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

Kunz R, Vist GE, Oxman AD

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
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.


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 (Eclampsia 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, 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 who 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,td
- [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,td,ut
- Randomised Controlled Trial (exp) AND Selection Bias
- Randomised Controlled Trial (exp) AND Follow-Up Studies
- Randomised Controlled Trial /mt,sn,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:

- Did the study control for 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

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample</th>
<th>Confounding control?</th>
<th>Reproducibility</th>
<th>Outcome measure</th>
<th>Overall judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Consecutive or complete sample?</td>
<td>1. Other methodological criteria: 2. Participants + clinical interventions</td>
<td>1. Explicit description 2. Assessment of agreement</td>
<td></td>
</tr>
<tr>
<td>Aronson 1996</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important - major flaw</td>
</tr>
<tr>
<td>Benson 2000</td>
<td>Yes</td>
<td>1 Partly 2 Partly</td>
<td>1 No 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Bhansali 1996</td>
<td>Yes</td>
<td>1 No 2 Partly</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Carroll 1996</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 partly 2 No</td>
<td>p-value or vote counting</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Chalmers 1977</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Major flaw =&gt; over-estimation</td>
</tr>
<tr>
<td>Chalmers 1983</td>
<td>Yes</td>
<td>1 No 2 Partly</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw =&gt; overestimation</td>
</tr>
<tr>
<td>Study 1989</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 Yes 2 partly</td>
<td>Standardized outcome measure</td>
<td>Possibly important flaw =&gt; underestimation</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>---------------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Diehl 1986</td>
<td>No</td>
<td>1 No 2 Yes</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw -- direction of bias unclear</td>
</tr>
<tr>
<td>Emerson 1990</td>
<td>No</td>
<td>1 Yes 2 No</td>
<td>1 Yes 2 partly</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Forgie 1998</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Guyatt 2000</td>
<td>Yes</td>
<td>1 No 2 Partly</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Imperiale 1990</td>
<td>Yes</td>
<td>1 Yes 2 Yes</td>
<td>1 Yes 2 Yes</td>
<td>Same outcome measure</td>
<td>No important flaw</td>
</tr>
<tr>
<td>Khan 1996</td>
<td>Yes</td>
<td>1 Yes 2 Yes</td>
<td>1 Yes 2 partly</td>
<td>Same outcome measure</td>
<td>No important flaw</td>
</tr>
<tr>
<td>Lipsey 1993</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 No 2 No</td>
<td>Standardized outcome measure</td>
<td>Major flaws</td>
</tr>
<tr>
<td>Miller 1989</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 Yes 2 partly</td>
<td>Standardized outcome measure</td>
<td>Possibly important flaw =&gt; underestimation</td>
</tr>
<tr>
<td>Moher 1998</td>
<td>No</td>
<td>1 Yes 2 No</td>
<td>1 Yes 2 No</td>
<td>Standardized outcome measure</td>
<td>No - Possibly important flaw</td>
</tr>
<tr>
<td>Mullen 1997</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Standardized outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Nurmohamed 1992</td>
<td>Yes</td>
<td>1 Yes 2 Yes</td>
<td>1 Yes 2 partly</td>
<td>Same outcome measure</td>
<td>No important flaw</td>
</tr>
<tr>
<td>Ortiz 1998</td>
<td>Yes</td>
<td>1 Yes 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Same outcome measure</td>
<td>No - possibly important flaw</td>
</tr>
<tr>
<td>Ottenbacher 1991</td>
<td>No</td>
<td>1 No 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Standardized outcome measure</td>
<td>Major flaw</td>
</tr>
<tr>
<td>Ottenbacher 1992</td>
<td>No</td>
<td>1 No 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Standardized outcome measure</td>
<td>Possibly important flaw =&gt; any direction</td>
</tr>
</tbody>
</table>
### Table 1. Critical appraisal of the methodological quality (Continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Conclusion</th>
<th>Randomization/Allocation Concealment</th>
<th>Correct is patients included</th>
<th>Standardized outcome measure</th>
<th>Important Flaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ottenbacher 1993</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 Yes 2 Yes</td>
<td>Standardized outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Potter 1998</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important - major flaw</td>
</tr>
<tr>
<td>Pyorala 1995</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 partly 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important - major flaw</td>
</tr>
<tr>
<td>Reimold 1992</td>
<td>Yes</td>
<td>1 Partly 2 Partly</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>RMIT Group 1994</td>
<td>Yes</td>
<td>1 partly 2 Yes</td>
<td>1 Yes 2 partly</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Sacks 1982</td>
<td>No</td>
<td>1 No 2 Partly</td>
<td>1 Yes 2 No</td>
<td>p-value or vote counting</td>
<td>Major flaw =&gt; over-estimation</td>
</tr>
<tr>
<td>Schulz 1995</td>
<td>Yes</td>
<td>1 Yes 2 Yes</td>
<td>1 Yes 2 partly</td>
<td>Same outcome measure</td>
<td>No important flaw</td>
</tr>
<tr>
<td>Shadish 1996</td>
<td>Yes</td>
<td>1 Yes 2 Partly</td>
<td>1 Yes 2 Yes</td>
<td>Standardized outcome measure</td>
<td>No important flaw</td>
</tr>
<tr>
<td>Stanton 1997</td>
<td>Yes</td>
<td>1 Partly 2 No</td>
<td>1 No 2 No</td>
<td>Standardized outcome measure</td>
<td>No - possibly important flaw</td>
</tr>
<tr>
<td>Watson 1994</td>
<td>Yes</td>
<td>1 Partly 2 Partly</td>
<td>1 Yes 2 No</td>
<td>Same outcome measure</td>
<td>Possibly important flaw</td>
</tr>
<tr>
<td>Wortman 1983</td>
<td>Yes</td>
<td>1 No 2 No</td>
<td>1 No 2 No</td>
<td>Same + standardized outcome measure</td>
<td>Major flaws</td>
</tr>
</tbody>
</table>

**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
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; Steib 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 (Imperi$ale 1990; Khan 1996; Nurme$amed 1992; Schultz 1995; Shadish 1996). However, three of those (Imperi$ale 1990; Khan 1996; Nurme$amed 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 Schultz (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 Schultz 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.

**Implication for methodological research**

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.

**Acknowledgements**

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]

Benson 2000  [published data only]

Bhansali 1996  [published data only]

Carroll 1996  [published data only]

Chalmers 1977  [published data only]

Chalmers 1983  [published data only]

Colditz 1989  [published data only]

Diehl 1986  [published data only]

Emerson 1990  [published data only]

Forgie 1998  [published data only]

Guyatt 2000  [published data only]

Imperiale 1990  [published data only]

Khan 1996  [published data only]

Lipsy 1993  [published data only]

Miller 1989  [published data only]

Moher 1998  [published data only]

Mullen 1997  [published data only]

Nurmohamed 1992  [published data only]

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

Ottenbacher 1991 [published data only]

Ottenbacher 1992 [published data only]

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]

Pyorala 1995 [published data only]

Reimold 1992 [published data only]

RMIT Group 1994 [published data only]

Sacks 1982 [published data only]

Schulz 1995 [published data only]

Shadish 1996 [published data only]

Stanton 1997 [published data only]

Watson 1994 [published data only]

Wortman 1983 [published data only]

References to studies excluded from this review

Colditz 1994 [published data only]

Concato 2000 [published data only]

Gilbert 1977 [published data only]

Hovell 1982 [published data only]

Hutchinson 1999 [published data only]

Kerlikowske 1995 [published data only]

Koes 1994 [published data only]

Kownacki 1996 [published data only]

MacArthur 1995 [published data only]

Mehta 1999 [published data only]

Rosenberg 1999 [published data only]

Steiber 1990 [published data only]
References to studies awaiting assessment

Abrutyn 1991 [published data only]


Black 1996 (P) [published data only]

Brosseau 2002 [published data only]

Callahan 1991 [published data only]

Camma 1996 [published data only]

Chinoy 1999 [published data only]

Cook 1998 [published data only]

Flor 1992 [published data only]

Fouque 1992 [published data only]

Gifford 1995 [published data only]

Glassziou 1993 [published data only]

Grullon 1997 [published data only]
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

Black 1996

Britton 1999

Chalmers 1997

Cochrane 1972

Counsell 1994

Dans 2002

Dickersin 1993

Dickersin 1997

ECIC Bypass 1985

Echt 1991

Eclampsia 1995

### Characteristics of excluded studies  
*ordered by study ID*

<table>
<thead>
<tr>
<th>Study</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colditz 1994</td>
<td>Comparison of randomised trials versus case-control studies</td>
</tr>
<tr>
<td>Concato 2000</td>
<td>Comparison of randomised trials versus cohort studies</td>
</tr>
<tr>
<td>Gilbert 1977</td>
<td>Narrative assessment only</td>
</tr>
<tr>
<td>Hovell 1982</td>
<td>No systematic review: no search strategy, no explicit inclusion or exclusion criteria; no explicit research question; narrative assessment</td>
</tr>
<tr>
<td>Hutchinson 1999</td>
<td>Comparison of randomised trials versus case-control studies</td>
</tr>
<tr>
<td>Kerlikowske 1995</td>
<td>Comparison of randomised trials versus case-control studies</td>
</tr>
<tr>
<td>Koes 1994</td>
<td>Descriptive assessment of individual studies, no summary comparison provided</td>
</tr>
<tr>
<td>Kownacki 1996</td>
<td>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.</td>
</tr>
<tr>
<td>MacArthur 1995</td>
<td>Narrative assessment only</td>
</tr>
<tr>
<td>Mehta 1999</td>
<td>Descriptive assessment of individual studies, no summary comparison provided</td>
</tr>
<tr>
<td>Rozenberg 1999</td>
<td>Narrative assessment only</td>
</tr>
<tr>
<td>Stieb 1990</td>
<td>Comparison of randomised trials versus case-control studies</td>
</tr>
<tr>
<td>Weisburd 2001</td>
<td>Non-healthcare interventions</td>
</tr>
</tbody>
</table>
## Data and Analyses

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

<table>
<thead>
<tr>
<th>Outcome or subgroup title</th>
<th>No. of studies</th>
<th>No. of participants</th>
<th>Statistical method</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined</td>
<td>Other data</td>
<td></td>
<td></td>
<td>No numeric data</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Outcome or subgroup title</th>
<th>No. of studies</th>
<th>No. of participants</th>
<th>Statistical method</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined</td>
<td>Other data</td>
<td></td>
<td></td>
<td>No numeric data</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Outcome or subgroup title</th>
<th>No. of studies</th>
<th>No. of participants</th>
<th>Statistical method</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined</td>
<td>Other data</td>
<td></td>
<td></td>
<td>No numeric data</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Outcome or subgroup title</th>
<th>No. of studies</th>
<th>No. of participants</th>
<th>Statistical method</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined</td>
<td>Other data</td>
<td></td>
<td></td>
<td>No numeric data</td>
</tr>
</tbody>
</table>
Analysis 1.1. **Comparison 1 Studies of randomised compared with non-randomised trials of the same intervention.** Outcome 1 **Undefined.**

<table>
<thead>
<tr>
<th>Study</th>
<th>Design and Trials</th>
<th>Comparison</th>
<th>Outcome</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aronson 1996</strong></td>
<td>4 RCTs and 4 HCTs</td>
<td>Comparison of RCTs vs. HCTs on response rate to tricyclic antidepressant drugs with or without triiodothyronine augmentation in the treatment of refractory depression.</td>
<td>The response rate on T3-supplementation observed in HCTs was exaggerated by 76% compared to the response rate observed in RCTs.</td>
<td>Overestimation of the effect</td>
</tr>
<tr>
<td><strong>Benson 2000</strong></td>
<td>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</td>
<td>Comparison of RCTs and non-RCTs on 1) % change in bone density 2) graft survival 3) infection rates</td>
<td>Only graphical display of results was provided. There, similar effects in RCTs and non-RCTs were detected in all three comparisons.</td>
<td>Similar effects</td>
</tr>
<tr>
<td><strong>Bhansali 1996</strong></td>
<td>12 RCTs and 8 HCTs</td>
<td>Comparison of RCTs and HCTs on patient survival at 1, 2 resp. 3 years</td>
<td>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&lt;0.0001)</td>
<td>Overestimation of the effect</td>
</tr>
</tbody>
</table>
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
### Forgie 1998

| 6 RCTs and 9 non-RCTs including cohort studies identified by a systematic search | Comparison 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 | Comparison 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

<p>| 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 in comparison 2. |</p>
<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Study Design</th>
<th>Comparison</th>
<th>Result</th>
<th>Bias Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyorala 1995</td>
<td>11 RCTs and 22 non-RCTs</td>
<td>Comparison of RCTs and non-RCTs on the descent of testes after LHRH/hCG therapy</td>
<td>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.</td>
<td>Overestimation of the effect</td>
<td></td>
</tr>
<tr>
<td>Reimold 1992</td>
<td>6 RCTs and 6 CCTs</td>
<td>Comparison of RCTs and CCTs on maintenance of sinus rhythm at 3, 6 and 12 months after cardioversion.</td>
<td>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.</td>
<td>Underestimation of the effect</td>
<td></td>
</tr>
<tr>
<td>RMIT Group 1994</td>
<td>9 RCTs and 6 CCTs</td>
<td>Comparison of RCTs and CCTs on live birth rate</td>
<td>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.</td>
<td>Underestimation of the effect when all women were considered, similar effect for pregnant women</td>
<td></td>
</tr>
<tr>
<td>Sacks 1982</td>
<td>Sample of 50 RCTs and 56 HCTs</td>
<td>Comparison of RCTs with HCTs on the frequency of detecting statistically significant results (p &lt; 0.05) of the primary outcome and reduction of mortality.</td>
<td>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/-</td>
<td>Overestimation of the effect</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Study</th>
<th>Design Details</th>
<th>Comparison Details</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watson 1994</td>
<td>4 RCTs and 6 CCTs/HCTs of oil-soluble contrast media during hysterosalpingography in infertile couples, identified by a systematic search</td>
<td>Comparison of RCTs and CCTs/HCTs on the rate of pregnancies</td>
<td>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).</td>
</tr>
<tr>
<td>Wortman 1983</td>
<td>9 RCTs and 16 concurrent quasi-experimental studies (QES) of medical and surgical therapy in coronary artery disease identified through a systematic search</td>
<td>Comparison of RCTs and QES on severity of disease and mortality at 2 and 4 years</td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Undefined</strong></th>
<th><strong>Colditz 1989</strong></th>
<th><strong>Lipsey 1993</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>113 studies published in 1980 comparing new interventions with old, identified in leading cardiology, neurology, psychiatry and respiratory journals by a systematic search</td>
<td>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.</td>
<td>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</td>
</tr>
<tr>
<td>Colditz 1989</td>
<td>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)</td>
<td>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.</td>
</tr>
</tbody>
</table>
One group pre-, post-design resulted in 62% larger treatment effect compared to parallel group design.

<table>
<thead>
<tr>
<th>Miller 1989</th>
<th>188 studies comparing new surgical interventions with old, published in 1983 and identified in leading surgical journals by a systematic search</th>
<th>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.</th>
<th>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).</th>
<th>Inconclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ottenbacher 1991</td>
<td>22 RCTs and 22 non-RCTs on the effectiveness of occupational therapy-interventions, identified by two occupational therapy journals.</td>
<td>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.</td>
<td>Non-RCTs had more often significant test results compared to RCTs (p&lt;0.01). However, the d-index revealed similar effect sizes for both designs (non-RCTs: 0.36 +/-0.18; RCTs: 0.35 +/-0.14).</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Ottenbacher 1992</td>
<td>Sample of 30 RCTs and 30 non-RCTs from a systematic search in NEJM and JAMA across a variety of medical spe-</td>
<td>RCTs were compared with non-RCTs on treatment effects as measured by standardised mean differ-</td>
<td>No difference in treatment effect was found between non-RCTs (0.23) and RCTs (0.21).</td>
<td>Similar effects</td>
</tr>
<tr>
<td>Source</td>
<td>Studies</td>
<td>Characteristics</td>
<td>Comparison</td>
<td>Effect Size</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>----------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Ottenbacher 1993</td>
<td>36 clinical trials on the effectiveness of stroke rehabilitation programmes on functional outcome, identified by a systematic search</td>
<td>Comparison 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.</td>
<td>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.</td>
<td>Overestimation of the effect</td>
</tr>
<tr>
<td>Shadish 1996</td>
<td>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</td>
<td>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.</td>
<td>The overall effect observed in non-RCTs was 87% smaller than the one observed in RCTs (p &lt; 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).</td>
<td>Underestimation of the effect</td>
</tr>
</tbody>
</table>
### Analysis 3.1. Comparison 3 Studies of trials with adequately compared with inadequately concealed allocation., Outcome 1 Undefined.

<table>
<thead>
<tr>
<th>Study</th>
<th>Citation</th>
<th>Trials</th>
<th>Comparison</th>
<th>Outcome</th>
<th>Effect Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chalmers 1983</td>
<td>145 controlled trials of the treatment of acute myocardial infarction, identified by a systematic search</td>
<td>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</td>
<td>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.</td>
<td>Overestimation of the effect</td>
<td></td>
</tr>
<tr>
<td>Moher 1998</td>
<td>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</td>
<td>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)</td>
<td>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</td>
<td>Overestimation of the effect in one comparison, similar effects in 2 other comparisons</td>
<td></td>
</tr>
<tr>
<td>Schulz 1995</td>
<td>250 RCTs from 33 meta-analyses from The Cochrane Pregnancy and Childbirth Database</td>
<td>Association between methodological features of controlled trials (allocation concealment, double blinding and follow-up) and the treatment effect (measured by the odds ratio)</td>
<td>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-</td>
<td>Overestimation of the effect in 2 comparisons, similar effects in 1 comparison</td>
<td></td>
</tr>
</tbody>
</table>
### Analysis 4.1. Comparison 4 Studies of high compared with low quality trials, Outcome 1 Undefined.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample</th>
<th>Assessment</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emerson 1990</td>
<td>Sample of 7 meta-analyses at hand with 107 primary studies where full information about quality scores was available.</td>
<td>Assessment of the relationship between the quality score and a) the observed treatment difference and b) the variation of the observed treatment difference</td>
<td>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).</td>
</tr>
<tr>
<td>Imperiale 1990</td>
<td>Meta-analysis of 11 RCTs of steroids in alcoholic hepatitis identified by a systematic search</td>
<td>Comparison of short-term mortality in studies with higher and lower methodological quality</td>
<td>In studies with lower quality the RRR on mortality was 86% smaller compared to the reduction observed in high quality studies. In studies with lower quality and hepatic encephalopathy no effect was observed, while the RRR of mortality in high quality studies was 55%</td>
</tr>
<tr>
<td>Author</td>
<td>Year</td>
<td>Methodology</td>
<td>Findings</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Khan 1996</td>
<td></td>
<td>Meta-analysis of 9 RCTs (parallel or cross-over design) evaluating the effect of antiestrogen treatment in male infertility</td>
<td>Comparison of pregnancy rates in studies with higher and lower methodological quality</td>
</tr>
<tr>
<td>Lipsey 1993</td>
<td></td>
<td>137 / 302 meta-analyses on mental health interventions, counseling, psychoeducational or special therapy, identified by a systematic search (165 meta-analyses on general educational issues were excluded).</td>
<td>After conversion of the results to mean treatment effect sizes, high vs. low quality studies were compared</td>
</tr>
<tr>
<td>Moher 1998</td>
<td></td>
<td>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.</td>
<td>Comparison of the impact of studies with low vs. high quality scores on treatment effect (measured as odds ratio)</td>
</tr>
<tr>
<td>Nurmohamed 1992</td>
<td></td>
<td>Meta-analysis of 35 surgical and orthopaedic RCTs on low-molecular weight heparin as thromboprophylaxis</td>
<td>Comparison of relative risk reduction for deep vein thrombosis (DVT) and pulmonary embolism (PE) in studies of higher and lower methodological quality</td>
</tr>
<tr>
<td>study</td>
<td>design</td>
<td>intervention</td>
<td>outcome measure</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Ortiz 1998</td>
<td>Meta-analysis of 7 RCTs on the effect of folic or folinic acid vs. placebo identified by a systematic search</td>
<td>Comparison of the frequency of gastrointestinal side effects in studies with higher and lower methodological quality</td>
<td>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)</td>
</tr>
<tr>
<td>Potter 1998</td>
<td>Meta-analysis of 30 RCTs on protein supplementation in change of weight, anthropometry and case fatality rate</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>
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

**WHAT'S NEW**

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

| 27 December 2007 | Amended | Converted to new review format. |

**HISTORY**


Review first published: Issue 3, 2002

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

**CONTRIBUTIONS OF AUTHORS**

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.

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