Citation:

Schöchlin, C., Klein, J., Abraham-Rudolf, D. & Engel, R. R. (2003). The influence of design variables on the results of controlled clinical trials on antidepressants: A meta-analysis. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 207-218). Hogrefe & Huber.

14

The Influence of Design Variables on the Results of Controlled Clinical Trials on Antidepressants: A Meta-Analysis

Claudia Schöchlin
Jürgen Klein
Department of Clinical Psychology and Psychophysiology
Hospital for Psychiatry and Psychotherapy
Ludwig-Maximilians-University of Munich

Dorothee Abraham-Rudolf
Hospital for Psychiatry
Technical University of Munich

Rolf R. Engel
Department of Clinical Psychology and Psychophysiology
Hospital for Psychiatry and Psychotherapy
Ludwig-Maximilians-University of Munich

Summary

Bias in clinical trials can be investigated by studying correlations between design variables and outcome of clinical trials. 72 studies on the antidepressant effect of imipramine and amitriptyline as well as four serotonin reuptake inhibitors were analyzed in a publication based meta-analysis. Treatment outcome was operationalized as an effect size in the basis of response rate differences between active drug and placebo or active drug and active drug. It was found that the number of treatment cells included in a study, the existence of a placebo cell as well as the severity of depression at inclusion and placebo response rate are associated with study outcome, and that they may interact with each other, presumably because

of differences in drop-out handling. Main conclusion is that the existence and the clinical results of a placebo cell are moderating variables for the results of clinical trials.

14.1 INTRODUCTION

Meta-analysis is a tool to statistically integrate results of empirical studies from publications or, less often, from raw data (Dickersin & Berlin, 1992). Two forms of empirical studies can be distinguished: Experimental or quasi-experimental studies with an independent and a dependent variable, such as clinical trials, and epidemiological studies, such as case-control studies, in which there is no independent variable that is manipulated (Petitti, 1994). Both kinds of study are subject to methodological flaws which, in turn, affect interpretation of meta-analytical results. A critical appraisal of these problems is presented by Feinstein (1995). On the other hand, meta-analysis can serve as a tool to investigate methodological questions arising from empirical studies; an example of this is furnished by Gotzsche (1990), who presents an investigation of methodological problems of trials in rheumatoid arthritis.

The present meta-analysis investigates the effects of design variables of clinical antidepressant trials. Design variables are variables that are voluntarily or involuntarily induced by all aspects concerning the design and the realization of clinical trials. They are not welcome, as they may restrict the ability to generalize study results, and therefore risk impeding the interpretation of data. Design effects can be caused by blindness, number of treatment conditions, handling of placebo-responders, or the baseline severity of illness.

The term "Depression" is frequently used in colloquial language and means sorrow or despair. As a psychiatric diagnosis it is actually operationalized by the persistence of a number of certain, well-defined symptoms (depressive mood, loss of interest, loss of weight, sleep disturbances, psychomotor agitation or retardation, fatigue, cognitive feelings of guilt or worthlessness, disturbances of thought, concentration, memory, decisiveness, suicidality) over a given time period (American Psychiatric Association, 1994). It is one of the most frequent diseases and has large economic consequences for a society as well as, with its infringing symptoms and finally its high suicide rate, important consequences for sufferers' lives (Goodwin & Jamison, 1990). Therefore, it constitutes one of the main psychiatric research areas, in which a large number of clinical trials has been conducted that need to be meta-analytically integrated and investigated.

Concerning the drugs used in antidepressive therapy, one can distinguish different groups: The first generation drugs were the tricyclic antidepressants (TCA), which are considered as a standard treatment in depression drug therapy. In the 1980s, the selective serotonin reuptake inhibitors (SSRI), which possess more specific receptor activity, were developed. They are considered to be an alternative to their predecessors, as they induce less side effects (Möller & Volz, 1996). Although a very large number of substances has been investigated

in clinical trials on drug therapy of depression, most of the studies included TCA or SSRI.

14.2 BACKGROUND

We started from the following result of a former meta-analysis on the tricyclic antidepressant imipramine, compared to placebo: Studies that compared imipramine only to placebo, and not to a third or fourth treatment condition, yielded considerably higher effect sizes ($r = .34$; $N = 703$)[1] than studies which included further treatment arms ($r = .17$; $N = 4673$). Thus, the difference between imipramine and placebo was higher if only imipramine and placebo were investigated ($z = 4.6$; $p < .001$). Greenberg, Bornstein, Greenberg, and Fisher (1992) proposed that a study is less susceptible to unblinding if more than two treatment conditions are investigated. We dared not to join this interpretation as there was a confounding with other variables we considered to have potential weight, that is, the year of publication and the status of a substance as a new or as a control substance. Among the 11 studies published before 1978, only 3 included more than three treatment arms, compared to 60 among the 66 studies published after 1977. Nearly all studies with two treatment conditions are older studies and may therefore be more prone to bias, as methodology in the early years of clinical trials was less elaborate than nowadays. Imipramine was investigated as the substance of interest mainly in studies with only two treatment cells (14 out of 15), whereas it served as a control substance (logically) only in studies with more than two treatment cells. Most of the studies with imipramine as a control substance were published after 1978 (59 out of 62), whereas it was investigated as a substance of interest before and after 1978.

We nevertheless considered the difference between studies with two and those with more than two treatment cells being important and took a further look at it by comparing placebo-controlled SSRI-studies. All studies were published after 1977. There were no publications on placebo-controlled studies with an SSRI as a control substance; all studies investigated the SSRI as the substance of interest. No difference was found among studies with two versus those with more than two treatment arms ($r = .17$, $N = 696$ vs. $r = .18$, $N = 3155$; $z = 0.28$, $p = .78$). Thus, the proposal based on the imipramine results that the number of treatment cells is a decisive factor for the effect size of a study was not affirmed by the SSRI data.

14.3 AIMS OF THE META-ANALYSIS

In the literature, further variables characterizing the design of clinical studies were identified that might influence effect sizes of antidepressant studies.

[1]Here, and in the following text, N always indicates the (total) number of patients.

We investigated correlations between study characteristics and effect size in a larger sample of controlled clinical trials. As outcome, the effect sizes for active medication vs. placebo or standard medication were chosen; they represent differences in efficacy between the two treatment cells. Predictor variables were the number of study centers, a placebo run-in vs. no placebo run-in, the study duration and mean patients' severity of depression at baseline.

14.4 METHODS

Controlled clinical trials on acute treatment of depression with imipramine, amitriptyline, fluoxetine, paroxetine, or sertraline were included if one of these substances was compared either to placebo and/or to one of the substances of the other group. Studies had to be published between January 1979 and April 1997, and had to indicate response rates.

The difference in efficacy between two treatment cells was expressed by the correlation coefficient r, based on the fourfold table φ, which corresponds to the response rate difference and thus can be interpreted as being quite close to clinical practice. Effect sizes were computed on an intent-to-treat-basis; all randomized patients were included in the effect size calculation. It has been discussed which effect size is preferable for clinical trials (Dickersin & Berlin, 1992; Fleiss, 1993), especially odds ratios are commonly used. We consider response rate differences the appropriate measure in our case, for the very reason that the original studies do not use odds ratios but response rate differences, which are comparable to φ in a meta-analysis. Effect sizes were computed by:

$$\varphi = \sqrt{\frac{\chi^2}{N}}.$$

They were weighted for sample size of the studies and averaged via Z-transformation. Homogeneity of study effect sizes was assessed by the usual chi-square test for homogeneity of correlation coefficients (Rosenthal, 1991; Mantel & Haenszel, 1959). Explorative comparisons of specific effect sizes were performed by means of contrasts for group comparisons (for details see Rosenthal, 1991; Rosenthal & Rubin, 1982), according to the following formula:

$$z = \frac{\sum_{j=1}^{k} \lambda_j Z_j}{\sqrt{\sum_{j=1}^{k} (\lambda_j^2 \, (N_j - 3))}}$$

with Z_j being the Fisher-Z-transformed effect size coefficients for the jth of k studies to be compared and λ_j being the orthogonal contrast coefficients summing up to zero. z is the standard normal deviate. For quantitative data, Pearson correlations were used; they were calculated with weights, but significance was judged referring to the number of studies, not the number of patients. Given p values are two-tailed.

Table 14.1 Comparisons Included in the Meta-Analysis

			Year of publication			BL-Ham-D	
	Number of comparisons	*N*	Min	Max	*M*	*M*	*SD*
Drug – Placebo							
Imipramine	46	6176	79	96	87.7	25.0	4.9
Amitriptyline	16	2604	79	95	87.8	25.5	5.1
Fluoxetine	11	1247	85	95	89.4	24.7	3.8
Fluvoxamine	8	1043	83	96	90.9	25.8	2.0
Paroxetine	5	1052	89	93	91.0	28.4	1.5
Sertraline	4	954	90	96	93.8	22.0	6.8
Drug – Drug							
Imipramine-SSRI	25	3380	83	96	90	26.15	3.2
Amitriptyline-SSRI	19	2337	85	96	91	26.09	2.9

Note. *N* = Number of patients; BL-Ham-D = Hamilton at baseline

14.5 DATA

Table 14.1 gives an overview of the comparisons on which the following analysis is based. Sixty-two of the comparisons refer to tricyclic drugs vs. placebo, 28 to SSRI vs. placebo, and 44 to tricyclic drugs vs. SSRI.

14.6 RESULTS

The Funnel plots in Figure 14.1 show that with increasing *sample sizes* effect sizes approach the mean effect sizes. Among the placebo-controlled studies (Figure 14.1, Panel A and B) there are more studies with higher than with lower effect sizes. This was to be expected, as it can be presumed that small positive studies are more likely to be published than small negative studies. This is an indicator of publication bias, which is less clear-cut among the drug-drug-comparisons (Figure 14.1, Panel C).

Data stemming from *single center studies* yielded higher effect sizes than data from multicenter studies (see Table 14.2). This means that active drugs show their superiority to placebo more clearly if the data are collected in only one center. The same effect can be found when comparing SSRI vs. TCA: Multicenter studies show slight differences between substance classes; in single center studies, there is a tendency for the SSRI to yield better results.

If a *placebo run-in* or placebo washout is included in a study, patients receive placebo during one or two weeks before randomization and are excluded if they respond during this time period. This placebo run-in aims at increasing effect sizes by reducing the number of responders of one group, the placebo

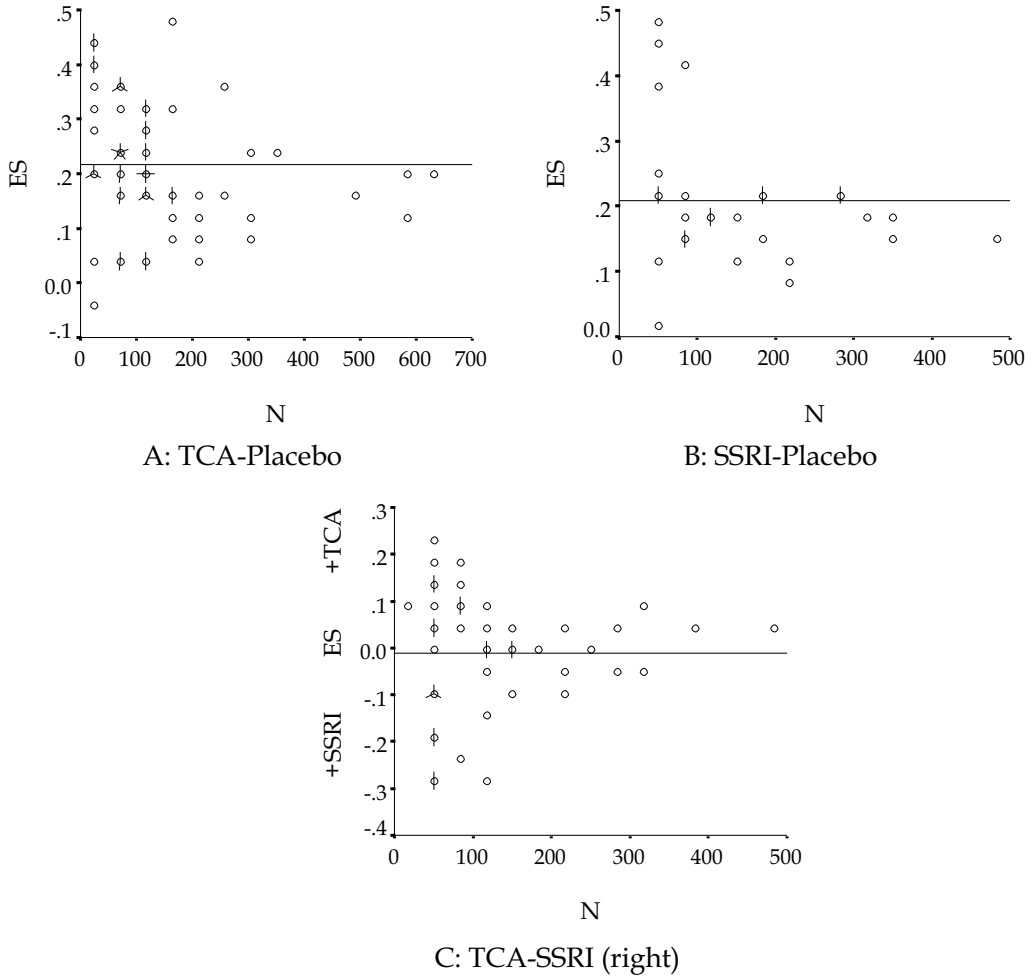

Figure 14.1 Funnel plots (sunflower plot): Number of patients and effect sizes (ES), line represents unweighted mean of effect size).

Table 14.2 Effect Sizes From Single- and Multicenter Studies

	Single-center Studies	Multicenter Studies		
	r	r	z	$p(z)$
TCA-PL	.25 (32)	.18 (30)	2.91	.004
SSRI-PL	.26 (11)	.17 (17)	2.07	.040
TCA-SSRI	−.08 (15)	.01 (28)	−2.51	.010

Note. For each r, the number of studies is given in brackets.

group (FDA, 1978; Feinberg, 1992). Only sparse data exist about the effects of this procedure on study results, and these are ambiguous (Khan, Cohen, Dager, Avery, & Dunner, 1989; Trivedi & Rush, 1994). The data presented here revealed no difference between studies with and without a placebo run-in, neither for drug-placebo nor for drug-drug differences (see Table 14.3). Although

Table 14.3 Effect Sizes of Studies With and Without Placebo Run-In

	No placebo run-in r	Placebo run-in r	z	$p(z)$
VERUM-PL	.18 (28)	.20 (62)	1.2	.23
TCA-SSRI	.01 (9)	−.01 (35)	−.50	.63

Note. For each r, the number of studies is given in brackets.

Table 14.4 Response Rates With and Without Placebo Run-In (Mean $\pm SD$)

	% Response Drug	% Response Placebo
Drug – Placebo		
No placebo run-in	52 ± 12	35 ± 10
Placebo run-in	48 ± 12	29 ± 11
	% Response Imipramine	% Response SSRI
Drug – Drug		
No placebo run-in	64 ± 9	63 ± 10
Placebo run-in	45 ± 15	46 ± 14

the number of responders was reduced among the studies with a placebo run-in, it was reduced in all treatment groups (see Table 14.4).

No correlation was found between effect size and *study duration* (see Figure 14.2, Panel A and B). The longer studies did not show higher effect sizes than the shorter ones. It must be noted, however, that there are only a few really short studies (< 4 weeks). Thus, we can neither conclude that longer studies show higher effect sizes, nor that there is no association.

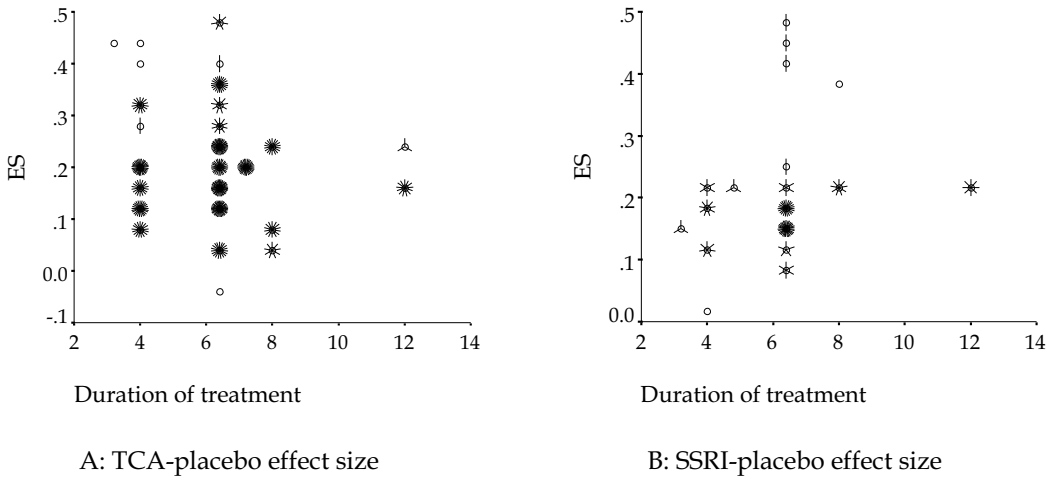

Figure 14.2 Study duration (weighted by sample size) and effect size (ES).

Patients' *severity of depression* at inclusion was not correlated with effect size among the drug-placebo comparisons ($r = -.03$; n.s.; see Figure 14.3, Panel A). This finding is in contradiction to the widespread assumption, based on the concept of endogenous versus neurotic depression, that the superiority of drugs over placebo is more pronounced if patients show higher levels of symptoms (Feinberg, 1992). It may be the result of a missing sensitivity of a group mean score for baseline depression, used in meta-analysis; this would call for raw data analysis. On the other hand, there seem to be only few empirical proofs for a differential efficacy of antidepressants, except psychotic depression, as Montgomery and Lecrubier (1999) conclude in their review. If we consider drug-drug comparisons within our data, there is a low but significant correlation of .34. If a study included patients with a higher Hamilton baseline score, it was more probable to show a (minor) superiority of the TCA, whereas a lower baseline severity was rather associated with a better result of the SSRI (see Figure 14.3, Panel B).

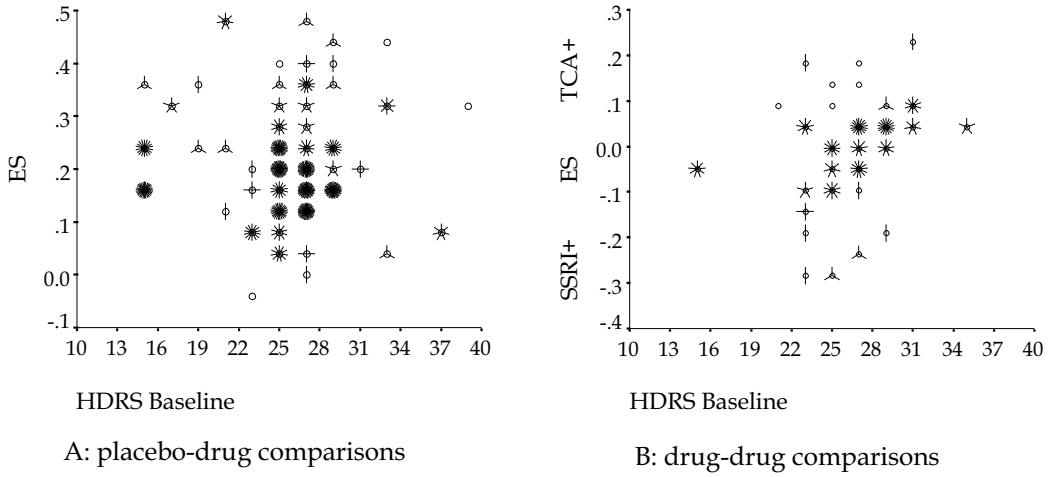

Figure 14.3 Sunflower plot of Hamilton Depression Rating Scale at inclusion (weighted by sample size) and effect size (ES).

These different results may be due to the presence or absence of a placebo cell. In fact, among the drug-drug-comparisons, there is a positive correlation of $r = .36$ between TCA-response and the Hamilton baseline score, whereas the SSRI response was not correlated with the Hamilton score ($r = .07$). Effect sizes of two cell studies (drug-drug comparisons) showed a correlation of .57 with the baseline Hamilton, compared to the zero correlation in drug-drug comparisons stemming from placebo-controlled studies ($r = .08$). In these studies, weak negative correlations with the Hamilton score can be found for all response rates (TCA $r = -.28$; SSRI $r = -.31$; Pl $r = -.29$). Thus, if a placebo is included in a study, the response in the single treatment cells is more likely to be negatively correlated with the severity of depression, if patients show lower levels of symptomatology, response occurs more often. We propose that this is the consequence of differences in drop-out handling. If there is no placebo cell in the protocol, it is likely that if symptomatology continues

to be present, severe patients also stay longer on study medication because the treating physician knows that it is highly probable that the patients are on an active drug and not on a placebo. The shorter a patients adheres to the protocol, the shorter the time during which the drug has the opportunity to unfold its action.

The *response rate* under placebo is also a possible design variable, even if, logically seen, it belongs to the outcome variables of a study. Its correlation with effect size was $r = -.41$. The lower the placebo response, the higher the difference was between drug and placebo (see Figure 14.4).

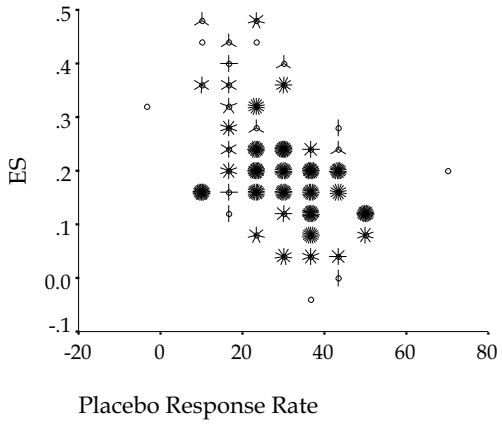

Figure 14.4 Sunflower plot: Placebo response (weighted by sample size) and effect size (ES) of drug-placebo comparisons.

14.7 DISCUSSION

This chapter focused on methodological aspects of controlled clinical trials on antidepressants. It presented the results of a publication-based meta-analysis of studies on acute therapy of depression with TCA (imipramine, amitriptyline) and/or SSRI (fluoxetine, fluvoxamine, paroxetine, sertraline) and/or placebo. Outcome was the effect size coefficient φ, which was based on intent-to-treat response rate differences.

Funnel plots indicate a publication bias, which is weak and probably does not affect the analysis of associations. We wish to consider our interpretations as hypothetical, as they are not based on a prospective design which aims at testing hypotheses and are subject to the problems of post hoc analysis by integration of different studies. Moreover, our interpretations base on data that were gained with a specific meta-analytical procedure and should be verified by other research strategies. Nevertheless, some statements can be made that may serve to better understand empirical results.

- Studies on imipramine yield higher effect sizes if they include only two treatment cells; this is in line with the results of Greenberg et al. (1992). We are inclined to attribute this to the earlier year of publication of the

two cell studies and their less sophisticated research methodology; alternatively one can think of the status of an active or a control substance – or some other variables – being responsible for this difference in effect sizes.

- Smaller studies, or single center studies, have a higher probability of yielding or publishing higher effect sizes that decline if larger studies are performed.
- A placebo run-in with exclusion of placebo responders does not seem to have any effect on the outcome of a study and therefore becomes ethically questionable unless another argument is advanced for placebo run-in.
- In acute therapy of depression, there is no reliable association between the length of a study and its outcome.
- The correlation between severity of depression and response to treatment is linked to the design of a study: A placebo cell seems to decrease response in all treatment cells with increasing severity of illness. Presumably, if a placebo cell exists, patients drop out earlier, especially if symptomatology is more impairing.
- A negative correlation was found between placebo response and effect size. It is obvious that the correlation is not caused by a ceiling effect, as it is not only present in the margin values of the placebo response. For its interpretation, statistical or content aspects can be referred to. Statistically seen, the correlation between placebo response and effect size meets the expectation, as the difference between two sizes always correlates with both sizes at about .70. This so-called a (b-a) effect (van der Bijl, 1951) was discussed in psychophysiology in the context of the law of initial value (Myrtek & Foerster, 1986). Some authors (e.g., Curnow, 1987; Thompson, Smith, & Sharp, 1997) proposed methods to correct statistically for this problem, which is linked to the regression to the mean. One can, however, also find a non-statistical interpretation: There is one group of patients which responds well to placebo and another group who does not respond to placebo. The drug responders remain the same in both samples; this leads to varying differences in response rates, depending on the placebo response rate (Montgomery, 1999). Whatever the reason for this correlation, it should call into question the concept of additivity of the placebo- and drug-effect because this implies an independence of the response difference from the initial value.

Moreover, some hypotheses can be generated which could be prospectively investigated by empirical studies or raw data. It would be interesting, for example, to design a trial in which one part of the study is conducted within a placebo-controlled design while the other one consists only of a drug-drug comparison, if possible with documentation of the patients' and doctors' estimate of treatment allocation. Systematic differences of studies with and without a placebo control, which may be the consequence of differences in inclusion, medication and drop-out handling, could be investigated, as well as determinants and consequences of blindness. The comparability of placebo- and

drug-controlled study designs is relevant for ethical reasons, since a placebo control is refused if a standard medication exists whose efficacy has been scientifically proven. This is only useful if a body of knowledge exists on the consequences of different study designs. The data presented here reveal the importance of this question, at least for antidepressant medication: If the SSRI had never been tested against placebo, and if statements on efficiency of the TCA had only been based on studies that investigated primarily TCA, the efficacy of the SSRI would be judged as higher, since older two cell studies on TCA yielded higher effect sizes.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM–IV)* (4th ed.). Washington, DC.

Curnow, R. M. (1987). Correcting for regression in assessing the response to treatment in a selected population. *Statistics in Medicine*, *6*, 113–117.

Dickersin, K., & Berlin, J. A. (1992). Meta-analysis: State-of-the-science. *Epidemiologic Reviews*, *14*, 154–176.

FDA. (1978). Guidelines for clinical evaluation of psychotropic drugs – antidepressant and antianxiety drugs. *Psychopharmacological Bulletin*, *14*, 45–53.

Feinberg, M. (1992). Comment: Subtypes of depression and response to treatment. *Journal of Consulting and Clinical Psychology*, *60*, 670–674.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71–79.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, *2*, 121–145.

Goodwin, F. K., & Jamison, K. R. (1990). *Manic-depressive illness.* New York: Oxford University Press.

Gotzsche, P. C. (1990). Bias in double-blind trials. *Danish Medical Bulletin*, *37*, 329–336.

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

Khan, A., Cohen, S., Dager, S., Avery, D. H., & Dunner, D. L. (1989). Onset of response in relation to outcome in depressed outpatients with placebo and imipramine. *Journal of Affective Disorders*, *17*, 33–38.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *2*, 719–748.

Möller, H. J., & Volz, H. P. (1996). Drug treatment of depression in the 1990s. An overview of achievements and future possibilities. *Drugs*, *52*, 625–638.

Montgomery, S. A. (1999). The failure of placebo-controlled studies. *European Neuropsychopharmacology*, *9*, 271–276.

Montgomery, S. A., & Lecrubier, Y. (1999). Is severe depression a separate indication? *European Neuropsychopharmacology*, *9*, 259–264.

Myrtek, M., & Foerster, F. (1986). The law of initial value: A rare exception. *Biological Psychology*, *22*, 227–237.

Petitti, D. B. (1994). Of babies and bathwather. *American Journal of Epidemiology*, *140*, 779–782.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Thompson, S. G., Smith, T. C., & Sharp, S. J. (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, *16*, 2741–2758.

Trivedi, M., & Rush, H. (1994). Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology*, *11*, 33–43.

van der Bijl, W. (1951). Fünf Fehlerquellen in wissenschaftlicher statistischer Forschung [Five sources of error in scientific statistical research]. *Annalen der Meteorologie*, *4*, 183–212.