The final, definitive version of this paper has been published in *Journal of Psychopharmacology*, vol. 22 no. 4, June 2008 by SAGE Publications Ltd. All rights reserved. © SAGE Publications Ltd.

Establishing non-inferiority in treatment trials in psychiatry – guidelines from an Expert Consensus Meeting

D. Nutt, Psychopharmacology Unit, University of Bristol, Bristol, UK.

C. Allgulander, Karolinska Institute, Karolinska University Hospital Huddinge, Huddinge, Sweden.

Y. Lecrubier, La Salpêtrière – Pavillon Clérambault, Boulevard de l'Hôpital, Paris, France.

T. Peters, Department of Community Based Medicine, Academic Unit of Primary Health Care, University of Bristol, Bristol, UK.

U. Wittchen, Fachrichtung Psychologie, TU Dresden, Dresden, Germany.

Abstract

Comparing the efficacy of different treatments in psychiatry is difficult for many reasons, even when they are investigated in 'head-to-head' studies. A consensus meeting was, therefore, held to produce best practice guidelines for such studies. This article presents the conclusions of this consensus and illustrates it using published data in the field of antidepressant treatment of generalized anxiety disorder.

Key words clinical trials; duloxetine; GAD; non-inferiority; venlafaxine

Introduction

The issue of non-inferiority when comparing different drugs or other treatment modalities in medicine has for many years been very contentious. Most previous claims of comparable efficacy between treatments have been complicated by the fact that a lack of statistical power can lead to apparent non-difference in outcome that may well be falsely assumed to imply non-inferiority. Moreover, there is no generally accepted methodology for establishing claims of non-inferiority in such comparative studies. In order to overcome these problems, in September 2005 the authors of this article met as a panel of experts in the relevant areas, namely anxiety disorders, clinical trials and statistics. The objective was to provide a set of consensus guidelines on how to conduct non-inferiority trials.

To take advantage of the current practice and clinical evidence literature, the group focused on one psychiatric disorder, generalized anxiety disorder (GAD). The published data from GAD placebo-controlled trials provided established efficacy estimates to guide discussions regarding non-inferiority comparisons (Appendix). In addition, there were comparative studies of two drug treatments soon to be forthcoming, thus providing a prospective test of the utility and practicality of the consensus output (Allgulander, et al., 2001). The principles, however, are relevant to all psychiatric disorders and treatments, and therefore this paper should have wide applicability.

A non-inferiority trial provides the opportunity to evaluate the efficacy of a new treatment vs. a well-known treatment with established efficacy. Typically, a 'test' treatment is compared with a treatment that is already approved and/or widely used by practicing physicians. The objective of a non-inferiority clinical trial is to establish that the new treatment is not worse than the active control by more than a pre-specified non-inferiority margin (also referred to as the maximum clinically irrelevant difference or delta). In other words, the new treatment is

shown to be as effective or no worse than the active control based upon an a priori defined clinically appropriate response measure. It is also important that the comparator should be superior to placebo in the corresponding trial.

The consensus meeting and subsequent discussions reviewed a series of issues that bear on the validation of a non-inferiority trial. These are discussed separately below and include consideration of regulatory, statistical and clinical issues, such as the sensitivity of the measure, clinically meaningful differences between treatments and sample size (Gotzsche 2006).

Identification of the non-inferiority margin

The EMEA (The European Medicines Agency) requires clinical trials with the comparison of a test product with an authorized standard product in an adequate dose (EMA guidance for GAD). For our purposes, we considered the development of guidelines for non-inferiority with a newly approved treatment, duloxetine, with the established or reference treatment of venlafaxine XR, an antidepressant that has been approved for the treatment of GAD in the dose range used in these duloxetine comparator studies.

The issue here is: what is an appropriate clinically meaningful margin of non-inferiority for duloxetine vs. venlafaxine in the treatment of GAD? This margin is the largest difference that is judged to be clinically acceptable and cannot be greater than the smallest effect size that the active drug would be expected to have compared with placebo in the setting of a planned trial (ICH E10 guidelines). The choice of measure for this is open – it could be either doctor- or patient-related scales or differences in response or remission rates.

Given that the preponderance of efficacy data in GAD trials has been collected using the Hamilton Anxiety Rating Scale (HAM-A), those present at the consensus meeting agreed to use these scores in the worked example.

In placebo-controlled (superiority) trials, a statement is made in advance regarding the minimum difference between the active treatment and placebo that would be considered clinically important. The sample size calculation for a placebo-controlled superiority trial requires a reasonable chance (power) of detecting this ‘target’ difference in the primary efficacy measure for the test treatment vs. placebo. The observed difference together with the 95% confidence interval for the difference would then be used to provide an estimate of the improvement in efficacy seen with the test treatment. The observed data should then be compared with the clinically important value to determine whether the active treatment provides at least this minimum improvement, with in particular the confidence interval clinically interpreted in the light of the pre-specified target difference.

In non-inferiority trials, the ‘target’ difference is much more difficult to specify because the comparison is between two active treatments, and therefore what constitutes minimal improvement is more subjective. In other words, determining the non-inferiority margin is much more difficult and potentially contentious. In the past, the non-inferiority margin was sometimes set to be a proportion (one-third to one-half) of the observed difference between the active comparator and placebo (possibly that found from meta-analysis of placebo-controlled trials). This was based on the reasoning that the test treatment would then be shown to be closer to the comparator than to placebo. However, the establishment of a non-inferiority margin cannot be based purely on statistical criteria because the uncertainty associated with the selected margin is not statistically measurable (Hung, et al., 2003). The

EMA Points to Consider on the Choice of Non-Inferiority Margin (page 2) also states: ‘Such an approach [taking a proportion of the superiority margin] does not necessarily even ensure superiority over placebo and has no clinical rationale to support it. The choice of delta must always be justified on both clinical and statistical grounds’. Therefore, clinical relevance together with statistical reasoning is essential to establish the non-inferiority margin. The ICH E9 guidelines state that the choice of a clinically acceptable non-inferiority margin needs justification with respect to its meaning for future patients and should be smaller than differences observed in superiority trials of the active comparator vs. placebo.

Although there is no gold standard in determining a non-inferiority margin, the approach taken in the consensus meeting was to first identify an appropriate clinically meaningful drug-placebo difference on the primary efficacy variable. Once the drug-placebo difference is identified, this information can be used to assist in the identification of a ‘no less effective or no worse than’ difference between two active treatments.

Statistical methods

In the analysis of a non-inferiority trial, a point estimate is calculated for the difference in efficacy between the test product and the active comparator. This represents the best estimate of the true difference. The 97.5% one-sided confidence interval is used to both provide an interval estimate for this difference and to investigate the null hypothesis that the difference in efficacy between the test product and the active comparator (in this case, duloxetine vs. venlafaxine) is no worse than the prespecified margin, vs. the alternative hypothesis that the true difference is greater than this target. Figure 1 represents the confidence interval approach to analysis of non-inferiority trials. The lower limit of the 97.5% confidence interval is used to determine whether a difference between the two products greater than the pre-set non-inferiority margin can be excluded with sufficient confidence. If the non-inferiority margin is excluded from the 97.5% one-sided confidence interval, then it can be concluded that the test product (duloxetine) is not inferior to the active comparator (venlafaxine).

The point estimate may even indicate that the test product (duloxetine) has better efficacy than the active comparator (venlafaxine); however, if the confidence interval is wide as result of high variability and/or insufficient sample size and includes the non-inferiority margin, then inferiority of the test product cannot be excluded.

Trial criteria that will allow a fair assessment of non-inferiority

There are many aspects of trial design and analysis that can influence a non-inferiority assessment. These include: the nature of the comparator studies; whether a placebo group is included; the relative outcome of each active arm compared with placebo and the overall response rate in the trial. Following discussion, the consensus committee’s view was that for a test intervention to be considered as potentially non-inferior (NI) to an active (licensed) comparator, the following conditions are all required to be met:

- 1) There is at least one three-arm double-blind placebo-controlled comparison trial for the test intervention with an active comparator.
- 2) Both test intervention and active comparator must be superior to placebo by a clinically meaningful difference [For the HAMA this would be at least a (baseline corrected) change of ≥ 2 points at endpoint compared with placebo at 8 weeks. This figure of 2 is based on the range of deltas in the venlafaxine GAD studies see Figure 2].

3) The 'responder rates' [defined as a >50% reduction in HAMA] should be different by at least 10% points, e.g. if 30% for placebo then they would need to be >40% for active comparator and test intervention.

4) There should be reasonable evidence that both the active comparator and the test intervention are better than placebo on the primary outcome measure in the trial ($P < 0.05$ in each case). (In particular, if the test intervention does not separate from placebo but active comparator does, then the test intervention has failed).

Once these initial criteria have been met, it is then possible to consider the test intervention to show non-inferiority compared with the active comparator, if the following additional two conditions are both met:

1) The difference between the active comparator and the test intervention is <50% of difference between the active comparator and placebo, and this difference is not clinically meaningful [e.g. not more than 1.5 HAMA points].

2) The responder rate of the test intervention is not more than 5% points lower than that of the active comparator. [For the HAMA, the rationale for the delta of 1.5 is on three grounds: firstly, it is (just) below the lower limit of the range of previously observed differences in mean HAMA between venlafaxine and placebo; secondly, figures below this would be virtually impossible to justify on clinical grounds and thirdly, from Figure 2 of the Appendix corresponds approximately to a 5% point difference in responder rates, which in turn is also considered a minimum for NI as stated above.]

Other necessary criteria

In addition, there are a number of other criteria that must be met:

1) Dosing must be proportionate to the known efficacy of each drug (noting that any conclusions drawn from a non-inferiority analysis only apply to the doses of drugs tested).

2) Treatment duration must be appropriate to the known pharmacological actions of the drug (see comment on benzodiazepines in point 6 below).

3) The sample size of the study must be such that it has a reasonable power of obtaining a lower confidence limit that exceeds delta [see Statistics section above].

4) There are no age, gender or other major prognostic imbalances between the groups (despite randomization).

Additional considerations

In addition to these main criteria, a number of other considerations and observations were made:

1) Either a flexible or fixed dose treatment design could be used.

2) Distortions due to markedly different speed of therapeutic response between different classes of drug should be avoided as far as possible. Thus comparing a fast-acting drug, such

as a benzodiazepine with an antidepressant in GAD would be unduly influenced if the outcome measures were taken at the time of the onset of benzodiazepine efficacy.

- 3) Minimal mean entry level of severity should be at least in the moderately ill range. A suggestion of a score of >20, ideally at least 22, on the HAMA was recommended.
- 4) One could use for the outcome comparisons a change score on questions that measure core symptoms of the illness e.g. item 1 on the HAMA. It would also be acceptable to use changes in a subscale – e.g. that for worry anxiety tension, although a lesser range of scores may reduce power and may compromise reliability.
- 5) Disorder-specific scales may offer greater power – e.g. GADI for GAD (Argyropoulos, et al., 2007). Trials in other disorders may of course use other scales that are specifically relevant to that disorder.
- 6) Remission should be considered as an alternative and potentially more secure endpoint (but this is hard to reach with GAD though longer trial duration may achieve this).
- 7) The choice of analytical models may well be influential – for example, a ‘last observation carried forward’ (LOCF) approach to missing data will widen confidence intervals due to high scoring losses to follow-up, and observed/complete case analysis may overestimate the efficacy of interventions with high attrition rates. Sensitivity analyses employing various approaches including multiple imputation may be preferable.
- 8) Studies should include compliance evaluations to ensure that compliance levels in the different groups cannot explain any observed difference in efficacy and to ensure that compliance is sufficiently good to avoid a general bias.
- 9) Consideration should be given to methods to minimize measurement error in the primary efficacy measure. Methods to minimize measurement error in the HAMA, such as careful training and certification of HAMA raters, should always be considered.

Conclusions

Non-inferiority studies are likely to assume greater importance in psychiatry as healthcare providers seek to optimize treatment provision. This article sets out the results of an expert consensus meeting that addressed this issue using the worked example of antidepressant efficacy in GAD. The conclusions will hopefully provide a basis for other such studies involving various treatments and disorders.

Acknowledgements

The consensus meeting was supported by an unrestricted educational grant from Eli Lilly and Company, who also provided copies of the EMEA guidelines on non-inferiority trials and the data from the published antidepressant GAD trials. There were no company members present at the consensus meeting and unless stated otherwise the contents of this paper are solely the product of the independent members of the consensus committee.

Appendix Published data on approved GAD medications

One approach to identify a non-inferiority margin is to assess the full range of established differences in mean changes of the HAMA total score vs. placebo that were used as evidence

of a clinically relevant improvement in efficacy for the medications approved for the treatment of GAD. Three medications have been approved in Europe for the treatment of GAD (venlafaxine, paroxetine and escitalopram). The published literature on these medications in GAD has shown difference between active drug and placebo in mean change in HAMA total score from baseline ranging from 1.1 to 5.4 (Tables 1 and 2). After 8 weeks of treatment, venlafaxine differences in mean values ranged from 1.7 to 4.4 with an overall mean difference of 2.7 (all doses). With paroxetine differences ranged from 2.1 to 3.2 with an overall mean difference of 2.7, escitalopram differences ranged from 2.0 to 3.9 with an overall mean difference of 2.8. While most of the studies showed evidence of differences compared with placebo, the range of differences in mean values across the studies is large. Focusing on venlafaxine, these studies have shown differences in mean values in HAMA total score ranging from 1.7 to 5.4 points. The difference in responder rates (based on the reduction in HAMA total score) ranged from 8% to 29%.

Another approach to identifying a non-inferiority margin would be to examine the meaning of differences in HAMA change scores of various magnitudes. Logistic regression analysis was used to predict drug-placebo differences in HAMA response rates utilizing drug-placebo HAMA change score differences. The regression analysis was performed because response is a dichotomous variable and this type of data tends to be nonlinear. The varying sample sizes in these three studies are also taken into account in the analysis.

Included in this analysis were acute phase drug-placebo HAMA response rates and HAMA mean changes (from baseline to endpoint) for all active drug comparisons with placebo in three venlafaxine studies (Davidson, et al., 1999; Allgulander, et al., 2001; Gelenberg, et al., 2000). Unfortunately, HAMA response rate data was not provided for paroxetine and escitalopram studies. With various dosages of venlafaxine included in these studies, a total of six comparisons with placebo were available. The Gelenberg study used a 40% or greater decrease from baseline to define clinical response, whereas the other two studies used a 50% or greater decrease from baseline. Placebo responder rates across the three studies were 36%, 40% and 42%. To perform the regression analysis, a fixed value of 40% for the placebo response rate was assumed.

Figure 2 displays the actual study results, the predicted values based on the regression model, and 95% confidence intervals around the predicted values. Using this regression model, we can estimate HAMA response rate differences at various levels of HAMA change differences. For example, a HAMA change difference between two treatments of three points would correspond to approximately a 17% difference in the response rates between the treatments. A HAMA change difference of two points corresponds to a HAMA responder rate difference of about 10%.

A similar analysis was performed using HAMA changes and response rates based on the Clinician's Global Impression – Improvement (CGI-I) scale (CGI-I response of 1 or 2) using data from the three approved medications. Data from six studies based upon the three approved medications were included in the analysis (Davidson, et al. 1999; Allgulander, et al. 2001; Lenox-Smith & Reynolds, 2003; Pollack, et al. 2001; Rickels, et al. 2003; Davidson et al. 2004). Placebo response rates across the six studies were 39%, 44%, 37%, 47%, 46% and 38%. To perform the analysis, a fixed value of 40% for the placebo response rate was assumed.

Figure 3 displays the actual study results, the predicted values based on the logistic regression model, and 95% confidence intervals around the predicted values. Using this regression

model, we can estimate CGI-I response rate differences at various levels of HAMA change differences. For example, a HAMA change difference between two treatments of three points would correspond to approximately a 23% difference in the response rates between the treatments. A HAMA change difference of two points corresponds to a HAMA responder rate difference of about 19%. Note that the association between the difference in HAMA mean changes and the difference in CGI-I response rates was not very strong so the analysis of the CGI-I data is not of great help in identifying the non-inferiority margin for the HAMA mean difference.

References

- Allgulander, C, Hackett, D, Salinas, E (2001) Venlafaxine extended release (ER) in the treatment of generalized anxiety disorder. *Br J Psychiatry* 179: 15–22.
- Argyropoulos, SV, Ploubidis, GB, Wright, TS, Palm, ME, Hood, SD, Nash, JR, Taylor, AC, Forshall, SW, Anderson, IM, Nutt, DJ, Potokar, JP (2007) Development and validation of the Generalized Anxiety Disorder Inventory (GADI). *J Psychopharmacol* 21: 145–152.
- CHMP Guideline on the Clinical investigation of medicinal products indicated for Generalised Anxiety Disorder CPMP/EWP/4284/02. Published January 2005 Effective July 2005. Access at <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>
- CHMP Points to Consider on the Choice of Non-inferiority Margin. CPMP/EWP/2158/99. Published Jul 2005. Effective Jan 2006. Access at <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>
- CHMP Points to Consider on Switching Between Superiority and Non-inferiority. CPMP/EWP/482/99. Published July 2000. Effective February 2001. Access at <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>
- Davidson, JRT, Bose, A, Su, G (2002a, b) Escitalopram in the treatment of generalized anxiety disorder. Poster presented at the 42nd Annual New Clinical Drug Evaluation (NCDEU) Meeting. Boca Raton, FL, USA, 10-13 June 2002.
- Davidson, JR, Bose, A, Korotzer, A, Zheng, H (2004) Escitalopram in the treatment of generalized anxiety disorder: a double-blind, placebo-controlled, flexible dose study. *Depress Anxiety* 19: 234–240.
- Davidson, JR, DuPont, RL, Hedges, D, Haskins, JT (1999) Efficacy, safety, and tolerability of venlafaxine extended release and buspirone in outpatient with GAD. *J Clin Psychiatry* 60: 528–535.
- Gelenberg, AJ, Lydiard, RB, Rudolph, RL, Aguiar, L, Haskins, JT, Salinas, E (2000) Efficacy of venlafaxine extended release in nondepressed outpatients with generalized anxiety disorder: a 6 month randomized controlled trial. *JAMA* 283: 3082–3088.
- Goodman, WK, Bose, A, Wang, Q (2003) Escitalopram 10 mg/day is effective in the treatment of generalized anxiety disorder. Poster presented at: 23rd Annual Conference of the Anxiety Disorders Association of America; March 27-30, 2003; Toronto, Canada.
- Gotzsche, PC (2006) Noninferiority and equivalence randomized trials: lessons and reservations-secondary publication. *Ugeskr Laeger* 168:3137–3139.
- Hung, HMJ, Wang, SJ, Tsong, Y, Lawrence, J, O’Neil, RT (2003) Some fundamental issues with non-inferiority testing in active controlled trials. *Statist Med* 22: 213–225.
- ICH Note for Guidance E9 (Statistical Principles for Clinical Trials). Adopted by CPMP, March 1998, issued as CPMP/ICH/363/96. Access at <http://www.ich.org/cache/compo/475-272-1.html>
- ICH Note for Guidance E10 (Choice of Control Group). Adopted by CPMP, July 2000, issued as CPMP/ICH/364/96. Access at <http://www.ich.org/cache/compo/475-272-1.html>
- Lenox-Smith, AJ, Reynolds, A (2003) A double-blind, randomized, placebo-controlled study of venlafaxine XL in patients with generalized anxiety disorder in primary care. *Br J Gen Pract* 53: 772–777.
- Pollack, MH, Zaninelli, R, Goddard, A, McCafferty, JP, Bellew, KM, Burnham, DB, Iyengar, MK (2001) Paroxetine in the treatment of generalized anxiety disorder: results of a placebo-controlled, flexible dosage trial. *J Clin Psychol* 62: 350–357.

Prescrire International Paroxetine in generalized anxiety: uncertain efficacy. *Prescrire Int* 2003; 12: 175–177.

Rickels, K, Pollack, MH, Sheehan, DV, Haskins, JT (2000) Efficacy of extended-release venlafaxine in nondepressed outpatients with GAD. *Am J Psychiatry* 157: 968–974.

Rickels, K, Zaninelli, R, McCafferty, J, Bellew, K, Iyengar, M, Sheehan, D (2003) Paroxetine treatment of generalized anxiety disorder: a double-blind, placebo controlled study. *Am J Psychiatry* 160: 749–756.