Meta-analysis of trials comparing antidepressants with active placebos.

J Moncrieff, S Wessely and R Hardy

BJP 1998, 172:227-231.
Access the most recent version at DOI: 10.1192/bjp.172.3.227

Meta-analysis of trials comparing antidepressants with active placebos†

JOANNA MONCRIEFF, SIMON WESSELY and REBECCA HARDY

Background Unblinding effects may introduce bias into clinical trials. The use of active placebos to mimic side-effects of medication may therefore produce more rigorous evidence on the efficacy of antidepressants.

Method Trials comparing antidepressants with active placebos were located. A standard measure of effect was calculated for each trial and weighted pooled estimates obtained. Heterogeneity was examined and sensitivity analyses performed. A subgroup analysis of in-patient and out-patient trials was conducted.

Results Only two of the nine studies examined produced effect sizes which showed a consistent significant difference in favour of the active drug. Combining all studies produced pooled effect size estimates of between 0.41 (0.27–0.56) and 0.46 (0.31–0.60) with high heterogeneity due to one strongly positive trial. Sensitivity analyses excluding this and one other trial reduced the pooled effect to between 0.21 (0.03–0.38) and 0.27 (0.10–0.45).

Conclusions Meta-analysis is very sensitive to decisions about exclusions. Previous general meta-analyses have found combined effect sizes in the range 0.4–0.8. The more conservative estimates produced here suggest that unblinding effects may inflate the efficacy of antidepressants in trials using inert placebos.

†See Commentary, pp. 232–234 this issue.

'Unblinding' effects represent a source of potential bias in controlled trials. These occur when a supposed double-blind design is subverted because the different physiological experiences associated with ingestion of an active drug and an inert placebo lead subjects and assessors to suspect the identity of the medication (Greenberg & Fischer, 1994). It has been confirmed that antidepressants, among other drugs, can be distinguished from placebo (White *et al*, 1992). Some research indicates that unblinding can produce spurious positive results in the absence of a real effect (Engelhardt *et al*, 1969; Toneatto & Sellers, 1992). Placebos containing active substances have been used to address this problem, and drugs with anticholinergic actions (most commonly atropine) have typically been employed to mimic side-effects of tricyclic antidepressants (TCAs). Ethical consensus would prevent the execution of another such study at present and so meta-analysis of controlled trials using such 'active placebos' provides an opportunity to investigate the efficacy of antidepressants under conditions of greater blindness.

METHOD

Electronic searches were performed using the databases Medline, Embase and PsychLit and key terms 'active placebos' and 'atropine'. Trials identified from hand searches of major psychiatric journals were scanned, as well as reference lists of previous published reviews. Inclusion criteria, in addition to the use of active placebos, were that the trial was concerned with the treatment of depression, an antidepressant currently regarded as efficacious was used, allocation was random and some outcome assessment of mood was made.

A variety of different outcome measures were used in the trials identified. These were converted to effect sizes to obtain a standard measure across trials. Change in mood at the end of treatment was defined as the outcome of interest. This was obtained either from change in scores on rating scales pre- and post-treatment or from direct measures of improvement or change. Where there was a choice, the observer-rated measure indicated by the authors as the one of principal importance was selected, or if none was specified, priority was given to instruments that have been widely used and subject to reliability testing. Where different measures or ratings within the same study disagreed substantially, separate effect sizes were calculated. 'Intention to treat' data were used where possible. In one trial, with a large number of early withdrawals, this was calculated by assigning a poor outcome to drop-outs (Daneman, 1961). Results consisting only of categorical ratings of degree of improvement were weighted and mean scores and standard deviations obtained as described in a previous meta-analysis in this area (Quality Assurance Project, 1983). Results adjusted for baseline values were used where they were presented.

Effect sizes were calculated by subtracting the mean score in the placebo group from that of the group on the antidepressant and dividing by the pooled standard deviation. A number of papers did not report standard deviations and so estimates were obtained from other trials using the same outcome measures and similar subject groups. Methods described by Hedges & Olkin (1985) were used to calculate the overall effect size using a fixed effects model and weighting each individual effect size by the inverse variance. An approximation to the variance which does not depend on the effect size was used ($(n_1+n_2)/n_1n_2$) to avoid the problem of including an estimate of effect size in the calculation of each weight. All pooled calculations included a test of heterogeneity. A subgroup analysis of in-patients and out-patients, defined *a priori*, was also performed. Sensitivity analyses were conducted using various combinations of trials and estimates.

RESULTS

Individual studies

Nine trials were identified which satisfied inclusion criteria (see Table 1). All compared TCAs, at a minimum dose of 100 mg amitriptylene, with placebos containing atropine. The effect sizes calculated for each study in units of standard deviation

Table 1 Characteristics of trials

Trial	Design	Antidepressant	Placebo	Source of subjects	*n*	Female subjects (%)	Mean age of subjects (range)	Duration	Outcome measure used to calculate effect size	Number of subjects analysed drug/placebo
Daneman, 1961	parallel groups	imipramine, mean dose 133 mg	atropine, 1.25 mg	depressed out-patients	195	69	NR (17–75)	Variable, evaluations at 1 and 2 months	4 'response to treatment' categories scored	94/101
Uhlenthuth & Park, 1963	crossover[1]	imipramine, 150 mg	atropine, 0.6 mg	depressed out-patients	50	76	42 (22–71)	4 weeks	TDS change	22/20
Weintraub & Aronson, 1963	parallel groups	imipramine, 150 mg	atropine, 0.6 mg	depressed in-patients	89	60	51 (19–73)	4 weeks	3 categories of improvement scored Ratings by hospital director and ward doctor	36/33
Wilson *et al*, 1963	parallel groups factorial design (also evaluating ECT)	imipramine, 150–220 mg	atropine, dose NR	depressed in-patients	24	100	NR (40–59)	5 weeks	HDRS change	10/12
Hollister *et al*, 1964	parallel groups	imipramine, mean dose 171 mg and amitriptyline, mean dose 157 mg	atropine, 1 mg	depressed in-patients (veterans)	110	0	median 43 (26–72)	3 weeks	IMPS change	62/31
Friedman *et al*, 1966	parallel groups	imipramine, 150–200 mg	atropine, dose NR	depressed in-patients	78	about 66	60	3 weeks	Global Clinical Improvement Scale	36/26
Hussain, 1970	parallel groups	amitriptyline and amitriptyline–perphenazine combination (doses NR)	atropine, dose NR	depressed patients from 'hospital practice'	34	NR	NR	(assessed on 3 occasions)	5 categories of improvement scored	15/19
Friedman, 1975	parallel groups factorial design (also evaluating marital therapy)	amitriptyline, 100 mg	atropine, 0.4 mg	depressed married out-patients	196	79	36 (21–67)	12 weeks	Global Clinical Improvement Scale	98/98
Murphy *et al*, 1984	parallel groups with adjunctive cognitive therapy	nortriptyline, 100–150 mg	atropine, 0.1–0.15 mg+ phenobarbital sodium 10–15 mg	depressed out-patients	39	66 (completers)	34 (19–59) completers	12 weeks	HRSD change	22/17

1. Data for the first four weeks was presented as for a parallel group design.
NR, not reported; TDS, Total Distress Score (developed from Symptom Check List); HRSD, Hamilton Rating Scale for Depression; IPMS, In-Patient Multidimensional Psychiatric Scale; 1 s.d. estimated from Hollister *et al* (1963), 2 s.d. estimated from Friedman (1975), 3 s.d. estimated from Rush *et al* (1977).

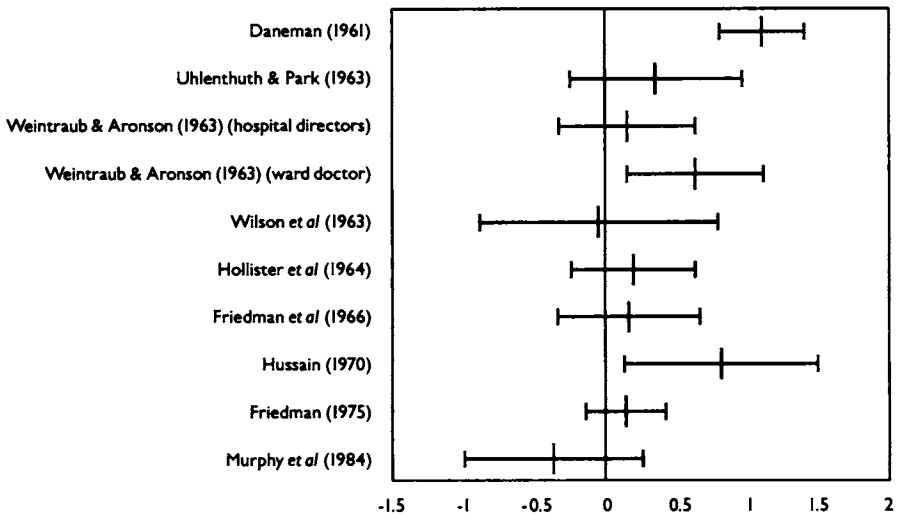

Fig. 1 Effect size.

are shown in Fig. 1. Ratings by the two observers in the trial of Weintraub & Aronson (1963) yielded discrepant estimates of effect size, and meta-analysis was conducted separately using both estimates. In three trials (Hollister *et al*, 1964; Friedman *et al*, 1966; Murphy *et al*, 1984) standard deviations for the relevant measures were not reported and estimates were taken from studies by the same authors or, in one case (Murphy *et al* 1984), from the study that the authors referenced as their blueprint (Rush *et al*, 1977). Two trials showed a consistent and statistically significant difference favouring the antidepressant drug over placebo (Hussain, 1970; Daneman, 1961), although only one of these authors (Daneman, 1961) concluded that an effect had been demonstrated.

Combined analyses

Combining effect sizes from all nine trials, using the more conservative estimate from Weintraub & Aronson (1963), yielded a pooled estimate of 0.41 (95% CI 0.27–0.56, see Table 2). However, a high degree of heterogeneity was revealed. Inspection of the results (see Fig. 1) indicated that the source of heterogeneity was likely to be one trial by Daneman (1961), with other results being reasonably consistent. This trial produced a large positive effect size of 1.1 (0.8–1.4) despite assuming a poor outcome in subjects lost to follow-up. It yielded an even larger estimate of 2.80 (2.41–3.19) when these assumptions were not made, and the improvement rate in the placebo group was unusually poor (9% at eight weeks). Closer inspection revealed the possibility that rating of response was not blind and that selective reporting of outcomes had occurred. It was therefore decided to repeat the analysis excluding this study. There were also grounds for excluding the study by Murphy *et al* (1984) since all subjects received cognitive therapy, which may have reduced the likelihood of finding differences between the effects of drug and placebo. Meta-analysis with the seven remaining trials reduced heterogeneity to a non-significant level and produced a smaller overall estimate of effect of 0.21 (0.03–0.38).

Repeating these analyses with the higher estimate from the trial by Weintraub & Aronson (1963) marginally increased the size of the overall estimates, but did not influence heterogeneity findings.

In-patient trials predominantly involved people with endogenous or severe depression. The majority of people in out-patient trials were diagnosed as having neurotic or moderate depression. Subgroup analysis in in-patients produced a small pooled effect size of 0.15 (−0.12–0.41) using the lower of the two estimates from Weintraub & Aronson (1963), which increased and became significant at the 5% level using the higher estimate from this trial. Combining out-patient trials again revealed significant heterogeneity due to

Table 2 Results of meta-analysis

Combination of studies used	Number of studies used in analysis	Combined effect size(95% CI)	Heterogeneity χ^2 (d.f.)
Using Weintraub & Aronson (1963) hospital director's assessment			
All studies included	9	0.41 (0.27–0.56) $P<0.001$	38.0 (8) $P<0.001$
Daneman (1961) and Murphy *et al* (1984) excluded	7	0.21 (0.03–0.38) $P=0.02$	3.9 (6) NS
In-patient trials[1]	4	0.15 (−0.12–0.41) NS	0.26 (3) NS
Using Weintraub & Aronson (1963) ward doctor's assessment			
All studies included	9	0.46 (0.31–0.60) $P<0.001$	37.3 (8) $P<0.001$
Daneman (1961) and Murphy *et al* (1984) excluded	7	0.27 (0.10–0.45) $P=0.002$	6.4 (6) NS
In-patient trials[1]	4	0.29 (0.04–0.55) $P=0.03$	3.0 (3) NS
Out-patient trials[2]	5	0.54 (0.36–0.71) $P<0.001$	31.9 (4) $P<0.001$
Out-patient trials excluding Daneman (1961) and Murphy *et al* (1984)	3	0.26 (0.02–0.49) $P=0.04$	3.3 (2) NS

1. Weintraub & Aronson, 1963; Wilson *et al*, 1963; Hollister *et al*, 1964; Friedman *et al*, 1966.
2. Daneman, 1961; Ulenthuth & Park, 1963; Hussain, 1970; Friedman, 1975; Murphy *et al*, 1984.

the trial by Daneman (1961). Including this trial produced a large estimate of effect of 0.55 (0.38–0.73). Excluding this trial and that by Murphy *et al* (1984) again reduced heterogeneity and produced a considerably smaller overall estimate of effect.

Quality of studies

Despite the age of most of the trials their quality was judged to be reasonable. Inclusion criteria ensured that they were conducted double-blind and had taken measures to strengthen this procedure by using an active placebo. They all used random allocation and although only two did an explicit intention-to-treat analysis (Friedman, 1975; Murphy *et al*, 1984), all but one (Daneman, 1961) of the others documented only small numbers of early withdrawals. Two studies tested the integrity of the blind in assessors by asking for guesses of medication group; although the guesses were more accurate than would be predicted by chance, the effect was not statistically significant in either trial (Uhlenthuth & Park, 1963; Weintraub & Aronson, 1963). However, in the Weintraub & Aronson trial it was found that both raters assessed those they guessed to be on the active drug as more improved. One other trial reported that side-effects had been more prominent in people on antidepressants (Hollister *et al*, 1964), indicating the possibility that residual unblinding effects may have occurred despite the use of active placebos.

DISCUSSION

Summary of results

All except one of the individual studies were fairly consistent in finding a small, and in most cases non-significant, difference between antidepressant drugs and an active atropine placebo. The pooled estimates of effect varied according to which combination of studies was used. The most conservative estimate was 0.21 standard deviations and the least conservative was 0.47. Assuming a normal response to treatment, these estimates indicate that between 58 and 68% of people on antidepressant drugs would respond better than people on placebo. Alternatively, using the standard deviations reported by Friedman (1975), the estimates would translate into a difference of between 0.5 and 1.0 on the six-point Clinical Global Improvement Scale. The more conservative estimates might be preferred because of the reasons given for exclusion of the trial by Daneman (1961), and because the findings about unblinding and rating bias in the trial by Weintraub & Aronson (1963) might favour the selection of the lower of the two estimates of effect in this trial. However, the higher figures are more consistent with other estimates of the effects of antidepressants. Subgroup analyses did not confirm the prevalent view that severe depression is more responsive to antidepressants than milder forms.

Comparisons with other meta-analyses

Previous meta-analyses have produced diverse estimates of effect size. The largest estimates of 0.81 (95% CI 0.65–0.97) for endogenous depression and 0.55 (95% CI 0.43–0.67) for neurotic depression were found in the Quality Assurance Project (1983). Other general samples of trials produced effect sizes of 0.4 (Smith *et al*, 1980) and 0.67 (Steinbrueck *et al*, 1983). The smallest estimate came from a review of trials comparing a new antidepressant with both a standard drug and a placebo. It was hypothesised that this design would reduce the influence of expectation on the performance of the standard drug. 'Older' antidepressants yielded a combined effect size of 0.25 ($P<0.001$) using observer-rated measures and 0.06 (NS) with subject-rated measures (Greenberg *et al*, 1992). The more conservative estimates from the present study are similar in magnitude to the pooled observer-rated outcomes in Greenberg *et al*, 1992. This would be consistent with the hypothesis that effect sizes in antidepressant trials are inflated by the expectations of participants. However, confidence intervals were wide and the less conservative estimates, which included the Daneman (1961) trial, were closer to combined results obtained from unselected analyses of antidepressant trials.

Limitations of results

This study demonstrates the difficulty of performing meta-analysis with small numbers of trials because of the sensitivity of the results to the inclusion or exclusion of individual studies. For this reason, decisions about which studies to include in the analysis and which estimates of effect to use should be explicit, and results of sensitivity analyses should be presented. The exclusion of the large trial by Daneman (1961), which was the source of significant heterogeneity, had the most substantial impact on this meta-analysis. It is generally recommended that the source of heterogeneity should be investigated rather than proceeding with a combined analysis of discrepant results (Abramson, 1991). In this case it was apparent that the results of the Daneman (1961) study were inconsistent with the other studies in this review, as well as with well-known trials using inert placebos (Medical Research Council, 1965).

In addition, calculating effect size was rarely straightforward, involving conversion of categorical ratings to continuous data and the use of estimated standard deviations in some cases. These problems are endemic to meta-analysis in the absence of standard forms of measurement and reporting. They limit the accuracy of the results but should not alter their general interpretation. However, the results of a meta-analysis are only as good as the trials on which it is based. Most trials in this review were conducted before operationalised diagnostic criteria were available and when standardised outcome measures were still being developed. Methodological concerns that have only recently had widespread publicity, such as randomisation and blinding, were addressed in these studies. However, the short duration of most of the studies should be noted.

An alternative explanation of the present findings is that atropine itself has antidepressant properties and hence acts not as a placebo in these trials, but as a specific therapeutic agent. Although some open studies have suggested that this may be the case (Kasper *et al*, 1981), this was not confirmed in a randomised controlled trial comparing centrally and peripherally acting anticholinergic agents, which found no difference in their effect on mood (Gillin *et al*, 1995).

Implications

This review suggests that unblinding effects and expectations of treatment may influence the results of antidepressant trials. The specific effects of antidepressants may therefore be smaller than is generally believed, with the placebo effect accounting for more of the clinical improvement observed than is already known to be the case. However, the age and quality of the studies and the problems of meta-analysis in this situation should not be disregarded

and mean that these conclusions must remain tentative. The findings constitute a cause for concern about the potential effects of unblinding in psychiatric trials. This should encourage researchers to include a test of the integrity of the double-blind, as well as seeking to identify safe active placebos, which are important means of improving the validity of antidepressant research. Results of trials comparing newer antidepressants with an active placebos would be particularly interesting.

CLINICAL IMPLICATIONS

- Active placebos are necessary to provide comparable conditions for control groups in clinical drug trials.
- Trials of tricyclic antidepressants using inert placebos may have overestimated their efficacy.
- New antidepressants should be compared with inert and active placebos to obtain reliable evidence of their effects.

LIMITATIONS

- Only a small number of studies using active placebos could be located.
- The diversity of outcome measures used and data presented meant that the calculation of standard 'effect sizes' was imprecise.
- The results of the meta-analysis were sensitive to selection of studies for inclusion.

ACKNOWLEDGEMENT

We thank South Thames NHS Executive who have funded J.M.

REFERENCES

Abramson, J. H. (1991) Meta-analysis: a review of pros and cons. *Public Health Reviews*, **18**, 1–47.

Daneman, E. A. (1961) Imipramine in office management of depressive reactions. *Diseases of the Nervous System*, **22**, 213–217.

Engelhardt, D. M., Margolis, R. A., Rudorfer, L., *et al* (1969) Physician bias and the double-blind. *Archives of General Psychiatry*, **20**, 315–320.

Friedman, A. S. (1975) Interaction of drug therapy with marital therapy in depressive patients. *Archives of General Psychiatry*, **32**, 619–637.

—, Granick, S., Cohen, H. W., *et al* (1966) Imipramine (tofranil) vs. placebo in hospitalised psychotic depressives. *Journal of Psychiatric Research*, **4**, 13–36.

Gillin, J. C., Laurellio, J., Kelsoe, J. R., *et al* (1995) No antidepressant effect of biperiden compared with placebo in depression: A double-blind 6-week clinical trial. *Psychiatric Research*, **58**, 99–105.

Greenberg, R. P., Bornstein, R. F., Greenberg, M. D., *et al* (1992) A meta-analysis of antidepressant outcome under "blinder" conditions. *Journal of Consulting and Clinical Psychology*, **60**, 664–669.

— & Fisher, S. (1994) Suspended judgement. Seeing through the double-masked design: a commentary. *Controlled Clinical Trials*, **15**, 244–246.

Hedges, L. V. & Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press.

Hollister, L. E., Overall, J. E., Johnson, M., *et al* (1963) Evaluation of desipramine in depressive states. *Journal of New Drugs*, **3**, 161–166.

—, —, —, *et al* (1964) Controlled comparison of imipramine, amitriptyline and placebo in hospitalised depressed patients. *Journal of Nervous and Mental Disease*, **139**, 370–375.

Hussain, Z. (1970) Drugs in depressive illness. *British Medical Journal*, *i*, 482.

Kasper, S., Moises, H-W. & Beckmann, H. (1981) The anticholinergic biperiden in depressive disorders. *Pharmacopsychiatry*, **14**, 195–198.

Medical Research Council (1965) Clinical trial of the treatment of depressive illness. *British Medical Journal*, *i*, 881–886.

Murphy, G. E., Simons, A. D., Wetzel, R. D., *et al* (1984) Cognitive therapy and pharmacotherapy. *Archives of General Psychiatry*, **41**, 33–41.

Quality Assurance Project (1983) A treatment outline for depressive disorders. *Australian and New Zealand Journal of Psychiatry*, **17**, 129–146.

Rush, A. J., Beck, A. T., Kovacs, M., *et al* (1977) Comparative effects of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Research and Therapy*, **1**, 17–37.

Smith, M. C., Glass, G. V. & Miller, T. I. (1980) Reviews of research on the effects of drug therapy on psychological disorders. In *The Benefits of Psychotherapy* (eds M. C. Smith, G. V. Glass & T. I. Miller), pp. 127–261. Baltimore, MD: Johns Hopkins University Press.

Steinbrueck, S. M., Maxwell, S. E. & Howard, G. S. (1983) A meta-analysis of psychotherapy and drug therapy in the treatment of unipolar depression with adults. *Journal of Consulting and Clinical Psychology*, **51**, 856–863.

Toneatto, T. & Sellers, E. M. (1992) Perception of whether drug or placebo has been administered is a determinant of drinking reduction. In *Novel Pharmacological Interventions for Alcoholism* (eds E. M. Sellers & C. A. Naranjo). Berlin: Springer Verlag.

Uhlenthuth, E. H. & Park, L. C. (1963) The influence of medication (imipramine) and doctor in relieving depressed psychoneurotic outpatients. *Journal of Psychiatric Research*, **2**, 101–122.

Weintraub, W. & Aronson, H. (1963) Clinical judgement in psychopharmacological research. *Journal of Neuropsychiatry*, **5**, 65–70.

White, K., Kando, J., Park, T., *et al* (1992) Side effects and the "blindability" of clinical drug trials. *American Journal of Psychiatry*, **149**, 1730–1761.

Wilson, I. C., Vernon, J. T., Guin, T., *et al* (1963) A controlled study of treatments of depression. *Journal of Neuropsychiatry*, **4**, 331–337.

JOANNA MONCRIEFF, MRCPsych, Section of Epidemiology and General Practice, Institute of Psychiatry, London; SIMON WESSELY, MRCPsych, Department of Psychological Medicine, Kings College School of Medicine and Dentistry, London; REBECCA HARDY, PhD, Department of Epidemiology and Public Health, University College London Medical School, London

Correspondence: Joanna Moncrieff, Section of Epidemiology and General Practice, Institute of Psychiatry, De Crespigny Park, London SE5 8AF

(First received 28 April 1997, final revision 17 September 1997, accepted 18 September 1997)