Introduction

One of the key theoretical principles of meta-analyses is that all data must be treated equally with precision. In recent years, however, the quality of the reporting of data in primary studies, often used as a proxy measure for methodological quality, has been shown to affect estimates of intervention efficacy reported in meta-analyses (Schulz et al., 1995; Moher et al., 1999; Tierney & Stewart, 2005; Gluud, 2006), although data are still controversial Emerson et al., 1990; Kjaergard et al., 2001; Balk et al., 2002; Juni et al., 2001). Meta-analysts need to take quality into consideration to reduce heterogeneity and to provide unbiased treatment estimates (Moher et al., 1999). In order to investigate whether different methods of quality assessment provide different estimates of intervention efficacy, Moher and colleagues randomly selected 11 meta-analyses (127 RCTs, mostly placebo-controlled) dealing with different medical areas (digestive diseases, circulatory diseases, mental health, neurology and pregnancy and childbirth) (Moher et al., 1999). A statistically significant exaggeration of treatment efficacy was found when results of lower-quality trials were pooled whether the trial quality assessments were made by a scale approach or by an individual component approach. However, generalisability of findings can be limited by whether or not there is an active comparator (heterogeneity of intervention, population and outcome) and furthermore sensitivity analyses can miss to find possible confounding variables, apparently not related to trial quality.

In the field of meta-analyses of data extracted from antidepressant (AD) RCTs, quality remains a hot issue. It is unclear whether in this specific field a relationship exists between quality measures and treatment estimates and, additionally, it is unclear whether different quality measures provide different estimates of treatment efficacy. Furthermore, to reliably inform clinical practice there is the need for grading the evidence coming from systematic reviews (and meta-analyses) in the field of AD treatment for major depression. To answer these questions, we therefore investigated the following issues in a step-wise approach:

(1) whether RCT quality, assessed by either validated rating scales or individual components, influenced treatment estimates in a homogeneous sample of AD RCTs. An ongoing Cochrane review concerned with fluoxetine included published clinical trials comparing fluoxetine to other ADs, offered an opportunity for this analysis (Cipriani et al., 2006).

(2) whether it is possible to find a validated way of grading the quality of systematic reviews (and meta-analyses) in order to have an explicit hierarchy of robustness and reliability of findings. An ongoing multiple treatment meta-analysis (MTM) was used to test this hypothesis.

Overview of the scientific literature

Although RCTs provide the best evidence of the efficacy of medical interventions, they are not immune to bias (Easterbrook et al., 1991). Studies relating methodological features of trials to their results have shown that trial quality influences effect sizes and conclusions exclusively based on published studies, therefore, can be misleading (Egger & Smith, 1995). Quality is complex and difficult to define, because it could encompass the design, conduct, analysis, and external validity, as well as the reporting of a clinical experiment.

For populations of trials examining treatments in myocardial infarction (Chalmers et al., 1983), perinatal medicine (Schultz et al., 1995), and various disease areas (Moher et al., 1998), it has consistently been shown that inadequate concealment of treatment allocation, resulting, for example, from the use of open random-number tables, is associated on average with larger treatment effects. Schultz and colleagues found larger average effect sizes if trials were not double-blind (Schultz et al., 1995).

Analyses of individual trials suggest that in some instances effect sizes are also overestimated if some participants, for example, those not adhering to study medications, were excluded from the analysis (Sackett & Gent, 1979; May et al., 1981; Peduzzi et al., 1993). Informal qualitative research has indicated that investigators sometimes undermine the random allocation of study participants, for example, by opening assignment envelopes or holding translucent envelopes up to a light bulb (Schultz, 1995).

In response to this situation, guidelines on the conduct and reporting of clinical trials and scales to measure the quality of published trials have been developed (Begg et al., 1996; Moher et al., 1995). RCTs provide the best test of the efficacy of preventive or therapeutic interventions because they can separate the effects of the intervention from those of extraneous factors such as natural recovery and statistical regression.

When more than one trial has examined a particular intervention, systematic reviews potentially provide the best summaries of the available evidence. Systematic reviewers can summarise findings of randomised trials using an impressionistic approach (qualitative synthesis) or they can produce quantitative syntheses by statistically combining the results from several studies (meta-analysis).

Early reports on the quality of reporting for systematic reviews indicated that many reviews have serious flaws. Jadad and colleagues reported on the quality of reporting in 50 systematic reviews (38 paper-based and 12 Cochrane) that examined the treatment of asthma (Jadad e al., 2000). Of these reviews, 58% were published in 1997 or 1998. The authors found that 80% had serious or extensive flaws; however, they found that the Cochrane reviews were more rigorous and better reported than the paper-based publications. In contrast, other researchers found only minor or minimal flaws in the quality of reporting in nearly half of 82 systematic reviews of perioperative medicine (Choi et al., 2001). This latter study suggests that there may be an association between quality of reporting and the content area of the systematic review.

The quality of trials is of obvious relevance to meta-analysis. If the raw material used is flawed, then the conclusions of meta-analytic studies will be equally invalid. Meta-analysis is widely used to summarize the evidence on the benefits and risks of medical interventions. However, the findings of several meta-analyses of small trials have been contradicted subsequently by large controlled trials (Egger et al., 1997a; LeLorier et al., 1997). The fallibility of meta-analysis is not surprising, considering the various biases that may be introduced by the process of locating and selecting studies, including publication bias (Easterbrook et al., 1991), language bias (Egger et al., 1997b), and citation bias (Gøtzsche, 1987). Low methodological quality of component studies is another potential source of systematic error. The critical appraisal of trial quality is therefore widely recommended and a large number of different instruments are currently in use. However, the method of assessing and incorporating the quality of clinical trials is a matter of ongoing debate (Moher et al., 1996).

This is reflected by the plethora of available instruments. In a search covering the years up to 1993, Moher and colleagues identified 25 different quality assessment scales (Moher et al., 1996) (Table I). More recently, in a hand search of 5 general medicine journals dating 1993 to 1997 (*Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine*) Juni and colleagues identified 37 meta-analyses using 26 different instruments to assess trial quality (Juni et al., 1999).

Table I. Characteristics of 25 Scales for quality assessment of clinical trials

		Weight given to methodological key domains (%)*		
Scale	No. of Items	Randomisation	Blinding	Withdrawals
Andrew 1984	11	9.1	9.1	9.1
Beckerman 1992	24	4.0	12.0	16.0
Brown 1991	6	14.3	4.8	0
Chalmers 1990	3	33.3	33.3	33.3
Chalmers 1981	30	13.0	26.0	7.0
Cho & Bero 1994	24	14.3	8.2	8.2
Colditz 1989	7	28.6	0	14.3
Detsky	14	20.0	6.7	0
Evans & Pollock 1985	33	3.0	4.0	11.0
Goodman 1994	34	2.9	2.9	5.9
Gotzsche 1989	16	6.3	12.5	12.5
Imperiale 1990	5	0	0	0
Jadad 1996	3	40.0	40.0	20.0
Jonas 1993	18	11.1	11.1	5.6
Kleijnen 1991	7	20.0	20.0	0
Koes 1991	17	4.0	20.0	12.0
Levine 1991	29	2.5	2.5	3.1
Linde 1991	7	28.6	28.6	28.6
Nurmohamed 1992	8	12.5	12.5	12.5
Onghena 1992	10	5.0	10.0	5.0
Poynard 1988	14	7.7	23.1	15.4
Reisch 1989	34	5.9	5.9	2.9
Smith 1992	8	0	25.0	12.5
Spitzer 1990	32	3.1	3.1	9.4
ter Riet 1990	18	12.0	15.0	5.0

*Weight of methodological domains most relevant to the control of bias, expressed as percentage of maximum scores.

Most of these scoring systems lack a focused theoretical basis and their objectives are unclear. The scales differ considerably in terms of dimensions covered, size, and complexity, and the weight assigned to the key domains most relevant to the control of bias (randomization, blinding, and withdrawals) varies widely. Many meta-analysts assess the quality of trials and exclude trials of low methodological quality in sensitivity analyses. Medical literature can provide us a famous example to clarify clinical correlates of such a problematic issue. In a meta-analysis of trials comparing low-molecular-weight heparin (LMWH) with standard heparin for thrombo-prophylaxis in general surgery, Nurmohamed and colleagues found a significant reduction of 21% in the risk of deep vein thrombosis (DVT) with LMWH (p = 0.012) (Nurmohamed et al., 1992). However, when the analysis was limited to trials with strong methods, as assessed by a scale consisting of 8 criteria, no significant difference between the 2 heparins remained (relative risk [RR] reduction, 9%; p = 0.38). The authors therefore concluded that "there is at present no convincing evidence that in general surgery patients LMWHs, compared with standard heparin, generate a clinically important improvement in the benefit to risk ratio." By contrast, another group of meta-analysts did not consider the quality of trials and concluded that "LMWHs seem to have a higher benefit to risk ratio than unfractionated heparin in preventing perioperative thrombosis (Leizorovicz et al., 1992)." Juni and colleagues repeated the meta-analysis of Nurmohamed using 25 different scales examining whether the type of scale used for assessing the quality of trials affects the conclusions of meta-analytic studies (Juni et al., 1999). This study showed that the type of scale used to assess trial quality could dramatically influence the interpretation of meta-analytic studies. (Figure 1).

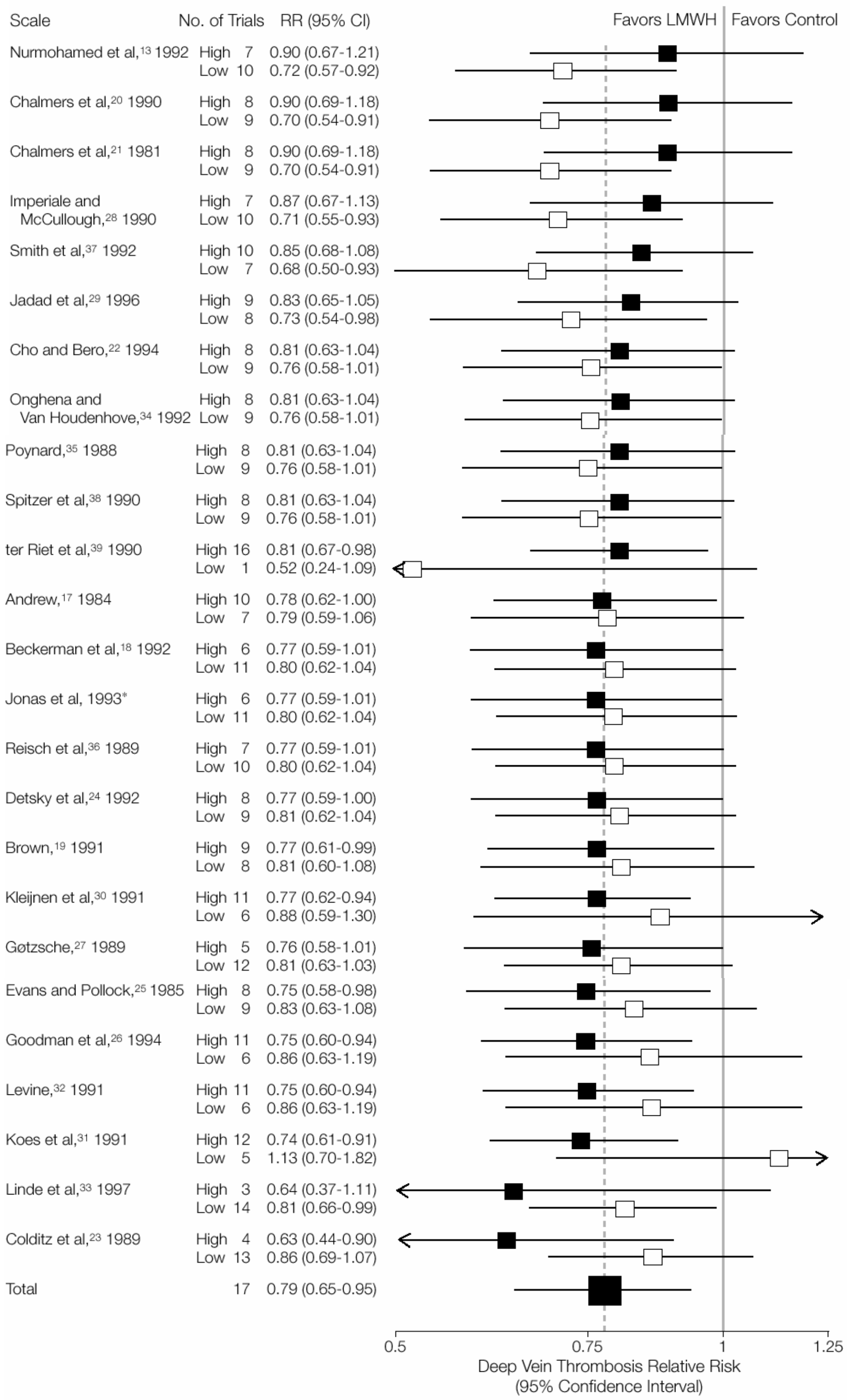


Figure 1. Results from sensitivity analyses dividing trials in high- and low-quality strata.

Whereas for some scales these findings were confirmed, the use of others would have led to opposite conclusions, indicating that the beneficial effect of LMWH was particularly robust for trials deemed to be of high quality. Similarly, in meta-regression analysis effect size was negatively associated with some quality scores, but positively associated with others. Accordingly, RRs estimated for hypothetical trials of maximum or minimum quality varied widely between scales.

In Juni and colleagues' review, blinding of outcome assessment was the only factor significantly associated with effect size, with RRs on average being exaggerated by 35% if outcome assessment was open (Juni et al., 1999). The importance of blinding could have been anticipated considering that the interpretation of the test (fibrinogen leg scanning) used to detect DVT can be subjective (Lensing & Hirsh, 1993); in other situations, blinding of outcome assessment may be irrelevant, such as when examining the effect of an intervention on overall mortality.

In contrast to studies including large numbers of trials (Moher et al., 1998), in this meta-analysis there was not a significant association of concealment of treatment allocation with effect estimates. This meta-analysis could have been too small to show this effect, or, alternatively, concealment of treatment allocation may not have been relevant in the context of this study. The importance of allocation concealment may to some extent depend on whether strong beliefs exist among investigators regarding the benefits or risks of assigned treatments or whether equipoise of treatments is accepted by all investigators involved (Schultz, 1995). Strong beliefs are probably more prevalent in trials comparing an intervention with placebo than in trials comparing two similar, active interventions.

The fact that the type of scale used to assess trial quality could dramatically influence the interpretation of meta-analytic studies is not surprising when considering the heterogeneous nature of the instruments (Moher et al., 1996). Many scales include items that are more closely related to reporting quality, ethical issues, or to the interpretation of results rather than to the internal validity of trials. For example, some scales assessed whether the rationale for conducting the trial was clearly stated, whether the trialists' conclusions were compatible with the results obtained, or whether the report stated that participants provided written informed consent.

Important differences also exist between scales that focus on internal validity. For example, the scale developed by Jadad and colleagues gives more weight to the quality of reporting than to actual methodological quality (Jadad et al., 1996). A statement on withdrawals and dropouts earns the point allocated to this domain, independently of whether the data were analyzed according to the intention-to-treat principle. The instrument addresses randomization but does not assess allocation concealment. The use of an open random-number table would thus be considered equivalent to concealed randomization using a telephone or computer system and earn the maximum points foreseen for randomization.

Conversely, the scale developed by Chalmers and collaborators allocates 0 points for unconcealed but the maximum of 3 points for concealed randomization (Chalmers et al., 1990). The authors of the different scales clearly had different perceptions of trial quality, but definitions were rarely given, and the ability of the scales to measure what they are supposed to measure remains unclear.

Interestingly, in a review of treatment effects from trials deemed to be of high or low quality, Kunz and Oxman found that in some meta-analyses there were no differences whereas in other meta-analyses high-quality trials showed either larger or smaller effects (Kunz & Oxman, 1998). Different scales had been used for assessing quality and it is possible that the choice of the scale contributed to the discrepant associations observed in these meta-analyses.

Although improved reporting practices should facilitate the assessment of methodological quality in the future, incomplete reporting continues to be an important problem when assessing trial quality. Because small single-centre studies may be more likely to be of inadequate quality and more likely to be reported inadequately than large multi-centre studies, the sample size and number of study centres may sometimes be useful proxy variables for study quality (Begg et al., 1996). Confounding could exist between measures of trial quality and other characteristics of trials, such as the setting, the characteristics of the participants, or the treatments (Egger et al., 1997a).

The assessment of the methodological quality of randomized trials and the conduct of sensitivity analyses should be considered routine procedures in meta-analysis.

To summarise:

- Although composite quality scales may provide a useful overall assessment when comparing populations of trials, for example, trials published in different languages or disciplines, such scales should not generally be used to identify trials of apparent low quality or high quality in a given meta-analysis (Greenland, 1994).
- the relevant methodological aspects should be identified, ideally a priori, and assessed individually.
- this should always include the key domains of concealment of treatment allocation, blinding of outcome assessment or double blinding, and handling of withdrawals and dropouts.
- the lack of well-performed and adequately sized trials cannot be remedied by statistical analyses of small trials of questionable quality.

The quality of reporting is therefore often used as a proxy measure for methodological quality; however, similar quality of reporting may hide important differences in methodological quality (Huwiler-Muntener et al., 2002). Meta-analysts need to take this information into consideration to reduce or avoid bias whenever possible.

Although has been pointed out previously by Detsky and colleagues that the incorporation of quality scores as weights lacks statistical or empirical justification (Detsky et al., 1992), it has been suggested that estimates of the quality of reports of clinical trials should be taken into account in the synthesis of evidence from these reports (Moher et al., 1998). The aim of this study is to investigate whether the method of quality assessment of RCTs and of systematic reviews by a validated approach influences estimates of intervention efficacy.

Materials and Methods

1. *Quality of RCTs*

RCTs were identified by searching the Cochrane Collaboration Depression, Anxiety and Neurosis Controlled Trials Register (CCDANCTR) and the Cochrane Central Register of Controlled Trials (CENTRAL). The following terms were used: FLUOXETIN* OR *adofen* or *docutrix* or *erocap* or *fluctin* or *fluctine* or *fluoxeren* or *fontex* or *ladose* or *lorien* or *lovan* or *mutan* or *prozac* or *prozyn* or *reneuron* or *sanzur* or *saurat* or *zactin*. MEDLINE (1966–2004) and EMBASE (1974–2004) were searched using the terms *fluoxetine* and *randomized controlled trial* or *random allocation* or *double-blind method*. Non–English language publications were included. Reference lists of relevant papers and previous systematic reviews were hand-searched for published reports up to March 2006.

Selection and study characteristics

Only RCTs which presented results on efficacy and dropouts, and compared fluoxetine with any other antidepressant agent, including St John's wort, in the acute treatment of major depression in patients aged more than 18 years were eligible for inclusion. Crossover studies and trials in depressed patients with a concurrent medical illness were excluded.

Data abstraction

Two reviewers independently extracted data; any disagreement was solved by discussion and consensus with a third member of the team. Reviewers were not blinded to the journal name and authors. All reviewers underwent training in evaluating trial quality. Before training, the definition of each item was discussed. Inter-rater agreement was checked by calculating a correlation coefficient (k coefficient); as stated elsewhere, values above 0.60 were taken to indicate substantial agreement (Landis & Koch, 1977a). The inter-rater reliability was also evaluated by Analysis of Variance Intraclass Correlation Coefficient (ANOVA-ICC). The ANOVA-ICC assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects and in general, an ANOVA ICC above 0.7 indicates good reliability (Landis & Koch, 1977b).

Quality assessment

The quality of RCTs was assessed using the Jadad scale (Jadad et a., 1996) and the CCDAN quality assessment instrument (Moncrieff et al., 2001). Additionally, the Consolidated Standards of Reporting Trials (CONSORT) statement was employed to assess reports of RCTs (Moher et al., 2001).

The Jadad scale consists of three items pertaining to descriptions of randomization, masking, dropouts and withdrawals. The scale ranges from 0 to 5, with higher scores indicating better reporting.

The CCDAN instrument, specifically developed for trials of treatments for depression and neurosis, consists of 23 items covering a wide range of aspects of quality including objective formulation, design, presentation of results, analysis and quality of conclusions (http://web1.iop.kcl.ac.uk/IoP/ccdan/qrs.htm for full details) (Moncrieff et al., 2001). It covers aspects of both internal validity (or control of bias) and external validity (or generalisability). Each item can score 0 to 2 and all items equally contribute to the final score. The final score ranges from 0 to 46, with higher scores indicating better quality.

The revised CONSORT statement, primarily intended for use in writing, reviewing or assessing reports of simple two-group parallel RCTs, consists of a checklist of 22 items. It's not a rating scale and has been endorsed by many medical journals (Moher et al., 2001; Altman, 2005). Among the overall 22 items, we selected randomisation, allocation concealment and power calculation as proxy measures of trial quality, according to Schulz and Grimes (Grimes & Schulz, 1996; Schulz & Grimes, 2002; Schulz & Grimes, 2005).

Statistical analysis

Efficacy was defined as the number of patients who failed to respond. Tolerability was defined as the number of patients who failed to complete the study due to any cause. Efficacy and tolerability outcomes were pooled across studies to produce overall estimates of treatment effect. We pre-planned to compare fluoxetine against tricyclics (TCAs) and against selective serotonin reuptake inhibitors (SSRIs). Newer ADs were not considered because they are not considered an homogeneous group. Medium/high quality RCTs were defined as those scoring more than 2 out of a maximum of 5 at the Jadad scale; this threshold was derived from Moher and colleagues. Overall CCDAN quality score was categorized according to a final score of more than 20 as a cut-off value for high quality studies.[6] According to the CONSORT statement instructions, each of the three items was assigned a "yes/no" response depending on whether the authors had reported appropriate performance on the required quality parameter (instructions can be accessed at www.consort-statement.org). Studies reporting at least one "yes" in one of the three items were considered high quality RCTs.

We used Review Manager 4.2.10 (http://www.cc-ims.net/RevMan) to pool data for summary estimates. We expressed results for dichotomous outcomes as risk ratio (Peto Odds Ratio (OR)), with values of <1 favouring fluoxetine, and continuous efficacy outcomes as standardised mean difference, both with 95% confidence intervals. Efficacy and tolerability outcomes were calculated for the overall sample of included trials and for the subgroup of high-quality trials according to the Jadad scale, the CCDAN checklist and the three items of the CONSORT statement.

Heterogeneity among trials was assessed by using a Cochran *Q* test and calculating I^2 to measure the proportion of total variation due to heterogeneity beyond chance (Higgins et al., 2003). Publication bias was assessed by using funnel plots of the log OR (Egger et al., 1997). After potential confounding factors not strictly related to trial quality were controlled for, a meta-regression technique was employed in order to ascertain whether RCT quality influences treatment estimates. STATA 9.0 software was used to perform the meta-regression analysis on the log OR scale, with each trial weighting equal to the inverse of the variance of the estimate for that study and between study variance estimated with the restricted maximum likelihood method. Meta-regression is a useful tool for analysing the associations between treatment effect and study characteristics, and is particularly useful where heterogeneity in the effect of treatment between studies is found (Sterne et al., 2002).

Efficacy and tolerability outcomes were used as dependent variables and the Jadad, CCDAN and CONSORT scores were used as continuous predictive variables. The following independent variables were controlled for (Thompson & Higgins, 2002; Barbui et al., 2004): year of publication (continuous variable), age (1=adults only; 0=other), study setting (1=inpatients; 0=outpatients), fluoxetine dose (continuous outcome) and fluoxetine used as the experimental rather than comparator drug (1=yes; 0=no). Sample size was not inserted into the model because this was one item of the CCDAN rating scale.

2. *Quality of systematic reviews (and meta-analyses)*

Up to now current quality measures are not related with treatment estimates in AD trials and may not be useful weighting tools when meta-analyses of data extracted from AD RCTs are carried out. To overcome this problem, we tried to assess quality of groups of studies instead of focusing on individual trials.

Firstly, we reviewed (searching PubMed and Medline up to October 2007) the scientific literature to identify some important issues strictly related to quality of research findings. At the end of the reviewing process, we identified the following five issues: randomization, overall sample size, number of included studies, sponsorship, internal and external validity, missing data/imputation.

Secondly, we analyzed a homogeneous group of studies, to avoid the confounding bias possibly related to study design. We therefore chose a set of systematic reviews on antidepressants and ran a multiple treatment meta-analysis (MTMC). This set of systematic reviews is part of the Meta-Analyses of New Generation Antidepressants (MANGA) project in which a group of researchers within the Cochrane Collaboration Depression, Anxiety and Neurosis Group agreed to systematically review all available evidence for each specific newer antidepressant, in order to inform clinical practice and mental health policies.

Important issues strictly related to quality of RCTs and systematic reviews

Randomisation

The simplest approach to evaluating a new treatment is to compare a single group of patients given the new treatment with a group previously treated with an alternative treatment (Altman, 2005). Usually such studies compare two consecutive series of patients in the same hospital. This approach is seriously flawed. Problems will arise from the mixture of retrospective and prospective studies, and we can never satisfactorily eliminate possible biases due to other factors (apart from treatment) that may have changed over time. Sacks et al compared trials of the same treatments in which randomised or historical controls were used and found a consistent tendency for historically controlled trials to yield more optimistic results than randomised trials. The use of historical controls can be justified only in tightly controlled situations of relatively rare conditions, such as in evaluating treatments for advanced cancer. The need for contemporary controls is clear, but there are difficulties. If the clinician chooses which treatment to give each patient there will probably be differences in the clinical and demographic characteristics of the patients receiving the different treatments. Much the same will happen if patients choose their own treatment or if those who agree to have a treatment are compared with refusers. Similar problems arise when the different treatment groups are at different hospitals or under different consultants. Such systematic differences, termed bias, will lead to an overestimate or underestimate of the difference between treatments. Bias can be avoided by using random allocation.

A well known example of the confusion engendered by a non-randomised study was the study of the possible benefit of vitamin supplementation at the time of conception in women at high risk of having a baby with a neural tube defect. The investigators found that the vitamin group subsequently had fewer babies with neural tube defects than the placebo control group. The control group included women ineligible for the trial as well as women who refused to participate. As a consequence the findings were not widely accepted, and the Medical Research Council later funded a large randomised trial to answer to the question in a way that would be widely accepted. The main reason for using randomisation to allocate treatments to patients in a controlled trial is to prevent biases of the types described above. We want to compare the outcomes of treatments given to groups of patients which do not differ in any systematic way. Another reason for randomising is that statistical theory is based on the idea of random sampling. In a study with random allocation the differences between treatment groups behave like the differences between random samples from a single population. We know how random samples are expected to behave and so can compare the observations with what we would expect if the treatments were equally effective.

The term *random* does not mean the same as hap-hazard but has a precise technical meaning (Schulz & Grimes, 2002). By random allocation we mean that each patient has a known chance, usually an equal chance, of being given each treatment, but the treatment to be given cannot be predicted. If there are two treatments the simplest method of random allocation gives each patient an equal chance of getting either treatment; it is equivalent to tossing a coin. In practice most people use either a table of random numbers or a random number generator on a computer. This is simple randomisation. Possible modifications include block randomisation, to ensure closely similar numbers of patients in each group, and stratified randomisation, to keep the groups balanced for certain prognostic patient characteristics. Fifty years after the publication of the first randomised trial the technical meaning of the term randomisation continues to elude some investigators. Journals continue to publish "randomised" trials which are no such thing. One common approach is to allocate treatments according to the patient's date of birth or date of enrolment in the trial (such as giving one treatment to those with even dates and the other to those with odd dates), by the terminal digit of the hospital number, or simply alternately into the different treatment groups. While all of these approaches are in principle unbiased—being unrelated to patient characteristics—problems arise from the openness of the allocation system. Because the treatment is known when a patient is considered for entry into the trial this knowledge may influence the decision to recruit that patient and so produce treatment groups which are not comparable. Of course, situations exist where randomisation is simply not possible. The goal here should be to retain all the methodological features of a well conducted randomised trial other than the randomisation.

Regardless of how the allocation sequence has been generated—such as by simple or stratified randomisation—there will be a pre-specified sequence of treatment allocations. In principle, therefore, it is possible to know what treatment the next patient will get at the time when a decision is taken to consider the patient for entry into the trial. The strength of the randomised trial is based on aspects of design which eliminate various types of bias.

Randomisation of patients to treatment groups eliminates bias by making the characteristics of the patients in two (or more) groups the same on average, and stratification with blocking may help to reduce chance imbalance in a particular trial. All this good work can be undone if a poor procedure is adopted to implement the allocation sequence. In any trial one or more people must determine whether each patient is eligible for the trial, decide whether to invite the patient to participate, explain the aims of the trial and the details of the treatments, and, if the patient agrees to participate, determine what treatment he or she will receive. Suppose it is clear which treatment a patient will receive if he or she enters the trial (perhaps because there is a typed list showing the allocation sequence). Each of the above steps may then be compromised because of conscious or subconscious bias. Even when the sequence is not easily available, there is strong anecdotal evidence of frequent attempts to discover the sequence through a combination of a misplaced belief that this will be beneficial to patients and lack of understanding of the rationale of randomisation. How can the allocation sequence be concealed?

Firstly, the person who generates the allocation sequence should not be the person who determines eligibility and entry of patients. Secondly, if possible the mechanism for treatment allocation should use people not involved in the trial (Schulz et al., 1995a). A common procedure, especially in larger trials, is to use a central telephone randomisation system. Here patient details are supplied, eligibility confirmed, and the patient entered into the trial before the treatment allocation is divulged (and it may still be blinded). Another excellent allocation concealment mechanism, common in drug trials, is to get the allocation done by a pharmacy. The interventions are sealed in serially numbered containers (usually bottles) of equal appearance and weight according to the allocation sequence. If external help is not available the only other system that provides a plausible defence against allocation bias is to enclose assignments in serially numbered, opaque, sealed envelopes. Apart from neglecting to mention opacity, this is the method used in the famous 1948 streptomycin trial. This method is not immune to corruption, particularly if poorly executed. However, with care, it can be a good mechanism for concealing allocation. We recommend that investigators ensure that the envelopes are opened sequentially, and only after the participant's name and other details are written on the appropriate envelope. If possible, that information should also be transferred to the assigned allocation by using pressure sensitive paper or carbon paper inside the envelope. If an investigator cannot use numbered containers, envelopes represent the best available allocation concealment mechanism without involving outside parties, and may sometimes be the only feasible option.

Sponsorship

Investigators who contribute to clinical trials often receive funding, either directly or indirectly, from sponsors with an interest in the outcome and reporting of these trials (Schulz et al., 1995b). Such a relationship may create conflict of interest for these authors, in which their interest in an objective description of outcomes competes with their obligation, perceived or real, to the sponsor. This concern is more than hypothetical: industry-sponsored trials may be more likely to report favourable outcomes, raising the possibility of influence on study design or publication bias.

To address this potential bias, journals typically require disclosure of conflict of interest by authors, although journal policies on disclosure have been suggested to be inconsistent and prone to abuse. The potential consequences of financial conflict of interest in medicine as a whole have raised substantial concern in both the medical literature and the lay press. However, the prevalence and implications of conflict of interest in psychiatry have received relatively little attention. This is particularly notable given the extent of industry involvement in drug development in psychiatry, the rapid growth in pharmacotherapies in psychiatry approved by the Food and Drug Administration, and recent calls for the establishment of a clinical trial registry to ensure the fair reporting of the results of clinical trials.

Imputation and dealing with missing data

a) Imputing standard deviation

Conduct of a systematic review or a meta-analysis involves comprehensive search of relevant RCTs and their quantitative or qualitative synthesis. To pool results on a continuous outcome measure of the identified RCTs quantitatively, one needs both means and standard deviations (SDs) on that outcome measure for each RCT (Furukawa et al., 2006). Many reports of RCTs, however, fail to provide SDs for their continuous outcomes. It is sometimes possible to use P or t or F values, reported in the original RCTs, to calculate exact SDs. When none of these is available, it is recommended that one should contact primary authors. However, the yield is very often very low; some are incontactable, some never respond, and others report that the data are discarded, lost or irretrievable because there are no longer any computers to read the tapes. Some meta-analysts then resort to substitution of SDs of known outcome measures by those reported in other studies, either from another meta-analysis or from other studies in the same meta-analysis. But the validity of such practices has never been empirically examined.

One study therefore aimed at examining empirically the validity of borrowing SDs from other studies when individual RCTs fail to report SDs in a meta-analysis, by simulating the above-mentioned two imputation methods for SDs in two meta-analyses on antidepressants that have been previously conducted (Furukawa et al., 2006). Systematic reviews for depression are particularly suitable for this purpose, because Hamilton Rating Scale for

Depression (HRSD) is the de facto standard in symptom assessment and is used in many depression trials identified for overviews. The degree of concordance of the actual effect sizes and the imputed effect sizes was gratifying both on individual trial basis and on aggregate basis. Strictly speaking, it is not straightforward to generalize the current findings beyond pharmacologic trials for depression with regard to the Hamilton Rating Scale for Depression. However, the good to excellent correspondence between the actual SMDs and imputed SMDs of individual RCTs, and the virtual agreement between the actual meta-analyzed SMDs and the imputed meta-analyzed SMDs strongly argue for the appropriateness of both imputation methods.

One must also remember that the present simulation study borrowing SDs from a previous meta-analysis represents the worst-case scenario, where none of the included trials had reported SDs, and therefore, the observed discrepancy, if any, would correspond with the biggest difference possible. In actuality, at least some of the identified trials do report SDs, and the resultant pooled estimates of the SMD would be less subject to the imputation assumption. Leaving out, for example, five of the included trials would be closer to borrowing from a different meta-analysis than the leaving-one-out method, which we employed in this article, but we felt that we did not need to simulate the former, as we had already examined the ''worst case.'' At the moment we do not have much ground to choose between the two imputation methods. We would, therefore, like to recommend, in the case of systematic reviews where some of the identified trials do not report SDs:

□ When the number of RCTs with missing SDs is small and when the total number of RCTs is large, to use the pooled SDs from all the other available RCTs in the same meta-analysis. It is possible and recommended in this case to examine the appropriateness of the imputation by comparing the SMDs of those trials that had reported SDs against the hypothetical SMDs of the same trials based on the imputed SDs. If they converge, we can be more confident in the meta-analytic results.

□ When the number of RCTs with missing SDs is large or when the total number of RCTs is small, to borrow SDs from a previous systematic review, because the small sample size may allow unexpected deviation due to chance. One must remember, however, that the credibility of the meta-analytic findings will be less secure in this instance.

b) Imputing response rate

Much discussion and examination on how to deal with missing values can be found in the literature in the case of individual RCTs (Furukawa et al., 2005). By contrast, there is only limited literature about this problem in the case of meta-analysis. However, meta-analysts often try to perform the ITTanalysis, even when the original RCTs fail to do so. When the outcome is a dichotomous scale, one common approach is to assume either that all missing participants experienced the event or that all missing participants did not experience the event, and to test the impact of such assumptions by undertaking sensitivity analyses. If these worst case/best case analyses converge, then we can have more confidence in the obtained results (Pogue & Yusuf, 1998). On the other hand, approaches to impute missing continuous data in the context of a meta-analysis have received little attention in the methodological literature. One possible approach is to dichotomize the continuous values, so that the above worst case/best case analyses will be applicable. Although dichotomizing continuous outcomes decreases statistical power, it has the additional merit of being easier to interpret clinically. For example, in the case of depression trials, along with the means ± SDs of depression severity measures, some studies report the response rates, usually defined as a 50% or greater reduction in the depression severity from baseline, to assist the clinical interpretation of treatment magnitude.

When studies report rating scale scores only and fail to report response rates, it is theoretically possible to impute response rates, based on reported means ± SDs, by assuming a normal distribution of the rating scale. Some meta-analyses have employed this strategy (Furukawa et al., 2005) but its appropriateness has never been systematically examined. One study aimed to report the results of an empirical examination of such a procedure for depression and anxiety trials. When the response was defined as a more than 50% reduction from baseline depression or anxiety scores and was imputed assuming a normal distribution of the relevant outcome measure, the agreement between the actually observed and the imputed was surprisingly satisfactory not only for individual trials, but also for the meta-analytic summaries. It should be emphasized that the pooled RRs in systematic reviews were virtually identical, including even their 95% confidence intervals, regardless of whether they were based on actually observed values or on those imputed under the normal distribution assumption, and that the clinical conclusions to be drawn from these meta-analyses were therefore not at all affected, even when based on imputed values.

Mixed treatment comparison meta-analysis

It is noteworthy that some systematic reviews have found that certain second-generation antidepressants are more efficacious than other drugs both within and between classes (Hansen et al., 2005; Cipriani et al., 2006; Papakostas et al., 2007). However, these differences are inconsistent across different systematic reviews.

A systematic review conducted by the RTI International-University of North Carolina Evidence-based Practice Centre and the Agency for Healthcare Research and Quality (AHRQ) summarized the available evidence on the comparative efficacy, effectiveness, and harms of 12 second-generation antidepressants and conducted meta-analyses for four direct drug-drug comparisons 62 indirect comparisons between drugs (Gartlehner et al., 2007). Neither direct or indirect comparisons found substantial differences in efficacy among second-generation antidepressants. However, the main limitation of this review is that authors synthesized the literature qualitatively, augmenting findings with quantitative analyses only if head-to-head data were sufficient. By contrast, indirect evidence can be used not only *in lieu* of direct evidence, but also to supplement it (Song et al., 2003).

Moreover, Garlehner et al limited themselves to English language literature and consequently included only a subset of relevant RCTs. MTM is a statistical technique that allows both direct and indirect comparisons to be undertaken, even when two of the treatments have not been directly compared (Higgins et al., 1996; Hasselblad et al., 1998; Lumley, 2002). In other words, it is a generalisation of standard pair-wise meta-analysis for A vs B trials, to data structures that include, for example, A vs B, B vs C, and A vs C trials.

MTM (also known as *network meta-analysis*) can summarise RCTs of several different treatments providing point estimates (together with 95% confidence intervals [CIs]) for their association with a given endpoint, as well as an estimate of incoherence (that is, a measure of how well the entire network fits together, with small values suggesting better internal agreement of the model). MTM has already been used successfully in other fields of medicine (Psaty et al., 2003; Elliott et al., 2007) and two fruitful roles for MTC have been identified (Lu & Ades, 2004):

(i) to strengthen inferences concerning the relative efficacy of two treatments, by including both direct and indirect comparisons to increase precision and combine both direct and indirect evidence (Salanti et al., in press);

(ii) to facilitate simultaneous inference regarding all treatments in order for example to select the best treatment. Considering how important comparative efficacy could be for clinical practice and policy making, it is useful to use all the available evidence to estimate potential differences in efficacy among treatments.

Criteria for considering studies for this review

Types of studies

RCTs comparing one drug with another (head-to-head studies) within the same group of 12 second-generation antidepressants (namely, bupropion, citalopram, duloxetine, escitalopram, fluoxetine, fluvoxamine, milnacipran, mirtazapine, paroxetine, reboxetine, sertraline, and venlafaxine) as monotherapy in the acute phase treatment of depression will be included. We will include only head-to-head active comparisons, excluding placebo arms where present. Trials in which antidepressants were used as an augmentation strategy will be excluded. Quasi-randomized trials (such as those allocating by using alternate days of the week) will be excluded. For trials which have a crossover design only results from the first randomisation period will be considered.

Types of participants

Patients aged 18 or older, of both sexes with a primary diagnosis of depression. Studies adopting any standardised criteria to define patients suffering from unipolar major depression will be included. Most recent studies are likely to have used DSM-IV (APA 1994) or ICD-10 (WHO 1992) criteria. Older studies may have used ICD-9 (WHO 1978), DSM-III (APA 1980)/DSM-III-R (APA 1987) or other diagnostic systems. ICD-9 is not operationalised, because it has only disease names and no diagnostic criteria, so studies using ICD-9 will be excluded. On the other hand, studies using Feighner criteria or Research Diagnostic Criteria will be included. Studies in which less than 20% of the participants may be suffering from bipolar depression will be included.

A concurrent secondary diagnosis of another psychiatric disorder will not be considered as exclusion criteria. Trials in which all participants have a concurrent primary diagnosis of Axis I or II disorders will be excluded. Antidepressant trials in depressive patients with a serious concomitant medical illness will be excluded. RCTs of women with post-partum depression will be also excluded, because post-partum depression appears to be clinically different from major depression (Cooper & Murray, 1998).

Outcome ***measures***

(1) Response to antidepressant treatment

Response is defined as the proportion of patients who show at 8 weeks a reduction of at least 50% on Hamilton Depression Rating Scale (HDRS) (Hamilton, 1960) or Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery, 1979) or who will score 'much improved' or 'very much improved' at the or Clinical Global Impression (CGI) (Guy, 1970), out of the total number of patients randomly assigned to each antidepressant. When all the scores are provided, we will prefer the former measurement for judging response. Furukawa and colleagues have reported the possibility of underreporting the measured outcomes (reporting bias), therefore we will not employ the original author's definitions of response outcomes (Furukawa et al., 2007).

(2) Acceptability of treatment

Treatment discontinuation (acceptability) is defined as the proportion of patients who leave the study early for any reason during the first 8 weeks of treatment, out of the total number of patients randomly assigned to each antidepressant.

Search strategy

All published and unpublished randomized controlled trials that compared the efficacy and acceptability (dropout rate) of one second generation antidepressants with another (see the list of included antidepressants here above) in the treatment of major depression will be identified by searches of the Cochrane Collaboration Depression, Anxiety & Neurosis Review Group Controlled Trials Registers. This register is compiled from systematic and regularly updated searches of Cochrane Collaboration CENTRAL register, AMED, CINAHL, EMBASE, LiLACS, MEDLINE, UK National Research Register, PSYCINFO, PSYNDEX supplemented with hand searching of 12 conference proceedings (Scandinavian Society for Psychopharmacology, Association of European Psychiatrists, Academy of Psychosomatic Medicine, World Psychiatric Association, British Psychological Society, American Psychiatric Association, European College of Neuropsychopharmacology, Society for Psychosomatic Research, First International Symposium on Drugs as Discriminate Stimuli, Stanley Symposia (In Neuropsychobiology), International Society for Traumatic Stress Studies, British Association for Psychopharmacology).

Trial databases of the following drug-approving agencies - (the Food and Drug Administration (FDA) in the USA, the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK, the European Medicines Agency (EMEA) in the EU, the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan, the Therapeutic Goods Administration (TGA) in Australia) and ongoing trial registers (clinicaltrials.gov in the USA, ISRCTN and National Research Register in the UK, Netherlands Trial Register in the Netherlands, EUDRACT in the EU, UMIN-CTR in Japan and the Australian Clinical Trials Registry in Australia) will be hand-searched for published, unpublished and ongoing controlled trials.

No language restrictions will be applied. The following phrase will be used: [*depress** or *dysthymi** or *adjustment disorder** or *mood disorder** or *affective disorder* or *affective symptoms*] and combined with a list of 12 specific second-generation antidepressants (bupropion, citalopram, duloxetine, escitalopram, fluoxetine, fluvoxamine, milnacipran, mirtazapine, paroxetine, reboxetine, sertraline, and venlafaxine). All relevant authors will be contacted to supplement the incomplete report of the original papers. We are aware that there are many trials carried out in China (Chakrabarti et al., 2007). However, for many of these studies only incomplete or conflicting information is available. In an effort to avoid the potential biases that may be introduced by including these trials without further information, we listed them as "awaiting assessment" for transparency.

Study selection and data extraction

Two persons independently reviewed references and abstracts retrieved by the search. If both reviewers agreed that the trial didn't meet eligibility criteria, we excluded it. We obtained the full text of all remaining articles and used the same eligibility criteria to determine which, if any, to exclude at this stage. Any disagreements was solved via discussion with a third member of the reviewing team.

Two reviewers then independently read each article, evaluated the completeness of the data abstraction, and confirmed the quality rating. We designed and used a structured data abstraction form to ensure consistency of appraisal for each study. Information extracted included study characteristics (such as lead author, publication year, journal), participant characteristics (such as diagnostic criteria for depression, age range, setting, diagnosis of bipolar depression), intervention details (such as dose ranges, mean doses of study drugs) and outcome measures (such as the number of patients who responded to treatment and the number of patients who failed to complete the study by any cause). A double-entry procedure was employed by two reviewers.

Length of follow up

In many systematic reviews the ability to provide valid estimates of treatment effect, applicable to the real world, is limited because trials with different durations of follow-up have been combined (Edwards & Anderson, 1999; Geddes et al., 2000; Zimmerman et al., 2002). Clinically, the assessment of efficacy after 6 weeks of treatment or after 16 to 24 weeks or more may lead to wide differences in terms of treatment outcome.

Clinicians need to know whether (and to what extent) treatments work within a clinically reasonable period of time. One recent systematic review of AD clinical trial data, which investigated the issue of early response to ADs, employed a common definition of early response across all included studies (Taylor et al., 2006). Apart from this review however, no systematic reviews have studied the comparative efficacy of ADs in individuals with major depression employing a common definition of acute response that includes a pre-defined follow-up duration. In the present review, acute treatment will be defined as an 8-week treatment in both the efficacy and acceptability analyses (Bauer et al., 2002).

If 8-week data are not available, we used data ranging between 6 to 12 weeks, the time point given in the original study as the study endpoint is given preference.

Quality Assessment

To assess the quality (internal validity) of trials, we used predefined criteria based on those developed by the Cochrane Collaboration. Inadequate concealment undermines the principle of randomization, because participants may then be allocated to a treatment according to prognostic variables rather than by pure chance. Therefore, two independent review authors will independently assess trial quality in accordance with the Cochrane Handbook (Higgins & Green, 2005). This pays particular attention to the adequacy of the random allocation concealment and double blinding (6.11 of the Handbook). Studies will be given a quality rating of A (adequate), B (unclear), and C (inadequate) according to these two items. Studies which scored A or B on these criteria constitute the final list of included studies. In addition, a general appraisal of study quality was made by assessing key methodological issues such as completeness of follow-up and reporting of study withdrawals.

Where inadequate details of allocation concealment and other characteristics of trials were provided, the trial authors were contacted in order to obtain further information. If the raters disagreed, the final rating was made by consensus with the involvement (if necessary) of another member of the review group. Non-congruence in quality assessment was reported as percentage disagreement.

Comparability of dosages

In addition to internal and external validity, we assessed the comparability of dosages. Because we could not find any clear definitions about equivalence of dosages among second-generation antidepressants in the published literature, we used the same roster of low, medium, and high dosages for each drug as Gartlehner and colleagues used in their AHRQ report (Gartlehner et al., 2007) (Table II). This roster was employed to detect inequalities in dosing that could affect comparative effectiveness.

Drug	**Range**	**Low**	**Medium**	**High**
Bupropion	250-450 mg/d	< 300	300-400	> 400
Citalopram	20-60 mg/d	< 30	30-50	> 50
Duloxetine	60-100 mg/d	< 70	70-90	> 90
Escitalopram	10-30 mg/d	< 15	15-25	> 25
Fluoxetine	20-60 mg/d	< 30	30-50	> 50
Fluvoxamine	50-300 mg/d	< 75	75-125	> 125
Milnacipran	50-300 mg/d	< 75	75-125	> 125
Mirtazapine	15-45 mg/d	< 22.5	22.5-37.5	> 37.5
Paroxetine	20-60 mg/d	< 30	30-50	> 50
Reboxetine	4-12 mg/d	< 5	5-9	> 9
Sertraline	50-150 mg/d	< 75	75-125	> 125
Venlafaxine	125-250 mg/d	< 156.25	156.25-218.75	> 218.75

Table II. Dosing classification based on lower to upper dosing range quartiles

Statistical analysis

Considering that clinical trials of antidepressant drugs are usually small and that data distribution is difficult to assess for studies with small samples, in this review priority was be given to the use and analysis of dichotomous variables both for efficacy and acceptability. When dichotomous efficacy outcomes were not reported but baseline mean and endpoint mean and standard deviation of the depression rating scales (such as HDRS or MADRS) were provided, we calculated the number of responding patients at 8 weeks (range 6 to 12 weeks) employing a validated imputation method (Furukawa et al., 2005).

Even though the change scores give more precision (i.e. narrower 95% CI), we used for imputation the endpoint scores for the following reasons: (i) standardised mean difference should focus on standard deviation of endpoint scores (standard deviation of change does not represent population variation); (ii) reporting change may represent outcome reporting bias; (iii) we would need to make up more data to impute standard deviation of change scores; (iv) observed standard deviation of change is about the same as observed standard deviation of endpoint. Where outcome data or standard deviations were not recorded, authors were asked to supply the data. When only the standard error or t-statistics or p values were reported, standard deviations were calculated according to Altman (Altman, 1996). In the absence of data from the authors, the mean value of known standard deviations was calculated from the group of included studies according to Furukawa and colleagues (Furukawa et al., 2006).

We checked that the original standard deviations were properly distributed, so that the imputed standard deviation represented the average.

Responders to treatment were calculated on an intention-to-treat (ITT) basis: drop-outs were always included in this analysis. When data on drop-outs were carried forward and included in the efficacy evaluation (Last Observation Carried Forward, LOCF), they were analysed according to the primary studies; when dropouts were excluded from any assessment in the primary studies, they were considered as drug failures.

To synthesise results, we generated descriptive statistics for trial and study population characteristics across all eligible trials, describing the types of comparisons and some important variables, either clinical or methodological. For each pair-wise comparison between antidepressants, the odds ratio was calculated with a 95% CI. We first performed pair-wise meta-analyses by synthesizing studies that compared the same interventions using a random effects model (DerSimonian & Laird, 1986) to incorporate the assumption that the different studies were estimating different, yet related, treatment effects (Higgins & Green, 2005). Visual inspection of the forest plots was used to investigate the possibility of statistical heterogeneity. This was supplemented using, primarily, the I-squared statistic. This provides an estimate of the percentage of variability due to heterogeneity rather than a sampling error (Higgins et al., 2003).

We conducted a MTM. MTM is a method of synthesizing information from a network of trials addressing the same question but involving different interventions. For a given comparison, say A versus B, direct evidence is provided by studies that compare these two treatments directly. However, indirect evidence is provided when studies that compare A versus C and B versus C are analyzed jointly. The combination of the direct and indirect into a single effect size can increase precision while randomization is respected. The combination of direct and indirect evidence for any given treatment comparison can be extended when ranking more than three types of treatments according to their effectiveness: every study contributes evidence about a subset of these treatments. We performed MTM within a Bayesian framework (Ades et al., 2006). This enables us to estimate the probability for each intervention to be the best for each positive outcome, given the results of the MTM.

The analysis was performed using WinBUGS (MRC Biostatistics Unit, Cambridge, U.K., http://www.mrcbsu cam.ac.uk/bugs/winbugs/contents.shtml).

MTM should be used with caution, and the underlying assumptions of the analysis should be investigated carefully. Key among these is that the network is coherent, meaning that direct and indirect evidence on the same comparisons agree. Joint analysis of treatments can be misleading if the network is substantially incoherent, i.e., if there is disagreement between indirect and direct estimates. So, as a first step, we calculated the difference between indirect and direct estimates in each closed loop formed by the network of trials as a measure of incoherence and we subsequently examined whether there were any material discrepancies. In case of significant incoherence we investigated possible sources of it by means of subgroup analysis. Therefore, we investigated the distribution of clinical and methodological variables that we suspected to be potential sources of either heterogeneity or incoherence in each comparison-specific group of trials.

Results

1. RCT quality

A total of 39 RCTs were included in the efficacy analysis and 74 in the tolerability analysis (a QUOROM diagram is presented here below in Figure 2).

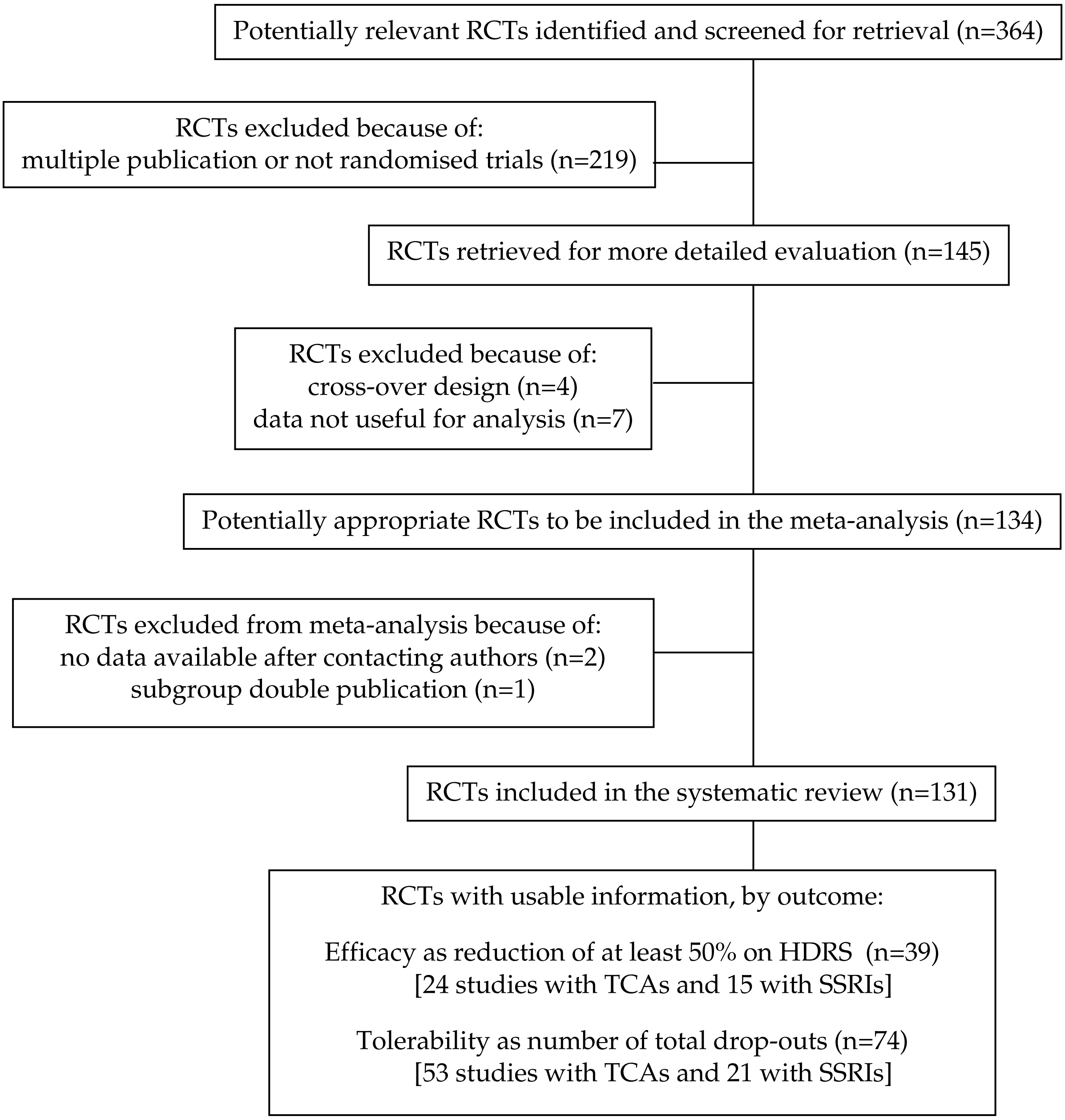

Potentially relevant RCTs identified and screened for retrieval (n=364)
RCTs excluded because of: multiple publication or not randomised trials (n=219)
RCTs retrieved for more detailed evaluation (n=145)
RCTs excluded because of: cross-over design (n=4) data not useful for analysis (n=7)
Potentially appropriate RCTs to be included in the meta-analysis (n=134)
RCTs excluded from meta-analysis because of: no data available after contacting authors (n=2) subgroup double publication (n=1)
RCTs included in the systematic review (n=131)
RCTs with usable information, by outcome:
Efficacy as reduction of at least 50% on HDRS (n=39) [24 studies with TCAs and 15 with SSRIs]
Tolerability as number of total drop-outs (n=74) [53 studies with TCAs and 21 with SSRIs]

Figure 2. Included and excluded studies with reasons (QUOROM flow-diagram).

In terms of efficacy, 24 reports compared flouxetine with TCAs (2256 participants) and 15 compared fluoxetine with other SSRIs (2328 participants). In terms of tolerability, 53 reports compared flouxetine with TCAs (4580 participants) and 21 with other SSRIs (3647 participants). The overall efficacy and tolerability estimates are presented in Table 1. Substantial agreement was found between raters for the Jadad scale (k values ranged from 0.74 to 1.0) and the 3 items of the CONSORT checklist (k values ranged from 0.79 to 1.0). However, only moderate agreement was found for the CCDAN scale, with k values ranging from 0.58 to 1.00. The ANOVA-ICC was 0.98. Funnel plots did not suggest evidence of publication bias. No statistically significant heterogeneity among trials was found.

Relationship between quality and efficacy

In the group of trials comparing fluoxetine with TCAs, the sensitivity analyses, which included high-quality trials according to the Jadad, CCDAN and CONSORT, provided treatment estimates similar to the overall estimate (Table III).

Table III: Subgroup analysis and meta-regression analysis.

Fluoxetine versus:	EFFICACY [failure to respond]						ACCEPTABILITY [failure to complete]					
	TCAs			Other SSRIs			TCAs			Other SSRIs		
	Peto OR* (95% CI)		RCTs [patients]	Peto OR* (95% CI)		RCTs [patients]	Peto OR* (95% CI)		RCTs [patients]	Peto OR* (95% CI)		RCTs [patients]
Overall estimate	.98 (.82 to 1.16)		24 [2256]	1.26 (1.05 to 1.50)		15 [2328]	.77 (.68 to .88)		53 [4580]	1.02 (.87 to 1.20)		21 [3647]
Jadad rating scale (high quality RCTs)	.98 (.82 to 1.18)		20 [2055]	1.25 (1.02 to 1.53)		10 [1614]	.81 (.69 to .94)		41 [3487]	1.01 (.85 to 1.20)		16 [2979]
CCDAN rating scale (high quality RCTs)	.96 (.78 to 1.20)		11 [1353]	1.24 (1.04 to 1.50)		12 [1995]	.65 (.53 to .80)		16 [1914]	1.00 (.85 to 1.18)		18 [3127]
CONSORT (Items 7-8-9) (high quality studies)	1.31 (.72 to 2.40)		2 [182]	**-**		0	.96 (.59 to 1.54)		4 [351]	.88 (.61 to 1.27)		3 [773]
META-REGRESSION**	**Coeff. (95% CI)**	**z**	**p**	**Coeff. (95% CI)**	**z**	**p**	**Coeff. (95% CI)**	**z**	**p**	**Coeff. (95% CI)**	**z**	**p**
JADAD rating scale (continuous variable)	.18 (- .43 to .80)	.58	.565	- .10 (- .63 to .43)	- .37	.712	- .03 (- .54 to .47)	- .13	.894	.09 (- .25 to .43)	.54	.593
CCDAN Rating Scale (continuous variable)	- .06 (- .13 to .006)	- 1.77	.077	- .01 (- .13 to .10)	- .23	.821	- .05 (- .11 to .002)	- 1.88	.060	- .03 (- .08 to .02)	-1.10	.271
CONSORT (Items 7-8-9) (continuous variable)	.26 (- .66 to 1.20)	.57	.570	-	-	-	.12 (- .43 to .69)	.45	.656	- .24 (- .78 to .28)	- .90	.367

TCAs = tricyclic antidepressants; SSRIs = selective-serotonin reuptake inhibitors; OR = odds ratio; CI = confidence interval.

* Peto OR (95% CI) < 1 favours fluoxetine; > 1 favours control antidepressants.

** Dependent variable: Peto OR (95% CI). Positive coefficients indicate that explanatory variables were correlated with higher treatment estimates; positive upper and lower limits of confidence intervals indicate a statistically significant positive association. Negative coefficients indicate that explanatory variables were correlated with lower treatment estimates; negative upper and lower limits of confidence intervals indicate a statistically significant negative association. Meta-regression adjusted for the following terms: year of publication (continuous variable), age (1 = adults; 0 = adults and/or elderly subjects), setting (1 = inpatients; 0 = outpatients), fluoxetine dose (continuous outcome) and wish bias (1 = experimental drug; 0 = reference drug).

An upside down pyramid-shaped trend was observed, in the sense that most RCTs were of high quality according to the Jadad scale, 11 RCTs were of high quality according to the CCDAN, and only 2 RCTs were of high quality according to the CONSORT items. In the group of trials comparing fluoxetine with other SSRIs, the sensitivity analyses, which included high-quality trials according to the Jadad and CCDAN scales, provided treatment estimates similar to the overall estimate, while no high-quality RCTs were detected according to the CONSORT items.

Relationship between quality and tolerability

In the group of trials comparing fluoxetine with TCAs, the sensitivity analyses, which included high-quality trials according to the Jadad, CCDAN and CONSORT, provided treatment estimates similar to the overall estimate (Table II). Similarly to the relationship between quality and efficacy, an upside down pyramid-shaped trend was observed: most of RCTs were of high quality according to the Jadad scale, 16 RCTs were of high quality according to the CCDAN, and only 4 RCTs were of high quality according to the CONSORT items. In the group of trials comparing fluoxetine with other SSRIs, the sensitivity analyses, which included high-quality trials according to the Jadad, CCDAN and CONSORT, provided treatment estimates similar to the overall estimate (Table II).

Meta-regression analysis

A meta-regression analysis was carried out to investigate whether quality of primary studies was associated with treatment effect, after possible confounders were controlled for (Table II). Negative estimates indicate that the covariates included in the meta-regression model were inversely correlated with outcome. The meta-regression analysis showed that quality, measured with the Jadad, CCDAN and CONSORT, was not correlated with efficacy and tolerability outcomes (Table II).

2. Quality of systematic reviews (and meta-analyses)

Description of the available data

Twelve systematic reviews dealing with twelve different antidepressant treatments were included in this study. In the reporting of the results, these have been coded in different ways (to facilitate the various methods of analysis).

The codes were as follows:

1	A	201	paroxetine
2	B	202	sertraline
3	C	203	citalopram
4	D	206	escitalopram
5	E	207	fluoxetine
6	F	208	fluvoxamine
7	G	302	milnacipran
8	H	303	venlafaxine
9	I	307	reboxetine
10	J	308	bupropion
11	K	311	mirtazapine
12	L	314	duloxetine

There were 111 trials in total for outcome ‘response”, (109 two-arm trials, 2 with three arms). Equivalently, there were 112 studies for outcome “dropouts” (110 two-arm trials, 2 with three arms). Figures "network R" and "network D" show the networks for each outcome (Figures 3 and 4).

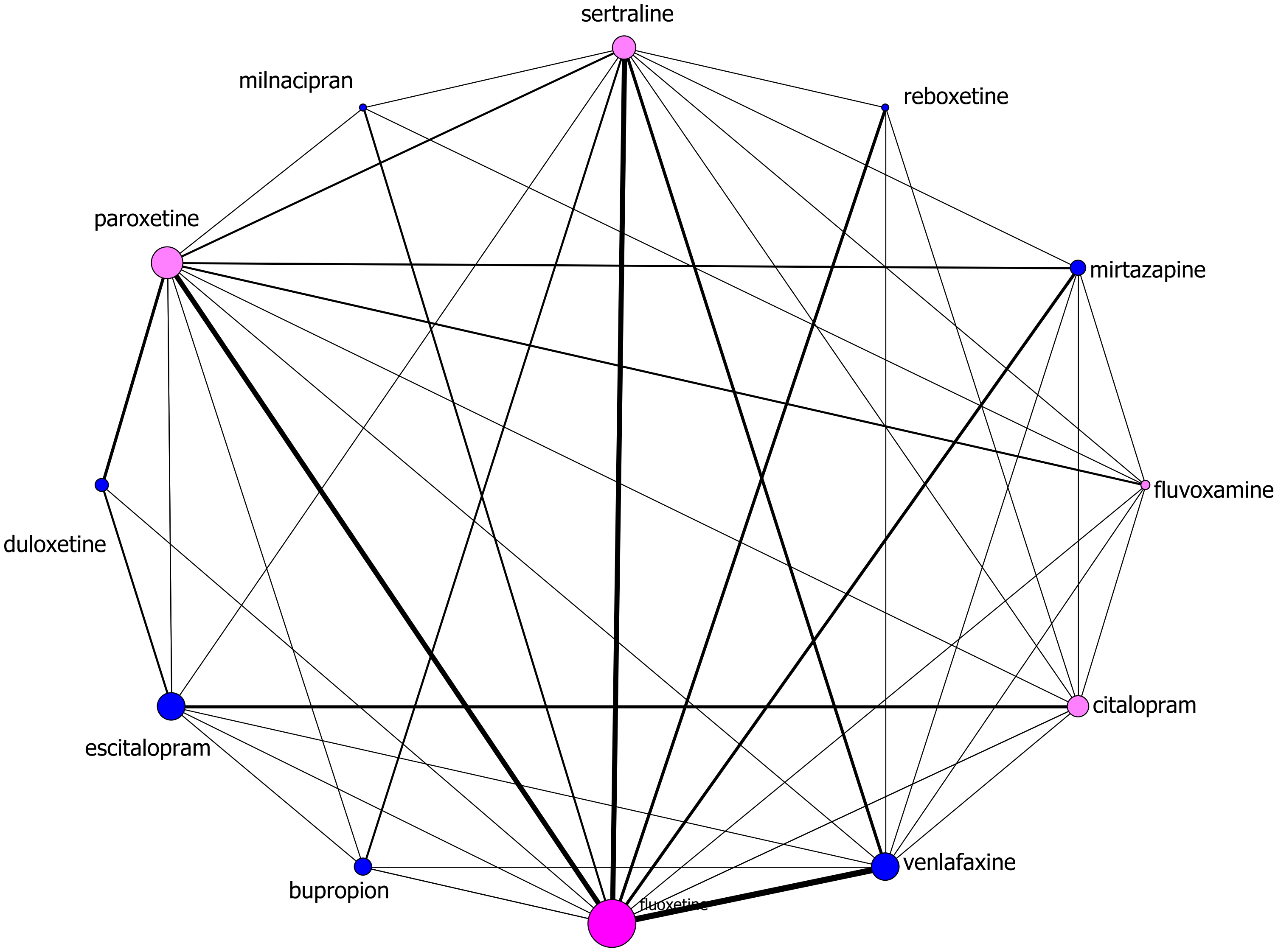


Figure 3: "Network R", network of trials reporting response rates

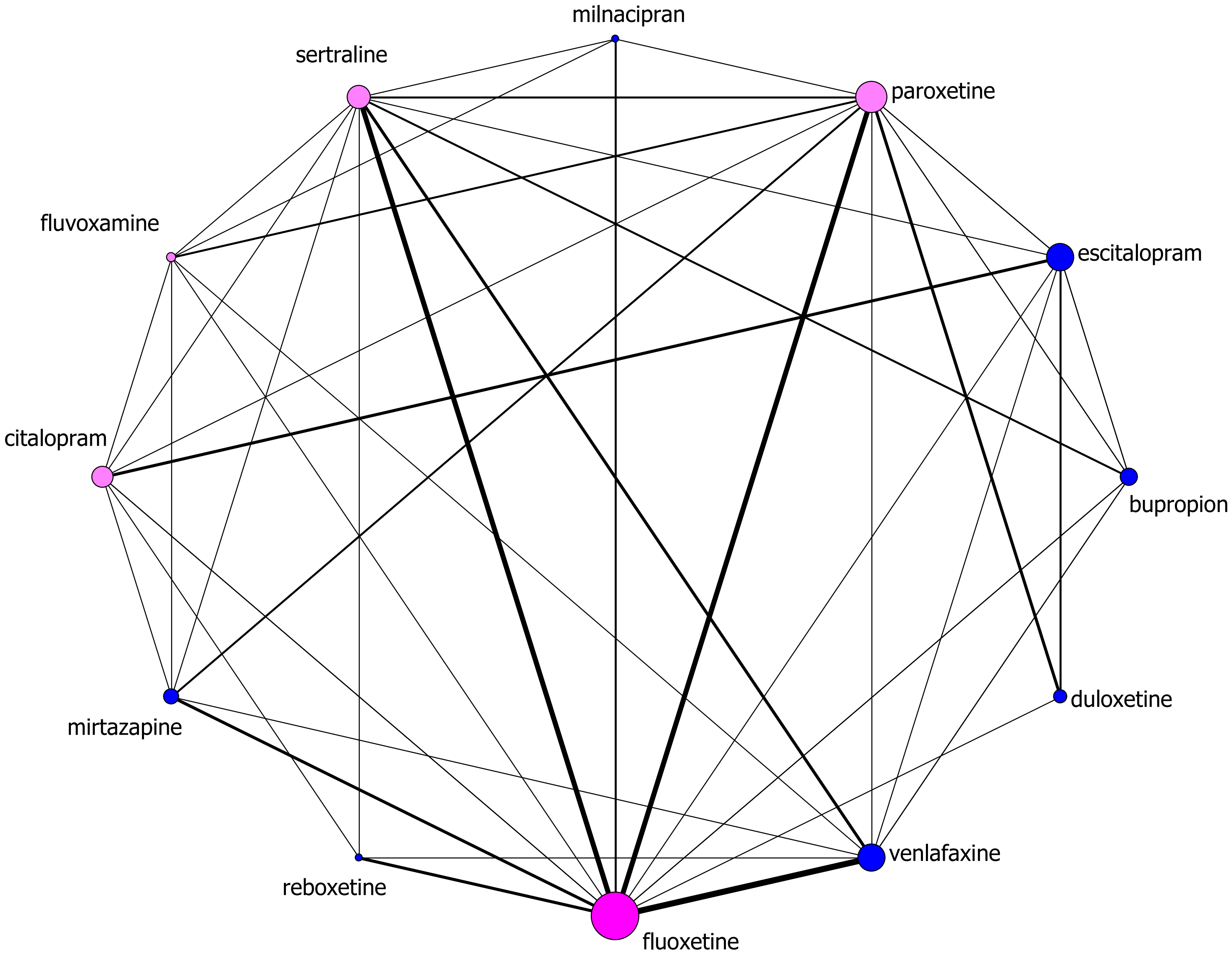


Figure 4: "Network D", network of trials reporting dropout rates

The width of lines is proportional to the number of trials comparing pairs of treatments and the size of each node is proportional to the sample size (participants). In blue are the nodes that are believed to be favored by sponsorship bias (we decided to add a score of -1, 0 or +1 if sponsorship bias is absent, unclear or present, respectively). Here below there is a more detailed description of the distribution of year and sponsorship bias (Table IV and V).

ID	NAMES	MEDIAN YEAR	SPONSORSHIP
201	paroxetine	2000.0	- 1
202	sertraline	2000.0	- 4
203	citalopram	2002.0	- 5
206	escitalopram	2006.0	13
207	fluoxetine	1999.0	- 42
208	fluvoxamine	1998.0	- 2
302	milnacipran	2000.5	5
303	venlafaxine	2001.5	7
307	reboxetine	2003.5	1
308	bupropion	2006.0	10
311	mirtazapine	2002.0	12
314	duloxetine	2006.5	4

Table IV: distribution of year and sponsorship bias per treatment.

TREATMENT	MIN.	1ST QU.	MEDIAN	MEAN	3RD QU.	MAX.
Paroxetine	1993	1998	2000	2001	2006	2007
Sertraline	1993	1998	2000	2000	2003	2007
Citalopram	1993	1999	2002	2002	2005	2007
Escitalopram	2000	2005	2006	2005	2007	2007
Fluoxetine	1991	1997	1999	2000	2003	2007
Fluvoxamine	1993	1995	1998	1998	2002	2006
Milnacipran	1994	1999	2001	2000	2002	2003
Venlafaxine	1994	1999	2002	2002	2005	2007
Reboxetine	1997	2001	2004	2003	2005	2006
Bupropion	1991	1999	2006	2003	2007	2007
Mirtazapine	1997	2000	2002	2002	2003	2005
Duloxetine	2002	2004	2007	2006	2007	2007

Table V: Distribution of the year per treatment.

Analysis of coherence

The analysis of coherence indicated that there were 3 incoherent loops (for full details on analysis of coherence - see Appendix)

□ *For Response (70 loops)*

201-203-206 (paroxetine – citalopram – escitalopram)

208-303-311 (fluvoxamine – venlafaxine – mirtazapine)

202-207-308 (sertraline – fluoxetine – bupropion)

□ *For Dropouts (63 loops)*

208-303-311 (fluvoxamine – venlafaxine – mirtazapine)

202-203-207 (sertraline – citalopram - fluoxetine)

202-203-206 (sertraline – citalopram – escitalopram)

1. *Multiple-treatments meta-analysis: original data*

Table VI shows the relative ORs (and standard deviations) for both outcomes (response and dropout), using fluoxetine as reference drug.

	Low	**OR**	**high**
Paroxetine	0,86	0,98	1,12
Sertraline	**0,69**	**0,80**	**0,93**
Citalopram	0,76	0,91	1,08
Escitalopram	**0,65**	**0,76**	**0,89**
Fluvoxamine	0.80	1.02	1.29
Milnacipran	0.74	0.99	1.311
Venlafaxine	**0.68**	**0.78**	**0.90**
Reboxetine	**1.16**	**1.48**	**1.90**
Bupropion	0.77	0.92	1.107
Mirtazapine	**0.60**	**0.73**	**0.87**
Duloxetine	0.80	1.01	1.27

Figure 5 shows the ranking distribution for response (solid lines) and for dropout (dotted lines) for each treatment.

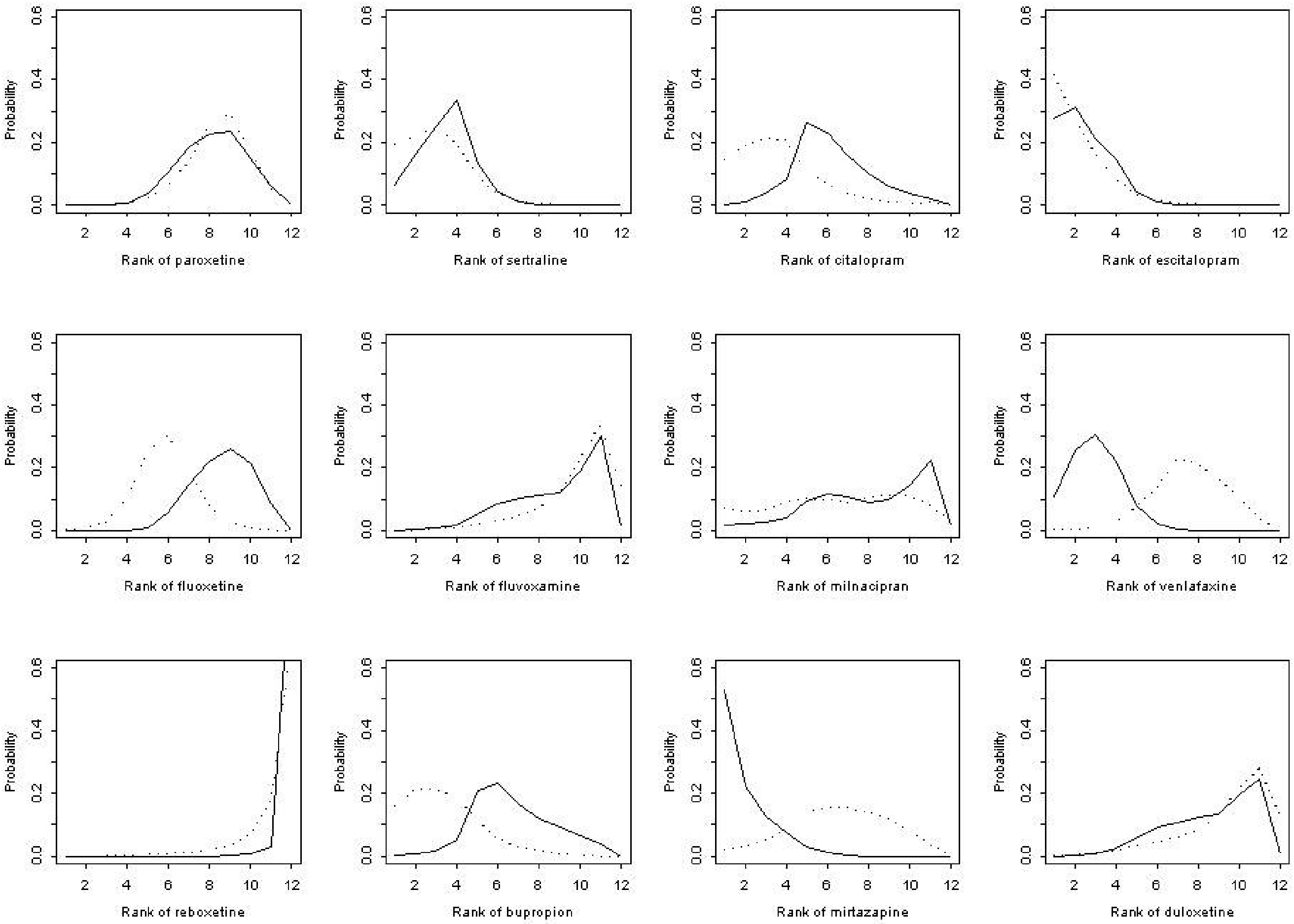


Figure 5: Ranking distribution for response (solid lines) and for dropout (dotted lines) for each treatment

The cumulative ranking which makes the comparison of the treatments possible, is presented in Figure 6.

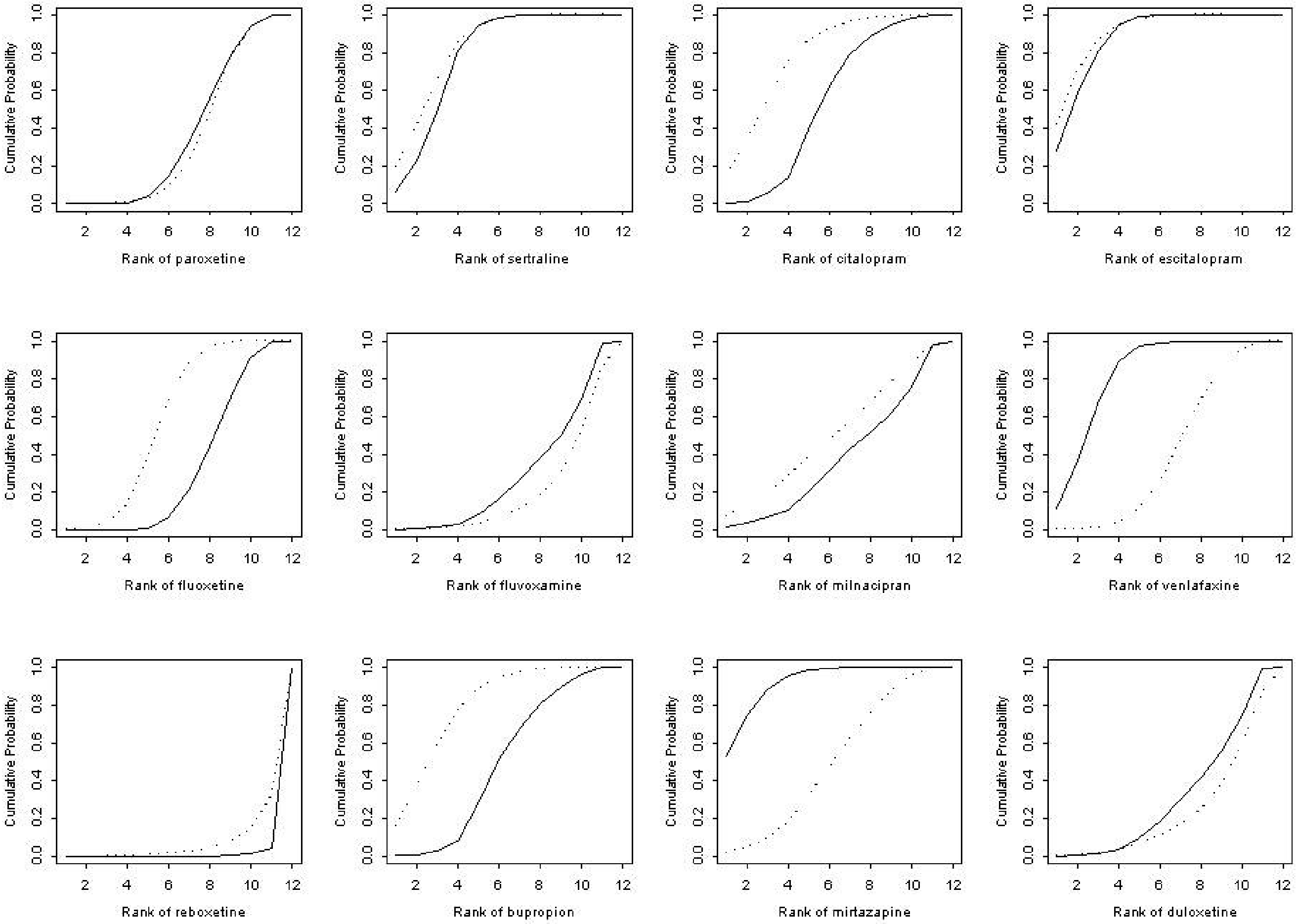


Figure 6: Cumulative ranking distribution for response (solid lines) and for dropout (dotted lines) for each treatment

2. *Multiple-treatments meta-analysis: adjusting for sponsorship bias*

An arm-specific variable "sponsorship" (denoted as *S*) has been added to the data taking values 1 (for drug sponsored), -1 (for drug being the comparator) and 0 if there is no sponsorship in the trial.

Then, the success probabilities for two drugs A and B in a study *i* the model has been modified as

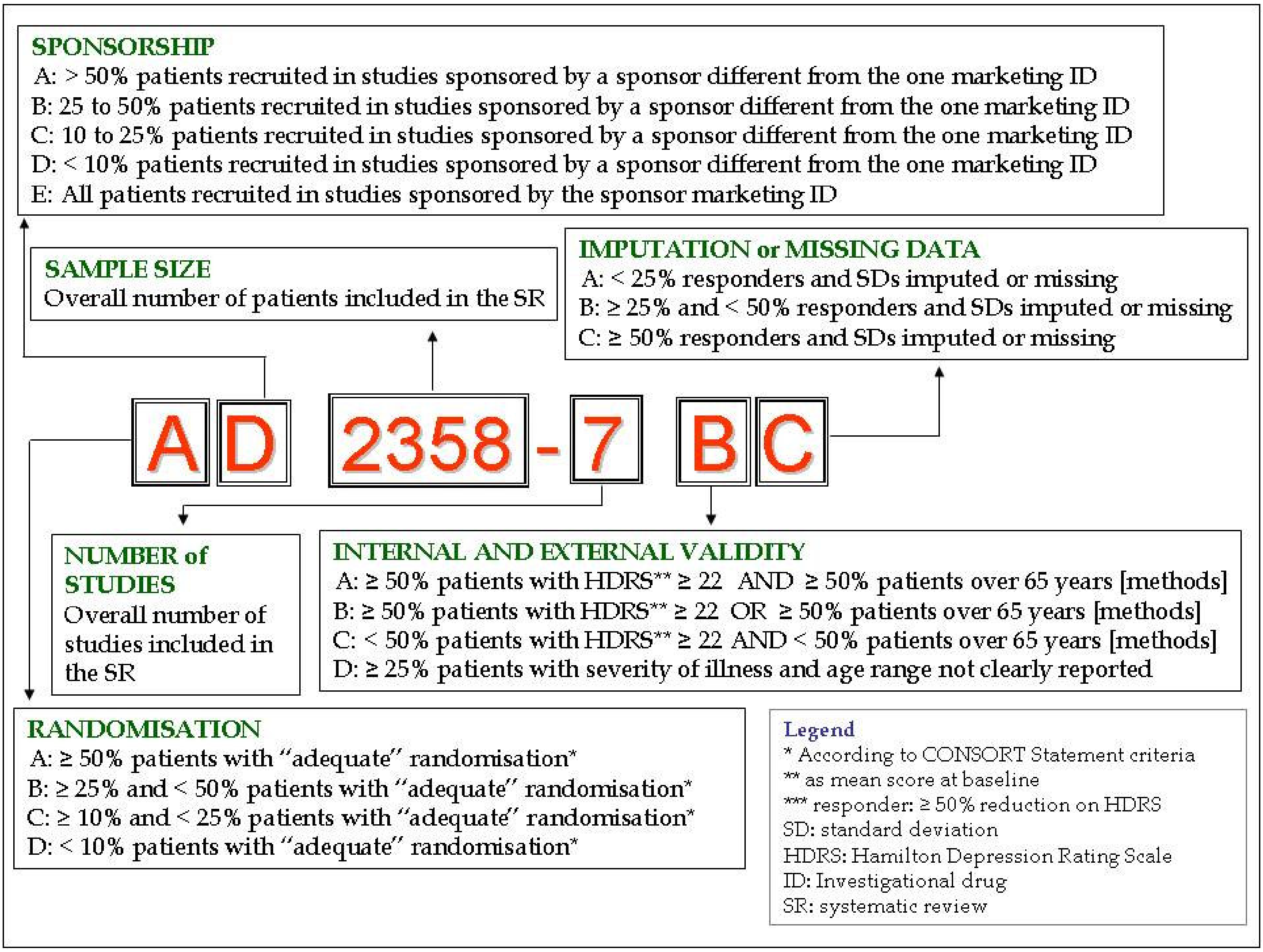
$$\log\text{it}(p_{Ai}) = u_i + \frac{\beta}{2} \cdot S_{Ai}$$

$$\log\text{it}(p_{Bi}) = u_i + \delta_{BvsA} + \frac{\beta}{2} \cdot S_{Bi}$$

Say a trial is sponsored by the manufacturer of drug B. Then S_{Bi}=1 and S_{Ai}= -1, Therefore, Log Odds Ratio (LOR) is a s follows: $LOR_{BvsA} = \delta_{BvsA} + \beta$. Inversely, if drug A is sponsored, $LOR_{BvsA} = \delta_{BvsA} - \beta$ I placed a vague normal prior on β, truncated on zero (to reflect the strong belief that there is bias). The ORs did not change all that much.

The surface under the curve becomes smaller for those drugs that are sponsored and the comparators do better in the ranking. However, if the coefficient β is allowed to take negative values, the posterior credible interval contains the zero value, so there is no clear statistical evidence of bias.

3. *Graphical representation of the five important issues strictly related to quality of RCTs and systematic reviews*



SPONSORSHIP
A: > 50% patients recruited in studies sponsored by a sponsor different from the one marketing ID
B: 25 to 50% patients recruited in studies sponsored by a sponsor different from the one marketing ID
C: 10 to 25% patients recruited in studies sponsored by a sponsor different from the one marketing ID
D: < 10% patients recruited in studies sponsored by a sponsor different from the one marketing ID
E: All patients recruited in studies sponsored by the sponsor marketing ID

SAMPLE SIZE
Overall number of patients included in the SR

IMPUTATION or MISSING DATA
A: < 25% responders and SDs imputed or missing
B: ≥ 25% and < 50% responders and SDs imputed or missing
C: ≥ 50% responders and SDs imputed or missing

NUMBER of STUDIES
Overall number of studies included in the SR

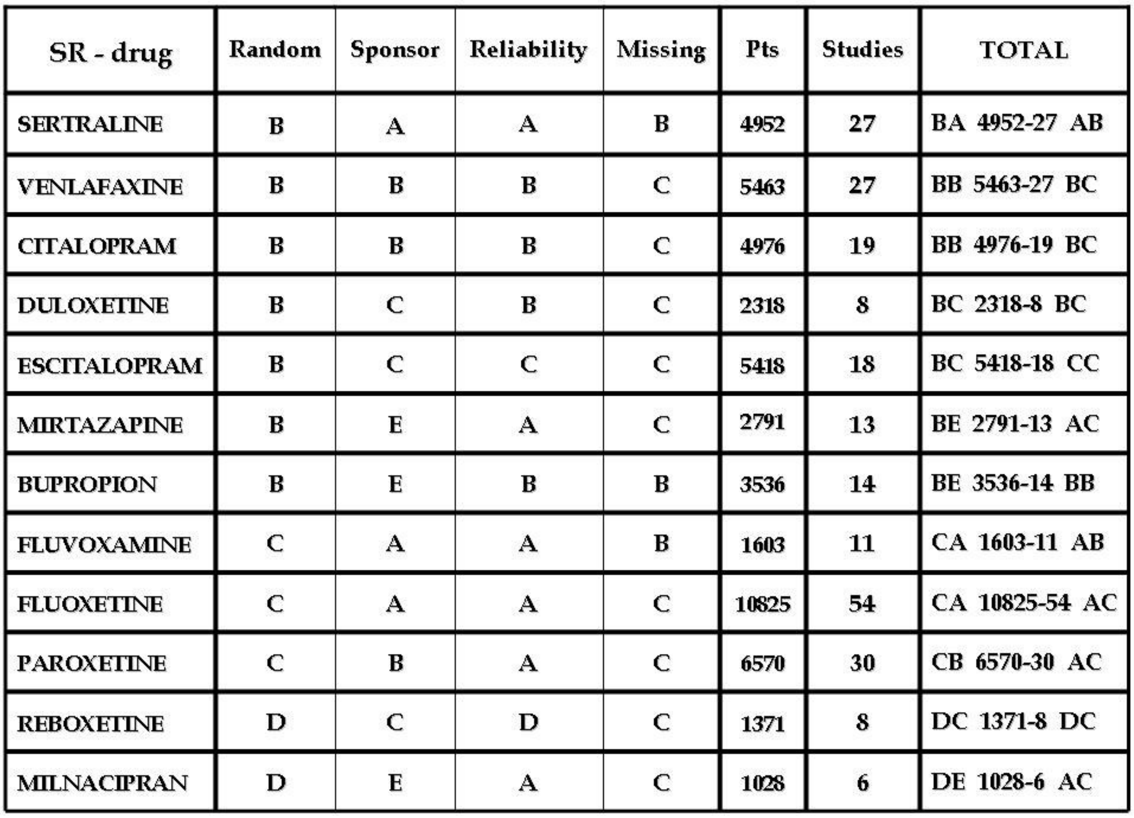
INTERNAL AND EXTERNAL VALIDITY
A: ≥ 50% patients with HDRS** ≥ 22 AND ≥ 50% patients over 65 years [methods]
B: ≥ 50% patients with HDRS** ≥ 22 OR ≥ 50% patients over 65 years [methods]
C: < 50% patients with HDRS** ≥ 22 AND < 50% patients over 65 years [methods]
D: ≥ 25% patients with severity of illness and age range not clearly reported

RANDOMISATION
A: ≥ 50% patients with "adequate" randomisation*
B: ≥ 25% and < 50% patients with "adequate" randomisation*
C: ≥ 10% and < 25% patients with "adequate" randomisation*
D: < 10% patients with "adequate" randomisation*

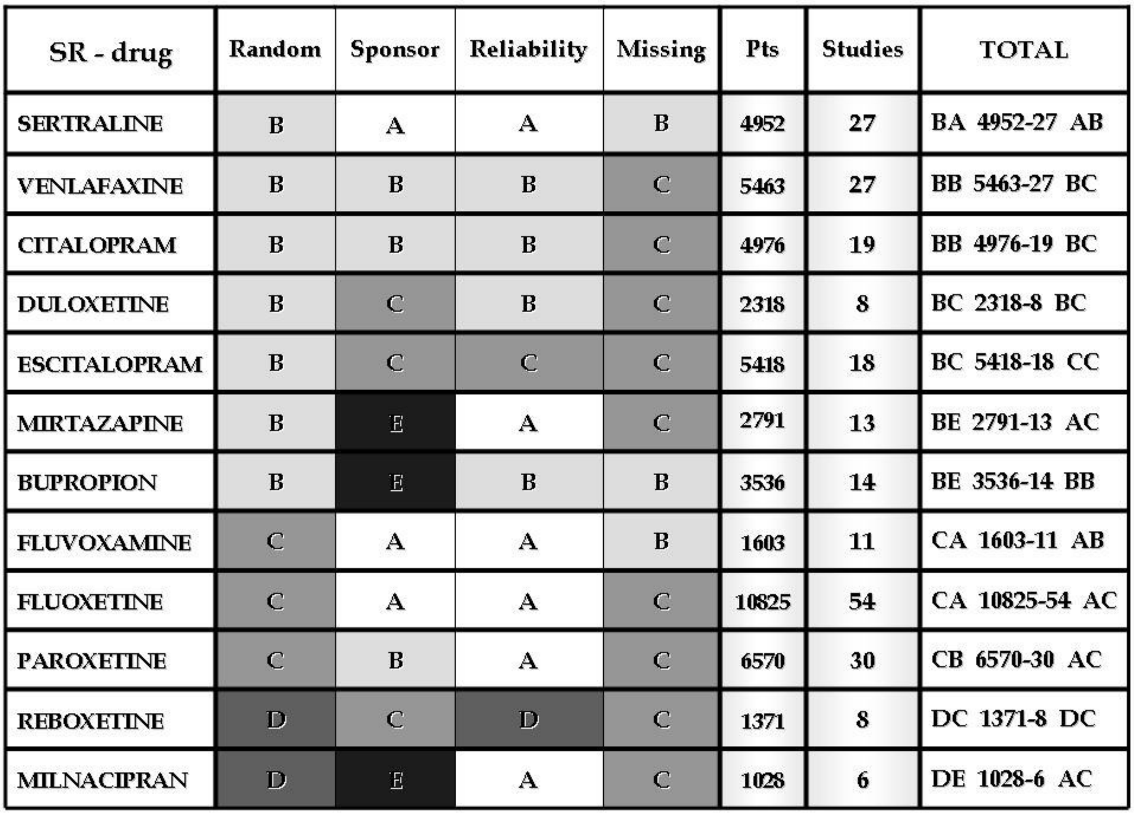
Legend
* According to CONSORT Statement criteria
** as mean score at baseline
*** responder: ≥ 50% reduction on HDRS
SD: standard deviation
HDRS: Hamilton Depression Rating Scale
ID: Investigational drug
SR: systematic review

Figure 7: Graphical representation of the five important issues (see Methods)

Figure 7 shows the Graphical representation of the five important issues discussed in the Method section of the present study. Following these criteria, the scoring of the 12 systematic reviews included in the present study are as follows in Table VII.



SR - drug	Random	Sponsor	Reliability	Missing	Pts	Studies	TOTAL
SERTRALINE	B	A	A	B	4952	27	BA 4952-27 AB
VENLAFAXINE	B	B	B	C	5463	27	BB 5463-27 BC
CITALOPRAM	B	B	B	C	4976	19	BB 4976-19 BC
DULOXETINE	B	C	B	C	2318	8	BC 2318-8 BC
ESCITALOPRAM	B	C	C	C	5418	18	BC 5418-18 CC
MIRTAZAPINE	B	E	A	C	2791	13	BE 2791-13 AC
BUPROPION	B	E	B	B	3536	14	BE 3536-14 BB
FLUVOXAMINE	C	A	A	B	1603	11	CA 1603-11 AB
FLUOXETINE	C	A	A	C	10825	54	CA 10825-54 AC
PAROXETINE	C	B	A	C	6570	30	CB 6570-30 AC
REBOXETINE	D	C	D	C	1371	8	DC 1371-8 DC
MILNACIPRAN	D	E	A	C	1028	6	DE 1028-6 AC



SR - drug	Random	Sponsor	Reliability	Missing	Pts	Studies	TOTAL
SERTRALINE	B	A	A	B	4952	27	BA 4952-27 AB
VENLAFAXINE	B	B	B	C	5463	27	BB 5463-27 BC
CITALOPRAM	B	B	B	C	4976	19	BB 4976-19 BC
DULOXETINE	B	C	B	C	2318	8	BC 2318-8 BC
ESCITALOPRAM	B	C	C	C	5418	18	BC 5418-18 CC
MIRTAZAPINE	B	E	A	C	2791	13	BE 2791-13 AC
BUPROPION	B	E	B	B	3536	14	BE 3536-14 BB
FLUVOXAMINE	C	A	A	B	1603	11	CA 1603-11 AB
FLUOXETINE	C	A	A	C	10825	54	CA 10825-54 AC
PAROXETINE	C	B	A	C	6570	30	CB 6570-30 AC
REBOXETINE	D	C	D	C	1371	8	DC 1371-8 DC
MILNACIPRAN	D	E	A	C	1028	6	DE 1028-6 AC

Table VII: Scoring of the 12 systematic reviews according to quality criteria.

Discussion

This study found no correlation between the methodological quality of reports of RCTs and treatment estimates of efficacy and tolerability. The subgroup analyses, which included high-quality trials only, provided treatment estimates that did not materially change from overall estimates. This finding was further confirmed by the meta-regression analysis, which indicated that measures of quality, after potential confounders were controlled for, were not correlated with treatment estimates. While high quality reports, according to the Jadad and CCDAN scales, tended to replicate overall estimates, the CONSORT component approach to quality assessment was able to identify only a selected minority of studies, making this way of sensitivity analysis less meaningful.

The main limitation is that quality of reporting is often used as a proxy measure for methodological quality, although similar quality of reporting may hide important differences in methodological quality (Huwiler-Muntener, 2002). By contrast, absence of association between quality scores and treatment estimates may have several interpretations. It is possible that no association exists between any of the components of the score and treatment effect, or an association with a single component might have been diluted by lack of association with other components (Greenland, 1994). Lastly, two components might be associated with treatment effect but in opposite directions (Greenland, 1994). For this reason, in this systematic review quality of primary studies was assessed by either validated rating scales or individual items, taking into account all other possible estimate confounders.

Another limitation of this study refers to the possibility that the rigorous Cochrane procedure for systematic reviews, which presumes the exclusion of RCTs reporting no outcome information, systematically selected a sample of trials quite homogeneous from a qualitative viewpoint. This might partly explain the difficulty of establishing a clear association between trial quality and outcome. However, the included trials have involved different comparator drugs, different doses, different follow-up periods. We found not statistically significant heterogeneity and estimates were controlled for possible confounding variables.

Although from a theoretical viewpoint meta-analyses should take quality into consideration to provide less biased treatment estimates, it is additionally possible that, in this specific field of medicine, these quality measures may not be suitable when quality needs to be incorporated into the meta-analytical process of summarising trial results (Juni et al., 2001; Juni et al., 1999). The Jadad scale is very focused on key trial characteristics, such as randomization, masking, dropouts and withdrawals, but obviously does not cover other trial features leading to important methodological flaws (above all allocation concealment which has been most consistently found to be associated with exaggeration of treatment effect estimates) (Moher et al., 1999). According to JADAD scores, we showed that almost all AD trials fell in the high-quality category, and therefore meta-analysts can hardly use this scale as a weighting tool.

The CCDAN quality score has the positive characteristic of covering a very wide range of aspects associated with the conduct and reporting of clinical trials, representing this way a suitable instrument when a general description of quality is warranted. However, it is striking to note that while a substantial proportion of RCTs comparing fluoxetine versus tricyclic ADs were of low methodological quality on the basis of the CCDAN rating, the majority of recent trials, comparing fluoxetine versus other SSRIs, were of high methodological quality on the basis of the CCDAN rating. This was explained by better reporting of some ancillary information on study design and trial characteristics, and not by better reporting of key details on randomization and its concealment. Unfortunately, the CCDAN checklist does not adopt any weighting procedure in the calculation of the overall quality score, i.e. all items equally contribute to the final score. Therefore, items investigating the randomization procedure or the concealment of allocation are given the same weight received by items investigating side-effect reporting or evaluating the reporting of patients' demographic characteristics. Finally, the CONSORT component approach does not differentiate high- from low-quality studies among AD trials. In this sample of trials we found no evidence of a significant association between quality and treatment estimate using three of many possible quality measures. The most likely explication of these findings is that up to now current quality measures are not related with treatment estimates in AD trials and may not be useful weighting tools when meta-analyses of data extracted from AD RCTs are carried out.

Regarding sponsorship, interestingly some authors attempted to quantify the extent of industry sponsorship and financial conflict of interest in reports of clinical trials in the four general psychiatric journals with the greatest citation impact factors that commonly publish such studies (Perlis et al., 2005). This is one of the first recent examinations of conflict of interest specifically in the psychiatric literature and the authors also assessed the possible relationship between such conflict and study design and reporting. They found that financial conflict of interest is prevalent among clinical trials published in four widely cited general psychiatric journals. The study identified industry funding in 60% of the trials; studies of general medical journals have revealed rates of 40% to 66%. The prevalence of studies with author conflict of interest in psychiatry journals (47%) was slightly higher than the rates found in general medical journals (34%–43%).

The relationship between financial conflict of interest and positive outcome is consistent with prior reports in the general medical literature. One previous report did note differences in articles about sertraline written by medical communications companies compared to those without this affiliation, providing some support for the hypothesis that industry involvement influences reporting. Industry sponsorship and author conflict of interest are prevalent and do appear to affect study outcomes. Given this prevalence and the potential influence on the general psychiatric literature, it will be critical to obtain a better understanding of the ways in which industry funding or the presence of conflict of interest influences the design, conduct, and/or reporting of clinical trials. Strategies to ensure that conflict of interest is disclosed consistently and completely and registries to ensure that all clinical trials, regardless of outcome, are reported should be considered in psychiatry as in other areas of medicine.

Empirical evidence of the bias associated with failure to conceal the allocation and explicit requirement to discuss this issue in the CONSORT statement seem to be leading to wider recognition that allocation concealment is an essential aspect of a randomised trial. Allocation concealment is completely different from (double) blinding (Cipriani et al., in press). It is possible to conceal the randomisation in every randomised trial. Also, allocation concealment seeks to eliminate selection bias (who gets into the trial and the treatment they are assigned). By contrast, blinding relates to what happens after randomisation, is not possible in all trials, and seeks to reduce ascertainment bias (assessment of outcome).

Results form this study highlighted once more the need for reliable tools to assess quality of the retrieved evidence, incorporating quality into the assessment of treatment effects. No standard procedures are available up to now, however the field of antidepressant trials has shown to be a good example on how to build an hierarchical pattern of summarised evidence to better inform clinical practice.

References

Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, Lu G. (2006). Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics* 24, 1-19.

Altman DG, Bland JM. (1996). Detecting skewness from summary information. *British Medical Journal* 313, 1200.

Altman DG. (2005). Endorsement of the CONSORT statement by high impact medical journals: survey of instructions for authors. *British Medical Journal* 330, 1056-7.

Balk EM, Bonis PA, Moskowitz H. (2002). Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *Journal of American Medical Association* 287, 2973–82.

Barbui C, Cipriani A, Brambilla P, Hotopf M. (2004). "Wish bias" in antidepressant drug trials? *Journal of Clinical Psychopharmacology* 24,126-30.

Bauer M, Whybrow PC, Angst J, Versiani M, Moller HJ; World Federation of Societies Biological Psychiatry Task Force on Treatment Guidelines for Unipolar Depressive Disorders. World Federation of Societies of Biological Psychiatry (2002). Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 1: Acute and continuation treatment of major depressive disorder. *World Journal of Biological Psychiatry* 3, 5-43.

Begg C, Cho M, Eastwood S. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Journal of American Medical Association* 276, 637-639.

Chalmers TC, Celano P, Sacks HS, Smith H Jr. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* 309, 1358-1361.

Chalmers I, Adams M, Dickersin K. (1990). A cohort study of summary reports of controlled trials. *Journal of American Medical Association* 263, 1401-1405.

Chakrabarti A, Adams CE, Rathbone J, Wright J, Xia J, Wong W, Von Reibnitz P, Koenig C, Baier S, Pfeiffer C, Blatter J, Mantz M, Kloeckner K. (2007). Schizophrenia trials in China: a survey. *Acta Psychiatrica Scandinavica* 116, 6-9.

Choi PT, Halpern SH, Malik N, Jadad AR, Tramer MR, Walder B. (2001). Examining the evidence in anesthesia literature: a critical appraisal of systematic reviews. *Anesthesiology and Analgesia* 92, 700–9.

Cipriani A, Barbui C, Brambilla P, Geddes J. (2006). Are all antidepressants really the same? The case of fluoxetine: a systematic review. *Journal of Clinical Psychiatry* 67, 850-864.

Cipriani A, Nosè, Barbui C. Allocation concealment and blinding in clinical trials. *Epidemiologia e Psichiatria Sociale*, in press.

Cooper PJ, Murray L. (1998). Postnatal depression. *British Medical Journal* 316, 1884-6.

DerSimonian R, Laird N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* 7, 177-88

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology* 45, 255-265.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. (1992). Publication bias in clinical research. *Lancet* 337, 867-72.

Edwards JG, Anderson I. (1999). Systematic review and guide to selection of selective serotonin reuptake inhibitors. *Drugs* 57, 507-33.

Egger M & Smith GD. (1995). Misleading meta-analysis. *British Medical Journal* 310, 752-4.

Egger M, Davey Smith G, Schneider M, Minder C. (1997a) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 315, 629-34.

Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C & Antes G. (1997b). Language bias in randomised controlled trials published in English and German. *Lancet* 350, 326-9.

Elliott WJ, Meyer PM. (2007). Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *Lancet* 369, 201-7.

Emerson JD, Burdick E, Hoaglin DC. (1990). An empirical study of the possible relation of treatment and standard differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials* 11, 339-52

Furukawa TA, Cipriani A, Barbui C, Brambilla P, Watanabe N. (2005). Imputing response rates from means deviations in meta-analyses. *International Clinical Psychopharmacology* 20, 49-52

Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology* 59, 7-10.

Furukawa TA, Watanabe N, Omori IM. (2007). Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. *Journal of American Medical Association* 297, 468-70.

Gartlehner G, Hansen RA, Thieda P, DeVeaugh-Geiss AM, Gaynes BN, Krebs EE, Lux LJ, Morgan LC, Shumate JA, Monroe LG, Lohr KN. (2007). Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Geddes JR, Freemantle N, Mason J, Eccles MP, Boynton J. (2000). Selective serotonin reuptake inhibitors (SSRIs) versus other antidepressants for depression. The Cochrane Library (Cochrane Review). Issue 2.

Gluud LL. (2006). Bias in clinical intervention research. *American Journal of Epidemiology* 163, 493-501.

Goodman SN, Berlin JA, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. 1994;121:11-21.

Gotzsche PC. (1987). Reference bias in reports of drug trials. *British Medical* 295, 654-6.

Gøtzsche PC. (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials* 10, 31-56.

Greenland S. (1994a). Can meta-analysis be salvaged? *American Journal of Epidemiology* 140, 783-7.

Greenland S. (1994b). Invited commentary: A critical look at some popular meta-analytical methods. *American Journal of Epidemiology* 140, 290-296.

Grimes DA, Schulz KF. (1996). Determining sample size and power in clinical trials: the forgotten essential. *Seminars on Reproductive Endocrinology* 14, 125-31.

Guy W, Bonato RR. (1970). *Manual for the ECDEU Assessment Battery.2*. Chevy Chase, MD: National Institute of Mental Health.

Hamilton M. (1960). A rating scale for depression. *Journal of Neurology Neurosurgery and Psychiatry* 23, 56-62.

Hansen RA, Gartlehner G, Lohr KN, Gaynes BN, Carey TS. (2005). Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. *Annals of Internal Medicine* 143, 415-26.

Hasselblad V. (1998). Meta-analysis of multi-treatment studies. *Medical Decision Making* 18, 37-43.

Higgins JP, Whitehead A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 15, 2733-49.

Higgins JPT, Thompson SG, Deeks JJ. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* 327, 557-60.

Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 In: The Cochrane Library, Issue 3, 2005 Chichester, UK: John Wiley & Sons, Ltd., May 2005.

Huwiler-Muntener K, Juni P, Junker C. (2002). Quality of reporting of randomized trials as a measure of methodologic quality. *Journal of American Medical Association* 287, 2801-4.

Jadad AR, Moore RA, Carroll D. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials* 17, 1-12.

Jadad AR, Moher M, Browman GP, Booker L, Sigouin C, Fuentes M, Stevens R. (2000). Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *British Medical Journal* 320, 537-40.

Juni P, Witschi A, Bloch R, Egger M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of American Medical Association* 282, 1054-60.

Juni P, Altman DG, Egger M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal* 323, 42-6.

Kjaergard LL, Villumsen J, Gluud C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 135, 982–9

Kunz R, Oxman AD. (1998). The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal* 317, 1185-90.

Landis RJ, Koch GG. (1977a). The measurement of observer agreement for categorical data. *Biometrics,* 33, 159–74.

Landis JR, Koch GG. (1977b). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 363-374.

Leizorovicz A, Haugh MC, Chapuis FR, Samama MM, Boissel JP. (1992). Low molecular weight heparin in prevention of perioperative thrombosis. *British Medical Journal* 305, 913-920.

LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine* 337, 536-42.

Lensing AW, Hirsh J. (1993). 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. *Thrombolysis and Haemostasis*. 69, 2-7.

Lu G, Ades AE. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 23, 3105–3124.

Lumley T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* 21, 2313-24.

May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. (1981). The randomized clinical trial: bias in analysis. *Circulation* 64, 669-673.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. (1995). Assessing the quality of randomized controlled trials. *Controlled Clinical Trials* 16, 62-73.

Moher D, Jadad AR, Tugwell P. (1996). Assessing the quality of randomized controlled trials. *International Journal of Technology Assessment for Health Care* 12, 195-208.

Moher D, Pham B, Jones A. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 352, 609-13.

Moher D, Cook JC, Eastwood S. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 354, 1896–1900.

Moher D, Schulz KF, Altman DG. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 357, 1191-4.

Moncrieff J, Churchill R, Drummond DC. (2001). Development of a quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods on Psychiatric Resesearch* 2001; 10: 126-33

Montgomery SA & Asberg M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* 134, 382-389.

Nurmohamed MT, Rosendaal FR, Buller HR. (1992). Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *Lancet* 340, 152-156.

Papakostas GI, Thase ME, Fava M, Nelson JC, Shelton RC. (2007). Are antidepressant drugs that combine serotonergic and noradrenergic mechanisms of action more effective than the selective serotonin reuptake inhibitors in treating major depressive disorder? A meta-analysis of studies of newer agents. *Biological Psychiatry* 62, 1217-27

Peduzzi P, Wittes J, Detre K, Holford T. (1993). Analysis as-randomized and the problem of non-adherence. *Statistics in Medicine* 12, 1185-1195.

Perlis RH, Perlis CS, Wu Y, Hwang C, Joseph M, Nierenberg AA. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry* 162, 1957-60.

Pogue J, Yusuf S. (1998). Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 351, 47-52.

Psaty BM, Lumley T, Furberg CD, Schellenbaum G, Pahor M, Alderman MH, Weiss NS. (2003). Health outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. *Journal of American Medical Association* 289, 2534-44.

Sackett DL, Gent M. (1979). Controversy in counting and attributing events in clinical trials. *New England Journal of Medicine* 301, 1410-1412.

Salanti G, Higgins J, Ades AE, Ioannidis JP. (in press). Evaluation of networks of randomized trials. *Statistical Methods and Medical Research*.

Schulz KF. (1995). Subverting randomization in controlled trials. *Journal of American Medical Association*. 274, 1456-1458.

Schulz KF, Chalmers I, Hayes RJ & Altman DG. (1995a). Empirical evidence of bias. *Journal of American Medical Association* 273, 408-412.

Schulz KF, Chalmers I, Hayes RJ. (1995b). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association* 273: 408-12

Schulz KF, Grimes DA. (2002). Allocation concealment in randomised trials: defending against deciphering. *Lancet* 359, 614-8.

Schulz KF, Grimes DA. (2005). Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 365, 1348-53.

Song F, Altman DG, Glenny AM, Deeks JJ. (2003). Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *British Medical Journal* 326, 472.

Sterne JA, Juni P, Schulz KF. (2002). Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statistics in Medicine* 21, 1513-24.

Taylor MJ, Freemantle N, Geddes JR, Bhagwagar Z. (2006). Early onset of selective serotonin reuptake inhibitor antidepressant action: systematic review and meta-analysis. *Archives of General Psychiatry* 63, 1217-23.

Thompson SG, Higgins PT. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 21, 1559–1573.

Tierney JF, Stewart LA. (2005). Investigating patient exclusion bias in meta-analysis. *International Journal of Epidemiology* 34, 79–87.

Zimmerman M, Posternak MA, Chelminski I. (2002). Symptom severity and exclusion from antidepressant efficacy trials. *Journal of Clinical Psychopharmacology* 22, 610-4.

APPENDIX

ANALYSIS OF COHERENCE

1	a	201	paroxetine
2	b	202	sertraline
3	c	203	citalopram
4	d	206	escitalopram
5	e	207	fluoxetine
6	f	208	fluvoxamine
7	g	302	milnacipran
8	h	303	venlafaxine
9	i	307	reboxetine
10	j	308	bupropion
11	k	311	mirtazapine
12	l	314	duloxetine

```
> cohR <- MTcoherence.fun(coherenceMANGA[outRthere,  - c(3, 4)])


 *-----  Evaluating the coherence of the network ------*

 Nr of treatments:  12
 Nr of all possible first order loops (triangles):  660
 Nr of available first order loops:  70


1 : Evaluation of the loop abc
Direct comparisons in the loop:
ab bc ac
 4  1  1

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Indirect comparison for the ac arm
  Mean(se)= -0.533(0.421)

  Incoherence within the loop:  Mean(se)= -0.101(0.467)


2 : Evaluation of the loop abd
Direct comparisons in the loop:
ab bd ad
 4  2  2

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bd arm
  mean(se)= 0.109(0.188)
  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Indirect comparison for the ad arm
  Mean(se)= -0.493(0.408)

  Incoherence within the loop:  Mean(se)= -0.381(0.453)


3 : Evaluation of the loop abe
Direct comparisons in the loop:
ab be ae
 4  8 12

The study with id=41 has more than two treatments in this loop (out is row nr 18 for comparison ae)
The study with id=42 has more than two treatments in this loop (out is row nr 19 for comparison ae)
  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the ae arm
  mean(se)= 0.002(0.119)
  Indirect comparison for the ae arm
  Mean(se)= -0.252(0.38)

  Incoherence within the loop:  Mean(se)= -0.254(0.398)


4 : Evaluation of the loop abf
Direct comparisons in the loop:
ab bf af
 4  2  3

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Indirect comparison for the af arm
  Mean(se)= -0.792(0.553)

  Incoherence within the loop:  Mean(se)= -0.98(0.606)


5 : Evaluation of the loop abh
Direct comparisons in the loop:
ab bh ah
 4  5  1

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Indirect comparison for the ah arm
  Mean(se)= -0.751(0.4)

  Incoherence within the loop:  Mean(se)= -0.636(0.454)


6 : Evaluation of the loop abg
Direct comparisons in the loop:
ab bg ag
 4  1  1

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bg arm
  mean(se)= -0.736(0.914)
  Meta-analysis for the ag arm
  mean(se)= 0.053(0.23)
  Indirect comparison for the ag arm
  Mean(se)= -1.338(0.983)

  Incoherence within the loop:  Mean(se)= -1.391(1.009)


7 : Evaluation of the loop abj
Direct comparisons in the loop:
ab bj aj
 4  3  1

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bj arm
  mean(se)= -0.068(0.155)
  Meta-analysis for the aj arm
  mean(se)= 0.317(0.457)
  Indirect comparison for the aj arm
  Mean(se)= -0.67(0.394)

  Incoherence within the loop:  Mean(se)= -0.987(0.603)


8 : Evaluation of the loop abk
Direct comparisons in the loop:
ab bk ak
 4  1  3

  Meta-analysis for the ab arm
  mean(se)= -0.602(0.362)
  Meta-analysis for the bk arm
  mean(se)= 0.026(0.228)
  Meta-analysis for the ak arm
  mean(se)= -0.237(0.15)
  Indirect comparison for the ak arm
  Mean(se)= -0.576(0.428)

  Incoherence within the loop:  Mean(se)= -0.339(0.453)


9 : Evaluation of the loop acd
Direct comparisons in the loop:
ac cd ad
 1  4  2

  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Meta-analysis for the cd arm
  mean(se)= -0.39(0.125)
  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Indirect comparison for the ad arm
  Mean(se)= -0.821(0.237)

  Incoherence within the loop:  Mean(se)= -0.709(0.308)


10 : Evaluation of the loop ace
Direct comparisons in the loop:
ac ce ae
 1  3 12

  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Indirect comparison for the ae arm
  Mean(se)= -0.38(0.256)

  Incoherence within the loop:  Mean(se)= -0.374(0.277)


11 : Evaluation of the loop acf
Direct comparisons in the loop:
ac cf af
 1  1  3

  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Meta-analysis for the cf arm
  mean(se)= 0.102(0.298)
  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Indirect comparison for the af arm
  Mean(se)= -0.329(0.359)

  Incoherence within the loop:  Mean(se)= -0.517(0.436)
```

```
12 : Evaluation of the loop ach
Direct comparisons in the loop:
ac ch ah
 1  1  1

  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Indirect comparison for the ah arm
  Mean(se)= -0.334(0.397)

  Incoherence within the loop:  Mean(se)= -0.219(0.452)


13 : Evaluation of the loop ack
Direct comparisons in the loop:
ac ck ak
 1  1  3

  Meta-analysis for the ac arm
  mean(se)= -0.431(0.201)
  Meta-analysis for the ck arm
  mean(se)= 0.281(0.357)
  Meta-analysis for the ak arm
  mean(se)= -0.237(0.15)
  Indirect comparison for the ak arm
  Mean(se)= -0.151(0.41)

  Incoherence within the loop:  Mean(se)= 0.086(0.436)


14 : Evaluation of the loop ade
Direct comparisons in the loop:
ad de ae
 2  2 12

  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Indirect comparison for the ae arm
  Mean(se)= 0.097(0.265)

  Incoherence within the loop:  Mean(se)= 0.104(0.286)


15 : Evaluation of the loop adh
```

```
Direct comparisons in the loop:
ad dh ah
 2  2  1

  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Meta-analysis for the dh arm
  mean(se)= 0.192(0.285)
  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Indirect comparison for the ah arm
  Mean(se)= 0.08(0.347)

  Incoherence within the loop:  Mean(se)= 0.195(0.408)


16 : Evaluation of the loop adj
Direct comparisons in the loop:
ad dj aj
 2  2  1

  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Meta-analysis for the dj arm
  mean(se)= 0.07(0.224)
  Meta-analysis for the aj arm
  mean(se)= 0.317(0.457)
  Indirect comparison for the aj arm
  Mean(se)= -0.042(0.299)

  Incoherence within the loop:  Mean(se)= -0.358(0.546)


17 : Evaluation of the loop adl
Direct comparisons in the loop:
ad dl al
 2  3  4

  Meta-analysis for the ad arm
  mean(se)= -0.112(0.197)
  Meta-analysis for the dl arm
  mean(se)= 0.262(0.194)
  Meta-analysis for the al arm
  mean(se)= 0.097(0.201)
  Indirect comparison for the al arm
  Mean(se)= 0.15(0.277)

  Incoherence within the loop:  Mean(se)= 0.053(0.342)


18 : Evaluation of the loop aef
Direct comparisons in the loop:
ae ef af
12  2  3
```

```
  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Indirect comparison for the af arm
  Mean(se)= -0.04(0.264)

  Incoherence within the loop:  Mean(se)= -0.228(0.361)


19 : Evaluation of the loop aeh
Direct comparisons in the loop:
ae eh ah
12 11  1

  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Indirect comparison for the ah arm
  Mean(se)= -0.314(0.141)

  Incoherence within the loop:  Mean(se)= -0.198(0.257)


20 : Evaluation of the loop aeg
Direct comparisons in the loop:
ae eg ag
12  3  1

  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the eg arm
  mean(se)= 0.144(0.249)
  Meta-analysis for the ag arm
  mean(se)= 0.053(0.23)
  Indirect comparison for the ag arm
  Mean(se)= 0.137(0.272)

  Incoherence within the loop:  Mean(se)= 0.084(0.356)


21 : Evaluation of the loop aej
Direct comparisons in the loop:
ae ej aj
12  3  1

  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the ej arm
  mean(se)= 0.194(0.148)
  Meta-analysis for the aj arm
  mean(se)= 0.317(0.457)
  Indirect comparison for the aj arm
  Mean(se)= 0.188(0.183)

  Incoherence within the loop:  Mean(se)= -0.129(0.493)


22 : Evaluation of the loop aek
Direct comparisons in the loop:
ae ek ak
12  5  3

  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the ek arm
  mean(se)= -0.415(0.17)
  Meta-analysis for the ak arm
  mean(se)= -0.237(0.15)
  Indirect comparison for the ak arm
  Mean(se)= -0.421(0.201)

  Incoherence within the loop:  Mean(se)= -0.184(0.251)


23 : Evaluation of the loop ael
Direct comparisons in the loop:
ae el al
12  1  4

  Meta-analysis for the ae arm
  mean(se)= -0.007(0.108)
  Meta-analysis for the el arm
  mean(se)= -0.01(0.424)
  Meta-analysis for the al arm
  mean(se)= 0.097(0.201)
  Indirect comparison for the al arm
  Mean(se)= -0.017(0.437)

  Incoherence within the loop:  Mean(se)= -0.114(0.482)


24 : Evaluation of the loop afh
Direct comparisons in the loop:
af fh ah
 3  1  1

  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Meta-analysis for the fh arm
  mean(se)= -0.861(0.42)
  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Indirect comparison for the ah arm
  Mean(se)= -0.673(0.488)

  Incoherence within the loop:  Mean(se)= -0.557(0.533)


25 : Evaluation of the loop afg
Direct comparisons in the loop:
af fg ag
 3  1  1

  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Meta-analysis for the fg arm
  mean(se)= -0.568(0.396)
  Meta-analysis for the ag arm
  mean(se)= 0.053(0.23)
  Indirect comparison for the ag arm
  Mean(se)= -0.38(0.467)

  Incoherence within the loop:  Mean(se)= -0.433(0.52)


26 : Evaluation of the loop afk
Direct comparisons in the loop:
af fk ak
 3  1  3

  Meta-analysis for the af arm
  mean(se)= 0.188(0.247)
  Meta-analysis for the fk arm
  mean(se)= -0.13(0.204)
  Meta-analysis for the ak arm
  mean(se)= -0.237(0.15)
  Indirect comparison for the ak arm
  Mean(se)= 0.058(0.32)

  Incoherence within the loop:  Mean(se)= 0.295(0.354)


27 : Evaluation of the loop ahj
Direct comparisons in the loop:
ah hj aj
 1  3  1

  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Meta-analysis for the hj arm
  mean(se)= 0.159(0.155)
  Meta-analysis for the aj arm
  mean(se)= 0.317(0.457)
  Indirect comparison for the aj arm
  Mean(se)= 0.043(0.265)
  Incoherence within the loop:  Mean(se)= -0.273(0.529)


28 : Evaluation of the loop ahk
Direct comparisons in the loop:
ah hk ak
 1  2  3

  Meta-analysis for the ah arm
  mean(se)= -0.115(0.215)
  Meta-analysis for the hk arm
  mean(se)= -0.423(0.199)
  Meta-analysis for the ak arm
  mean(se)= -0.237(0.15)
  Indirect comparison for the ak arm
  Mean(se)= -0.538(0.293)

  Incoherence within the loop:  Mean(se)= -0.301(0.329)


29 : Evaluation of the loop bcd
Direct comparisons in the loop:
bc cd bd
 1  4  2

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the cd arm
  mean(se)= -0.39(0.125)
  Meta-analysis for the bd arm
  mean(se)= 0.109(0.188)
  Indirect comparison for the bd arm
  Mean(se)= -0.32(0.249)

  Incoherence within the loop:  Mean(se)= -0.429(0.312)


30 : Evaluation of the loop bce
Direct comparisons in the loop:
bc ce be
 1  3  8

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Indirect comparison for the be arm
  Mean(se)= 0.121(0.267)

  Incoherence within the loop:  Mean(se)= -0.229(0.291)
```

```
31 : Evaluation of the loop bcf
Direct comparisons in the loop:
bc cf bf
 1  1  2

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the cf arm
  mean(se)= 0.102(0.298)
  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Indirect comparison for the bf arm
  Mean(se)= 0.172(0.368)

  Incoherence within the loop:  Mean(se)= 0.361(0.557)


32 : Evaluation of the loop bch
Direct comparisons in the loop:
bc ch bh
 1  1  5

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Indirect comparison for the bh arm
  Mean(se)= 0.167(0.405)

  Incoherence within the loop:  Mean(se)= 0.316(0.439)


33 : Evaluation of the loop bck
Direct comparisons in the loop:
bc ck bk
 1  1  1

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the ck arm
  mean(se)= 0.281(0.357)
  Meta-analysis for the bk arm
  mean(se)= 0.026(0.228)
  Indirect comparison for the bk arm
  Mean(se)= 0.35(0.417)

  Incoherence within the loop:  Mean(se)= 0.324(0.475)


34 : Evaluation of the loop bci
Direct comparisons in the loop:
```

```
bc ci bi
 1  2  1

  Meta-analysis for the bc arm
  mean(se)= 0.07(0.216)
  Meta-analysis for the ci arm
  mean(se)= 0.543(0.272)
  Meta-analysis for the bi arm
  mean(se)= 0.312(0.613)
  Indirect comparison for the bi arm
  Mean(se)= 0.613(0.347)

  Incoherence within the loop:  Mean(se)= 0.301(0.704)


35 : Evaluation of the loop bde
Direct comparisons in the loop:
bd de be
 2  2  8

  Meta-analysis for the bd arm
  mean(se)= 0.109(0.188)
  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Indirect comparison for the be arm
  Mean(se)= 0.319(0.257)

  Incoherence within the loop:  Mean(se)= -0.031(0.282)


36 : Evaluation of the loop bdh
Direct comparisons in the loop:
bd dh bh
 2  2  5

  Meta-analysis for the bd arm
  mean(se)= 0.109(0.188)
  Meta-analysis for the dh arm
  mean(se)= 0.192(0.285)
  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Indirect comparison for the bh arm
  Mean(se)= 0.301(0.341)

  Incoherence within the loop:  Mean(se)= 0.45(0.382)


37 : Evaluation of the loop bdj
Direct comparisons in the loop:
bd dj bj
 2  2  3
```

```
  Meta-analysis for the bd arm
  mean(se)= 0.109(0.188)
  Meta-analysis for the dj arm
  mean(se)= 0.07(0.224)
  Meta-analysis for the bj arm
  mean(se)= -0.068(0.155)
  Indirect comparison for the bj arm
  Mean(se)= 0.18(0.292)

  Incoherence within the loop:  Mean(se)= 0.247(0.331)


38 : Evaluation of the loop bef
Direct comparisons in the loop:
be ef bf
 8  2  2

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Indirect comparison for the bf arm
  Mean(se)= 0.317(0.267)

  Incoherence within the loop:  Mean(se)= 0.506(0.497)


39 : Evaluation of the loop beh
Direct comparisons in the loop:
be eh bh
 8 11  5

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Indirect comparison for the bh arm
  Mean(se)= 0.043(0.147)

  Incoherence within the loop:  Mean(se)= 0.192(0.225)


40 : Evaluation of the loop beg
Direct comparisons in the loop:
be eg bg
 8  3  1

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the eg arm
  mean(se)= 0.144(0.249)
  Meta-analysis for the bg arm
  mean(se)= -0.736(0.914)
  Indirect comparison for the bg arm
  Mean(se)= 0.494(0.275)

  Incoherence within the loop:  Mean(se)= 1.229(0.954)


41 : Evaluation of the loop bej
Direct comparisons in the loop:
be ej bj
 8  3  3

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the ej arm
  mean(se)= 0.194(0.148)
  Meta-analysis for the bj arm
  mean(se)= -0.068(0.155)
  Indirect comparison for the bj arm
  Mean(se)= 0.544(0.188)

  Incoherence within the loop:  Mean(se)= 0.612(0.244)


42 : Evaluation of the loop bek
Direct comparisons in the loop:
be ek bk
 8  5  1

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the ek arm
  mean(se)= -0.415(0.17)
  Meta-analysis for the bk arm
  mean(se)= 0.026(0.228)
  Indirect comparison for the bk arm
  Mean(se)= -0.065(0.206)

  Incoherence within the loop:  Mean(se)= -0.091(0.307)


43 : Evaluation of the loop bei
Direct comparisons in the loop:
be ei bi
 8  4  1

  Meta-analysis for the be arm
  mean(se)= 0.35(0.116)
  Meta-analysis for the ei arm
  mean(se)= 0.323(0.169)
  Meta-analysis for the bi arm
  mean(se)= 0.312(0.613)
  Indirect comparison for the bi arm
  Mean(se)= 0.673(0.205)

  Incoherence within the loop:  Mean(se)= 0.361(0.646)


44 : Evaluation of the loop bfh
Direct comparisons in the loop:
bf fh bh
 2  1  5

  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Meta-analysis for the fh arm
  mean(se)= -0.861(0.42)
  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Indirect comparison for the bh arm
  Mean(se)= -1.05(0.593)

  Incoherence within the loop:  Mean(se)= -0.901(0.617)


45 : Evaluation of the loop bfg
Direct comparisons in the loop:
bf fg bg
 2  1  1

  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Meta-analysis for the fg arm
  mean(se)= -0.568(0.396)
  Meta-analysis for the bg arm
  mean(se)= -0.736(0.914)
  Indirect comparison for the bg arm
  Mean(se)= -0.758(0.576)

  Incoherence within the loop:  Mean(se)= -0.022(1.08)


46 : Evaluation of the loop bfk
Direct comparisons in the loop:
bf fk bk
 2  1  1

  Meta-analysis for the bf arm
  mean(se)= -0.19(0.418)
  Meta-analysis for the fk arm
  mean(se)= -0.13(0.204)
  Meta-analysis for the bk arm
  mean(se)= 0.026(0.228)
  Indirect comparison for the bk arm
  Mean(se)= -0.32(0.466)
  Incoherence within the loop:  Mean(se)= -0.346(0.519)


47 : Evaluation of the loop bhj
Direct comparisons in the loop:
bh hj bj
 5  3  3

  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Meta-analysis for the hj arm
  mean(se)= 0.159(0.155)
  Meta-analysis for the bj arm
  mean(se)= -0.068(0.155)
  Indirect comparison for the bj arm
  Mean(se)= 0.01(0.23)

  Incoherence within the loop:  Mean(se)= 0.077(0.278)


48 : Evaluation of the loop bhk
Direct comparisons in the loop:
bh hk bk
 5  2  1

  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Meta-analysis for the hk arm
  mean(se)= -0.423(0.199)
  Meta-analysis for the bk arm
  mean(se)= 0.026(0.228)
  Indirect comparison for the bk arm
  Mean(se)= -0.572(0.262)

  Incoherence within the loop:  Mean(se)= -0.598(0.347)


49 : Evaluation of the loop bhi
Direct comparisons in the loop:
bh hi bi
 5  1  1

  Meta-analysis for the bh arm
  mean(se)= -0.149(0.171)
  Meta-analysis for the hi arm
  mean(se)= 0.799(0.419)
  Meta-analysis for the bi arm
  mean(se)= 0.312(0.613)
  Indirect comparison for the bi arm
  Mean(se)= 0.65(0.452)

  Incoherence within the loop:  Mean(se)= 0.338(0.761)
```

```
50 : Evaluation of the loop cde
Direct comparisons in the loop:
cd de ce
 4  2  3

  Meta-analysis for the cd arm
  mean(se)= -0.39(0.125)
  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Indirect comparison for the ce arm
  Mean(se)= -0.181(0.216)

  Incoherence within the loop:  Mean(se)= -0.232(0.267)


51 : Evaluation of the loop cdh
Direct comparisons in the loop:
cd dh ch
 4  2  1

  Meta-analysis for the cd arm
  mean(se)= -0.39(0.125)
  Meta-analysis for the dh arm
  mean(se)= 0.192(0.285)
  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Indirect comparison for the ch arm
  Mean(se)= -0.198(0.311)

  Incoherence within the loop:  Mean(se)= -0.295(0.463)


52 : Evaluation of the loop cef
Direct comparisons in the loop:
ce ef cf
 3  2  1

  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the cf arm
  mean(se)= 0.102(0.298)
  Indirect comparison for the cf arm
  Mean(se)= 0.018(0.288)

  Incoherence within the loop:  Mean(se)= -0.084(0.414)


53 : Evaluation of the loop ceh
Direct comparisons in the loop:
ce eh ch
```

```
 3 11  1

  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Indirect comparison for the ch arm
  Mean(se)= -0.256(0.182)

  Incoherence within the loop:  Mean(se)= -0.353(0.388)


54 : Evaluation of the loop cek
Direct comparisons in the loop:
ce ek ck
 3  5  1

  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the ek arm
  mean(se)= -0.415(0.17)
  Meta-analysis for the ck arm
  mean(se)= 0.281(0.357)
  Indirect comparison for the ck arm
  Mean(se)= -0.364(0.232)

  Incoherence within the loop:  Mean(se)= -0.644(0.425)


55 : Evaluation of the loop cei
Direct comparisons in the loop:
ce ei ci
 3  4  2

  Meta-analysis for the ce arm
  mean(se)= 0.051(0.157)
  Meta-analysis for the ei arm
  mean(se)= 0.323(0.169)
  Meta-analysis for the ci arm
  mean(se)= 0.543(0.272)
  Indirect comparison for the ci arm
  Mean(se)= 0.374(0.231)

  Incoherence within the loop:  Mean(se)= -0.169(0.357)


56 : Evaluation of the loop cfh
Direct comparisons in the loop:
cf fh ch
 1  1  1

  Meta-analysis for the cf arm
  mean(se)= 0.102(0.298)
  Meta-analysis for the fh arm
  mean(se)= -0.861(0.42)
  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Indirect comparison for the ch arm
  Mean(se)= -0.759(0.515)

  Incoherence within the loop:  Mean(se)= -0.856(0.619)


57 : Evaluation of the loop cfk
Direct comparisons in the loop:
cf fk ck
 1  1  1

  Meta-analysis for the cf arm
  mean(se)= 0.102(0.298)
  Meta-analysis for the fk arm
  mean(se)= -0.13(0.204)
  Meta-analysis for the ck arm
  mean(se)= 0.281(0.357)
  Indirect comparison for the ck arm
  Mean(se)= -0.028(0.361)

  Incoherence within the loop:  Mean(se)= -0.309(0.508)


58 : Evaluation of the loop chk
Direct comparisons in the loop:
ch hk ck
 1  2  1

  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Meta-analysis for the hk arm
  mean(se)= -0.423(0.199)
  Meta-analysis for the ck arm
  mean(se)= 0.281(0.357)
  Indirect comparison for the ck arm
  Mean(se)= -0.326(0.396)

  Incoherence within the loop:  Mean(se)= -0.606(0.533)


59 : Evaluation of the loop chi
Direct comparisons in the loop:
ch hi ci
 1  1  2

  Meta-analysis for the ch arm
  mean(se)= 0.097(0.343)
  Meta-analysis for the hi arm
  mean(se)= 0.799(0.419)
  Meta-analysis for the ci arm
  mean(se)= 0.543(0.272)
  Indirect comparison for the ci arm
  Mean(se)= 0.896(0.541)

  Incoherence within the loop:  Mean(se)= 0.354(0.606)


60 : Evaluation of the loop deh
Direct comparisons in the loop:
de eh dh
 2 11  2

  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the dh arm
  mean(se)= 0.192(0.285)
  Indirect comparison for the dh arm
  Mean(se)= -0.098(0.198)

  Incoherence within the loop:  Mean(se)= -0.29(0.347)


61 : Evaluation of the loop dej
Direct comparisons in the loop:
de ej dj
 2  3  2

  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the ej arm
  mean(se)= 0.194(0.148)
  Meta-analysis for the dj arm
  mean(se)= 0.07(0.224)
  Indirect comparison for the dj arm
  Mean(se)= 0.404(0.23)

  Incoherence within the loop:  Mean(se)= 0.333(0.321)


62 : Evaluation of the loop del
Direct comparisons in the loop:
de el dl
 2  1  3

  Meta-analysis for the de arm
  mean(se)= 0.209(0.176)
  Meta-analysis for the el arm
  mean(se)= -0.01(0.424)
  Meta-analysis for the dl arm
  mean(se)= 0.262(0.194)
  Indirect comparison for the dl arm
  Mean(se)= 0.199(0.459)

  Incoherence within the loop:  Mean(se)= -0.064(0.499)


63 : Evaluation of the loop dhj
Direct comparisons in the loop:
dh hj dj
 2  3  2

  Meta-analysis for the dh arm
  mean(se)= 0.192(0.285)
  Meta-analysis for the hj arm
  mean(se)= 0.159(0.155)
  Meta-analysis for the dj arm
  mean(se)= 0.07(0.224)
  Indirect comparison for the dj arm
  Mean(se)= 0.351(0.325)

  Incoherence within the loop:  Mean(se)= 0.28(0.395)


64 : Evaluation of the loop efh
Direct comparisons in the loop:
ef fh eh
 2  1 11

  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the fh arm
  mean(se)= -0.861(0.42)
  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Indirect comparison for the eh arm
  Mean(se)= -0.894(0.485)

  Incoherence within the loop:  Mean(se)= -0.587(0.493)


65 : Evaluation of the loop efg
Direct comparisons in the loop:
ef fg eg
 2  1  3

  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the fg arm
  mean(se)= -0.568(0.396)
  Meta-analysis for the eg arm
  mean(se)= 0.144(0.249)
  Indirect comparison for the eg arm
  Mean(se)= -0.601(0.463)

  Incoherence within the loop:  Mean(se)= -0.745(0.526)
```

```
66 : Evaluation of the loop efk
Direct comparisons in the loop:
ef fk ek
 2  1  5

  Meta-analysis for the ef arm
  mean(se)= -0.033(0.241)
  Meta-analysis for the fk arm
  mean(se)= -0.13(0.204)
  Meta-analysis for the ek arm
  mean(se)= -0.415(0.17)
  Indirect comparison for the ek arm
  Mean(se)= -0.163(0.316)

  Incoherence within the loop:  Mean(se)= 0.251(0.359)


67 : Evaluation of the loop ehj
Direct comparisons in the loop:
eh hj ej
11  3  3

  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the hj arm
  mean(se)= 0.159(0.155)
  Meta-analysis for the ej arm
  mean(se)= 0.194(0.148)
  Indirect comparison for the ej arm
  Mean(se)= -0.148(0.18)

  Incoherence within the loop:  Mean(se)= -0.343(0.233)


68 : Evaluation of the loop ehk
Direct comparisons in the loop:
eh hk ek
11  2  5

  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the hk arm
  mean(se)= -0.423(0.199)
  Meta-analysis for the ek arm
  mean(se)= -0.415(0.17)
  Indirect comparison for the ek arm
  Mean(se)= -0.73(0.218)

  Incoherence within the loop:  Mean(se)= -0.315(0.277)


69 : Evaluation of the loop ehi
```

```
Direct comparisons in the loop:
eh hi ei
11  1  4

  Meta-analysis for the eh arm
  mean(se)= -0.307(0.091)
  Meta-analysis for the hi arm
  mean(se)= 0.799(0.419)
  Meta-analysis for the ei arm
  mean(se)= 0.323(0.169)
  Indirect comparison for the ei arm
  Mean(se)= 0.492(0.428)

  Incoherence within the loop:  Mean(se)= 0.17(0.461)


70 : Evaluation of the loop fhk
Direct comparisons in the loop:
fh hk fk
 1  2  1

  Meta-analysis for the fh arm
  mean(se)= -0.861(0.42)
  Meta-analysis for the hk arm
  mean(se)= -0.423(0.199)
  Meta-analysis for the fk arm
  mean(se)= -0.13(0.204)
  Indirect comparison for the fk arm
  Mean(se)= -1.284(0.465)

  Incoherence within the loop:  Mean(se)= -1.153(0.508)
```

Dropouts

```
> cohD <- MTcoherence.fun(coherenceMANGA[outDthere,  - c(1, 2)])


 *-----  Evaluating the coherence of the network ------*

 Nr of treatments:  12
 Nr of all possible first order loops (triangles):  660
 Nr of available first order loops:  63


 1 : Evaluation of the loop abc
Direct comparisons in the loop:
ab bc ac
 4  2  1

  Meta-analysis for the ab arm
 mean(se)= 0.424(0.453)
 Meta-analysis for the bc arm
 mean(se)= 0.401(0.193)
 Meta-analysis for the ac arm
 mean(se)= -0.01(0.245)
 Indirect comparison for the ac arm
 Mean(se)= 0.825(0.493)

 Incoherence within the loop:  Mean(se)= 0.835(0.551)


2 : Evaluation of the loop abd
Direct comparisons in the loop:
ab bd ad
 4  2  2

 Meta-analysis for the ab arm
 mean(se)= 0.424(0.453)
 Meta-analysis for the bd arm
 mean(se)= -0.212(0.237)
 Meta-analysis for the ad arm
 mean(se)= 0.284(0.227)
 Indirect comparison for the ad arm
 Mean(se)= 0.212(0.512)

 Incoherence within the loop:  Mean(se)= -0.072(0.56)


3 : Evaluation of the loop abe
Direct comparisons in the loop:
ab be ae
 4  7 13

The study with id=41 has more than two treatments in this loop (out is row nr 19 for comparison ae)
The study with id=42 has more than two treatments in this loop (out is row nr 20 for comparison ae)
 Meta-analysis for the ab arm
 mean(se)= 0.424(0.453)
 Meta-analysis for the be arm
 mean(se)= -0.215(0.168)
 Meta-analysis for the ae arm
 mean(se)= 0.046(0.087)
 Indirect comparison for the ae arm
 Mean(se)= 0.208(0.483)

 Incoherence within the loop:  Mean(se)= 0.162(0.491)


4 : Evaluation of the loop abf
Direct comparisons in the loop:
ab bf af
 4  2  3

 Meta-analysis for the ab arm
 mean(se)= 0.424(0.453)
```

```
  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Meta-analysis for the af arm
  mean(se)= -0.073(0.279)
  Indirect comparison for the af arm
  Mean(se)= 0.04(1.127)

  Incoherence within the loop:  Mean(se)= 0.113(1.161)


5 : Evaluation of the loop abh
Direct comparisons in the loop:
ab bh ah
 4  5  1

  Meta-analysis for the ab arm
  mean(se)= 0.424(0.453)
  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Indirect comparison for the ah arm
  Mean(se)= -0.159(0.645)

  Incoherence within the loop:  Mean(se)= -0.337(0.687)


6 : Evaluation of the loop abg
Direct comparisons in the loop:
ab bg ag
 4  1  1

  Meta-analysis for the ab arm
  mean(se)= 0.424(0.453)
  Meta-analysis for the bg arm
  mean(se)= -0.533(0.555)
  Meta-analysis for the ag arm
  mean(se)= 0.129(0.285)
  Indirect comparison for the ag arm
  Mean(se)= -0.11(0.716)

  Incoherence within the loop:  Mean(se)= -0.239(0.771)


7 : Evaluation of the loop abj
Direct comparisons in the loop:
ab bj aj
 4  2  2

  Meta-analysis for the ab arm
  mean(se)= 0.424(0.453)
  Meta-analysis for the bj arm
  mean(se)= 0.41(0.286)
  Meta-analysis for the aj arm
  mean(se)= 0.152(0.325)
  Indirect comparison for the aj arm
  Mean(se)= 0.833(0.536)

  Incoherence within the loop:  Mean(se)= 0.681(0.627)


8 : Evaluation of the loop abk
Direct comparisons in the loop:
ab bk ak
 4  1  3

  Meta-analysis for the ab arm
  mean(se)= 0.424(0.453)
  Meta-analysis for the bk arm
  mean(se)= -0.27(0.265)
  Meta-analysis for the ak arm
  mean(se)= 0.173(0.166)
  Indirect comparison for the ak arm
  Mean(se)= 0.154(0.525)

  Incoherence within the loop:  Mean(se)= -0.019(0.551)


9 : Evaluation of the loop acd
Direct comparisons in the loop:
ac cd ad
 1  5  2

  Meta-analysis for the ac arm
  mean(se)= -0.01(0.245)
  Meta-analysis for the cd arm
  mean(se)= 0.148(0.156)
  Meta-analysis for the ad arm
  mean(se)= 0.284(0.227)
  Indirect comparison for the ad arm
  Mean(se)= 0.137(0.29)

  Incoherence within the loop:  Mean(se)= -0.147(0.369)


10 : Evaluation of the loop ace
Direct comparisons in the loop:
ac ce ae
 1  3 13

  Meta-analysis for the ac arm
  mean(se)= -0.01(0.245)
  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Meta-analysis for the ae arm
  mean(se)= 0.074(0.082)
  Indirect comparison for the ae arm
  Mean(se)= 0.144(0.311)
  Incoherence within the loop:  Mean(se)= 0.07(0.322)


11 : Evaluation of the loop acf
Direct comparisons in the loop:
ac cf af
 1  1  3

  Meta-analysis for the ac arm
  mean(se)= -0.01(0.245)
  Meta-analysis for the cf arm
  mean(se)= -0.349(0.323)
  Meta-analysis for the af arm
  mean(se)= -0.073(0.279)
  Indirect comparison for the af arm
  Mean(se)= -0.359(0.405)

  Incoherence within the loop:  Mean(se)= -0.286(0.492)


12 : Evaluation of the loop ack
Direct comparisons in the loop:
ac ck ak
 1  1  3

  Meta-analysis for the ac arm
  mean(se)= -0.01(0.245)
  Meta-analysis for the ck arm
  mean(se)= -0.86(0.444)
  Meta-analysis for the ak arm
  mean(se)= 0.173(0.166)
  Indirect comparison for the ak arm
  Mean(se)= -0.87(0.507)

  Incoherence within the loop:  Mean(se)= -1.043(0.534)


13 : Evaluation of the loop ade
Direct comparisons in the loop:
ad de ae
 2  2 13

  Meta-analysis for the ad arm
  mean(se)= 0.284(0.227)
  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the ae arm
  mean(se)= 0.074(0.082)
  Indirect comparison for the ae arm
  Mean(se)= 0.262(0.541)

  Incoherence within the loop:  Mean(se)= 0.188(0.547)
```

```
14 : Evaluation of the loop adh
Direct comparisons in the loop:
ad dh ah
 2  2  1

  Meta-analysis for the ad arm
  mean(se)= 0.284(0.227)
  Meta-analysis for the dh arm
  mean(se)= -0.11(0.223)
  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Indirect comparison for the ah arm
  Mean(se)= 0.175(0.318)

  Incoherence within the loop:  Mean(se)= -0.003(0.396)


15 : Evaluation of the loop adj
Direct comparisons in the loop:
ad dj aj
 2  3  2

  Meta-analysis for the ad arm
  mean(se)= 0.284(0.227)
  Meta-analysis for the dj arm
  mean(se)= 0.019(0.159)
  Meta-analysis for the aj arm
  mean(se)= 0.152(0.325)
  Indirect comparison for the aj arm
  Mean(se)= 0.304(0.277)

  Incoherence within the loop:  Mean(se)= 0.151(0.427)


16 : Evaluation of the loop adl
Direct comparisons in the loop:
ad dl al
 2  2  4

  Meta-analysis for the ad arm
  mean(se)= 0.284(0.227)
  Meta-analysis for the dl arm
  mean(se)= -0.66(0.342)
  Meta-analysis for the al arm
  mean(se)= 0.094(0.158)
  Indirect comparison for the al arm
  Mean(se)= -0.376(0.41)

  Incoherence within the loop:  Mean(se)= -0.47(0.439)


17 : Evaluation of the loop aef
Direct comparisons in the loop:
```

```
ae ef af
13  2  3

  Meta-analysis for the ae arm
  mean(se)= 0.074(0.082)
  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the af arm
  mean(se)= -0.073(0.279)
  Indirect comparison for the af arm
  Mean(se)= -0.084(0.306)

  Incoherence within the loop:  Mean(se)= -0.011(0.414)


18 : Evaluation of the loop aeh
Direct comparisons in the loop:
ae eh ah
13 12  1

  Meta-analysis for the ae arm
  mean(se)= 0.074(0.082)
  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Indirect comparison for the ah arm
  Mean(se)= 0.009(0.126)

  Incoherence within the loop:  Mean(se)= -0.169(0.268)


19 : Evaluation of the loop aeg
Direct comparisons in the loop:
ae eg ag
13  3  1

  Meta-analysis for the ae arm
  mean(se)= 0.074(0.082)
  Meta-analysis for the eg arm
  mean(se)= -0.016(0.186)
  Meta-analysis for the ag arm
  mean(se)= 0.129(0.285)
  Indirect comparison for the ag arm
  Mean(se)= 0.059(0.203)

  Incoherence within the loop:  Mean(se)= -0.071(0.35)


20 : Evaluation of the loop aej
Direct comparisons in the loop:
ae ej aj
13  3  2
```

```
 Meta-analysis for the ae arm
 mean(se)= 0.074(0.082)
 Meta-analysis for the ej arm
 mean(se)= -0.007(0.154)
 Meta-analysis for the aj arm
 mean(se)= 0.152(0.325)
 Indirect comparison for the aj arm
 Mean(se)= 0.067(0.174)

 Incoherence within the loop:  Mean(se)= -0.085(0.369)


21 : Evaluation of the loop aek
Direct comparisons in the loop:
ae ek ak
13  4  3

 Meta-analysis for the ae arm
 mean(se)= 0.074(0.082)
 Meta-analysis for the ek arm
 mean(se)= -0.05(0.315)
 Meta-analysis for the ak arm
 mean(se)= 0.173(0.166)
 Indirect comparison for the ak arm
 Mean(se)= 0.025(0.326)

 Incoherence within the loop:  Mean(se)= -0.148(0.366)


22 : Evaluation of the loop ael
Direct comparisons in the loop:
ae el al
13  1  4

 Meta-analysis for the ae arm
 mean(se)= 0.074(0.082)
 Meta-analysis for the el arm
 mean(se)= 0.091(0.441)
 Meta-analysis for the al arm
 mean(se)= 0.094(0.158)
 Indirect comparison for the al arm
 Mean(se)= 0.165(0.448)

 Incoherence within the loop:  Mean(se)= 0.071(0.475)


23 : Evaluation of the loop afh
Direct comparisons in the loop:
af fh ah
 3  1  1

 Meta-analysis for the af arm
 mean(se)= -0.073(0.279)
 Meta-analysis for the fh arm
```

```
  mean(se)= 0.708(0.444)
  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Indirect comparison for the ah arm
  Mean(se)= 0.635(0.524)

  Incoherence within the loop:  Mean(se)= 0.457(0.575)


24 : Evaluation of the loop afg
Direct comparisons in the loop:
af fg ag
 3  1  1

  Meta-analysis for the af arm
  mean(se)= -0.073(0.279)
  Meta-analysis for the fg arm
  mean(se)= 0.199(0.418)
  Meta-analysis for the ag arm
  mean(se)= 0.129(0.285)
  Indirect comparison for the ag arm
  Mean(se)= 0.126(0.503)

  Incoherence within the loop:  Mean(se)= -0.003(0.578)


25 : Evaluation of the loop afk
Direct comparisons in the loop:
af fk ak
 3  1  3

  Meta-analysis for the af arm
  mean(se)= -0.073(0.279)
  Meta-analysis for the fk arm
  mean(se)= -0.186(0.241)
  Meta-analysis for the ak arm
  mean(se)= 0.173(0.166)
  Indirect comparison for the ak arm
  Mean(se)= -0.259(0.368)

  Incoherence within the loop:  Mean(se)= -0.432(0.404)


26 : Evaluation of the loop ahj
Direct comparisons in the loop:
ah hj aj
 1  3  2

  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Meta-analysis for the hj arm
  mean(se)= 0.006(0.14)
  Meta-analysis for the aj arm
  mean(se)= 0.152(0.325)
  Indirect comparison for the aj arm
  Mean(se)= 0.184(0.275)

  Incoherence within the loop:  Mean(se)= 0.032(0.426)


27 : Evaluation of the loop ahk
Direct comparisons in the loop:
ah hk ak
 1  2  3

  Meta-analysis for the ah arm
  mean(se)= 0.177(0.236)
  Meta-analysis for the hk arm
  mean(se)= 0.41(0.213)
  Meta-analysis for the ak arm
  mean(se)= 0.173(0.166)
  Indirect comparison for the ak arm
  Mean(se)= 0.587(0.318)

  Incoherence within the loop:  Mean(se)= 0.415(0.359)


28 : Evaluation of the loop bcd
Direct comparisons in the loop:
bc cd bd
 2  5  2

  Meta-analysis for the bc arm
  mean(se)= 0.401(0.193)
  Meta-analysis for the cd arm
  mean(se)= 0.148(0.156)
  Meta-analysis for the bd arm
  mean(se)= -0.212(0.237)
  Indirect comparison for the bd arm
  Mean(se)= 0.549(0.248)

  Incoherence within the loop:  Mean(se)= 0.76(0.344)


29 : Evaluation of the loop bce
Direct comparisons in the loop:
bc ce be
 2  3  7

  Meta-analysis for the bc arm
  mean(se)= 0.401(0.193)
  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Indirect comparison for the be arm
  Mean(se)= 0.555(0.272)
  Incoherence within the loop:  Mean(se)= 0.771(0.32)


30 : Evaluation of the loop bcf
Direct comparisons in the loop:
bc cf bf
 2  1  2

  Meta-analysis for the bc arm
  mean(se)= 0.401(0.193)
  Meta-analysis for the cf arm
  mean(se)= -0.349(0.323)
  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Indirect comparison for the bf arm
  Mean(se)= 0.052(0.376)

  Incoherence within the loop:  Mean(se)= 0.436(1.098)


31 : Evaluation of the loop bck
Direct comparisons in the loop:
bc ck bk
 2  1  1

  Meta-analysis for the bc arm
  mean(se)= 0.401(0.193)
  Meta-analysis for the ck arm
  mean(se)= -0.86(0.444)
  Meta-analysis for the bk arm
  mean(se)= -0.27(0.265)
  Indirect comparison for the bk arm
  Mean(se)= -0.459(0.484)

  Incoherence within the loop:  Mean(se)= -0.189(0.552)


32 : Evaluation of the loop bci
Direct comparisons in the loop:
bc ci bi
 2  2  1

  Meta-analysis for the bc arm
  mean(se)= 0.401(0.193)
  Meta-analysis for the ci arm
  mean(se)= -0.146(0.708)
  Meta-analysis for the bi arm
  mean(se)= -0.56(0.794)
  Indirect comparison for the bi arm
  Mean(se)= 0.255(0.734)

  Incoherence within the loop:  Mean(se)= 0.814(1.081)
```

```
33 : Evaluation of the loop bde
Direct comparisons in the loop:
bd de be
 2  2  7

  Meta-analysis for the bd arm
  mean(se)= -0.212(0.237)
  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Indirect comparison for the be arm
  Mean(se)= -0.234(0.545)

  Incoherence within the loop:  Mean(se)= -0.019(0.57)


34 : Evaluation of the loop bdh
Direct comparisons in the loop:
bd dh bh
 2  2  5

  Meta-analysis for the bd arm
  mean(se)= -0.212(0.237)
  Meta-analysis for the dh arm
  mean(se)= -0.11(0.223)
  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Indirect comparison for the bh arm
  Mean(se)= -0.321(0.325)

  Incoherence within the loop:  Mean(se)= 0.262(0.563)


35 : Evaluation of the loop bdj
Direct comparisons in the loop:
bd dj bj
 2  3  2

  Meta-analysis for the bd arm
  mean(se)= -0.212(0.237)
  Meta-analysis for the dj arm
  mean(se)= 0.019(0.159)
  Meta-analysis for the bj arm
  mean(se)= 0.41(0.286)
  Indirect comparison for the bj arm
  Mean(se)= -0.193(0.286)

  Incoherence within the loop:  Mean(se)= -0.602(0.405)


36 : Evaluation of the loop bef
Direct comparisons in the loop:
be ef bf
```

```
 7  2  2

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Indirect comparison for the bf arm
  Mean(se)= -0.374(0.339)

  Incoherence within the loop:  Mean(se)= 0.01(1.086)


37 : Evaluation of the loop beh
Direct comparisons in the loop:
be eh bh
 7 12  5

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Indirect comparison for the bh arm
  Mean(se)= -0.281(0.193)

  Incoherence within the loop:  Mean(se)= 0.302(0.498)


38 : Evaluation of the loop beg
Direct comparisons in the loop:
be eg bg
 7  3  1

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the eg arm
  mean(se)= -0.016(0.186)
  Meta-analysis for the bg arm
  mean(se)= -0.533(0.555)
  Indirect comparison for the bg arm
  Mean(se)= -0.231(0.25)

  Incoherence within the loop:  Mean(se)= 0.302(0.609)


39 : Evaluation of the loop bej
Direct comparisons in the loop:
be ej bj
 7  3  2

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the ej arm
  mean(se)= -0.007(0.154)
  Meta-analysis for the bj arm
  mean(se)= 0.41(0.286)
  Indirect comparison for the bj arm
  Mean(se)= -0.223(0.227)

  Incoherence within the loop:  Mean(se)= -0.632(0.366)


40 : Evaluation of the loop bek
Direct comparisons in the loop:
be ek bk
 7  4  1

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the ek arm
  mean(se)= -0.05(0.315)
  Meta-analysis for the bk arm
  mean(se)= -0.27(0.265)
  Indirect comparison for the bk arm
  Mean(se)= -0.265(0.357)

  Incoherence within the loop:  Mean(se)= 0.005(0.445)


41 : Evaluation of the loop bei
Direct comparisons in the loop:
be ei bi
 7  4  1

  Meta-analysis for the be arm
  mean(se)= -0.215(0.168)
  Meta-analysis for the ei arm
  mean(se)= -0.385(0.163)
  Meta-analysis for the bi arm
  mean(se)= -0.56(0.794)
  Indirect comparison for the bi arm
  Mean(se)= -0.601(0.234)

  Incoherence within the loop:  Mean(se)= -0.041(0.828)


42 : Evaluation of the loop bfh
Direct comparisons in the loop:
bf fh bh
 2  1  5

  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Meta-analysis for the fh arm
  mean(se)= 0.708(0.444)
  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Indirect comparison for the bh arm
  Mean(se)= 0.324(1.123)

  Incoherence within the loop:  Mean(se)= 0.907(1.213)


43 : Evaluation of the loop bfg
Direct comparisons in the loop:
bf fg bg
 2  1  1

  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Meta-analysis for the fg arm
  mean(se)= 0.199(0.418)
  Meta-analysis for the bg arm
  mean(se)= -0.533(0.555)
  Indirect comparison for the bg arm
  Mean(se)= -0.184(1.113)

  Incoherence within the loop:  Mean(se)= 0.349(1.244)


44 : Evaluation of the loop bfk
Direct comparisons in the loop:
bf fk bk
 2  1  1

  Meta-analysis for the bf arm
  mean(se)= -0.384(1.032)
  Meta-analysis for the fk arm
  mean(se)= -0.186(0.241)
  Meta-analysis for the bk arm
  mean(se)= -0.27(0.265)
  Indirect comparison for the bk arm
  Mean(se)= -0.57(1.06)

  Incoherence within the loop:  Mean(se)= -0.3(1.092)


45 : Evaluation of the loop bhj
Direct comparisons in the loop:
bh hj bj
 5  3  2

  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Meta-analysis for the hj arm
  mean(se)= 0.006(0.14)
  Meta-analysis for the bj arm
  mean(se)= 0.41(0.286)
  Indirect comparison for the bj arm
  Mean(se)= -0.577(0.48)

  Incoherence within the loop:  Mean(se)= -0.986(0.559)


46 : Evaluation of the loop bhk
Direct comparisons in the loop:
bh hk bk
 5  2  1

  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Meta-analysis for the hk arm
  mean(se)= 0.41(0.213)
  Meta-analysis for the bk arm
  mean(se)= -0.27(0.265)
  Indirect comparison for the bk arm
  Mean(se)= -0.173(0.506)

  Incoherence within the loop:  Mean(se)= 0.097(0.571)


47 : Evaluation of the loop bhi
Direct comparisons in the loop:
bh hi bi
 5  1  1

  Meta-analysis for the bh arm
  mean(se)= -0.583(0.459)
  Meta-analysis for the hi arm
  mean(se)= 0.151(0.574)
  Meta-analysis for the bi arm
  mean(se)= -0.56(0.794)
  Indirect comparison for the bi arm
  Mean(se)= -0.432(0.735)

  Incoherence within the loop:  Mean(se)= 0.127(1.082)


48 : Evaluation of the loop cde
Direct comparisons in the loop:
cd de ce
 5  2  3

  Meta-analysis for the cd arm
  mean(se)= 0.148(0.156)
  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Indirect comparison for the ce arm
  Mean(se)= 0.125(0.515)

  Incoherence within the loop:  Mean(se)= -0.029(0.549)
```

```
49 : Evaluation of the loop cef
Direct comparisons in the loop:
ce ef cf
 3  2  1

  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the cf arm
  mean(se)= -0.349(0.323)
  Indirect comparison for the cf arm
  Mean(se)= -0.004(0.352)

  Incoherence within the loop:  Mean(se)= 0.344(0.477)


50 : Evaluation of the loop cek
Direct comparisons in the loop:
ce ek ck
 3  4  1

  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Meta-analysis for the ek arm
  mean(se)= -0.05(0.315)
  Meta-analysis for the ck arm
  mean(se)= -0.86(0.444)
  Indirect comparison for the ck arm
  Mean(se)= 0.105(0.369)

  Incoherence within the loop:  Mean(se)= 0.965(0.577)


51 : Evaluation of the loop cei
Direct comparisons in the loop:
ce ei ci
 3  4  2

  Meta-analysis for the ce arm
  mean(se)= 0.154(0.191)
  Meta-analysis for the ei arm
  mean(se)= -0.385(0.163)
  Meta-analysis for the ci arm
  mean(se)= -0.146(0.708)
  Indirect comparison for the ci arm
  Mean(se)= -0.231(0.251)

  Incoherence within the loop:  Mean(se)= -0.085(0.751)


52 : Evaluation of the loop cfk
```

```
Direct comparisons in the loop:
cf fk ck
 1  1  1

  Meta-analysis for the cf arm
  mean(se)= -0.349(0.323)
  Meta-analysis for the fk arm
  mean(se)= -0.186(0.241)
  Meta-analysis for the ck arm
  mean(se)= -0.86(0.444)
  Indirect comparison for the ck arm
  Mean(se)= -0.535(0.403)

  Incoherence within the loop:  Mean(se)= 0.326(0.599)


53 : Evaluation of the loop deh
Direct comparisons in the loop:
de eh dh
 2 12  2

  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the dh arm
  mean(se)= -0.11(0.223)
  Indirect comparison for the dh arm
  Mean(se)= -0.088(0.5)

  Incoherence within the loop:  Mean(se)= 0.022(0.547)


54 : Evaluation of the loop dej
Direct comparisons in the loop:
de ej dj
 2  3  3

  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the ej arm
  mean(se)= -0.007(0.154)
  Meta-analysis for the dj arm
  mean(se)= 0.019(0.159)
  Indirect comparison for the dj arm
  Mean(se)= -0.03(0.514)

  Incoherence within the loop:  Mean(se)= -0.049(0.538)


55 : Evaluation of the loop del
Direct comparisons in the loop:
de el dl
 2  1  2
```

```
  Meta-analysis for the de arm
  mean(se)= -0.023(0.491)
  Meta-analysis for the el arm
  mean(se)= 0.091(0.441)
  Meta-analysis for the dl arm
  mean(se)= -0.66(0.342)
  Indirect comparison for the dl arm
  Mean(se)= 0.068(0.66)

  Incoherence within the loop:  Mean(se)= 0.729(0.743)


56 : Evaluation of the loop dhj
Direct comparisons in the loop:
dh hj dj
 2  3  3

  Meta-analysis for the dh arm
  mean(se)= -0.11(0.223)
  Meta-analysis for the hj arm
  mean(se)= 0.006(0.14)
  Meta-analysis for the dj arm
  mean(se)= 0.019(0.159)
  Indirect comparison for the dj arm
  Mean(se)= -0.103(0.263)

  Incoherence within the loop:  Mean(se)= -0.122(0.307)


57 : Evaluation of the loop efh
Direct comparisons in the loop:
ef fh eh
 2  1 12

  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the fh arm
  mean(se)= 0.708(0.444)
  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Indirect comparison for the eh arm
  Mean(se)= 0.549(0.533)

  Incoherence within the loop:  Mean(se)= 0.615(0.542)


58 : Evaluation of the loop efg
Direct comparisons in the loop:
ef fg eg
 2  1  3

  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the fg arm
  mean(se)= 0.199(0.418)
  Meta-analysis for the eg arm
  mean(se)= -0.016(0.186)
  Indirect comparison for the eg arm
  Mean(se)= 0.041(0.512)

  Incoherence within the loop:  Mean(se)= 0.057(0.545)


59 : Evaluation of the loop efk
Direct comparisons in the loop:
ef fk ek
 2  1  4

  Meta-analysis for the ef arm
  mean(se)= -0.158(0.295)
  Meta-analysis for the fk arm
  mean(se)= -0.186(0.241)
  Meta-analysis for the ek arm
  mean(se)= -0.05(0.315)
  Indirect comparison for the ek arm
  Mean(se)= -0.344(0.381)

  Incoherence within the loop:  Mean(se)= -0.295(0.494)


60 : Evaluation of the loop ehj
Direct comparisons in the loop:
eh hj ej
12  3  3

  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the hj arm
  mean(se)= 0.006(0.14)
  Meta-analysis for the ej arm
  mean(se)= -0.007(0.154)
  Indirect comparison for the ej arm
  Mean(se)= -0.059(0.17)

  Incoherence within the loop:  Mean(se)= -0.052(0.229)


61 : Evaluation of the loop ehk
Direct comparisons in the loop:
eh hk ek
12  2  4

  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the hk arm
  mean(se)= 0.41(0.213)
  Meta-analysis for the ek arm
  mean(se)= -0.05(0.315)
  Indirect comparison for the ek arm
  Mean(se)= 0.345(0.234)

  Incoherence within the loop:  Mean(se)= 0.394(0.392)


62 : Evaluation of the loop ehi
Direct comparisons in the loop:
eh hi ei
12  1  4

  Meta-analysis for the eh arm
  mean(se)= -0.065(0.096)
  Meta-analysis for the hi arm
  mean(se)= 0.151(0.574)
  Meta-analysis for the ei arm
  mean(se)= -0.385(0.163)
  Indirect comparison for the ei arm
  Mean(se)= 0.085(0.582)

  Incoherence within the loop:  Mean(se)= 0.471(0.604)


63 : Evaluation of the loop fhk
Direct comparisons in the loop:
fh hk fk
 1  2  1

  Meta-analysis for the fh arm
  mean(se)= 0.708(0.444)
  Meta-analysis for the hk arm
  mean(se)= 0.41(0.213)
  Meta-analysis for the fk arm
  mean(se)= -0.186(0.241)
  Indirect comparison for the fk arm
  Mean(se)= 1.118(0.492)

  Incoherence within the loop:  Mean(se)= 1.304(0.548)
```

Final results:

Incoherent loops in R
"acd" "fhk" "bej"

Incoherent loops in D
"fhk" "bce" "bcd