Journal of Consulting and Clinical Psychology
1991, Vol 59, No 1, 184-187


0022-006X/91/$3 00

## BRIEF REPORTS

# How "Blind" Are Double-Blind Studies?

Jürgen Margraf and Anke Ehlers
Philipps-University Marburg
Marburg, Federal Republic of Germany

Walton T. Roth
Department of Psychiatry and Behavioral Sciences
Stanford University

Duncan B. Clark
Western Psychiatric Institute and Clinic
University of Pittsburgh

Javaid Sheikh, W. Stewart Agras, and C. Barr Taylor
Department of Psychiatry and Behavioral Sciences
Stanford University

Psychopharmacological studies usually attempt to eliminate "nonspecific" influences on outcome by double-blind designs. In a randomized, double-blind comparison of alprazolam, imipramine, and placebo, the great majority of panic disorder patients ($N = 59$) and their physicians were able to rate accurately whether active drug or placebo had been given. Moreover, physicians could distinguish between the two types of active drugs. Inasmuch as correct rating was possible halfway through treatment, concerns about the internal validity of the double-blind strategy arise.

In view of the ubiquity of placebo effects, it has been standard to evaluate treatment effects using the double-blind strategy (Klerman, 1986). In psychopharmacology such studies are seen as indispensable for the assessment of efficacy without which the Food and Drug Administration will not permit marketing of new drugs (see Leber, 1986). A precondition for the effectiveness of the double-blind strategy is that the participants really do not know whether the individual patients have been assigned to an active treatment or to a placebo control condition. In psychotherapy research this usually cannot be achieved because the "content" of the treatments cannot easily be hidden. In psychopharmacology, however, it is seen as an easy task to give patients an inactive substance that looks just like the active drug. However, this view neglects a potential threat to the internal validity of double-blind studies (Fisher & Greenberg, 1989). Several studies over the past years have indicated that participants in some double-blind studies may be able to determine whether active drugs or placebo were given (Hughes & Krahn, 1985; Marini, Sheard, Bridges, & Wagner, 1976; Rabkin et al., 1986; Rickels, Hesbacher, Weise, Gray, & Feldman, 1970; Stallone, Mendlewicz, & Fieve, 1975). A problem in these studies is that ratings were obtained at the end of treatment. This makes it probable that the responses of physicians and patients were influenced by the ultimate treatment outcome. In this case even correct ratings would represent only a limited threat to internal validity because treatment effects may have preceded accurate rating. To assess whether ratings about the treatment conditions can influence therapeutic outcome, it is necessary to investigate such ratings before the end of treatment. This was the goal of our study in which two active drugs were compared with a placebo in the treatment of panic disorder. Another new feature of this study was that compliance with the drug assignments was monitored by repeated blood screens. Only patients who took the prescribed medication and no additional psychotropic drugs were included. Because the drugs have typical side effects and panic disorder patients usually have an internal focus of attention as well as enhanced interoception (Ehlers, Margraf, Davies, & Roth, 1988), we expected the patients to be able to distinguish between active drugs and placebo even before the end of treatment.

This research was conducted at Stanford University. We would like to thank Sylvia Davies for help in data entry and running subjects and Hans-Henning Schulze for methodological consultation.

The study was supported in part by National Institute of Mental Health Grant 40118, by a gift from the Upjohn Company, and by a travel grant from the German Research Foundation (Eh 97/1-1).

Correspondence concerning this article should be addressed to Jürgen Margraf, Fachbereich Psychologie, Philipps-Universität Marburg, Gutenbergstrasse 18, W-D-3550 Marburg, Federal Republic of Germany.

## Method

Data collection was part of a study of drug treatments for panic disorder conducted at Stanford University. Patients met *Diagnostic and Statistical Manual of Mental Disorders*, (3rd ed., rev. *DSM-III-R*; American Psychiatric Association, 1987) criteria for panic disorder and had given written consent to participate after having received detailed written information about the study. They were randomly assigned to one of three treatment conditions: alprazolam (a triazolobenzodiazepine), imipramine (a tricyclic antidepressant), or placebo. All medica-

Table 1
*Physicians' and Patients' Assessment of Treatment Condition: Placebo vs. Active Drug*

| Rating (halfway through 8-week treatment) | Actual medication | | |
|---|---|---|---|
| | Placebo | Alprazolam | Imipramine |
| Physician[a] | | | |
| Placebo | | | |
| *N* (%) | 13 (72) | 0 | 2 (11) |
| Certainty | 4.7 | 0 | 2.5 |
| Active drug | | | |
| *N* (%) | 5 (28) | 19 (100) | 17 (89) |
| Certainty | 3.0 | 4.7 | 4.7 |
| Patient | | | |
| Placebo | | | |
| *N* (%) | 10 (56) | 1 (5) | 1 (5) |
| Certainty | 4.1 | 1.0 | 1.0 |
| Active drug | | | |
| *N* (%) | 8 (44) | 20 (95) | 19 (95) |
| Certainty | 4.1 | 5.2 | 4.4 |

[a] Because of missing data, the total number of physician ratings is 56 instead of 59.

tions were dispensed in identical-looking capsules, and neither patients nor physicians were told what medication was given to the individual patients. Participants knew which three treatment conditions were being compared and that each patient had about a one-third chance of being randomized to each condition. Treatment lasted 8 weeks. During this time, medication dose was increased until patients were free of panic attacks, suffered from unpleasant side effects, or were taking 10 capsules per day. Each capsule contained either 1 mg of alprazolam, 30 mg of imipramine, or placebo. The mean daily doses were 3.7 mg or capsules (range 1–8 mg) alprazolam and 147 mg or 4.9 capsules (range 30–270 mg) imipramine. Patients in the placebo condition took on the average 6.8 capsules (range 2–10) per day.

Halfway through treatment (Week 4) and at the end of treatment (Week 8), the treating physicians and the patients were asked to independently rate whether (a) the patient was receiving an active drug or a placebo and, if they believed an active drug was given, whether (b) it was alprazolam or imipramine. In addition, participants were asked to indicate how certain they were that their ratings were correct (on a scale ranging from *not at all* (1) to *absolutely* (6). Furthermore, the relationship between the accuracy of the ratings and treatment outcome and side effects as judged by the treating physicians in Weeks 4 and 8 was investigated. Outcome ratings included the categories *worsened* (1), *no change* (2), *minimally improved* (3), *much improved* (4), and *very much improved* (5). Side effects were rated as *none* (0), *do not significantly interfere with patient's functioning* (1), *significantly interfere with patient's functioning* (2), and *nullify therapeutic effect* (3). The results were computed separately for Weeks 4 and 8 and compared using chi-square tests with continuity correction or *t* tests.

All patients had been instructed to stop taking any psychotropic medication at least 2 weeks before participating in the study. Compliance with these instructions and with the prescribed drug treatments was assessed at baseline and Weeks 4 and 8 of treatment by blood screens for benzodiazepines, tricyclic antidepressants, and beta-blocking agents. Of the 79 outpatients who originally entered the treatment study, only 59 patients (44 women, 15 men) could be used for the present analyses because 20 patients dropped out of treatment, their blood

Table 2
*Physicians' and Patients' Assessment of Drug Used: Imipramine or Alprazolam*

| Rating (halfway through 8-week treatment) | Actual medication | | |
|---|---|---|---|
| | Placebo | Alprazolam | Imipramine |
| Physician | | | |
| Alprazolam | | | |
| *N* (%) | 0 | 15 (79) | 2 (12) |
| Certainty | 0 | 4.9 | 3.5 |
| Imipramine | | | |
| *N* (%) | 5 (100) | 4 (21) | 15 (88) |
| Certainty | 3.4 | 4.5 | 4.0 |
| Patient[a] | | | |
| Alprazolam | | | |
| *N* (%) | 3 (27) | 7 (50) | 5 (29) |
| Certainty | 2.0 | 3.1 | 1.8 |
| Imipramine | | | |
| *N* (%) | 8 (73) | 7 (50) | 19 (71) |
| Certainty | 3.3 | 2.4 | 3.3 |

[a] Because of missing data, the total number of patient ratings is 39 instead of 47 (for active drugs).

Table 3
*Relationship Between Treatment Outcome and Side Effects (Means and Standard Deviations)*

| Ratings | Condition (actual drug/rated drug) | | | |
|---|---|---|---|---|
| | Placebo/ active | Placebo/ placebo | Active/ active | Active/ placebo[a] |
| Physician | | | | |
| *N* | 5 | 13 | 36 | 2 |
| Outcome | 3.6 (1.5) | 2.4 (0.8) | 4.1 (0.9) | 3.5 |
| Side effects | 1.0 (0) | 0.2 (0.4) | 1.0 (0.5) | 0.5 |
| Patient | | | | |
| *N* | 8 | 10 | 39 | 2 |
| Outcome | 3.5 (0.9) | 2.1 (0.9) | 4.1 (0.8) | 3.0 |
| Side effects | 0.6 (0.5) | 0.3 (0.5) | 1.0 (0.5) | 0.3 |

*Note.* Outcome was rated on a 1–5 scale; side effects were rated on a 0–3 scale.
[a] No standard deviations computed because of too few cases.

screens revealed noncompliance with the drug instructions, or both. Median age was 35 years (range 19–62). Median duration of panic disorder was 7 years (range 1–40). More detailed information about the recruitment procedures, sample, and efficacy of the drugs is given in Taylor et al. (1990).

## Results

Physician and patient ratings during the fourth week of treatment are shown in Tables 1 and 2. Chi-square tests (actual vs. rated drug, active drug vs. placebo) showed that actual and rated medications were significantly related to each other: physicians, $\chi^2(1, N = 56) = 24.6$, $p < .001$; patients, $\chi^2(1, N = 59) = 16.8$, $p < .001$. If participants had simply been guessing, the best strategy would have been to always guess an active drug inasmuch as they knew that patients had only about a one-third chance for receiving placebo. Under this assumption, we would expect 38 correct ratings (68% of 56) for the physicians and 41 (69% of 59) for the patients. One sample chi-square test showed that the observed values of 49 (88%, physicians) and 49 (83%, patients) significantly exceeded these levels: physicians, $\chi^2(1, N = 56) = 9.9$, $p < .002$; patients, $\chi^2(1, N = 59) = 5.1$, $p < .03$. Thus, as early as halfway through treatment, physicians and patients were able to discriminate between active drug and placebo at a level beyond chance. Moreover, physicians correctly distinguished between the two active drugs as is shown in Table 2. Without the placebo cases, the proportion of correct ratings was 83% for the physicians and 61% for the patients (chance levels: 50%). Chi-square tests (actual vs. rated drug, alprazolam vs. imipramine) showed that actual and rated medications were significantly related to each other for the physicians $\chi^2(1, N = 41) = 13.7$, $p < .001$, but not the patients, $\chi^2(1, N = 39) = .6$, $p > .40$.

We also tested whether subjects were more certain of their ratings when they rated correctly than when they rated incorrectly. For this purpose "certainty ratings" for these two conditions were compared using *t* tests. With the Bonferroni correction for multiple comparisons, the level of significance was adjusted to .0125. With respect to the distinction between placebo and active drug, both physicians and patients were significantly more certain when they rated correctly: physicians, $t(54) = -3.85$, $p < .001$; patients, $t(39) = -3.52$, $p < .002$. With respect to the distinction between the two types of active drugs, no significant differences in the certainty ratings emerged.

Two potential sources of information for accurate ratings are treatment outcome and side effects. Table 3 shows the physicians' global ratings of these variables for correct and incorrect ratings of active drug versus placebo. Because of the low incidence of incorrect ratings, no overall statistical test was possible. Therefore, we compared only the correct ratings for placebo with those for active drug (the two middle data columns of Table 3) using *t* tests. The level of significance was adjusted to .0125 after a Bonferroni correction. Patients as well as physicians gave lower ratings for treatment success and side effects if they thought (correctly) that placebo was given than if they rated (correctly) that an active drug was given (all *t* values between −4.49 and −5.66, *df*s between 45 and 49, all *p* values < .001).

The results for the ratings at the end of treatment closely resemble the findings for Week 4. Because of space limitations they are not presented in detail.

## Discussion

As predicted in our hypothesis, panic disorder patients as well as their physicians were able to rate very accurately whether an active drug or a placebo was given. In addition, physicians could distinguish between different types of active drugs. Moreover, the simple classification of the ratings as true or false may have led to an underestimation of the actual accuracy, because patients and physicians were more certain of their ratings if they distinguished correctly between placebo and active drug. Our results go beyond earlier findings in showing that correct rating is possible long before the end of treatment and therefore cannot simply be a consequence of the outcome after treatment. However, it is possible that treatment effects already evident halfway through treatment might have contributed to rating accuracy. On the other hand, an effect in the opposite direction cannot be excluded. It remains therefore desirable for future studies to assess patients' and physicians' ratings about the treatment conditions as early as the 1st week of treatment.

Regardless of the direction of the possible causal relation-

ships, our results show that double-blind studies of these pharmacological treatments for panic disorder are not really "blind." As was recently discussed by Fisher and Greenberg (1989), this poses a serious challenge to the internal validity of this control strategy. Possible solutions for this problem could theoretically be the use of so-called "active placebos" (see Fisher & Greenberg) or the experimental manipulation of expectancy effects and other nonspecific factors in order to gain insight into their contribution to the therapeutic effects of psychopharmacological agents. Both strategies pose obvious ethical (and practical) problems. The question is whether these problems are more serious than those of using ineffective drugs. An argument in favor of using control strategies for nonspecific factors is that long-term benefits can be expected from enhanced knowledge about the mechanisms of change in that better treatments may be developed. A thorough discussion of the ethical problems related to placebos can be found in White, Tursky, and Schwartz (1985).

## References


Ehlers, A., Margraf, J., Davies, S., & Roth, W. T. (1988). Selective processing of threat cues in subjects with panic attacks. *Cognition and Emotion, 2,* 201–219.

Fisher, S., & Greenberg, R. P. (Eds.). (1989). *The limits of biological treatments for psychological distress: Comparisons with psychotherapy and placebo.* Hillsdale, NJ: Erlbaum.

Hughes, J., & Krahn, D. (1985). Blindness and the validity of the double-blind procedure. *Journal of Clinical Psychopharmacology, 5,* 138–143.

Klerman, G. L. (1986). Scientific and ethical considerations in the use of placebo controls in clinical trials in psychopharmacology. *Psychopharmacology Bulletin, 22,* 25–29.

Leber, P. (1986). The placebo control in clinical trials (A view from the FDA). *Psychopharmacology Bulletin, 22,* 31–32.

Marini, J. L., Sheard, M., Bridges, C., & Wagner, E. (1976). An evaluation of the double-blind design in a study comparing lithium carbonate with placebo. *Acta Psychiatrica Scandinavica, 53,* 343–349.

Rabkin, J. G., Markowith, J. S., Stewart, J., McGrath, P., Harrison, W., Quitkin, F. M., & Klein, D. F. (1986). How blind is blind? Assessment of patient and doctor medication guesses in a placebo-controlled trial of imipramine and phenelzine. *Psychiatry Research, 19,* 75–86.

Rickels, K., Hesbacher, P. T., Weise, C. C., Gray, B., & Feldman, H. S. (1970). Pills and improvement: A study of placebo response in psychoneurotic outpatients. *Psychopharmacologia, 16,* 318–328.

Stallone, F., Mendlewicz, J., & Fieve, R. (1975). Double-blind procedure: An assessment in a study of lithium prophylaxis. *Psychological Medicine, 5,* 78–84.

Taylor, C. B., Hayward, C., King, R., Ehlers, A., Margraf, J., Maddock, R., Roth, W. T., & Agras, W. S. (1990). Cardiovascular and symptomatic reduction effects of alprazolam and imipramine in patients with panic disorder: Results of a double-blind placebo controlled trial. *Journal of Clinical Psychopharmacology, 10,* 112–118.

White, L., Tursky, B., & Schwartz, G. E. (Eds.). (1985). *Placebo: Theory, research, and mechanisms.* New York: Guilford Press.



Received March 28, 1990
Revision received August 6, 1990
Accepted August 6, 1990 ■