Eur Arch Psychiatry Clin Neurosci (2008) 258:451–455 DOI 10.1007/s00406-008-0836-5

REVIEW

Hans Jürgen Möller

Isn't the efficacy of antidepressants clinically relevant? A critical comment on the results of the metaanalysis by Kirsch et al. 2008

Received: 1 April 2008 / Accepted: 26 May 2008 / Published online: 19 November 2008

Abstract The metaanalysis of Kirsch (PLoS Med 5:e45, 2008) has (unfortunately!) attracted too much attention in the specialized press and especially in the lay press. Therefore, intensive critical commenting is necessary to not further alarm experts and health authorities as well as patients and family members. The specified commenting on these metaanalyses shall be prefaced with a short and critical commentary regarding the general significance of metaanalyses. The results of metaanalyses should not too naively be interpreted as the 'truth' as regards to the evidence based psychopharmacotherapy, but should be qualified in their significance due to principal methodological reasons Maier (Nervenarzt 78:1028–1036, 2007; Möller (Nervenarzt 78:1014–1027, 2007). Especially from these derived effect sizes should be interpreted carefully.

Key words antidepressants · effectiveness · meta-analyses

General comments on the metaanalytical approach

The specified commenting on these metaanalyses shall be prefaced with a short and critical commentary regarding the general significance of metaanalyses [10, 13]. The results of metaanalyses should not too naively be interpreted as the 'truth' as regards to the evidence based psychopharmacotherapy, but should be qualified in their significance due to principal methodological reasons [10]. Especially from these derived effect sizes should be interpreted carefully.

Great importance is attached to the results of metaanalyses regards the compilation of guidelines and textbooks on the basis of EBM [13] – possibly because the quantitative summary in effect sizes is easier to impart than differentiating qualitative conclusions on the basis of narrative review. Metaanalyses in comparison to systematic reviews do have in fact the advantage of being able to condense results into quantitative core values (effect sizes) while reviews can only draw qualitative conclusions. Even so, meta-analyses cannot replace the systematic reviews in their narrative form, which have the advantage of being able to consider in a differentiated way the special conditions of the individual studies with respect to study design, patient selection, drug dosing, etc. Yet this detailed analysis requires expert clinical-psychopharmacological knowledge and a highly detailed presentation.

The numerical value, often expressed as effect size, seems definite and significant, but is in fact full of ambiguity, resulting in basic methodical problems of meta-analyses. The seemingly succinct and pictorial value of the effect size can be too easily interpreted in a naïve and simplifying or deliberately tendentious way because of the immense and complex amount of clinical data behind it which is no longer apparent. Over-interpretation of effect size values, which can be often read, are inappropriate due to various basic problems of meta-analyses and should be questioned critically.

Due to different basic methodological problems, metaanalyses are not *via regia* for making statements regarding the efficacy or the tolerability of EBM. They are only one way among many and should be applied complementarily with other methods to summarise empirical knowledge, such as systematic reviews [13].

The metaanalysis of Kirsch has (unfortunately!) attracted too much attention in the specialized press and especially in the lay press. Therefore, intensive critical commenting is necessary to not further alarm experts and health authorities as well as patients and family members.

Prof. Dr. H.J. Möller (✉)
Chairman of the Department of Psychiatry
Ludwig-Maximilians-University Munich
Nussbaumstraße 7
80336 Munich, Germany
Tel.: +49-89/5160-5501
Fax: +49-89/5160-5522
E-Mail: hans-juergen.moeller@med.uni-muenchen.de

It should be advised against physicians or health authorities seeing the results of metaanalyses as the 'ultimate authority of truth' or over-interpreting the derived effect-sizes.

It should be underlined that major drug approval authorities, such as the American FDA and the European EMEA, are reluctant to accept metaanalyses as a primary decision basis for medication approval out of basic methodological considerations regarding confirmative hypotheses. They prefer to base their decisions on the results of methodologically adequate single studies of a confirmative character. The resulting conflicts are predictable: At worst, an approved substance could, on a metaanalytical basis in the context of EBM, be classified as ineffective. This is because not only the central phase III studies are evaluated but because other studies which are not primarily effectiveness oriented and have different objectives are also taken into account. Or because 'failed studies' are recognised as assessable, which is contrary to the practice of the drug approval authorities.

The metaanalysis by Kirsch et al. [9] does not offer any noteworthy new findings, contrary to the presentation in the lay press. The in the article presented placebo-verum difference of 1.8 HAMD points for the antidepressive effect of some modern antidepressants (apart from SSRI studies, six studies on Venlafaxine and eight studies on Nefazodone were included) is similar to other publications, amongst others to an earlier metaanalysis by Kirsch et al. 2002 [8], which found a difference value of 2.0 HAMD points. This metaanalysis was critically commented on by Fritze et al. [5].

Despite including non-published studies, the metaanalyses by Kirsch et al. [9] found a value of 1.8 HAMD points which is statistically highly significant due to its huge sample size. This should be emphasised, since the recently appeared publication by Turner et al. [20] shows a selective publication of data regarding studies on antidepressives (primarily studies with positive results are published) and critically presumes a possible over-interpretation of the effect-size of antidepressives in metaanalyses.

All definitions of clinical relevance of antidepressants' efficacy are arbitrary

It should be considered that the mean value differences on a depression scale between placebo groups and verum groups cannot show the efficacy for special patient groups. The efficacy can be considerably higher [14], due to the high variance in different patient groups, e.g. with severe depression, than revealed by the shown mean value differences. This is also mentioned by Kirsch et al. [9], who have found the biggest effect in case of severe depression at a placebo-verum difference of 4 HAMD points. Thus, it is for principal reasons not acceptable to deduct conclusions from only such general placebo-verum differences regarding the clinical relevance the way the authors do. Additionally it should be emphasized from a clinical perspective, that the effectiveness of antidepressives in clinical practice is normally optimised by sequential and combined therapy approaches [4, 18].

A statistically significant difference between placebo and verum does not automatically result in a clinical relevance of the found differences. To assess the clinical relevance of the differences Kirsch et al. referred to a suggestion of NICE [16], which regards a placebo-verum difference of 3 HAMD points as clinically relevant. Based on this, Kirsch et al. [9] generally deny the clinical relevance of the found effects of SSRIs; at best they consider them clinically relevant for severe depression. This can be countered by the fact that the cited NICE criterion is downright arbitrary and not supported, neither by empirical findings nor by expert opinion. To emphasise this it should be pointed out, that all SSRIs included in the metaanalysis were approved, among others by the European drug approval authority (EMEA) and the US American drug approval authority (FDA), and their effectiveness was therefore obviously considered clinically relevant. Hence a strange 'circulus in definiendo' comes up in the argumentation of Kirsch et al. [9] who do not, by the way, have clinical expertise in medical treatment of depression: a high relevance criterion is applied with the intention to then show the irrelevancy of the effects of antidepressants.

There is no defined criterion for the clinical relevance of antidepressive effects, there are only different approaches to evaluate them [15]. For the drug approval authorities, apart from a consistent replication of positive study results, the placebo-verum difference of approved antidepressives (see above) is definitely of importance, ranging at 2.0 HAMD points. Such a placebo-verum difference is therefore to be considered as clinically relevant.

However, much more important for the evaluation of the clinical relevance (Table 1) are the responder/remitter analyses [11], which compared the relative frequency of these categories between placebo and verum groups. This approach is demanded as an addition to the mean value analyses by drug approval authorities, to determine the clinical benefit of the therapy with an antidepressant for each single case. Considering the responder-analyses, which Kirsch et al. have unfortunately not taken account of in their metaanalytical examination, and counting the patients, whose

Table 1 Clinical relevance of antidepressives [19]

20% PL—AD difference → NNT5
15% PL—AD difference → NNT7
Reduction of myocardial infarction
with aspirin: NNT = 40
with statines: NNT = 20
NNTs for highly efficient treatment range within approximately 2–4
NNTs of ≤ 10 show a strong evidence for clinically relevant effectiveness

depression values have been reduced by at least 50% of the primal values, placebo-verum differences ranging at 15–20% are the result. This amounts to a number needed to treat (NNT) of 5–7. Such a NNT is regarded as a moderate to strong effectiveness and corresponds to the referring values of many therapies, which are standard therapies in internal medicine (Table 1). This consideration equally proves the clinical relevance of SSRIs and respectively antidepressants.

All antidepressants have reached approval despite such low classification of placebo-verum difference by Kirsch et al. [9]. This approval has been enunciated by the American drug approval authority as well as by the European drug approval authority. To imply that the examined modern antidepressants or antidepressants in general have only an insufficient effectiveness is therefore a questionable and misleading argumentation, even if Kirsch puts his focus on an arbitrary NICE criterion of a placebo-verum difference at 3 HAMD points. It should be emphasised that some antidepressives such as TCA, Venlafaxine and Escitalopram had a higher efficiency than SSRIs, which was shown in metaanalyses [1, 14, 15].

If the results of placebo controlled studies regarding a continuation therapy with antidepressives (continuation of the response of 6–12 months of the acute therapy) are considered in the argumentation as well, the conclusion regarding the clinical relevance of antidepressants is even strengthened. Kirsch et al. [9] consider only short term studies (up to 8 weeks). Geddes et al. [7] in their metaanalysis of 31 randomised, double-blind, placebo-controlled studies find a highly significant ($P < 0.00001$) efficacy of the continuation therapy of relapse rates of 41% under placebo versus 18% under verum. Therefore NNTs result in a range of 4–5.

Other critical aspects

The interpretation of the Kirsch group that the higher placebo-verum difference for severe depression is 'only' a consequence of a lower placebo-response and not of an increase of the pre-post difference in the verum-group is descriptively correct. But it is also a one-sided statistical and furthermore tendentious interpretation of the data, which disregards the fact that the effectiveness can only be deducted from the placebo-verum difference. It is therefore completely irrelevant whether a higher difference is reached by an increase in the pre-post difference of the verum group or by a decrease of the pre-post difference of the placebo group. The traditional point of view which regards 'endogenous depression' as an indication for antidepressives, TCA at that time, fit this data analysis well: strong verum-effect at a low placebo-response. The broader indication 'depressive episode' may have caused a softening and consequently possibly also a thinning out of the effectiveness of antidepressives, due to the higher placebo-response in mild/medium degrees of depression.

When interpreting the metaanalysis by Kirsch or also other single placebo controlled studies, attention should be paid to the fact that the placebo arm in a randomised control group study of an antidepressive means more than the patients receiving a placebo. The patients of both groups are additionally receiving supportive psychotherapy, co-medication with benzodiazepines/hypnotics, etc. Respective differential distinctions between placebo and verum groups have to be considered [12], since patients in the placebo group might have a higher demand in co-medication due to insufficient recovery. In a placebo-verum study, these and other methodological problems can cause the impossibility to differentiate between the efficiency in the verum group and the placebo group. This happens more often now than in the past. Today, only about 1/3 of the placebo controlled randomised double-blind antidepressives' studies have a positive result. About 1/3 are negative, which means that the antidepressive does not differ from the placebo. And about 1/3 are 'failed studies', which means that neither the new antidepressive nor the standard antidepressive differs from the placebo. The latter 'failed studies' are not regarded as significant by the drug approval authorities, since they have obviously not been carried out in accordance with antidepressive sensitive sample studies (e.g. AD non-responders) and are consequently not able to testify the efficiency of antidepressives. It is therefore also questionable to include such 'failed studies' in a metaanalysis. They lack the precondition for an efficiency proof, the „assay sensitivity" [3, 6]. Behind this, a number of methodological problems are hidden, which cannot be dealt with all in particular and only a few shall be named here: the inclusion of 'symptom bearers' found through newspaper ads (not patients in the true sense of the meaning), especially in the USA; a too high rating of the level of depression before the admittance to a study, (also called 'overrating', for the patient to fulfil the admittance criteria); a high number of patients receiving additional psychopharmaca (hypnotics, anxiolytics); a 'too good' standard care (e.g. very good supportive psychotherapeutical treatment). All these measures lead to a reduction of the possibility to separate the verum from the placebo regarding the efficiency.

Taking the argument of the 'failed studies' and the lacking 'assay sensitivity' into account of this here general argumentation, the metaanalysis by Turner et al. [20] can be somewhat relativised in its critical significance. Turner criticises, that a partly selective non-publication of not-positive studies leads to an over-evaluation of the effectiveness of antidepressives. This effect is also shown by him in the metaanalysis. However, negative studies and 'failed studies' should generally be differentiated when including not-published studies. Were 'failed studies' neglected

from metaanalyses for good a reason, the artefact due to publication bias criticised by Turner would probably be much less.

Different metaanalyses of antidepressives can differ substantially when compared, depending on whether such 'failed studies' had been included or not. The same goes for negative studies, i.e. studies where the experimentally examined substances have not differed from the placebo and, in case a standard comparing substance was added, the standard substance has differed from the placebo. The results of metaanalyses always have to be questioned regarding the included studies: were only studies with positive results included? Were studies with negative results or 'failed studies' also included? The inclusion of the latter naturally leads to a decreased effect size in the metaanalysis.

The tendency not to publish rather unfavourable study results has a complex background. Regards the pharmaceutical industry there is understandably a tendency to initially show the positive studies. The negative studies are published later if need be, and then often only in form of a summarising overview and not as the original work. The same goes for 'failed studies', which to the pharmaceutical industry have a too minor value in result and, just as studies with negative results, are only reluctantly taken on for publication by editors of scientific journals. Since a number of years it is mandatory for psychopharmacological studies to be registered in an international database. This offers for the future the possibility to control whether the main results of studies have actually been published. This is certainly a good development which makes it possible to eliminate the publication bias to a large extent.

Conclusions

Altogether the results of all antidepressive studies are actually that robust [2, 4] that the critical objections cannot disprove the positive overall result. What we need to be aware of just because of the newer metaanalyses is the fact that the mean placebo-verum difference amounts to only about 2 HAMD points and should therefore not be over-interpreted. By interpreting this value it should be taken into consideration that the study conditions were highly artificial and vulnerable to bias and could possibly underestimate the actual therapy effect of the antidepressant. In the everyday clinical practice we regard the effectiveness of antidepressants as much more positive, especially in case of patients who have not been pre-treated and are not partial non-responders.

Since evidence graduation and effect sizes are now increasingly used for showing the empirical evaluation in psychotherapy/psychosocial therapy as well, there is a principal possibility to compare these with the evidence grades and effect sizes from those of the psychopharmacotherapy. This leads to the danger of meaninglessly comparing effect sizes or respectively evidence grades based on different methodologies of therapy evaluation [13]. Also the publication by Kirsch et al. [9] implies such an inappropriate comparison of effect sizes, if the authors consider alternative therapies. The evaluation of the psychotherapy procedures is not testified under placebo or respectively double-blind conditions. The different methodological basis on which the evidence grading in psychotherapy and psychopharmacology is built makes such a direct comparison simply impossible [17].

The lurid and tendentious makeup of the articles in the lay press about Kirsch's metaanalysis has caused a strong uncertainty among patients. This is regrettable since it is already known that many patients, who would in actual fact need an antidepressive treatment, do not take antidepressants because of miscellaneous attitudes and anxieties, now less than ever.

References

1. Anderson IM (2000) Selective serotonin reuptake inhibitors versus tricyclic antidepressants: a meta-analysis of efficacy and tolerability. J Affect Disord 58:19–36
2. Baghai TC, Volz HP, Moller HJ (2006) Drug treatment of depression in the 2000s: an overview of achievements in the last 10 years and future possibilities. World J Biol Psychiatry 7:198–222
3. Baldwin D, Broich K, Fritze J, Kasper S, Westenberg H, Möller HJ (2003) Placebo-controlled studies in depression: necessary, ethical and feasible. Eur Arch Psychiatry Clin Neurosci 253: 22–28
4. Bauer M, Bschor T, Pfennig A, Whybrow PC, Angst J, Versiani M, Moller HJ (2007) World federation of societies of biological psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders in primary care. World J Biol Psychiatry 8:67–104
5. Fritze J, Aldenhoff J, Bergmann F, Maier W, Möller HJ (2005) Antidepressiva: Lebensgefährliche Placebos? Arznei-Telegramm: fahrlässiges journal? Psychoneuro 31:480–484
6. Fritze J, Möller HJ (2001) Design of clinical trials of antidepressants. Should a placebo control arm be included? CNS Drugs 15(10):755–764
7. Geddes JR, Carney SM, Davies C, Furukawa TA, Kupfer DJ, Frank E, Goodwin GM (2003) Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. Lancet 361:653–661
8. Kirsch I, Moore TJ, Scoboria A, Nicholls S (2002) The emperor's new drugs: an analyses of antidepressant medication data submitted to the U.S. Food and Drug Administration. Prevention & Treatment 5 (Artikle 23)
9. 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 5:260–268
10. Maier W, Möller HJ (2007) Metaanalysen. Methoden zur Evidenzmaximierung von Therapiestudien. Nervenarzt 78: 1028–1036
11. Möller HJ (2008a) Is there a need for a new psychiatric classification at the current state of knowledge? World J Biol Psychiatry 9:82–85
12. Möller HJ (2008b) Methodik empirischer Forschung. In: Möller HJ, Laux G, Kapfhammer H-P (eds) Psychiatrie und psychotherapie. Springer, Heidelberg, pp 346–367

13. Möller HJ, Maier W (2007) Probleme der ''evidence-based medicine'' in der Psychopharmakotherapie. Problematik der Evidenzgraduierung und der Evidenzbasierung komplexer klinischer Entscheidungsprozesse. Nervenarzt 78:1014–1027
14. Montgomery SA, Kasper S (2007) Severe depression and antidepressants: focus on a pooled analysis of placebo-controlled studies on agomelatine. Int Clin Psychopharmacol 22:283–291
15. Montgomery SA, Möller HJ (2008) The clinical relevance of the significant advantage of escitaloprom compared to other antidepressants. (in preperation)
16. National Institute for Clinical Excellence (2004) Depression: management of depression in primary and secondary care. Clinical practice guideline no 23. National Institute for Clinical Excellence, London
17. Nutt DJ, Sharpe M (2008) Uncritical positive regard? Issues in the efficacy and safety of psychotherapy. J Psychopharmacol 22:3–6
18. Russh AJ (2007) STAR*D: what have we learned? Am J Psychiatry 164:201–204
19. Storosum JG, Elferink AJ, van Zwieten BJ, van den BW, Gersons BP, van Strik R, Broekmans AW (2001) Short-term efficacy of tricyclic antidepressants revisited: a meta-analytic study. Eur Neuropsychopharmacol 11:173–180
20. 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