ELSEVIER

European Neuropsychopharmacology 5 (1995) 531–533

EUROPEAN NEURO-PSYCHOPHARMACOLOGY

Report of Consensus Meeting

Clinical relevance of response and improvement in psychopharmacology

A statement from the European College of Neuropsychopharmacology, from an ECNP workshop, Jerusalem, October 1995[a]

1. Introduction

In October 1994 the European College of Neuropsychopharmacology held a workshop to address the issue of clinical relevance in efficacy studies in depression. This consensus statement has been produced following the deliberations of the panel.

2. The problem

Conventionally the efficacy of a new psychotropic drug is established by comparing the response of the active drug with the response seen on placebo using a randomised double-blind group comparison design. It has been a continuing point of discussion whether the demonstration of a significant difference in response between a drug and placebo is sufficient in itself to establish efficacy or whether some judgement is also required about the clinical relevance of the observed difference.

3. The context – regulatory or clinical

The decision about whether the efficacy of a drug is accepted is affected by the context in which the decision is made. From the perspective of licensing new drugs the question addressed is whether a drug is effective and safe for use in given conditions. A statistically significant difference between drug and placebo registered on a prospectively determined outcome measure is accepted as evidence of efficacy.

Clinical relevance is concerned with the related concept of effectiveness rather than efficacy and is the concern of clinical evaluation. The evaluation of how effective a drug is goes beyond the decision of whether or not efficacy has been demonstrated for a drug and may be affected by other factorsincluding socio-political and economic factors, and also, for example, by the medical standard prevailing at the time of registration.

The decision on effectiveness made by licensing authorities is informed by a legal framework and weighs demonstrated efficacy, the degree of effectiveness and side effects of a drug against the risks of the disorder in making a risk benefit assessment.

4. Changes in response rates in studies

Attention has recently focused on the question of clinical relevance because of the trend towards smaller differences between active drug and placebo observed in clinical studies to test efficacy. The difference between drug and placebo has appeared to be more difficult to demonstrate in recent studies compared with earlier efficacy studies. The number of patients included in studies is generally considerably larger in recent studies. This increase in size addresses the increased difficulty in demonstrating efficacy and may also appear to provide a test model that is closer to the way a drug will be used in the general population. However it is possible for a statistically significant difference to be achieved between drug and placebo in large studies even though the difference in response between groups appears very small in absolute numerical terms. The question has arisen whether very small differences in response, which may be statistically significant, are of relevance clinically. There is a risk that drugs which are efficacious but have only modest but important clinical effects may be overlooked because the difference from placebo appears too small. The similar risk is that drugs which have been

[a] Panel: S.A. Montgomery (chairman), J. Angst, R. Buller, A. Derivan, K. Evans, N. Fineberg, D. Hackett, D. Harrison, I. Henry, J. Jorgensen, S. Kasper, S. Kindler, R. Lane, P.D. Leber, Y. Lecrubier, A.J. Schutte, J.M.A. Sitsen, E. Steiness, J.G. Storosum, S.J. van den Berg, J.M. van Ree, B.J. van Zwieten-Boot. Rapporteur: D.A. Montgomery

SSDI 0924-977X(95)00060-7

shown to be efficacious compared with placebo but have unimportant clinical effects may be licensed. It is important to discover whether there is any agreement on the basis for deciding at what level a difference signifies clinical relevance or what criteria might be applied.

5. Towards a consensus on clinical relevance

A clear consensus is lacking on how one judges how effective a treatment is and decisions may be affected by the preconceptions of assessors. There is a desire for a consensus on clinical relevance, not as an additional criterion of efficacy, but in order to help decisions to be consistent.

6. Criteria for clinical relevance

A number of methods have been used to indicate whether a statistically significant difference in improvement is clinically relevant. These include:

(1) comparing the number of patients achieving responder/nonresponder status at a given point defined by a percentage improvement in the pivotal severity scale;
(2) comparing the number of patients in different response categories at a given point defined by clinical global improvement scores or cutoffs on the severity rating scales;
(3) establishing criteria for accepting a particular difference in mean change on severity scales as clinically relevant.

7. Responder categories

7.1. Strengths

A reduction of 50% in the score on the pivotal severity scale is an arbitrary but helpful measure that has been widely used to define responders as a measure of clinical relevance. It is an easily understood measure that is intuitively acceptable in that halving the symptom severity is an important clinical achievement both for the sufferer and for the clinician.

This pragmatic approach is to some extent supported by the correlation between responders defined by this cutoff and the Clinician's Global Impression which itself probably mirrors most closely the judgement made by the clinician in clinical practice.

7.2. Problems

The objections to a definition of responder as a reduction of 50% in severity scale score are strong. To call a 50% reduction an absolute responder is a misnomer since 50% reduction in severity is not the same as being relatively free of symptoms or having a symptom level compatible with the normal recovered population. However there are no agreed definitive cutoffs for euthymia and, moreover, patients who respond to treatment reach euthymia relatively late. To require euthymia as a measure of clinical relevance would be impractical in view of the much longer study duration that would be needed.

The 50% reduction in scale score is not a consistent measure across levels of severity. It works best in illness of moderate severity where a 50% reduction in score would fit reasonably a clinical judgement of a response. In severe depression a 50% reduction in score would represent a different level of outcome than in moderate depression with patients still recording relatively high severity scores in spite of the amelioration. In mild illness there would be less scope for movement on a severity scale and a reduction of 50% might be unrealistic to expect.

The 50% reduction in score is also affected by time since different rates of regression would be expected from differing levels of severity. For example the usefulness of this measure would be limited in severe illness as it is likely that too few patients would reach this level of improvement within the predefined time of the study for valid comparisons to be made.

This measure is also affected by the type of sample included in the study. For example in resistant illness a 50% reduction might not be achieved but a lower percentage improvement could represent a clinically useful reduction in illness in these patients. Similarly in some disorders a small percentage improvement might represent a significant improvement in function, as for example in obsessive compulsive disorder where a 25% reduction in severity is used to define a responder.

8. Response categories

Converting the improvement on a sensitive severity scale into responder/nonresponder categories does not make full use of the available information. A comparison would be more usefully made on the basis of severity categories following treatment, for example recovered, mild, moderate, severe or similarly percentage improvement categories. This approach more closely meets the clinical evaluation of efficacy which asks what proportion of patients will respond how much.

However, for a valid estimate of clinical relevance of the results of a study the distribution of change scores needs to be known for the condition as well as an estimate of the representativeness of the sample.

9. Effect sizes

9.1. Strength

The difference between a drug and placebo in absolute effect size difference measured on the mean amelioration of the pivotal severity scale has the apparent attraction of being a clearly defined 'hard' criterion. A number of criteria have been proposed, for example, 4 points on the Montgomery & Åsberg Depression Rating Scale or the Hamilton Rating Scale.

9.2. Problems

If an absolute effect size is to be used as a criterion of clinical relevance a reference of expected effect sizes is needed to provide a check on the fairness of the comparison. The range of effect sizes observed in previous clinical studies provides some reference but this has the drawback of being historical data which cannot be used as an adequate control. Current studies are not necessarily comparable with earlier studies, being larger, having different methodologies and there being apparent differences in the population samples included. The principle of using randomised double-blind head to head comparisons against placebo is abandoned if data from historical controls are used.

It is also clear from a review of published studies shows that placebo shows a distribution of effect sizes and therefore the stability of expected effect size estimates is questionable.

The effect size is likely to be affected by the side effect profile of a drug since drop-outs due to side effects would be likely to occur at an earlier point with higher illness scores than the later placebo responder drop-outs and would thus contribute to higher end severity scores compared with placebo. It becomes very important to know about censoring during the study.

Studies vary substantially in the size of the placebo response. In studies with a high placebo response there is less room for drug response to be observed. There is an upper ceiling on response and therefore if the placebo response rises there is less room for a difference between drug and placebo to emerge and the difference in effect size will be smaller.

Some of the problems can be overcome by the inclusion of a third reference treatment arm in studies of a drug compared with placebo, which can help put any observed drug placebo difference into perspective. However, the absence of accepted criteria for what is a clinically relevant difference in effect size, or indeed for what is a not clinically relevant effect size, makes the 'hardness' of effect size as a criterion more imagined than real.

10. Summary

The use of responder categories may provide more clinically relevant information than mean differences on a severity scale but is not a definitive solution. Responder analysis is more useful in moderate illness than in mild or severe illness, depends on the timing of the analysis, and on the variance in the sample. Nevertheless, a statistically significant difference between drug and placebo treatment with respect to the percentage responders defined as a pre-established degree of reduction on a pivotal severity scale may provide the most objective information about clinical relevance yet available.

Using distributions of categories of response may be less restrictive and provide a better description of results but more data are needed before conclusions can be reached.

Effect size is an attractive, apparently simple method of establishing clinical relevance but criteria are lacking on cutoffs for accepting an effect size as relevant.

There is a need for more information on patient samples, a better exploration of the conditions affecting response, and a better description of outcome than a single point.

There are no accepted firm criteria for the clinical relevance of differences observed. There is insufficient precision in outcome measures as they currently used to set up rules.

It is proposed that the criteria under discussion are interesting and need to be further investigated.

More data are needed on the expected range of response, the range of size of effect of drugs, the range of response categories to drugs, and the conditions affecting response. This information is available in the large databases from clinical trial programmes of the pharmaceutical industry and a coordinated examination of available databases would provide a fruitful basis for arriving at a worthwhile estimate of the clinical relevance of psychotropic drug effects.