The clinical significance of antidepressant treatment effects cannot be derived from placebo-verum response differences

U. Hegerl and R. Mergl

J Psychopharmacol 2010; 24; 445 originally published online Oct 13, 2009;
DOI: 10.1177/0269881109106930

Perspective

The clinical significance of antidepressant treatment effects cannot be derived from placebo-verum response differences

Journal of Psychopharmacology
24(4) 445–448

DOI: 10.1177/0269881109106930
jop.sagepub.com

U Hegerl and R Mergl

Abstract

A placebo–verum difference in antidepressant response of at least three points on the Hamilton Depression Rating Scale has been proposed as the threshold for clinical significance in antidepressant trials. Given the considerable clinical and regulatory consequences of such a definition of clinical relevance it deserves a critical discussion. Eight arguments are presented indicating that using this definition as a criterion for assessing the clinical utility of antidepressants in daily practice would risk erroneously discarding treatments with a clear benefit for patients.

Keywords

antidepressive agents, clinical trial, depression, effect size, Hamilton Depression Rating Scale

In order to balance the costs and benefits of antidepressant drug treatment, the first step is to show that the treatment works in principle by statistically proving superiority compared with a placebo (efficacy studies). The next question is whether the antidepressant effect is strong enough to be of clinical significance and to balance the risks and costs of the treatment. It has been proposed by the National Institute for Clinical Excellence (NICE) (2004) that the difference in improvement in depression severity measured with the Hamilton Depression Rating Scale (HAMD-17) (Hamilton, 1960) between an antidepressant and a placebo should be at least three points or consist of at least a standardized mean difference of 0.50 in order to be considered as clinically relevant.

This criterion for clinical relevance appears to have gained some acceptance by the scientific community. For example, based on this definition of clinical relevance, the conclusion of a recent meta-analysis on antidepressants was that these drugs are not clinically relevant even in patients with moderate and severe depression and should only be used for patients with severe depression (Kirsch et al., 2008). This statement triggered a loud public echo in many countries, questioning the clinical value of antidepressants. Questioning the value of antidepressants is not without risk because physicians might abstain from starting efficacious treatments, and patients' therapy motivation, as well as compliance, might be negatively affected. This has probably been a consequence of the broad discussion in 2004 of possible suicide-inducing effects of antidepressants in children and adolescents, which was followed by both a reduction in antidepressant prescriptions and an increase in suicide rates in these age groups (e.g. Gibbons et al., 2007; Katz et al., 2008). Far-reaching decisions by regulatory agencies or health insurance providers regarding antidepressants might also be influenced by this or similar criteria of clinical relevance.

In the following it will be argued that a delta of at least three points in HAMD-17, or similar criteria obtained from placebo-controlled trials (efficacy studies), cannot be used even as a rough estimation of the clinical relevance of antidepressant treatment effects, and use of such criteria is likely to be grossly misleading (see also Möller, 2008).

(1) In randomised controlled trials (RCTs) the placebo ('I will please') will induce hope because patients do not know whether they receive placebo or active drug (verum). This is not comparable with situations in daily practice when the physician may decide to propose, for example, watchful waiting instead of a specific treatment to the patient. Such a decision will not induce a similar amount of hope and might even be detrimental to the patient (nocebo effect). In addition, patients treated with antidepressants in daily practice can be sure that they receive an active treatment, whereas in a RCT it is possible they may receive only placebo. This will decrease the treatment effect with antidepressants in RCTs compared with daily practice because hope is mitigated by doubts about receiving antidepressants or placebos. The size of this effect is likely to be relevant as Sneed et al. (2008) reported that antidepressant response rates in trials with two active drugs (comparative trials) were on average 60%, but only 46% in placebo-controlled trials. In daily practice, hope-related non-specific effects will be

Department of Psychiatry, University of Leipzig, Leipzig, Germany.

Corresponding author:
Ulrich Hegerl, Department of Psychiatry, University of Leipzig, Semmelweisstrasse 10, D-04103 Leipzig, Germany.
Email: ulrich.hegerl@medizin.uni-leipzig.de

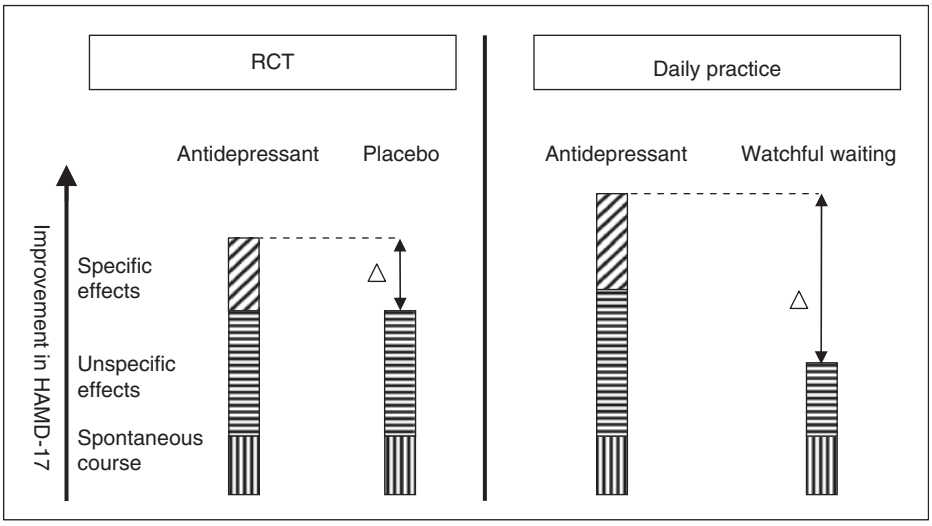

Figure 1. A simplified model considering spontaneous course, unspecific effects (e.g. induction of hope, care, activation) and specific effects of antidepressants as additive factors contributing to response. The response difference (△) between antidepressant and placebo within RCTs must be expected to be clearly smaller than that between receiving antidepressants versus watchful waiting in daily practice. Under the latter condition, the effect of antidepressants is likely to be larger than in RCTs (see arguments (1) to (5)) and the effect of watchful waiting smaller than that of placebo. HAMD-17: Hamilton Depression Rating Scale (17-item version) (Hamilton, 1960); RCT: randomised controlled trial; △: placebo–verum response difference.

larger for antidepressants and lower for the alternative ('watchful waiting'), resulting in larger delta between these two options (see Figure 1).

(2) The placebo effect in antidepressant RCTs has increased in the last decades and is around 30% (Walsh et al., 2002). This probably reflects in part the longer duration of acute antidepressant trials (up to 12 weeks, versus 4–6 weeks in earlier studies), resulting in an increased rate of remission due to spontaneous course of the disease. A further important factor is the increasingly high level of general care, activity and motivation associated with a modern RCT. Bearing in mind the short timescales available for the care of depressed patients in primary and secondary care, a comparable non-specific antidepressant effect cannot easily be induced in daily practice. It might be argued that such non-specific antidepressant effects inherent in modern RCTs add equally to the response in both the placebo and verum groups and therefore do not necessarily distort the placebo–verum difference in outcome. However, it is not known to what degree non-specific and specific effects in antidepressant treatment can be seen as independent and additive, or as interactive factors, nor how many of the patients in the verum group considered to improve due to non-specific factors would also be responders to specific effects of the antidepressant (see also Rihmer and Gonda, 2008).

(3) In daily practice, guidelines recommend changing treatment with antidepressants after 2–4 weeks if no clinical improvement is observed. Adjustment of dosage, changing to another antidepressant with different pharmacodynamic properties or augmentation strategies are good clinical practice. Although not proven by stringent studies it is likely that these established strategies increase the likelihood that the right dosage and drug will be found for the individual patient, and the chance to respond will be increased compared with the more rigid RCT, where one antidepressant is used for 4–10 weeks.

(4) Many of the multicentre RCTs are performed in countries in which the majority of patients do not have health insurance. Inclusion in such a study offers the chance to receive free medical treatment and other support. In addition, recruitment of patients is performed mostly by study centres which receive a certain amount of money for every included patient. Furthermore, many of these studies are performed under demanding time pressures. It would be surprising if such a constellation of pressures did not result in many patients being included who should not have been. For example, patients knowing that they will only be included in a trial if they reach a certain score in the depression rating could easily exaggerate problems with sleep or appetite in order to reach the cut-off level. After inclusion, exaggerations of depression severity are no longer necessary. Such mechanisms will lead to high placebo response rates as well as to a dilution of the effects of antidepressants, because the right patients for the antidepressants may not have been included. The placebo–verum difference in response will therefore shrink. A systematic comparison of placebo–verum differences between patients with and without health insurance, or between studies in countries with and without a general health insurance scheme would be of interest in this context.

(5) Many antidepressant RCTs are performed with outpatients. A high rate of non-compliance has to be expected in

these studies (Demyttenaere et al., 2008). Possibly, some patients may want to be included in RCTs for reasons other than receiving antidepressants (see above), and might not take the pills. Others stop taking their pills for other reasons, or simply forget. A control of adherence to dosage regimen, e.g. by measuring plasma levels, is seldom performed in phase III studies. In studies on outpatients, especially in those with milder depression, the percentage of patients with non-adherence can be estimated to be around 40% or higher (e.g. Demyttenaere et al., 1998; Peveler et al., 1999; Thompson et al., 2000). This will obviously lower the observed mean placebo–verum difference in outcome. In addition, drop out rates of around 30% (e.g. Hotopf et al., 1997) will further obscure the efficacy of the treatment, especially if intent-to-treat analyses with last observation carried forward methodology are applied. Using only completer or per protocol analyses for estimating clinical relevance would mitigate this source of bias. It is obvious that the clinical relevance of treatment cannot be calculated from the outcome of patients not receiving the treatment.

(6) A mean reduction of symptom severity of, say, two points, means that there are patients with even smaller reductions, and patients with greater reductions. Since markers for the identification of the subgroup of 'high responders' are presently not available, the question of whether or not a probationary treatment is justified has to be discussed. This would allow at least a subgroup to benefit from a treatment which works well for them, while treatment could be stopped in the others. Such a strategy is facilitated by the fact that the decision to continue or stop treatment can be made after two weeks of treatment. Patients without improvement after two weeks (e.g. less than 20% in HAMD-17) have a high chance of being non-responders after 10 weeks (79–94%) (Szegedi et al., 2003; Papakostas et al., 2006; Taylor et al., 2006; Stassen et al., 2007; Henkel et al., 2008; van Calker et al., 2009a; Tadić et al., 2009).

(7) Defining clinical relevance as the placebo–verum difference in acute treatment does not take into account additional benefits of antidepressant medication resulting from their clearly proven and strong effects concerning prevention of relapse and recurrence (e.g. Geddes et al., 2006; Reynolds et al., 2006; Blier et al., 2007; Kornstein, 2008). Compared with treatment discontinuation, antidepressants can reduce the risk of relapse or recurrence by about 50–70%. The mean rate of relapse on placebo was 41%, as compared with 18% on antidepressant drugs (Geddes et al., 2003).

(8) Independently of the arguments presented so far, defining two points in the HAMD-17 as not being clinically relevant is questionable. Such an improvement can, for example, signify the change from 'wishes s/he were dead' to 'suicidality absent' (Hamilton, 1960).

There are also factors which might introduce a bias in the opposite direction. For example, problems with blinding and exclusion of patients with co-morbidity and co-medication in RCTs might suggest better efficacy than that observable in daily practice. However, it is unlikely that these factors counterbalance the above-mentioned factors (arguments (1) to (7)).

For the clinical decision pro or con a certain treatment, many further individual aspects such as patients' preferences, unwanted side effects of treatments, or complicating factors (e.g. co-morbidity, multiple medications) come into play. Of major relevance for this decision process is also whether or not alternative evidence-based treatment options are available. In relation to depression, sceptical opinions about antidepressants are often combined with support for psychotherapy as a better alternative (e.g. Kirsch et al., 2008). This happens in spite of the fact that for psychotherapy the evidence base for efficacy is considerably weaker than that for antidepressants. This should be obvious because of difficulties in defining a control group and in blinding therapists as well as patients (e.g. Nutt and Sharpe, 2008). Often, even raters have not been blinded. Without blinding of patients, no placebo effect will be induced in the control group. Knowing that one is only in a 'waiting list group', 'treatment as usual group' or a 'moderated self-help group' might be unhelpful and could even result in nocebo effects. This was clearly shown by a recent five-arm study (MIND-study, $n = 368$; Hegerl et al. (2009)) where primary care patients with milder forms of depression were randomly assigned to treatment with: (1) an antidepressant (sertraline); (2) pill placebo; (3) cognitive behavioural therapy (CBT); (4) moderated self-help group (psychotherapy control condition); or (5) a free choice group (free choice of sertraline or CBT). The striking finding was that the outcome in the moderated self-help group was worse compared with the CBT group, but also compared with the pill placebo group and all other groups. Demonstrating superiority over such psychotherapy control groups obviously is not providing the same high level of evidence for efficacy as pill-placebo–antidepressant comparisons.

If the placebo–verum difference in depression ratings cannot be used to judge the clinical relevance of antidepressant effects in daily practice, what might be a better alternative? Comparing the improvement with antidepressants not to pill-placebo but to watchful waiting with supportive care within an open, randomised design might come closer to the real-life situation. In the MIND-study, the sertraline–placebo difference was 2.3 points in HAMD-17 and so would have been considered as clinically not relevant according to the NICE criteria. However, compared with the moderated self-help group (psychotherapy control group), the outcome difference was 4.87 points in HAMD-17 and therefore without doubt of clinical relevance. This was the case in spite of the fact that only patients with milder and even minor forms of depression were included.

In summary, according to arguments (1) to (5), the outcome difference between being treated with antidepressant versus watchful waiting or treatment as usual in daily practice is likely to be much larger than that between antidepressant versus placebo within RCTs. These and further arguments ((6) to (8)) show that it could be misleading to discard antidepressant treatments based on a criterion such as at least three points of placebo–verum difference in HAMD-17. Pill-placebo outcome differences in RCTs are best for proving the efficacy of a certain treatment; outcome differences between antidepressant treatment and watchful waiting in randomised

trials with blinding of only the raters might be best to judge the clinical significance of the treatment effects.

References

Blier P, Keller MB, Pollack MH, Thase ME, Zajecka JM, Dunner DL (2007) Preventing recurrent depression: long-term treatment for major depressive disorder. *Journal of Clinical Psychiatry* 68: e06.

Demyttenaere K, Adelin A, Patrick M, Walthère D, de Katrien B, Michèle S (2008) Six-month compliance with antidepressant medication in the treatment of major depressive disorder. *International Clinical Psychopharmacology* 23: 36–42.

Demyttenaere K, Van Ganse E, Gregoire J, Gaens E, Mesters P (1998) Compliance in depressed patients treated with fluoxetine or amitriptyline. *Belgian Compliance Study Group*. International Clinical Psychopharmacology 13: 11–17.

Geddes JR, Carney SM, Davies C, et al. (2003) Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *Lancet* 361: 653–661.

Geddes J, Freemantle N, Mason J, Eccles M, Boynton J (2006) Selective serotonin reuptake inhibitors (SSRIs) versus other antidepressants for depression. *Cochrane Database of Systematic Reviews* 3: CD001851.

Gibbons RD, Brown CH, Hur K, et al. (2007) Early evidence on the effects of regulators' suicidality warnings on SSRI prescriptions and suicide in children and adolescents. *American Journal of Psychiatry* 164: 1356–1363.

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

Hegerl U, Hautzinger M, Mergl R, et al. (2009) Effects of pharmaco- and psychotherapy in depressed primary care patients. A randomised, controlled trial including a patient choice arm. *International Journal of Neuropsychopharmacology* (doi: 10.1017/S1461145709000224).

Henkel V, Seemüller F, Obermeier M, et al. (2008) Does early improvement triggered by antidepressants predict response/remission? Analysis of data from a naturalistic study on a large sample of inpatients with major depression. *Journal of Affective Disorders* 115: 439–449.

Hotopf M, Hardy R, Lewis G (1997) Discontinuation rates of SSRIs and tricyclic antidepressants: a meta-analysis and investigation of heterogeneity. *British Journal of Psychiatry* 170: 120–127.

Katz LY, Kozyrskyj AL, Prior HJ, Enns MW, Cox BJ, Sareen J (2008) Effect of regulatory warnings on antidepressant prescription rates, use of health services and outcomes among children, adolescents and young adults. *Canadian Medical Association Journal* 178: 1005–1011.

Kirsch I, Deacon BJ, Hudeo-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 Medicine* 5: e45.

Kornstein SG (2008) Maintenance therapy to prevent recurrence of depression: summary and implications of the PREVENT study. *Expert Review of Neurotherapeutics* 8: 737–742.

NICE (National Institute for Clinical Excellence) (2004) *Depression: Management of Depression in Primary and Secondary Care. Clinical Practice Guideline No. 23*. London, UK: National Institute for Clinical Excellence.

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*. European Archives of Psychiatry and Clinical Neuroscience 258: 451–455.

Nutt DJ, Sharpe M (2008) Uncritical positive regard? Issues in the efficacy and safety of psychotherapy. *Journal of Psychopharmacology* 22: 3–6.

Papakostas GI, Perlis RH, Scalia MJ, Petersen TJ, Fava M (2006) A meta-analysis of early sustained response rates between antidepressants and placebo for the treatment of major depressive disorder. *Journal of Clinical Psychopharmacology* 26: 56–60.

Peveler R, George C, Kinmonth A-L, Campbell M, Thompson C (1999) Effect of antidepressant drug counselling and information leaflets on adherence to drug treatment in primary care: randomised controlled trial. *British Medical Journal* 319: 612–615.

Reynolds CF III, Dew MA, Pollock BG, et al. (2006) Maintenance treatment of major depression in old age. *New England Journal of Medicine* 354: 1130–1138.

Rihmer Z, Gonda X (2008) Is drug-placebo difference in short-term antidepressant drug trials on unipolar major depression much greater than previously believed? *Journal of Affective Disorders* 108: 195–198.

Sneed JR, Rutherford BR, Rindskopf D, Lane DT, Sackeim HA, Roose SP (2008) Design makes a difference: a meta-analysis of antidepressant response rates in placebo-controlled versus comparator trials in late-life depression. *American Journal of Geriatric Psychiatry* 16: 65–73.

Stassen HH, Angst J, Hell D, Scharfetter C, Szegedi A (2007) Is there a common resilience mechanism underlying antidepressant drug response? Evidence from 2848 patients. *Journal of Clinical Psychiatry* 68: 1195–1205.

Szegedi A, Müller MJ, Anghelescu I, Klawe C, Kohnen R, Benkert O (2003) Early improvement under mirtazapine and paroxetine predicts later stable response and remission with high sensitivity in patients with major depression. *Journal of Clinical Psychiatry* 64: 413–420.

Tadić A, Helmreich I, Mergl R, et al. (2009) Early improvement is a predictor of treatment outcome in patients with mild major, minor or subsyndromal depression. *Journal of Affective Disorders* (doi: 10.1016/j.jad.2009.04.014).

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: 1217–1223.

Thompson C, Peveler RC, Stephenson D, McKendrick J (2000) Compliance with antidepressant medication in the treatment of major depressive disorder in primary care: A randomized comparison of fluoxetine and a tricyclic antidepressant. *American Journal of Psychiatry* 157: 338–343.

van Calker D, Zobel I, Dykierek P, et al. (2009) Time course of response to antidepressants: Predictive value of early improvement and effect of additional psychotherapy. *Journal of Affective Disorders* 114: 243–253.

Walsh BT, Seidman SN, Sysko R, Gould M (2002) Placebo response in studies of major depression: variable, substantial, and growing. *Journal of the American Medical Association* 287: 1840–1847.