Eur Arch Psychiatry Clin Neurosci (2010) 260:17–23
DOI 10.1007/s00406-009-0068-3

SPECIAL ISSUE

Meta-analyses: a method to maximise the evidence from clinical studies?

Wolfgang Maier · Hans-Jürgen Möller

Published online: 26 November 2009

Abstract Evidence-based medicine (EBM) is becoming the guiding principle for clinical treatment decisions. But evidence remains a loosely defined term. Multiple criteria for evidence criteria have been proposed. Most influential evidence criteria give priority to meta-analyses because they promise an objective procedure to combine the outcomes of all informative, putatively conflicting studies on the same issue in an overall score. However, we claim that meta-analyses are of limited informative value for the following six reasons: (1) meta-analyses are often "overpowered" with clinically irrelevant results that might emerge as highly significant; (2) there is serious concern of publication biases with "negative" studies not being published; (3) meta-analyses consider the variation in the results of the empirical studies included to be random noise, however, the variability of results across studies can be informative; (4) the result of a meta-analysis depends on the strategy used to identify the included empirical studies; (5) the quality of conclusions from meta-analyses depends on the statistical tests used to combine the results of the separate studies; (6) the qualitative conclusions drawn from the meta-analytical combination of individual studies may depend on specific design aspects of the individual studies. Thus, meta-analyses are primarily a method to generate hypotheses through an a posteriori analysis of treatment effects.

Keywords Meta-analyses · Evidence-based medicine · Clinical studies

W. Maier (✉)
Department of Psychiatry and Psychotherapy,
Rheinische Friedrich-Wilhelms-University,
Sigmund-Freud-Str. 25, 53105 Bonn, Germany
e-mail: Wolfgang.Maier@ukb.uni-bonn.de

H.-J. Möller
Department of Psychiatry and Psychotherapy,
Ludwig-Maximilians-University, Nussbaumstr. 7,
80336 Munich, Germany
e-mail: hans-juergen.moeller@med.uni-muenchen.de

Introduction

It is widely accepted that treatment recommendations and decisions should be evidence-based although experience-based clinical practice is needed too for many occasions [30, 34]. Evidence relies on empirical studies but is an otherwise only loosely defined term; yet, it claims objectivity. The basis of evidence-based medicine (EBM) is the combination of findings from different studies to give a global conclusion.

Different individual studies on the same issue (hypothesis) often reach different conclusions. Evidence-based conclusions therefore require that the separate pieces of evidence are cumulated in order to come to an overall conclusion; but, of course, the rules for this approach do not exist naturally but have to be defined through conventions. To date, there are no generally accepted definitions for evidence criteria; in fact, different evaluation criteria have been proposed [26].

In principle, there are two basic approaches to systematically obtaining and presenting evidence:

(1) Comparative and cumulative assessment of individual studies in narrative reviews. The various studies are considered separately; their results are compared in a qualitative manner taking the peculiarities of each study into account. The result is a balanced conclusion. There is no specific statistical method for

combining evidence from different studies. The methodologically most advanced type is the systematic review, which aims to collect all available studies. It does so by critically evaluating the design of the individual studies (with respect to their internal and external validity [18]), weighting their special features (e.g. related to setting, treatment objective) and finally deducing qualitative conclusions (e.g. about effect/efficacy). Such conclusions are normally accompanied by caveats (e.g. there is still not enough evidence available, etc.). The subjective nature of this procedure is often critically discussed, particularly with regard to the final conclusion. Yet, drug authorities are working along those procedures and demonstrate continuously that the strategy works well in consensus processes [26].

(2) Cumulative evaluation of individual studies in meta-analyses, in which statistical tests are used to combine the respective evidence. Meta-analyses are preceded by a systematic search for informative studies. The methodological quality of studies can be assessed using evaluation instruments (scales). The final cumulative evaluation only includes studies of sufficient quality. Selected studies are summarized by the resulting quantitative effect size. An overall effect size is subsequently derived by weighing each individual effect size and combining the weighted effects into an overall effect size. This procedure allows derivation of the overall effect sizes, confidence-intervals and -values.

Yet, the strategy to select studies for meta-analyses has to be considered: the adequacy of the study quality is not a yes–no phenomenon but reveals a broader multifaceted variation which cannot be represented in a categorical decision; furthermore, incomplete and inaccurate reporting of randomized controlled trials (RCTs) is common with biased estimates of treatment effects as a consequence [1].

There is a growing tendency to prioritise meta-analyses. Thus, "evidence from meta-analyses of randomized controlled studies" is the highest level of evidence in the influential Agency of Health Care Policy and Research [28]. Because of the formalised way in which they are performed and the (apparently) clear statistical result, meta-analyses increasingly form the basis for treatment recommendations in guidelines. Despite the high level of acceptance there is a risk, through misconceptions about the limitations of their informative value, of incorrect application or use of meta-analyses. Later on, we will present seven critical points about the limited informative value of meta-analyses.

Methodological procedure in meta-analyses

Meta-analyses combine the quantitative results of different studies on the same topic and sum the results to give a global value (e.g. efficacy of A compared with efficacy of placebo in a defined timeframe). Individual studies comparing two or more treatment conditions test a priori specified hypotheses and should have sufficient statistical power (in the form of sample sizes) to do so. The statistical analysis calculates p values, which can prove or disprove hypotheses. But p values depend to a large degree on the characteristics of a study (above all on the formulation of the hypothesis and the sample sizes): for example "assay sensitivity" can fail and false negative results might be the consequence. But a properly designed and powered comparative study might become a "false negative" just because of random effects. Meta-analyses combine the results of individual studies in order to come to conclusions in case of divergent results of similarly designed studies on the same hypothesis.

Comparability between studies to be combined is needed, although it is frequently not granted. For e.g., the placebo response widely varies across RCT-efficacy trials of antidepressants [33] and the antidepressants trial outcome is affected by definition of inclusion criteria and by handling of dosages [19, 20]. Despite these differences various studies are compiled in meta-analyse (e.g. [21]). Available heterogeneity tests exploring the impact of differential study characteristics are underpowered (see below).

The main summary measure for meta-analyses is the overall effect size resulting from the combination of individual effect sizes. A confidence interval can be ascertained that also allows statements to be made about the statistical significance of the effects. Thus, even for an inconsistent set of results, a comprehensive, quantitative global evaluation can be derived which integrates all relevant statements from the individual studies. An alternative summary measure is the number needed to treat (NNT) (see below in more detail).

This in principle plausible strategy to summarise empirical evidence requires the study question to be specified a priori; the consistency of the questions being investigated is, however, sometimes not unambiguously determinable, so that a meta-analysis may include studies that consider a similar but not the same question. An example of such a case is the frequently cited meta-analysis on relapse prevention with antipsychotics by Leucht et al. [22]. This analysis combined studies on maintenance treatment and on relapse prevention, although these are related but distinct questions; studies on maintenance treatment require a different design than studies on relapse prevention.

Unresolved methodological insufficiencies

Despite these methodological advancements, several unresolved problems remain; e.g. results are limited by the following discrepancies:

(a) Clinically irrelevant effects can become highly significant. Meaningful comparative studies of clinical efficacy aim to test a qualitative hypothesis (A is superior to B) and are planned accordingly (including power analyses to estimate the required sample sizes); thus, these results can stand alone. On the other hand, meta-analyses extract quantitative effect sizes from individual studies that actually investigated a qualitative question (hypothesis) without being powered to answer a specific question properly. In other words, meta-analyses use a retrospective design, while the studies they include were prospective. For this reason, the method of 'meta-analysis' has been the subject of harsh criticism from leading biometricians [10].

While effect sizes usually do not increase with sample size their confidence intervals are getting smaller and empirical p values are becoming increasingly significant. The meta-analyses are regularly 'overpowered', i.e. it is highly probable that also clinically irrelevant differences in efficacy will be shown to be significant. For this reason, when applied to clinical practice meta-analyses can cause confusion [27].

One example is the effect of memantine on cognition in probable Alzheimer's disease. A meta-analysis by Winblad et al. 2007 [35] combined placebo-controlled RCTs funded by drug companies. Each of the trials was designed to be powerful enough to come to distinct conclusions with the following results: four studies were negative, two reveal a very small positive effect. The meta-analysis comes to a significant effect in favour of memantine with an effect size of 0.2 (ranging between 0.3 and 0.1). This effect size is below the level of clinical relevance (0.25) according to the criteria by Cohen.

On the other hand, the reverse can also occur; studies with inadequate methodology can hide verum–placebo differences so that the size of a true effect is underestimated in meta-analyses [1].

(b) Publication bias is difficult to be detected. The vast majority of meta-analyses refer to published studies. A precondition for meta-analyses is the unbiased selection of studies. Preferential selection of positive studies apparently enhances the positive effects of a specific substance. The identification of the included studies through a systematic literature search is insufficient. Publications are potentially dependent on the study outcome and, therefore, might not define an appropriate basis for an unbiased meta-analysis. Early concerns, e.g. voiced by Williams [34] and Melander [24] were countered by test statistic control of presence of a bias. Current state of the art is to perform an indirect control through statistical tests comparing the observed outcomes across the selected studies to a theoretically derived random outcome distribution across studies: the funnel plot and the heterogeneity test are the most commonly used instruments to "rule out" putative publication biases. Unfortunately, the discriminative power of those tests to uncover deviations between observation and expectation is seriously limited, and absence of significance is falsely interpreted as absence of "publication bias" [16, 17]. Therefore, relevant biases might remain undetected by these means; those can only be identified through direct exploration of all available data.

Indeed, although most of the meta-analyses include funnel plots and conclude no relevant deviations from the unbiased distribution, direct comparisons between conducted and published reports tell another story, e.g. Turner et al. [32] compared reviews from the FDA for placebo-controlled RCTs of 12 antidepressant substances for short-term treatment to the corresponding publications by cumulated outcomes and effect sizes: first, 31% of the studies in the FDA files were not published; second, a strong bias favouring the preferential publication of positive studies was observed; 94% of the trials conducted appeared to be positive according to publications; yet, according to the FDA files only 51% were definitely positive. Their derived combined effect size was overestimated for each of these drugs (which all received FDA approval) through the published reports ranging from 69% for nefazodone and 64% for sertraline to 11% for paroxetine: the overall effect size (across all 12 drugs) was, however, only 41% on the basis of FDA files (32% overestimation). These tendencies were consistently demonstrated across all fields of clinical studies [8, 29].

But even if unpublished studies are systematically included in the analysis considerable discrepancies remain still possible, because there is no unique comprehensive strategy to recruit all informative unpublished studies. This was shown by a recent analysis of the effects of selective serotonin reuptake inhibitors (SSRIs) and placebo on the number of suicide attempts. Two meta-analyses on this question were independently performed and simultaneously published. Gunnell et al. [15] identified all placebo-controlled studies in Medline and the Cochrane register, while Fergusson et al. [11] used all relevant studies reported to the Medicines and Health Care Production Regulation Agency. The different search strategies of the two meta-analyses generated different results and different qualitative conclusions: Gunnell et al. [15] found no significant difference in the frequency of suicide attempts with SSRIs and placebo, while Fergusson et al. reported significantly more suicide attempts with SSRIs. Although focus of meta-analysis on RCTs might fit methodological requirements, it does not guarantee a realistic

view on clinical practice. For example, an extensive "real world" study (on an observational basis) on the same topic clearly derived totally different conclusions: SSRIs reduce suicidality [14].

(c) Meta-analysis combines individual trials by ignoring relevant sources of variation within individual studies. Current diagnostic systems are far from identifying groups of patients with homogeneous response to specific treatments. Sometimes, there is the fortunate situation that the broad variation of inter-individual responses in a heterogeneous class is substantially reduced in specific subtypes. Whereas the mean placebo–verum difference for antidepressants in the whole group of unipolar depressed patients is very modest, the same difference is substantially higher for moderate and severe cases [21]. This observation is practically highly relevant and should be incorporated in guidelines and recommendations. While systematic reviews are always keen to detect factors accounting for the inter-individual differences in magnitude of drug response, meta-analyses are prone to merge all available studies together and to miss the patterns of variation in single studies. However, this concern can partly be circumvented by careful sensitivity analyses in the context of meta-analyses as it was recently performed by Kirsch et al. [21].

(d) Effect sizes instead of testing hypothesis: are those overall effect sizes of any practical use? Are these effects sizes in danger of misuse? The quantitative, overall effect sizes for a specific drug derived from placebo-controlled RCTs can be used to create a ranking scale which propose priorities for the choice of the most appropriate drug. Yet, this procedure is not justified, as long as it is not validated through an equal ranking emerging from comparisons between pairs of drugs in RCTs. SSRIs are the only group of psychotropic drugs with available substance-specific meta-analyses for placebo-controlled RCTs as well as for head to head RCTs [9, 32]. The priority lists which can be derived from both groups of RCTs are quite different, e.g. sertraline is recommended as the drug of first choice and turns out to be together with escitalopram the most efficacious among six investigated SSRIs when relying on direct comparisons between substances; on the contrary, on the basis of the effect sizes emerging from comparisons to placebo (FDA files) sertraline together with citalopram and fluoxetine turns out to be least efficacious among the six examined SSRIs. Thus, evidence-based psychiatry cannot be based on effect sizes.

Another type of misuse of overall effect sizes obtained by meta-analyses occurs in discussion on the relative efficacy of pharmacotherapy and psychotherapy [21]. The underlying individual studies evaluating psychotropic drugs follow in the vast majority other more rigid design paradigms than studies in the psychotherapy research: blinding as well as the application of inert drugs for control of "unspecific" therapeutic effects are standards in clinical psychopharmacology, but not implemented in psychotherapy research. Consequently, effects sizes for a specific treatment—despite of their persuasive quantitative appearance—can only be used in relation to the study context they were derived from, and to the selection and statistical meta-analytic strategy. A generalization or ignorance of differences of the study context is not justified and might induce wrong conclusions.

An alternative summary measure in meta-analyses derives from subdividing the sample into responders and non-responders (remitters and non-remitters) by applying specific categorical cutoff rules to the quantitative outcome measures: the number of patients needed to treat (abbreviation: NNT) in order to increase the number of responders (remitters) by 1. The responder analysis is usually considered to demonstrate clinical relevance in a more evident manner [7]; yet, this indicator is strongly dependent on the placebo-response rate which varies broadly across studies and is dependent on multiple influencing factors [7]. Thus, this summary measure cannot be interpreted in an unambiguous manner.

(e) The qualitative conclusions from the meta-analytical combination of separate studies may depend on the design of the studies. Meta-analyses combine studies with different designs. The conclusions of meta-analyses may therefore depend on the relative weight of the evaluation approaches most often chosen. Jüni et al. [18] considered meta-analytical comparisons for several methodological conditions (e.g. adequacy of the generation and blinding of randomization codes, the way in which study dropouts are dealt with) and found that shortcomings in the handling of these design criteria led to a falsification of the respective results. There are no clear solutions as to how meta-analyses should evaluate studies with such deficits. In any case, the shortcomings of each specific study should be evaluated individually from a qualitative and quantitative point of view (for which checklists are available). Meta-analyses can only deal with those insufficiencies by exclusion of studies failing minimal quality standards. Yet, methodological quality is not a yes–no phenomenon but a multifaceted, multidimensional phenomenon which requires study-specific consideration [18].

(f) Are meta-analyses robust? Dependency of statistical tests used to combine the results of the single studies. From the methodological point of view, meta-analyses represent a group of procedures to summarise statistically quantitative study results. The various procedures are interchangeable, although they are not sufficiently robust in this respect: different meta-analytical procedures reach different qualitative results from the same set of data. A striking example of this is the comparison of meta-analyses on the efficacy of different classes of antipsychotics: Leucht et al.

[23] compared the clinical efficacy of second generation antipsychotics (SGAs) with that of first generation antipsychotics (FGA). The authors found superior efficacy of the SGAs and concluded that the possibly better efficacy of SGAs should be considered in clinical treatment decisions and that they should rather be used than FGAs. This conclusion contradicts the results of a meta-analysis by Geddes et al. [12] on the same subject. A possible explanation for this discrepancy could be the more extensive study base used by Leucht et al. For this reason, Geddes et al. [13] re-analysed the study material used by Leucht et al.; they did so by applying the statistical analysis technique they had chosen in an earlier evaluation. However, surprisingly their results were not identical to those of Leucht et al.; again, they were not able to find a difference in the efficacy of the two substance classes. The only critical difference between the two contradictory meta-analyses was the statistical test used: risk-difference quotient versus log-odds ratio. This technical difference between the two analyses of the same material resulted in qualitatively different conclusions. Thus, the validity of the statistical methods of meta-analysis must be doubted and are not robust enough to draw valid conclusions.

(g) Meta-analyses treat the variation in the results of the included empirical studies as random noise. However, the variability of results across studies can be informative. Meta-analyses consider the included studies to be repeats of the same evaluation of the same hypothesis. Yet, this assumption is usually not appropriate. Qualitatively different results of empirical studies may be caused by different clinical conditions and might not present a random event. Knowledge about this source of variance can have significant practical consequences and should not be ignored. The meta-analysis comparing the antidepressant effects of SSRIs and tricyclics by Anderson [2] is a good example of such a case. Despite the apparently quite inconsistent findings across the different studies, a slight advantage for the tricyclics was concluded. The same group re-analysed the study data 2 years later by performing separate meta-analyses for the different study settings [3]: they found that each of the two substance classes showed a different relative efficacy in outpatient and inpatient treatment settings. Tricyclics were superior in inpatient, SSRIs in outpatient settings. The original analysis had included more inpatient than outpatient studies. Thus, a meta-analysis alone cannot guarantee evidence. Therefore, sensitivity analyses exploring the impact of defined subgroups of studies are an essential supplement for meta-analyses.

Thus, there are good reasons that expert panels advising the approval of drugs or the development of clinical guidelines are not relying on the meta-analyses method. They refer to the strategy of systematic review; they stick to the view that each properly designed study is informative by itself and requires detailed consideration [6, 7, 25]. Conflicting evidence between studies requires explanation. The final conclusions have to weight evidence and counterevidence in a rational manner (which is not necessarily quantitative by nature), and the weighting process should be guided by predefined rules (e.g. guidelines by the World Federation of Biological Psychiatry [4–6]. Hence, it does not come as a surprise that meta-analytic reviews of the available compiled evidence may come to qualitatively different conclusions than reviews of the same evidence through regulatory authorities; e.g. for selected SSRIs [7, 21].

These discrepancies in grading evidence for efficacy can create substantial confusion when conclusion from meta-analyses is compared to the decision of drug authorities. Inclusion of all the studies including those with failed "assay sensitivity" and other serious methodological weakness to the meta-analyses delivers only very modest overall effect sizes and call the efficacy of antidepressants into question. Those results are at variance to the admission practice of drug authorities and to the acceptance of those drugs by patients and physicians.

The terms 'evidence' or 'empirical evidence' suggest conclusiveness and clearness. As described above, however, there are different rational approaches to generating empirical evidence from different empirical studies. Inconsistent results of individual studies may lead to differing conclusions. Which of these two pragmatic approaches is the most convincing with respect to practical conclusions? It is not possible to give a clear answer to this question. It is possible, however, to present different arguments as to why meta-analyses are only of limited informative value.

Conclusions from these arguments

Empirical evidence clearly has to be based on controlled studies. Yet, the arguments (a–g) advocate not to consider meta-analyses automatically as the highest level of empirical evidence. Clinical trials which create evidence have to be designed suitably and powered for the detection of clinically relevant effects, interpreted in the framework of the statistical theory applied. The methodological foundations for this approach have been consistently developed over a century and have contributed to the central role of biostatistics in medicine. The amount or certainty of evidence available on a certain question (hypothesis) is determined by the quality and number of known, relevant and properly designed studies. Results from individual informative trials can only be generalized and transferred to clinical practice if they are replicated.

The minimal requirement is at least one replication of an initial positive study (practiced as minimal criterion by the FDA). A more detailed, still qualitative graduation of the degree of evidence was proposed by other expert groups [4, 6]. The diversity of those qualitative criteria is needed and rationale in order to serve different pragmatic purposes. Replicability cannot be quantified in a straightforward manner; particularly not through the overall effect sizes obtained by meta-analyses. In any case, the emphasis on meta-analyses oversimplify the complexity of individual informative studies. This is also true if conclusions from meta-analyses are transferred to clinical practice.

Meta-analyses, however, might serve other purposes beyond the derivation of evidence.

How can meta-analyses contribute to create clinical evidence?

The main use of differentiated meta-analyses is the systematic post hoc exploration of the source of variation of results across different studies on a certain question. Such 'moderator' variables can be identified in sensitivity analyses, either according to the hypotheses or empirically. For the latter purpose, meta-regression models are available that can determine clinically relevant influencing factors [32]. Subsequent studies can be performed with the specific aim of testing the hypotheses obtained a posteriori. Thus, meta-analyses can be used to derive differential indicators for specific treatment procedures (see the example above on the comparison of SSRIs and tricyclics).

Thus, meta-analyses are primarily a useful way to derive hypotheses but not so much a method to maximise evidence or to present a measure for reproducibility. The test of hypotheses (e.g. a specific substance is superior to placebo or not inferior to another drug) still require suitable, prospectively planned, randomized controlled studies to test efficacy and a sufficient number of replications to validate results. The global evaluation of the available empirical data for the benefits of a treatment option should not ignore the special features and limitations, weaknesses and strengths of the individual studies.

References

1. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134:663–694
2. Anderson IM (1998) SSRIS versus tricyclic antidepressants in depressed inpatients: a meta-analysis of efficacy and tolerability. Depress Anxiety 7(Suppl 1):11–17
3. Anderson IM (2000) Selective serotonin reuptake inhibitors versus tricyclic antidepressants: a meta-analysis of efficacy and tolerability. J Affect Disord 58:19–36
4. Bandelow B, Zohar J, Hollander E, Kasper S, Möller HJ et al (2008) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the pharmacological treatment of anxiety, obsessive–compulsive and post-traumatic stress disorders—first revision. World J Biol Psychiatry 9:248–312
5. Bandelow B, Zohar J, Kasper S, Möller HJ (2008) How to grade categories of evidence. World J Biol Psychiatry 9:242–247
6. Bauer M, Whybrow P, Angst J, Versiani M, Möller H-J (2004) Biologische Behandlung unipolarer depressiver Störungen. Behandlungsleitlinien der World Federation of Societies of Biological Psychiatry (WFSBP), Stuttgart, Wissenschaftliche Verlagsgesellschaft
7. Broich K (2009) Committee for Medicinal Products for Human Use (CHMP) assessment on efficacy of antidepressants. Eur Neuropsychopharmacol 19:305–308
8. Chan AW (2008) Bias, spin, and misreporting: time for full access to trial protocols and results. PLoS Med 5:e230
9. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JPT, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M, Barbui C (2009) Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 273:746–758
10. Feinstein AR (1995) Meta-analysis: statistical alchemy for the 21st century. J Clin Epidemiol 48:71–79
11. Fergusson D, Doucette S, Glass KC, Shapiro S, Healy D, Hebert P, Hutton B (2005) Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. BMJ 330:396. doi:10.1136/bmj.330.7488.396
12. Geddes J, Freemantle N, Harrison P, Bebbington P (2000) Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. BMJ 321:1371–1376
13. Geddes J, Harrison P, Freemantle N (2003) New generation versus conventional antipsychotics. Lancet 362:404–405
14. Gibbons RD, Hur K, Bhaumik DK, Mann JJ (2005) The relationship between antidepressant medication use and rate of suicide. Arch Gen Psychiatry 62:165–172
15. Gunnell D, Saperia J, Ashby D (2005) Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. BMJ 330:385. doi:10.1136/bmj.330.7488.385
16. Hayashino Y, Noguchi Y, Fukui T (2005) Systematic evaluation and comparison of statistical tests for publication bias. J Epidemiol 15:235–243
17. Ioannidis JP, Trikalinos TA (2007) The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ 176:1091–1096
18. Jüni P, Altman DG, Egger M (2001) Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 323:42–46
19. Khan A, Kolts RL, Thase ME, Krishnan KR, Brown W (2004) Research design features and patient characteristics associated with the outcome of antidepressant clinical trials. Am J Psychiatry 161:2045–2049
20. Khan A, Schwartz K (2005) Study designs and outcomes in antidepressant clinical trials. Essent Psychopharmacol 6:221–226
21. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT (2008) Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS Med 5:e45

22. Leucht S, Barnes TR, Kissling W, Engel RR, Correll C, Kane JM (2003) Relapse prevention in schizophrenia with new-generation antipsychotics: a systematic review and exploratory meta-analysis of randomized, controlled trials. Am J Psychiatry 160:1209–1222
23. Leucht S, Wahlbeck K, Hamann J, Kissling W (2003) New generation antipsychotics versus low-potency conventional antipsychotics: a systematic review and meta-analysis. Lancet 361:1581–1589
24. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B (2003) Evidence b(i)ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 326:1171–1173
25. Möller H-J (2008) Isn't the efficacy of antidepressants clinically relevant? A critical comment on the results of the metaanalysis by Kirsch et al. 2008. Eur Arch Psychiatry Clin Neurosci 258:451–455
26. Möller H-J, Maier W (2009) Evidence-based medicine in psychopharmacotherapy: possibilities, problems and limitations. Eur Arch Psychiatry Clin Neurosci [Epub ahead of print]
27. Moncrieff J, Kirsch I (2005) Efficacy of antidepressants in adults. BMJ 331:155–157
28. National Institute for Health and Clinical Excellence Depression Management of depression in primary and secondary care—NICE guidance (2009) Available at http://www.nice.org.uk/nicemedia/pdf/CG23fullguideline.pdf.Anonymous
29. Rising K, Bacchetti P, Bero L (2008) Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. PLoS Med 5:e217
30. Stahl SM (2002) Antipsychotic polypharmacy: evidence based or eminence based? Acta Psychiatr Scand 106:321–322
31. Sterne JA, Gavaghan D, Egger M (2000) Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 53:1119–1129
32. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R (2008) Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 358:252–260
33. Walsh BT, Seidman SN, Sysko R, Gould M (2002) Placebo response in studies of major depression: variable, substantial, and growing. JAMA 287:1840–1847
34. Williams DD, Garner J (2002) The case against "the evidence": a different perspective on evidence-based medicine. Br J Psychiatry 180:8–12
35. Winblad B, Jones RW, Wirth Y, Stoffler A, Mobius HJ (2007) Memantine in moderate to severe Alzheimer's disease: a meta-analysis of randomised clinical trials. Dement Geriatr Cogn Disord 24:20–27