

A Method to Quantify Rater Bias in Antidepressant Trials

Eva Petkova, Ph.D., Frederic M. Quitkin, M.D., Patrick J. McGrath, M.D., Jonathan W. Stewart, M.D., and Donald F. Klein, M.D.

Some studies indicate that the blind in clinical trials of the efficacy of antidepressant drugs is less than perfect. It has been suggested that, as a consequence of this incomplete blind, biased raters inflate efficacy and that, in fact, these drugs are relatively ineffective. However, in the literature, we could find no prior attempt to quantify rater bias and, thus, measure its contribution to claims of antidepressant efficacy. We used the distribution of SCL-90 (Symptom Check List) depression scale scores to derive a patient-based effect size, and contrasted this with the clinician-based effect size. We propose the difference between these two effect sizes (patient self-rating and clinician-derived) to be an indirect measure of bias. If patients had a prodrug bias, this method would be invalid. However the response rate from studies with active placebo suggest a patient prodrug bias is unlikely. The effect sizes derived from patient self-ratings are smaller than those derived from clinician ratings. This

allows for the possibility that some clinician ratings were biased. However, quantifying the effect of bias suggests that it was insufficient to invalidate the original study conclusions based on clinician ratings, because the proportion of responders, based on patient self-ratings, differed significantly between the two drugs and placebo. Their 95% confidence intervals (CI) did not overlap. This analysis allows that some clinician ratings may be biased. However, the extent of bias appears insufficient to alter conclusions based on clinician ratings regarding efficacy of antidepressant drugs in this trial. Application of our approach in other trials is necessary to establish generalizability. [Neuropsychopharmacology 22:559–565, 2000] © 2000 American College of Neuropsychopharmacology. Published by Elsevier Science Inc.

KEY WORDS: *Bias; Antidepressant trials; Study validity; Unblinded rater*

It has been suggested that an accurate assessment of antidepressant efficacy indicates that these drugs are barely distinguishable from placebo in alleviating depression (Greenberg et al. 1989a, 1992, 1994; Fisher and Greenberg 1993). This assertion is based on the observation that the ability of some blind raters to identify drug vs. placebo treatment exceeds chance (Rabkin et

al. 1986; Carrol et al. 1994; White et al. 1992). This is attributed to piercing the blind which biases the raters and, thus, inflates the appraisal of drug efficacy. However, if raters simply guessed drug wherever they saw improvement, and the drug was factually superior to placebo, guesses would exceed chance and this would not actually bias the effectiveness estimate as discussed below.

Further, drawing on studies in which imipramine and amitriptyline were used as active standard comparators in the evaluation of new antidepressants, Greenberg et al. (1992) suggest that standard drug efficacy was found to be only marginally superior to placebo because investigators were biased toward the newer drugs. It is difficult to understand how raters would effectively differentiate two active drugs or why they would risk classifying a study as "failed" by the FDA standards by rating the standard equivalent to placebo.

From the Department of Biostatistics and the Department of Therapeutics, Columbia University College of Physicians and Surgeons, New York State Psychiatric Institute, New York, NY.

Address correspondence to: Eva Petkova, Ph.D., 1051 Riverside Drive, New York, NY 10032.

Received June 22, 1999; revised November 9, 1999; accepted November 18, 1999.

The authors infer that all antidepressants would be found barely more effective than placebos without these biased ratings. Articles in both the popular and academic press cite this work as cogent (Goleman 1995; Muñoz et al. 1994; Young and Annable 1996).

These criticisms must be scientifically evaluated since, if they are invalid, they may discourage depressed people from seeking treatment that might, in fact, specifically help them. Repeated epidemiologic studies showing high rates of disorder and low rates of treatment in the general population suggest that these criticisms may contribute to a public health problem (Hirschfeld et al. 1997). Fisher and Greenberg's (1993) assumptions aside, it is still important to determine if there are clinician biased ratings that invalidate antidepressant studies (Rabkin et al. 1986; Carroll et al. 1994).

What factors contribute to bias in clinical drug trials? In one analysis of obstetrical clinical trials, multiple trials were grouped by study design and effect size was used as the dependent variable. In trials characterized by inadequate allocation concealment and lack of double blinding, effect sizes were larger than in trials with appropriate design (Schulz et al. 1995). Moher et al. (1998) conducted a similar meta-analysis of randomized clinical trials and also suggest that "low quality" trials inflate estimates of effect size. This is not definitive, because of other possible confounding factors, such as sample variations and effectiveness of treatment procedures.

We were unable to find reports assessing the effect of bias in psychopharmacologic clinical trials. Therefore, our goal was to develop an estimate to determine if bias was sufficient to invalidate clinician's rating of efficacy. A method that erred by overestimating true bias would be acceptable as a conservative error since the larger the bias estimate, the more likely the drug trial would be considered invalid. The proposed method contrasts the proportion of responders and effect sizes derived from patient self-ratings with those derived from clinician ratings. If outcomes based on patient ratings are equivalent to those based on clinician ratings, it suggests that the role of clinician bias is minimal. Confounding issues such as a patient equivalent pro-drug bias and the possibility that patients and clinicians focus on change in different aspects of psychopathology are critically reviewed in the discussion section.

Some definitions are in order. In a "blind" study, a rater uninformed of treatment assignment is considered to be effectively unblinded if the proportion of his or her correct treatment guesses exceeds chance. Such an unblinded rater may, wittingly or unwittingly, introduce bias by applying different criteria in evaluating patients guessed to be taking drug and those guessed to be taking placebo. Thus, ratings are biased if outcome criteria vary with treatment guess. That guesses about treatment identity exceed chance does not prove bias or

even that side-effects were identificatory cues, since correct guesses can be influenced by outcome. If 70% of patients on drug are correctly judged to respond, while 70% on placebo are correctly judged unresponsive, if the clinician rater guessed drug for all responders and placebo for nonresponders, the proportion of correct guesses would exceed chance. If the improvement criteria had not shifted with the guess, no bias is entailed.

Further, if the symptoms are either virtually gone or remain unchanged, the probability of a patient's being rated a responder should not be substantially affected by the rater's treatment guess, assuming the rater is trying to be objective. However, when assessing patients with intermediate improvement, where the outcome decision is not black or white, even a reasonable scientist could fall prey to criteria shifts. Therefore, assessing the magnitude of bias is relevant to the estimate of drug efficacy.

RATIONALE

The difference between the ratings of an unblind and a totally blind rater would estimate bias. This ideal exists in theory only, however, since there is reasonable doubt that any clinical rating can be totally blind (Carroll et al. 1994). Instead, the method reported here relies on comparing clinician assessments with those from a patient self-rating scale, the SCL-90 (Derogatis 1977). The utility of self-rating scales has received insufficient consideration in evaluation of non-demented, non-psychotic patients. Patient perception of treatment results is critical since, if a patient is not satisfied, why should he or she continue a treatment? The predominant reliance on clinician rather than patient ratings may be an anachronism resulting from the fact that early psychopharmacology research and scale development was conducted with hospitalized patients who were not considered to be good judges of their own behavior or reporters of their affective state.

Estimating Bias: Comparing Patient- and Clinician-Derived Proportions of Responders and Effect Sizes

If patient-derived outcomes are equivalent to clinician-derived outcomes, it suggests that clinician bias, if present is minimal and this supports study validity. The clinician-derived effect size, ES, is derived from the difference between the clinically evaluated proportions of responders taking drug (P_{drug}) and those taking placebo (P_{pbo}):

$$ES = P_{drug} - P_{pbo}$$

The patient-derived effect size, ES^* , is based on the SCL-90, a widely used self-rating scale. Counting the number of patients below a stipulated SCL-90 score, X^* ,

permits the calculation of proportions of responders, and P^*_{drug} and P^*_{pbo} are so defined. The patient derived effect size is:

$$ES^* = P^*_{drug} - P^*_{pbo}$$

The difference between patient- and clinician-derived effect sizes, $ES - ES^*$, may serve as an indirect estimate of clinician bias. Obviously, the SCL-90 score X^* , determining response must be selected *a priori*.

METHODS

Data from a double blind, placebo controlled study of the utility of imipramine and phenelzine in treating atypical depression were used. In this sample, the accuracy of clinicians' treatment guesses exceeded chance and, therefore, it is possible bias contributed to the observed treatment differences (Rabkin et al. 1986). Patient self ratings and clinician ratings of outcome are available for all participants. Extensive treatment outcome and clinician rater guess data have been published (Rabkin et al. 1986; Quitkin et al. 1993). The CGI clinician rating classified response: patients receiving scores of 1 (very much improved) or 2 (much improved) were judged responders; those with scores of 3 to 7 (3, minimally improved; 4, unchanged, etc.) were judged nonresponders. The CGI rating was made at the end of Week 6 by a nominally "blind" clinician who had seen the patient weekly. Patient completed the SCL 90 at baseline and study end (Derogatis 1977). Most often the patient completed their self ratings prior having a final clinical evaluation.

Of the 401 patients completing the study, 135 received placebo, 139 imipramine, and 127 phenelzine. Treatment guesses were collected on 136 virtually consecutive patients (51 receiving placebo, 48 imipramine, and 37 phenelzine). Because collection was dependent on the availability of research assistants, the series was not absolutely consecutive. The investigators' primary interest was to use psychopharmacologic dissection to establish the validity of a distinct depressive subtype. It was hypothesized that atypical depressives, in contrast to other depressive sub-groups, would have a superior response to the MAOI (phenelzine) thus supporting the validity of this nosologic distinction. Therefore, bias if present should favor the phenelzine group.

The items used on the SCL 90 were, feeling blue, feeling hopeless, feeling worthless, crying easily, thinking of ending life, blaming oneself, and lack of interest. Some items included in the original SCL depression factor were excluded. We excluded *a priori* feeling trapped, loss of sex interest, feeling lonely, and loss of appetite because they were either not characteristic of depressed patients (feeling trapped) or possible side effects (loss of

Table 1. Treatment Outcome Based on Clinician Ratings Using CGI Change Scale

Treatment	Number of Patients	Number of Responders	Percent of Responders and 95% CI
Placebo	139	35	25% (18, 32)
Imipramine	135	70	52% (43, 60)
Phenelzine	127	96	76% (68, 83)

$\chi^2 = 67.68, df = 2, p < .001.$

sex interest, loss of appetite), or too general (feeling lonely). This *a priori* exclusion probably was not necessary since the correlation between the 11 items included in the depression factor and the seven items used in this analysis is 96.6 (Pearson correlation coefficient $N = 354$) suggesting it is extremely unlikely outcome would be different with 11 items or seven items.

Clinicians "blind" treatment guesses were made, prior to completion of the final clinical evaluation and code breaking.

RESULTS

Study outcome based on clinicians nominally "blind" ratings is summarized in Table 1, the study hypotheses were supported, with phenelzine being superior to imipramine, which was superior to placebo. In Table 2, a cross-tabulation of actual and guessed treatment assignment is presented for the 136 patients for whom guess are available. The unweighted Kappa measuring exact agreement between guess and actual treatment, 0.58, $p < .01$ indicates that correct guesses exceeded chance.

Figure 1 depicts the distribution of post-treatment SCL-90 depression scale scores for patients in each treatment group. Scores from patients treated with phenelzine accumulated in the area characterized by the least post-treatment pathology (i.e., less than 2 on the SCL-90) than scores from those given placebo. The differences in the distributions of the scores for patients receiving imipramine and placebo is not as striking as the phenelzine-placebo contrast.

In Table 3, the percent of responders determined

Table 2. Cross-Tabulation of 136 Subjects, Actual Treatment and the Clinician Guess about the Treatment

Actual Treatment	Clinician Guess			Total
	Placebo	Imipramine	Phenelzine	
Placebo	42	4	5	51
Imipramine	4	36	8	48
Phenelzine	5	12	20	37
Total	51	52	33	136

$Kappa = 0.58, p < .01.$

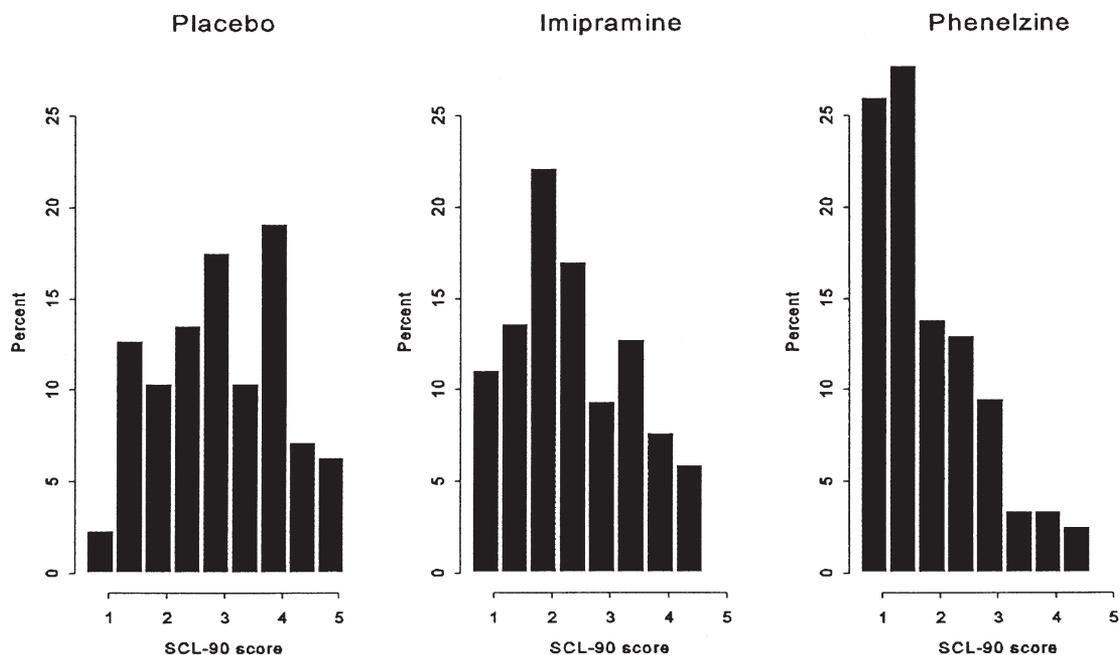


Figure 1. Distribution of post-treatments SCL-90 depression scale by treatment. Low scores of SCL-90 represent less pathology

from the patient self-rating scale are presented. Prior to examine the data three arbitrary SCL 90 scores were selected to identify responders and non responders, 1.5, 2.0, and 2.5. A separate analysis was done for the three arbitrary SCL 90 points (Table 3).

The original clinician-based ratings suggested that phenelzine is superior to both imipramine and placebo and imipramine is superior to placebo since the 95% confidence intervals (CI) for the three groups do not overlap (Table 1). Using the three SCL-90 response thresholds, the proportions of responders in the two drug and placebo conditions are similar to clinicians. The one difference is the overlap in the CI for placebo and imipramine using the 1.5 cut off. The relationship of all other proportion of responders and CI are similar to the clinicians with no overlap in CI for all phenelzine vs. placebo, phenelzine vs. imipramine, and imipramine vs., placebo (two contrasts). This supports the validity of clinician ratings.

Effect sizes are presented in Table 4 (from calculations using the data presented in Table 3). Clinician out-

come ratings suggest an effect size (ES) for the phenelzine-placebo difference of 51% [76–25%; 95% CI (41, 61)]. Patient ratings using the SCL-90 1.5 cut off result in an effect size (ES*) of 39% for phenelzine-placebo [51–12%; 95% CI (29, 49)]. The difference between these effect sizes, $ES - ES^*$, estimates the bias as 12% [95% CI (2, 22)] (see Table 5). The formula used to calculate the CI assumes correlated effect sizes, since they were derived from the same patient sample. Any bias in clinician ratings may have overestimated the superiority of phenelzine to placebo by 2–22% (the upper bound of the CI).

A similar calculation for imipramine-placebo gives effect sizes of 27% using clinician ratings (ES) and 9% using patient self-ratings ($ES^*_{1.5}$), for a difference ($ES - ES^*_{1.5}$) of 18% possibly due to bias. It should be noted that using the other two cut offs (2.0, 2.5) there is no overlap in the CI for the proportion of responders between placebo and imipramine, the effect sizes are 21% (10, 32) and 24% (12, 36) (Table 4), and the bias estimate is 6% or 3% (Table 5). For the phenelzine-imipramine comparisons, the clinician-derived effect size is 24%

Table 3. Proportion of Responders Using Three SCL 90 Cut Offs, and 95% CI

Treatment	Clinician's Rating by CGI	Criteria for Response by Self Rating Scale SCL 90		
		SCL 90 < 1.5	SCL 90 < 2	SCL 90 < 2.5
Placebo	25% (18, 32)	12% (6, 18)	22% (14, 30)	36% (28, 44)
Imipramine	52% (44, 60)	21% (13, 29)	44% (35, 53)	60% (51, 69)
Phenelzine	76% (69, 83)	51% (42, 60)	63% (54, 72)	80% (73, 87)

Table 4. Effect Sizes Derived from Clinician Ratings and Three SCL 90 Cut Offs

Comparisons	Effect Sizes			
	ES (Clinician)	ES* _{1.5} (SCL 90 < 1.5)	ES* ₂ (SCL 90 < 2)	ES* _{2.5} (SCL 90 < 2.5)
Phenelzine-Placebo	51% (41, 61)	39% (29, 49)	41% (30, 52)	44% (33, 55)
Imipramine-Placebo	27% (16, 38)	9% (0, 19)	21% (10, 32)	24% (12, 36)
Phenelzine-Imipramine	24% (13, 35)	30% (18, 42)	20% (7, 33)	20% (9, 31)

[95%CI (13, 35)], and the patient-derived effect size using the three cut offs are 30% (18, 42); 20 % (7, 33)], and 20% (9, 31).

For all estimates of bias, except the phenelzine -placebo and imipramine-placebo using 1.5 SCL 90 cut off, the CI include zero (see Table 5). Therefore, the null hypothesis that the imipramine-placebo and phenelzine-imipramine effect size estimates have zero bias can only be rejected for 2/9 bias estimates. Further, conclusions based on clinician ratings and patient self-ratings concur, i.e., phenelzine is superior to imipramine and placebo for all three cut offs and for two of three imipramine is superior to placebo. Clinicians’ estimates of treatment efficacy appear valid, since the proportion of parallel patient-derived responses to phenelzine, imipramine (2/3) , and placebo also have non-overlapping CI.

DISCUSSION

To the best of our knowledge, there have been no prior attempts to estimate the extent of bias due to penetration of the double blind in clinical trials of antidepressant drugs. Clearly, the method described here only makes indirect inferences about bias. We suspect that bias may exist but was of insufficient magnitude to invalidate the original comparative conclusions based on clinician ratings. Only one of nine estimates of effect size based on patients ratings haz CI, which includes zero, supporting the validity of clinicians ratings.

Possible Confounders: Patient Pro-Drug Bias

Since our goal is to err conservatively in estimating bias, a patient pro-drug bias could defeat our purpose

Table 5. Bias Estimates Using Three SCL 90 Cut Offs

Comparisons	Bias = ES – ES*		
	ES – ES* _{1.5}	ES – ES* ₂	ES – ES* _{2.5}
Phenelzine-Placebo	12% (2, 22) ^a	10% (0, 20) ^b	7% (-2, 16)
Imipramine-Placebo	18% (9, 27) ^a	6% (-3, 15)	3% (-7, 13)
Phenelzine-Imipramine	-6% (-17, 5)	4% (-7, 15)	4% (-6, 14)

^a The 95% confidence interval does not include 0.

^b The 95% confidence interval includes 0.

since this would erroneously narrow the gap between patient and clinician estimates of benefit. Is there evidence that patients manifest such a bias? Pooling data from all available studies in which active placebos producing side effects were used (Friedman et al. 1966; Hollister et al. 1964; Hussain 1970; Uhlenhuth and Park 1963; Wilson et al. 1963; Daneman 1961; Fahy et al. 1963; McLean and Hakstian 1979; Weintraub and Aronson 1963/1964; Friedman 1975) should help clarify the issue, since a patient pro-drug bias and active pharmacological effects should result in a high placebo response rate.

Of patients on active placebo, 22% (61/276) were rated as responders. Even removing an outlier with a low placebo response rate (6/90) (Daneman 1961) results in a response rate of 30%—an outcome quite comparable to those found using inactive placebo. In addition, merely the presence of side effects coupled with investigator enthusiasm, in the absence of antidepressant effect, do not result in patient improvement. A case in point , L-DOPA studied in the 1970s had considerable investigator enthusiasm because of its apparent heuristic relevance, stemming from the putative role of dopamine in models of depression current at that time. In spite of its side effects, L-DOPA was not found to be an effective antidepressant (Gershon et al. 1971; Bunney et al. 1971). Other studies contrasting patient self-ratings and clinician ratings support patients’ tendencies to minimize change (Hill et al. 1976). These considerations suggest that patient pro-drug bias is probably minimal. Unfortunately, since we were studying old drugs, introduced in the late 1950s (imipramine and phenelzine) whose side effects are well documented, we did not systematically record side effects. Therefore we could not assess the relationship of side effects to guess, although it would still have been virtually impossible to determine the contribution of side effects vs. outcome to any treatment guess.

Possible Confounders: Differences in Objects of Patient and Clinician Ratings

Lambert et al. (1986) suggest that patient and clinician may focus on different aspects of depression and that this could contribute to differences on self- and clinician-rated scales. For example, if the patient focuses on aspects of psychopathology referred to as A, while the clinician focuses on A + B, and if B is more responsive

to drug than to placebo, a portion of the difference between the two ratings may be due to a true treatment effect. This results in a conservative over-estimate of bias, which is acceptable for our purposes. Below, we examine the issue of differences and overlaps in clinician and patient biases.

Does a possible patient negative drug bias help explain the smaller effect size derived from patient ratings (vs. clinicians ratings). For instance, patients on medication may experience "drugged feeling" creating a "halo effect" that made all patient self-ratings more pathologic. In contrast, the clinician's assessment of mood improvement may be unaffected by patient report or display of side-effect mediated discomfort. If true, our bias estimate might actually be inflated because of the contribution of patient negatively biased self-ratings. Another way to phrase this is that global improvement is multidimensional, with various dimensions being differentially drug responsive. The obtained global improvement ratings are a weighted average taken over these dimensions, but clinicians and patients may use unequal weights. Clinicians may be more impressed than patients by drug-benefitted dimensions but this would increase our bias estimate and therefore be a conservative error.

CONCLUSIONS

In the absence of a "gold standard" for assessing mood, we cannot categorically state that invalidating clinician bias did not occur, but the similarity in patient and clinician ratings suggests that this is unlikely. Our approach allows for re-evaluation of potential bias in other previously conducted clinical trials, since it relies on commonly collected data.

Methods proposed to protect against breaks in the blind include use of independent assessors and active placebo, triple blind studies (i.e., keeping raters and patients blind to study design), and three-armed studies with two active drugs and placebo (Thompson 1982; Blumenthal et al. 1974; Henker et al. 1979). All have merits and limitations. The bias estimate method proposed here might be added to those safe guards. Confidence in clinician ratings is increased if patient ratings are similar since this is a second relatively independent measure of change. This is not an alternative to an independent assessor, which should be used universally.

Presentation of proportions of responders and effect sizes derived from both patient self-ratings and clinician ratings allows the reader to come to his/her own conclusions about bias. We believe that, in studies of non-psychotic and non-demented patient populations, the most valid estimate of response rate results from using both patient self-ratings and clinician ratings. If cli-

nician effect sizes substantially exceed patient based effect sizes suspicions concerning invalidating evaluation bias must be addressed by the investigators. Our results clearly apply only to this specific study. We hope that others will use this approach in a variety of studies to determine if this method has general utility.

ACKNOWLEDGMENTS

Supported by USPHS grant MH30906.

REFERENCES

- Blumenthal DS, Burke R, Shapiro AK (1974): The validity of "identical matching placebos." *Arch Gen Psychiatry* 31:214–215
- Bunney WE, Brodie KH, Murphy DL, Goodwin FK (1971): Studies of alpha-methyl-para-tyrosine, L-DOPA and L-Tryptophan in depression and mania. *Am J Psychiatry* 127:48–71
- Carroll KM, Rounsaville BJ, Nich C (1994): Blind man's bluff: Effectiveness and significance of psychotherapy and pharmacotherapy. Blinding procedures in a clinical trial. *J Cons Clin Psychol* 62(2):276–280
- Daneman E (1961): Imipramine in office management of depressive reactions (a double blind clinical study). *Dis Nerv System* 22(4):213–217
- Derogatis LR, Lipman RS, Covey L (1973): The SCL-90: An outpatient psychiatric scale. *Psychopharmacology Bulletin* 9:13–28
- Fahy P, Imlah N, Harrington J (1963): A controlled comparison of electroconvulsive therapy, imipramine and thiopentone sleep in depression. *Neuropsychiatry* 4:310–314
- Fisher S, Greenberg R (1993): How sound is the double-blind design for evaluating psychotropic drugs? *J Nerv Mental Dis* 181:345–350
- Friedman AS (1975): Interaction of drug therapy with marital therapy in depressive patients. *Arch Gen Psychiatry* 32:619–637
- Friedman AS, Granick S, Cohen HW, Cowitz B (1966): Imipramine (Tofranil) vs. placebo in hospitalized psychotic depressives (a comparison of patients' self-ratings, psychiatrists' ratings and psychological test scores). *J Psychiatric Res* 4:13–36
- Gershon ES, Bunney WE, Goodwin FK, Murphy D, Henry GM (1971): Catecholamines and affective illness. In Ho BT, McIsaac WM (eds), *Brain Chemistry and Mental Illness*. New York, NY, Plenum Press, pp 125–161
- Goleman P (1995): Psychologists dispute value of antidepressants. *New York Times*, November 29
- Greenberg RP, Bornstein RK, Greenburg MD, Fisher S (1992): A meta-analysis of antidepressant outcome under "blinder" conditions. *J Cons Clin Psychol* 60(5):664–669
- Greenberg RP, Bornstein RF, Zborowski MJ, Fisher S, Greenberg MD (1994): A meta-analysis of fluoxetine outcome in the treatment of depression. *J Nerv Mental Dis* 182:547–551

- Greenberg RP, Fisher S (1989a): Examining antidepressant effectiveness: Findings, ambiguities and some vexing puzzles. In Fisher S, Greenberg RP (eds), *The Limits of Biological Treatments for Psychological Distress: Comparisons with Psychotherapy and Placebo*. Hillsdale, NJ, Erlbaum, p 137
- Greenberg RP, Fisher S (1989b): A second opinion: Rethinking the claims of biological psychiatry. In Fisher S, Greenberg RP (eds), *The Limits of Biological Treatments for Psychological Distress: Comparisons with Psychotherapy and Placebo*. Hillsdale, NJ, Erlbaum, pp 309–336
- Henker B, Whalen CK, Collins BE (1979): Double-blind and triple-blind assessments of medication and placebo responses in hyperactive children. *J Abn Child Psychiatry* 7:1–13
- Hill LE, Nun AJ, Fox W (1976): Matching quality of agents employed in “double-blind” controlled clinical trials. *The Lancet* 352–356
- Hirschfeld R, Keller M, Panico S, Arons BS, Barlow D, Davidoff F, Endicott J, Froom J, Goldstein M, Gorman JM, Guthrie D, Marek RG, Maurer TA, Meyer R, Phillips K, Ross J, Schwenk TL, Sharfstein SS, Thase ME, Wyatt RJ (1997): The National Depressive and Manic-Depressive Association consensus statement on the under-treatment of depression. *JAMA* 277:333–340
- Hollister L, Overall J, Johnson M, Pennington V, Katz G, Shelton J (1964): Controlled comparison of amitriptyline, imipramine, and placebo in hospitalized depressed patients. *J Nerv Mental Dis* 139:370–375
- Hussain Z (1970): Drugs in depressive illness. *Br Med J* 2:482
- Lambert MJ, Doxey RH, Kingston MD, Edwards BC (1986): Zung, Beck, and Hamilton rating scales as measures of treatment outcome: A meta-analytic comparison. *J Cons Clin Psychol* 54(1):54–59
- McLean PD, Hakstian AR (1979): Clinical depression: Comparative efficacy of outpatient treatments. *J Cons Clin Psychol* 47:818–836
- Moher D, Pham B, Jones A, Cook DJ, Jadad AR (1998): Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analysis? *The Lancet* 352:609–613
- Muñoz RF, Hollon SD, McGrath E, Rehm LP, Van den Bos GR (1994): On the AHCPR depression in primary care guidelines. *Am Psychol* 49(1):42–61
- Quitkin FM, Stewart JW, McGrath PJ, Tricamo E, Rabkin JG, Ocepek-Welikson K, Nunes E, Harrison W, Klein DF (1993): Columbia atypical depression: A subgroup of depressives with better responses to MAOI than to tricyclic antidepressants or placebo. *Br J Psychiatry* 163(Suppl 21):30–34
- Rabkin JG, Markowitz JS, Stewart J, McGrath P, Harrison W, Quitkin FM, Klein DF (1986): Assessment of patient and doctor medication guesses in a placebo-controlled trial of imipramine and phenelzine. *Psychiatric Res* 19:75–86
- Schulz K, Chalmers I, Hayes R, Altman D (1995): Empirical evidence of bias. *JAMA* 273(5):408–412
- Thompson R (1982): Side-effects and placebo amplifications. *Br J Psychiatry* 140:64–68
- Uhlenhuth E, Park L (1963): The influence of medication (imipramine) and doctor in relieving depressed psychoneurotic outpatients. *J Psychiatric Res* 2:101–122
- Weintraub W, Aronson H (1963/1964): Clinical judgment in psychopharmacological research. *J Neuropsychiatry* 5:65–70
- White K, Kando J, Park T, Waternaux C, Brown WA (1992): Side-effects and the “blind ability” of clinical drug trials. *Am J Psychiatry* 149:1730–1731
- Wilson I, Vernon J, Sandiler M (1963): A controlled study of treatment of depression. *J Neuropsychiatry* 4:331–337
- Young SN, Annable L (1996): The use of placebos in psychiatry: A response to the draft document prepared by the TRI-Council Working Group. *J Psychiatry Neurosci* 21(4):235–238