the descent of testes after LHRH / hCG therapy	The success rate of descent of testes after LHRH therapy was 2.3 times larger in non-RCTs than in RCTs and 1.7 times larger after therapy with hCG.	Overestimation of the effect
Reimold 1992	6 RCTs and 6 CCTs of chinidine in atrial fibrillation, identified by a systematic search	Comparison of RCTs and CCTs on maintenance of sinus rhythm at 3, 6 and 12 months after cardioversion.	At 3 months, the beneficial effect of maintaining sinus rhythm with chinidine was 54% less in non-RCTs (compared to RCTs) and was 76% less at 12 months.	Underestimation of the effect
RMIT Group 1994	9 RCTs and 6 CCTs (with self-selected treatment) of allogenic leukocyte immunotherapy for recurrent miscarriage, identified by systematic search	Comparison of RCTs and CCTs on live birth rate	The beneficial effect of immunotherapy on birth rate among pregnant women was 9% larger in CCTs compared to RCTs, but was 63% lower in CCTs when all women were considered.	Underestimation of the effect when all women were considered, similar effect for pregnant women
Sacks 1982	Sample of 50 RCTs and 56 HCTs at hand, assessing 6 interventions (treatment of oesophageal varices, coronary artery surgery, anti-coagulation in myocardial infarction, chemotherapy for colon cancer and melanoma, and diethylenstilbestroel for recurrent miscarriage)	Comparison of RCTs with HCTs on the frequency of detecting statistically significant results ($p < 0.05$) of the primary outcome and reduction of mortality.	20% of the RCTs found a statistically significant benefit from the new therapy, compared with 79% of the HCTs. The RRR of mortality in HCTs vs RCTs (HCT/RCT) was 0.49/0.27 (1.8) for cirrhosis, 0.68/0.26 (2.6) for coronary artery surgery at 3 years, 0.49/0.22 (2.2) for anticoagulation in myocardial infarction, and 0.67/-	Overestimation of the effect

Undefined (*Continued*)

			0.02 for DES in recurrent miscarriage. A descriptive analysis for chemotherapy of melanoma and colon cancer confirmed the pattern, that larger effects were observed in HCTs compared to RCTs. Outcomes in the treatment groups were similar in both designs, but outcomes in the control groups were worse among historical controls.	
Watson 1994	4 RCTs and 6 CCTs /HCTs of oil-soluble contrast media during hysterosalpingography in infertile couples, identified by a systematic search	Comparison of RCTs and CCTs / HCTs on the rate of pregnancies	RCTs and CCTs / HCTs detected similar increases in pregnancy rates: ORRCT 1.92 (95% CI: 1.33-2.68)and ORCCT / HCT 1.92 (95% CI: 1.55-2.38).	Similar effect
Wortman 1983	9 RCTs and 16 concurrent quasi-experimental studies (QES) of medical and surgical therapy in coronary artery disease identified through a systematic search	Com-parison of RCTs and QES on severity of disease and mortality at 2 and 4 years.	The relative risk reduction of surgery on mortality was 36,7% in RCTs and 48,5% in QES at 2 years and 29% resp. 44,9% at 4 years. However, more patients with 3 vessel disease were included in QES, while more patients with less severe disease were included in RCTs. An analysis correcting for these differences had not been performed.	Overestimation of the effect

Analysis 2.1. Comparison 2 Studies of randomised compared with non-randomised trials across different interventions., Outcome I Undefined.

Undefined

Colditz 1989	113 studies published in 1980 comparing new interventions with old, identified in leading cardiology, neurology, psychiatry and respiratory journals by a systematic search	36 parallel RCTs, 29 randomized and 46 non-randomized COTs, 3 CCTs, 5 ECTs, 9 observational studies were compared on 'treatment gain' (Mann-Whitney statistic) and the relation between quality score and 'treatment gain' was assessed.	All but one design achieved similar 'treatment gains' (0.56-0.65). Overall 89% of new treatments were rated as improvements, but only non-randomized COTs detected a significant higher 'treatment gain' from the new treatment compared to RCTs ($p=0.004$). Within RCTs, there was no correlation between quality score and 'treatment gain' ($p=0.18$)	Inconclusive	
Lipsey 1993	137 / 302 meta-analyses on mental health interventions, counseling, psycho-educational or special therapy identified on a systematic search (165 meta-analyses on general non-medical educational issues were excluded).	After conversion of the results to mean treatment effect sizes, the following comparisons were included: overall effectiveness of psychological interventions; random vs. non-random treatment allocation; parallel-group vs. pre-post-comparison	Overall effectiveness of psychological interventions showed a mean effect size (MES) of $0.5 + 0.29$. Seventy four meta-analyses allowed further break-down of results according to random and non-random allocation. No difference in MES was detected (0.46 ± 0.28 , resp. 0.41 ± 0.36). However, in 20%, the difference of MES between RCTs and non-RCTs within an individual meta-analysis (MES (RCT) - MES(non-RCT)) was larger than $ 0.2 $ in both direction.	Inconclusive	

Undefined (*Continued*)

			One group pre-, post-design resulted in 62% larger treatment effect compared to parallel group design.		
Miller 1989	188 studies comparing new surgical interventions with old, published in 1983 and identified in leading surgical journals by a systematic search	81 RCTs, 15 CCTs, 27 HCTs, 91 observational studies, 7 BASs were compared on 'treatment gain' (Mann-Whitney statistic). The association between treatment success and study design and the relation between quality score and treatment gains were assessed.	There was a non-significant trend towards larger 'treatment gains' for new therapies on the principal disease in non-RCTs (0.56 - 0.78) compared with RCTs (0.56). For therapies on treatment complications the 'treatment gain' was similar across all study designs (0.54-0.55) except in BASs (0.90). Within RCTs, there was no correlation between quality scores and treatment gains (p=0.7).	Inconclusive	
Ottenbacher 1991	22 RCTs and 22 non-RCTs on the effectiveness of occupational therapy-interventions, identified by two occupational therapy journals.	Crude comparison of RCTs vs non-RCTs on the number of statistically significant results, and comparison of effect size estimates (d-index) weighted by sample size.	Non-RCTs had more often significant test results compared to RCTs (p<0.01). However, the d-index revealed similar effect sizes for both designs (non-RCTs: 0.36 +/- 0.18; RCTs: 0.35 +/- 0.14).	Inconclusive	
Ottenbacher 1992	Sample of 30 RCTs and 30 non-RCTs from a systematic search in NEJM and JAMA across a variety of medical spe-	RCTs were compared with non-RCTs on treatment effects as measured by standardised mean differ-	No difference in treatment effect was found between non-RCTs (0.23) and RCTs (0.21).	Similar effects	

Undefined (*Continued*)

	cialties	ences			
Ottenbacher 1993	36 clinical trials on the effectiveness of stroke rehabilitation programmes on functional outcome, identified by a systematic search	Compar-ison of mean effect size of all assessed outcome measures according to study design (RCTs, one group pre-post-design, quasi-experimental, cluster), randomization, matching and blinding of assessment.	The overall mean effect size in 36 trials was 0.40. Non-RCTs significantly inflated the treatment effect by 22% (pre-post design vs. RCT) resp. 24% (quasi-experimental allocation vs. RCT). Unblind assessment of outcome resulted in a 50% inflation of effect size compared to blind assessment. No difference was noted when type of assessment was unclear.	Overestimation of the effect	
Shadish 1996	100 comparative studies (34 published and 30 non-published RCTs; 17 published and 19 non-published non-RCTs) of marital and family psychotherapy identified through a systematic search	Comparison of the effect size of all RCTs vs non-RCTs; effect sizes present at pretest, publication status, level of attrition, matching and stratification as well as regression analysis including all important independent variables	The overall effect observed in non-RCTs was 87% smaller than the one observed in RCTs ($p < 0.05$). This difference was weaker but was maintained after control for other methodological features. Correlation between pre- and post-test effect size was significant in both designs, but much stronger in non-RCTs (0.84) than in RCTs (0.39).	Underestimation of the effect	

Analysis 3.1. Comparison 3 Studies of trials with adequately compared with inadequately concealed allocation., Outcome I Undefined.

Undefined

Chalmers 1983	145 controlled trials of the treatment of acute myocardial infarction, identified by a systematic search	Comparison of studies with different allocation schemes (non-random, non-concealed random, and concealed random allocation) on maldistribution of prognostic variables, frequency of significant outcomes and case fatality rates	In non-RCTs, non-concealed RCTs, and RCTs with concealed allocation, the maldistribution of prognostic factors was 34%, 7% and 3.5% respectively, the frequency of significant outcomes was 25%, 11% and 5% respectively, the average RRR for mortality was 33%, 23% and 3% respectively. The case-fatality rate for the control groups was 32%, 23% and 16% and for the treatment groups was 21%, 18% and 16% respectively.	Overestimation of the effect	
Moher 1998	127 RCTs from a randomly selected set of 11 meta-analyses (MA) on digestive, circulatory, mental diseases, stroke and infertility from a convenience database of MAs, resp. from the Cochrane Database of Systematic Reviews.	Comparison of the impact of RCTs with unclear / no reporting of allocation concealment, of double blinding, of random generation vs. clear reporting of these features (measured by odds ratio)	The treatment effect was significantly overestimated by 37% in studies lacking allocation concealment. Lack of reporting how double blinding or randomisation generation was achieved had no influence of treatment effect	Over-estimation of the effect in one comparison, similar effects in 2 other comparisons	
Schulz 1995	250 RCTs from 33 meta-analyses from The Cochrane Pregnancy and Childbirth Database	Association between methodological features of controlled trials (allocation concealment, double blinding and follow-up) and the treatment effect (measured by the odds ratio)	The treatment effect was overestimated by 41% in RCTs with inadequate concealment and by 30% in RCTs with unclear adequacy of concealment compared with those with adequate con-	Overestima-tion of the effect in 2 comparisons, similar effects in 1 compari-sion	

Undefined (*Continued*)

			cealment ($p < 0.001$) after adjustment for other methodological features. Studies with no double blinding overestimated the treatment effect by 17% compared with double blinded studies ($p = 0.01$). Lack of complete follow up had no influence on the treatment effect (7%, $p=0.32$).	
--	--	--	--	--

Analysis 4.1. Comparison 4 Studies of high compared with low quality trials., Outcome I Undefined.

Undefined

Emerson 1990	Sam- ple of 7 meta-analyses at hand with 107 primary studies where full information about quality scores was available.	Assessment of the relationship between the quality score and a) the observed treatment difference and b) the variation of the observed treatment difference	No correlation was detected between either quality score and treatment difference or variation of treatment difference within each meta-analysis or in the combined analysis ($p=0.29$).	Similar effects
Imperiale 1990	Meta-analysis of 11 RCTs of steroids in alcoholic hepatitis identified by a systematic search	Compar- ison of short-term mortality in stud- ies with higher and lower methodologi- cal quality	In studies with lower quality the RRR on mortality was 86% smaller compared to the reduction ob- served in high quality studies. In stud- ies with lower qual- ity and hepatic en- cephalopathy no ef- fect was observed, while the RRR of mortality in high quality studies was 55%	Underestimation of the effect

Undefined (*Continued*)

Khan 1996	Meta-analysis of 9 RCTs (parallel or cross-over design) evaluating the effect of antiestrogen treatment in male infertility	Comparison of pregnancy rates in studies with higher and lower methodological quality	In studies of lower quality, the pregnancy rate increased under treatment (OR = 2.6), while it declined under treatment in higher quality studies (OR = 0.5)	Reversal of the effect	
Lipsey 1993	137 / 302 meta-analyses on mental health interventions, counseling, psycho-educational or special therapy, identified by a systematic search (165 meta-analyses on general educational issues were excluded).	After conversion of the results to mean treatment effect sizes, high vs. low quality studies were compared	27 meta-analyses allowed further breakdown of results according to high and low quality studies. No difference in mean effect size was detected (0.40 ± 0.27 resp. 0.37 ± 0.29)	Similar results	
Moher 1998	127 RCTs from a randomly selected set of 11 meta-analyses on digestive, circulatory, mental diseases, stroke and infertility from a convenience database of meta-analyses, resp. from the Cochrane Database of Systematic Reviews.	Comparison of the impact of studies with low vs. high quality scores on treatment effect (measured as odds ratio)	Low quality RCTs resulted in a (significant) 34% greater estimate of treatment effect than high quality RCTs	Overestimation of the effect	
Nurmohamed 1992	Meta-analysis of 35 surgical and orthopaedic RCTs on low-molecular weight heparin as thromboprophylaxis	Comparison of relative risk reduction for deep vein thrombosis (DVT) and pulmonary embolism (PE) in studies of higher and lower methodological quality	In studies with lower quality, the relative risk reduction for venous thrombosis in surgical trials was 2.6 times and in orthopaedic trials 1.4 times larger compared to studies with higher	Overestimation of the effect	

Undefined (*Continued*)

			quality. The relative risk reduction for pulmonary embolus in surgical trials was 1.7 times and in orthopedic trials 2.8 times larger compared to studies with higher quality		
Ortiz 1998	Meta-analysis of 7 RCTs on the effect of folic or folinic acid vs. placebo identified by a systematic search	Comparison of the frequency of gastrointestinal side effects in studies with higher and lower methodological quality	In studies with lower quality there was a 43% reduction in the OR of side-effects (OR 0.57) compared to a 70% reduction in studies with higher quality (OR 0.3)	Underestimation of the effect	
Potter 1998	Meta-analysis of 30 RCTs on protein supplementation in change of weight, anthropometry and case fatality rate	Comparison of RCTs of different methodological quality (group A, B, and C based on concealment of allocation and follow up). Clinical endpoints were change of weight and case fatality rate.	The lowest change in weight was observed in group A (weighted mean difference 1.00%). The benefit of intervention was significantly exaggerated in lower quality studies: It was 3,5 times larger in group B-studies, and 2,5 times larger in group -C studies. Comparison the effect of quality on case fatality rate provided inconsistent results, observing no effect in group A-studies, a trend towards harm in group B-studies and a significant benefit in group C-studies.	Overestimation of the effect for change in weight and inconclusive results for case fatality rate	

Undefined (*Continued*)

Stanton 1997	15 RCTs on family - couples treatment for drug abuse, identified by a systematic search	Correlation between quality rating and size of reported effect size	There was a non-significant trend towards higher quality studies reporting somewhat lower effect sizes ($r = -0.41$)	Non-significant trend for overestimation of the effect	
--------------	---	---	--	--	--

W H A T ' S N E W

Last assessed as up-to-date: 19 February 2007.

27 December 2007 Amended Converted to new review format.

H I S T O R Y

Protocol first published: Issue 3, 2002

Review first published: Issue 3, 2002

20 February 2007 New citation required and conclusions have changed Substantive amendment

C O N T R I B U T I O N S O F A U T H O R S

Regina Kunz and Andy Oxman contributed to the preparation of the protocol and the final manuscript and assessed the relevance and methodological quality of retrieved reports. Regina Kunz prepared the first draft of the protocol and the paper, undertook the majority of the searches with help from David Cowan, Steve Halpern, Alex Jadad; and collected data from included studies. Gunn Vist and Andy Oxman checked the collected data against the original reports. Regina Kunz, Gunn Vist and Kirsty Olsen updated an earlier version of this review published in BMJ and converted the review to a Cochrane Methodology Review.

DECLARATIONS OF INTEREST

None known.

SOURCES OF SUPPORT

Internal sources

- German Cochrane Centre, University of Freiburg, Germany.
- Department of Nephrology, Charité-Mitte, Germany.
- Norwegian Directorate for Health and Social Welfare, Norway.

External sources

- Sonderprogramm zur Förderung von Nachwuchswissenschaftlerinnen, Land Berlin, Germany.

INDEX TERMS

Medical Subject Headings (MeSH)

*Random Allocation; *Selection Bias; Clinical Trials as Topic [*methods; standards; statistics & numerical data]; Controlled Clinical Trials as Topic [methods; standards; statistics & numerical data]; Randomized Controlled Trials as Topic [methods; standards; statistics & numerical data]; Treatment Outcome