Critical approach to antidepressant trials

Blindness protection is necessary, feasible and measurable

C. EVEN, E. SIOBUD-DOROCANT and R. M. DARDENNES

Background Double-blind placebo-controlled trials are the academic standard for clinical psychopharmacology research.

Aims To identify the potential defects of current double-blind procedures in trials involving antidepressants and to investigate whether safeguards for blindness protection are used.

Method We reviewed the literature and devised a short seven-item checklist for evaluating the quality of blindness protection. We performed a computerised search for 1998 to identify the placebo-controlled studies that evaluated the efficacy of an antidepressant. The checklist was used to assess all traceable antidepressant trials published in 1998.

Results Relevant criticisms question the blindness procedures. The available methods which may bolster blindness are very seldom used.

Conclusions Improvement in the blindness procedures used for antidepressant trials is necessary, feasible and measurable.

Declaration of interest None.

Double-blind placebo-controlled trials are the academic standard for clinical psychopharmacology research. Regulatory authorities agree on the necessity of double-blind studies for determining the efficacy and place of new psychoactive agents. However, such studies are open to criticism regarding the lack of safeguards in blindness procedures (e.g. Oxtoby *et al*, 1989). We review these criticisms and describe practical methodologies for renewing and improving the rigour of blindness procedures in controlled trials in psychopharmacology, especially antidepressant trials. We propose a short seven-item checklist for evaluating the quality of blindness procedures.

EXAMINATION OF THE DOUBLE BLIND

Is true blindness achieved?

Clinicians may discern whether a patient receives placebo or antidepressant in a trial, especially if the drug under study has recognisable adverse effects. However, the reasons for the unblinding are not straightforward and may involve the therapeutic effects as well as the adverse effects. Unblinding may even be linked to subtle cues that cannot be formalised and that only experienced clinicians can detect. Patients also wonder what treatment they receive. A number of them try to find out, while some are able to recognise a 'medicated' state, particularly when they have received antidepressants in the past. Some even deliberately try to break blindness by opening and tasting the capsules (Howard *et al*, 1982). Moreover, the requirement for describing side-effects before the patient gives informed consent enhances the risk of breaking the integrity of blindness (Brownell & Stunkard, 1982).

Many publications have reported the ability of patients and/or evaluators to distinguish placebo and active treatment in the area of psychotropic drugs (e.g. Brownell & Stunkard, 1982) and non-psychotropic drugs (e.g. Howard *et al*, 1982). Antidepressants are no exception. In a three-arm trial evaluating a tricyclic and a non-selective monoamine oxidase inhibitor *v.* placebo for depression, assessors and patients were able to predict the treatment received in 87% and 78% of cases respectively (Rabkin *et al*, 1986). In another three-arm trial testing alprazolam and imipramine *v.* placebo in panic disorder patients, assessors and patients were able to predict the treatment received in 88% and 83% of cases respectively (Margraf *et al*, 1991). Moreover, the assessors could distinguish alprazolam and imipramine significantly more accurately than would be predicted by chance. Bystritsky & Waikar (1994) reported data from two placebo-controlled trials, one evaluating etoperidone (a putative antidepressant that has not been marketed) in depression and the other evaluating clomipramine in obsessive–compulsive patients. Once again the patients were able to recognise their treatment more accurately than would be predicted by chance.

It therefore appears that blindness is a relative notion and that many patients in controlled antidepressant trials are not adequately blinded.

Potential consequences of breaching blindness

In any trial, both investigators and participants may breach blindness. Bias may occur if either side has expectations that the active treatment is effective.

When the physician breaches blindness his impartiality may decrease, and the conditions of a single-blind trial are then unwittingly achieved. It has been demonstrated that physicians' expectations influence outcome (Engelhardt *et al*, 1969). This is unwelcome, especially as in psychiatry the dependant variable is not a hard factual parameter but a subjective judgement, even when standardised psychometric instruments are used.

When a subject correctly guesses, from side-effects or the perception of a medicated state, that he is receiving the active treatment, an additional 'suggestion' bias may occur. Hence, the placebo response rate (the proportion of patients who respond owing to placebo effect) may be higher in the treated group than in the placebo group. It follows that the rate of true responses (those specifically due to the active

treatment) cannot simply be derived by subtracting the response rate in the placebo group from the one in the treated group. Doing so in the case of a double-blind breach may over-estimate the 'true' response rate.

Hence, controlled trials have the limitations and biases of open trials if there is significant unblinding.

Breaching blindness and apparent efficacy

Greenberg *et al* (1992, 1994) have hypothesised that results in favour of antidepressants are entirely attributable to biases linked to unblinding, namely evaluators' expectations and patients' suggestion. They put this provocative hypothesis to the test in their two studies as described below.

Greenberg *et al*'s first (1992) study is a meta-analysis involving two standard antidepressants, imipramine and amitriptyline. Only three-arm studies with one placebo group, one new antidepressant and one reference antidepressant (imipramine or amitriptyline) were included. The authors hypothesised that in such trials the evaluators' expectations are less focused on the issue of the efficacy of the reference antidepressant than in two-arm trials comparing placebo with the reference antidepressant. The distinction between the two active treatments is more difficult. This effect also lessens the potential suggestion of improvement linked to the perception of a 'medicated state'. Indeed, even if the patient guesses that he is receiving an active treatment, he may think that he is only receiving the putative treatment and not the reference one. Therefore, the authors state that three-arm studies are more blinded than two-arm studies. Cohen's effect size estimate was d=0.19, 2–4 times less than the effect size obtained in previous meta-analyses that included two-arm placebo-controlled trials evaluating the same compounds. An effect size of less than 0.20 reflects a small effect which will generally be of limited clinical significance and may be due to subtle residual biases. The authors concluded that the effect size relates more to the study design and non-specific factors associated with the study than to the specific drug effects. An alternative explanation may be that most of the three-arm trials were conducted in community samples of people with mild to moderate depression. Hence, the effect size was of necessity smaller than in the older two-arm trials which included people with more severe depression.

Greenberg *et al*'s second (1994) study was a meta-analysis of placebo-controlled trials evaluating fluoxetine. Whenever possible the authors abstracted the effect size estimate and the proportion of patients having side-effects. A significant correlation was demonstrated between these two variables. Since side-effects may lead to blindness penetration, the authors suggested that an apparent efficacy could be due to unblinded conditions. One could also argue that patients having side-effects are those having therapeutic effects. As it is, one cannot decide between these two explanations.

METHODS OF ASSESSING AND REDUCING UNBLINDING

Item 1: Assessment of true blindness

Prospective assessment of blindness

One method for checking whether blindness has been breached involves collecting both the patients' and assessors' guesses about their treatment and comparing them to chance.

The relationship of guesses to time of onset of action is an important consideration when trying to determine whether it is therapeutic or other effects which lead to breach of the blind. For this purpose, repeated evaluations of blindness would be useful when assessing whether blindness was breached before or after the onset of action for each patient. A second analysis could then be performed without the subjects who broke blindness before the onset of therapeutic effect (or without those for whom the evaluator broke blindness before the onset of therapeutic effect). The guesswork would therefore include a rating of the degree of certainty, so that 'breaking blindness' could be operationally defined as 'being rightfully certain of the treatment received'. Instead of assessing the degree of certainty in guesses, which may be influenced by the personality style, another approach is to collect forced guesses in order to elicit vague suspicions.

Retrospective assessment of blindness

White *et al* (1992) devised an original method for the retrospective evaluation of blindness. They re-analysed the data from a three-arm trial that compared placebo and low and high dosages of etoperidone. For each patient the data record included the initial clinical picture, the history and the side-effects. All data concerning therapeutic effects or conjunctural thoughts about the origin of side-effects were excluded. The evaluator was able to differentiate placebo and active treatment beyond chance. This indicates that the side-effects themselves may be a sufficient clue for the evaluator to break blindness. When the evaluators' guesses have not been recorded, this very simple method can provide information about the degree of blindness of a trial even if it has been completed for some time. It also allows the secondary re-analysis of the data using the method of Hughes & Krahn (1985), described below.

Item 2: Independent evaluation

Protocols that employ assessors who are not involved in patient care are clearly preferable, because contacts between patients and assessors are restricted to the evaluation interviews. Independent assessors may have expectations just as high as those involved in patient care but are probably less aware of the symptoms or signs that can be suggestive of the treatment received.

Item 3: Separate evaluation of adverse effects

Ideally the assessment of the therapeutic and the adverse effects should also be independent.

Item 4: Three-arm trials

As already suggested, three-arm trials (one placebo and two antidepressant arms) entail a higher degree of blindness. This study design is methodologically better, because the reference antidepressant can be validated in addition to the new antidepressant *v.* placebo. In other words, the reference antidepressant acts as an internal control as to the validity of the study (Leber, 1989). However, using a three-arm design alone does not provide full blindness protection (see above).

Item 5: Re-analysing results according to patients' and/or observers' guesses

Re-analysis requires that patients' and assessors' guesses are collected. The second analysis can then be performed taking into account the guesswork in order to appreciate whether unblinding has influenced the

results. This method has been used in a study testing nicotine gum in tobacco-withdrawal syndrome (Hughes & Krahn, 1985). The method consisted of classifying patients and assessors according to the accuracy of their answer (correct, incorrect or uncertain) and forming two tables (one for the patients and one for the assessors) based on the actual treatment and the accuracy of the answers. In this way it was possible to calculate three different effect sizes (one for each category of answer: correct, incorrect or uncertain) and thus discriminate between the specific effect of the medication and the unspecific unblinding effects. Moreover, variance analysis can be performed in order to test whether there is a significant interaction factor between the actual treatment and the opinion of the assessor or patient. Such an interaction would demonstrate the influence of the opinion of patients and/or assessors on the results. In addition, the efficacy of treatment within the truly blind group (either the group of patients who have an uncertain opinion or the group for which the assessor has an uncertain opinion) would not be questioned. In order to perform these calculations a large sample is required, but this is already necessary in antidepressant trials, which are often multi-centred (Healy, 1998).

Item 6: Use of non-inert placebos

Non-inert placebos have already been used in antidepressant trials. Thomson (1982) identified seven studies which compared tricyclics to atropine, a rather appropriate compound as some of the side-effects of tricyclics are due to anticholinergic effects. In only one of these seven studies was the antidepressant superior to atropine. A recent meta-analysis involving antidepressant trials using active placebo also puts forward the influence of unblinding effects on the results of antidepressant trials (Moncrieff *et al*, 1998). However, the poor methodology of the studies included lessens the significance of this meta-analysis. In addition, atropine and other anticholinergic agents may have antidepressant properties (e.g. Snyder & Yamamura, 1977). This alternative explanation cannot be ruled out.

Item 7: Triple-blind trials

In double-blind trials, the evaluators know which compounds are under investigation but do not know which component the patient receives. In triple-blind trials, the evaluators have no knowledge of the compounds under study and may even be unaware that a pharmacological investigation is under way (Henker *et al*, 1979). Although more difficult to implement than double-blind trials, this procedure has been used in several areas of medicine and in child psychiatry (e.g. Henker *et al*, 1979). To our knowledge no triple-blind trial involving an antidepressant has yet been conducted with adult patients.

Item 8: Selection of patients according to factors predicting placebo response

Patients who have already received one of the treatments under investigation or those who have received multiple antidepressant treatments in the past are more likely to subvert blindness. This is illustrated by data reporting a lower placebo response rate for patients with more previous episodes of depression and antidepressant treatments (Bialik *et al*, 1995), suggesting that these patients were able to identify the presence or absence of an active treatment. It has also been reported that placebo responders are characterised by a shorter duration of both the current episode and the mood disorder history (Fairchild *et al*, 1986). Obtaining samples of patients devoid of predicting factors for a placebo response would enhance a therapeutic trial by lowering the response rate in the placebo group. This procedure would take place before a possible unblinding and would therefore also make the blindness issue less significant. It may be important to identify factors which may predict blindness penetration, because if patients who are less likely to predict their treatment are selected, fewer unblinding problems are likely to occur. Factors predicting response to placebo and blindness penetration, which incidentally may overlap, have to be better delineated. Any research in this direction would be of great help. The issue of generalising findings obtained with selected samples of patients would of course have to be dealt with.

EVALUATION OF BLINDNESS PROCEDURES: THE BAPC

We have devised a short seven-point checklist (the Blindness Assessment and Protection Checklist (BAPC); see Appendix) for evaluating the degree to which supposedly blind trials are protected from blindness penetration. Among the methods described above to assess and protect blindness, those which appeared readily usable were selected by consensus among the authors to make up the checklist. Better knowledge of the factors predicting placebo response (item 8) would possibly have indirect but promising repercussions on the blindness issue, but current knowledge would not yet yield reliable techniques for protecting blindness that could be included in the BAPC. Accordingly, the BAPC consists of items 1–7 above.

As there is no evidence on the relative importance of the individual items to protect blindness, we arbitrarily gave an equal weight of one point to all items so that for a single study the score may range from 0 to 7. In the future, empirical data may suggest the use of another weighting scheme.

We performed a computerised search (via Medline) for the year 1998 using the keywords 'antidepressant', 'placebo' and 'double blind'. Forty double-blind placebo-controlled studies published in English that evaluated the efficacy of antidepressants in diverse disorders were identified. One of the authors (C.E.) reviewed the method section of the studies and rated them according to the BAPC. The average score of the studies was 0.70 (s.d.=0.82). None of the studies met the criteria in items 1, 5, 6 or 7. Only two studies were in accordance with item 3. Twelve studies used a multiple-arm design (item 4), and 14 used evaluators who were independent of patient care. We randomly selected 15 studies (using a table of randomised numbers); these were rated by another author (R.D.) in order to check the interrater reliability. The scores were considered as eight classes ranging from 0 to 7, so that a chance-corrected agreement (κ) could be calculated (Streiner, 1995). A weighted (using quadratic weights) κ value of 0.95 was found. This indicates an excellent agreement level, owing partly to the very close scores (which only ranged from 0 to 2). For a sample of studies with more dispersed scores, the interrater reliability may decrease to an unpredictable level.

DISCUSSION

Are all antidepressants equal?

Our intention is not to state that therapeutic antidepressant trials are no longer appropriate because of their weakness in maintaining blindness. However, serious defects implicate the blindness procedures of therapeutic trials which evaluate psychotropic agents, and particularly antidepressants.

This raises troublesome questions. For example, have all antidepressants consistently demonstrated their efficacy? Would the defects in design of therapeutic trials have smoothed out differences in strength of the available antidepressants? Might truly blind trials enable us to discriminate between efficacious and inefficacious antidepressants?

Need for systematic blindness checks

Relevant criticisms should not result in a sense of therapeutic nihilism. Conversely, we argue that further progress is both imperative and feasible. Double-blind controlled trials can remain the 'academic standard' provided that they systematically protect and at least check blindness with the appropriate available resources and strategies. We advocate a systematic incorporation of procedures (i.e. repeatedly collecting patients' and evaluators' guesswork) for checking true blindness for the treatment received as part of research protocols. Checking the degree of unblinding will not reduce this bias but will enable a more informative interpretation of the results to be made.

Strengths and weaknesses of the BAPC

Methods for assessing and strengthening blindness do exist, and may be appropriate. However, these sources should be better evaluated. For instance, attempts to check on true blindness may create another bias. Indeed, the very act of trying to guess whether a patient is taking a placebo or an active treatment may sensitise the evaluator (or even the patient) to medication–placebo differences and influence the estimates of drug effectiveness. This problem could be addressed by a study with two sets of evaluators: one that is asked to 'crack the code' and one that is not. Similarly, the actual merit of triple-blind procedures should also be assessed by comparing triple- and double-blind evaluators in a single trial.

The relationship between the time of unblinding and the time of onset of action is very important, but the determination of the time of onset of action still raises difficult methodological problems (Muller & Moller, 1998). Some of the remedial measures, especially three-arm trials, would raise the already high cost of randomised control trials. But as mentioned above, three-arm trials are far more advisable, even if the blindness issue is disregarded.

The re-analysis of a trial according to patients' and/or evaluators' guesses no longer compares randomised groups. It is therefore a secondary analysis: it can only strengthen or weaken the primary analysis, not replace it.

The use of active placebos appears appealing. Some authors promote their use (Moncrieff *et al*, 1998), despite being aware that ethical committees do not generally accept them. However, an empirical comparison between active and inert placebo would be needed to establish their relative values.

The empirical evaluation and broader use of the potential resources for improving blindness could be very rewarding. For now, the measures proposed in items 1–5 of the BAPC appear feasible.

Practical applications of the BAPC

Although not exhaustive, our literature search illustrates that the possibility of unblinding is usually disregarded and that the potential resources are under-utilised.

The BAPC allows an assessment of the quality of blindness in an individual trial to be made, as well as providing a standard assessment tool for the re-evaluation of the quality of blindness of antidepressant trials in the future.

Finally, the most widely used scales for evaluating the general methodological quality of trials do not attempt to assess blindness protection (e.g. Chalmers *et al*, 1981). We advocate the incorporation of one or more items from the BAPC (or at least one general item referring to blindness protection) into such scales.

CLINICAL IMPLICATIONS

- Serious defects in blindness procedures cast doubt on the trustworthiness of therapeutic antidepressant trials.
- Existing resources which can bolster blindness are under-utilised.
- Blindness protection is measurable. Instruments which aim to measure the methodological quality of clinical trials should incorporate items concerning blindness protection.

LIMITATIONS

- Methods that may improve blindness protection should be better tested. For instance, active placebo has not been shown to be superior to inert placebo.
- The interrater reliability of the Blindness Assessment and Protection Checklist (BAPC), a seven-item blindness protection measure, was found to be high. However, for samples of studies with more dispersed scores, interrater reliability may decrease to an unpredictable level.
- As there is no empirical evidence regarding the relative importance of the individual items of the BAPC, we arbitrarily gave an equal weight to all items. Future data may suggest another weighting scheme.

CHRISTIAN EVEN, MD, ERYC SIOBUD-DOROCANT, MD, ROLAND M. DARDENNES, MD, Clinique des Maladies Mentales et de l'Encéphale, Centre Hospitalier Sainte-Anne, Paris, France

Correspondence: Dr C. Even, Clinique des Maladies Mentales et de l'Encéphale, Centre Hospitalier Sainte-Anne, Paris, France

(First received 20 September 1999, final revision 10 January 2000, accepted 19 January 2000)

APPENDIX

Blindness assessment and protection checklist (BAPC) for double-blind placebo-controlled trials

(1) An assessment has been made of true blindness of patients and/or assessors for the treatment received False ☐ True ☐

(2) Study evaluators are independent of patient care False ☐ True ☐

(3) Study evaluators assessing drug efficacy do not assess drug side-effects False ☐ True ☐

(4) Three- or multiple-arm design (placebo and two or more active treatments) False ☐ True ☐

(5) Re-analysis of results according to patients' and/or observers' guesses False ☐ True ☐

(6) Non-inert placebo False ☐ True ☐

(7) Triple-blind design False ☐ True ☐

TOTAL SCORE (one point for true and zero for false) ☐

When item 2 is true or when there is no side-effect evaluation, item 3 should be omitted.

REFERENCES

Bialik, R. J., Ravindran, A. V., Bakish, D., *et al* (1995) A comparison of placebo responders and nonresponders in subgroups of depressive disorder. *Journal of Psychiatry and Neuroscience*, **20**, 265–270.

Brownell, K. & Stunkard, A. (1982) The double blind in danger: untoward consequences of informed consent. *American Journal of Psychiatry*, **139**, 1487–1489.

Bystritsky, A. & Waikar, S. V. (1994) Inert placebo versus active medication. Patient blindability in clinical pharmacological trials. *Journal of Nervous and Mental Disease*, **182**, 485–487.

Chalmers, T. C., Smith, H., Blackburn, B., *et al* (1981) A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, **2**, 31–49.

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

Fairchild, C. J., Rush, A. J., Vasavada, N., *et al* (1986) Which depressions respond to placebo? *Psychiatry Research*, **18**, 217–226.

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.

___, ___, Zborowski, M. J., *et al* (1994) A meta-analysis of fluoxetine outcome in the treatment of depression. *Journal of Nervous and Mental Disease*, **182**, 547–551.

Healy, D. (1998) Meta-analysis of trials comparing antidepressants with active placebos (commentary). *British Journal of Psychiatry*, **172**, 232–234.

Henker, B., Whalen, C. K. & Collins, B. E. (1979) Double-blind and triple-blind assessments of medication and placebo responses in hyperactive children. *Journal of Abnormal Child Psychology*, **7**, 1–13.

Howard, J., Whittemore, A. S., Hoover, J. J., *et al* (1982) How blind was the patient blind in AMIS? *Clinical Pharmacology and Therapeutics*, **32**, 543–553.

Hughes, J. & Krahn, D. (1985) Blindness and the validity of the double-blind procedure. *Journal of Clinical Psychopharmacology*, **5**, 138–142.

Leber, P. D. (1989) Hazards of inference: the active control investigation. *Epilepsia*, **30**, 57–63; discussion 64–68.

Margraf, J., Ehlers, A., Roth, W. T., *et al* (1991) How 'blind' are double-blind studies? *Journal of Consulting and Clinical Psychology*, **59**, 184–187.

Moncrieff, J., Wessely, S. & Hardy, R. (1998) Meta-analysis of trials comparing antidepressants with active placebos. *British Journal of Psychiatry*, **172**, 227–231.

Muller, H. & Moller, H. J. (1998) Methodological problems in the estimation of the onset of the antidepressant effect. *Journal of Affective Disorders*, **48**, 15–23.

Oxtoby, A., Jones, A. & Robinson, M. (1989) Is your 'double-blind' design truly double-blind? *British Journal of Psychiatry*, **155**, 700–701.

Rabkin, J., Markowitz, J., Stewart, J., *et al* (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.

Snyder, S. & Yamamura, H. (1977) Antidepressants and the muscarinic acetylcholine receptor. *Archives of General Psychiatry*, **2**, 236–239.

Streiner, D. L. (1995) Learning how to differ: agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry*, **40**, 60–66.

Thomson, R. (1982) Side effects and placebo amplification. *British Journal of Psychiatry*, **140**, 64–68.

White, K., Kando, J., Park, T., *et al* (1992) Side effects and the 'blindability' of clinical drug trials. *American Journal of Psychiatry*, **149**, 1730–1731.