Contemporary Clinical Trials 43 (2015) 60–62

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Short communication

Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences

Joanna Moncrieff [a,*], Irving Kirsch [b]

[a] University College London, Division of Psychiatry, Maple House, 149, Tottenham Court Road, London, W1T 7NF, UK
[b] Harvard Medical School, Program in Placebo Studies, Beth Israel Deaconess Medical Centre, 330 Brookline Avenue, Boston, MA 02215, United States

ARTICLE INFO

Article history:
Received 9 March 2015
Received in revised form 7 May 2015
Accepted 9 May 2015
Available online 12 May 2015

Keywords:
Antidepressant efficacy
Clinical relevance of antidepressant effects
Hamilton rating scale for depression
Measurement of depression
Effect size in depression

ABSTRACT

Meta-analyses indicate that antidepressants are superior to placebos in statistical terms, but the clinical relevance of the differences has not been established. Previous suggestions of clinically relevant effect sizes have not been supported by empirical evidence. In the current paper we apply an empirical method that consists of comparing scores obtained on the Hamilton rating scale for depression (HAM-D) and scores from the Clinical Global Impressions-Improvement (CGI-I) scale. This method reveals that a HAM-D difference of 3 points is undetectable by clinicians using the CGI-I scale. A difference of 7 points on the HAM-D, or an effect size of 0.875, is required to correspond to a rating of 'minimal improvement' on the CGI-I. By these criteria differences between antidepressants and placebo in randomised controlled trials, including trials conducted with people diagnosed with very severe depression, are not detectable by clinicians and fall far short of levels consistent with clinically observable minimal levels of improvement. Clinical significance should be considered alongside statistical significance when making decisions about the approval and use of medications like antidepressants.

Decisions about the approval and use of medications should not be based on statistical significance alone. Estimation of the clinical relevance of drug-placebo differences is also necessary, to balance the utility of a drug's effects against its side effects and health risks. Antidepressants have been compared with placebo in numerous randomised controlled trials. The methodological flaws of these studies have been widely discussed and include selective publication and outcome reporting, bias introduced by placebo washout procedures, infringement of the double bind, and inflation of drug-placebo differences through categorisation of data [1–3]. Despite these problems, there remains a consensus that antidepressants have worthwhile effects in people with more severe depression, at least. The difference between antidepressants and placebo in the treatment of major depression is small, however. Mean differences between antidepressants and placebo reported in meta-analyses of the Food and Drug Administration data set have ranged from 1.80 to 2.56 points on the widely used Hamilton rating scale for depression (HAM-D) [4], with effect sizes (d) ranging from 0.31 to 0.32 [5–8]. In some studies [6,9], effects varied as a function of baseline severity ranging from d = 0.11 for patients in the mild to moderate range (HAM-D ≤ 18) to 0.47 for patients with very severe depression (HAM-D ≥ 23) [9], although another study as failed to find a severity effect [8].

1. Defining clinical relevance

Until recently there have been no empirically validated criteria for establishing the clinical significance of change scores on scales measuring psychiatric symptoms. In the 2004 National Institute of Health and Clinical Excellence (NICE) guidelines on treating depression, it was suggested that differences of three points on the HAM-D and standardized mean differences of 0.50 might be clinically significant [10], but no evidence was cited to support these proposed cut-offs, and they were criticised as arbitrary [11]. The specification of criteria for clinical relevance was removed from the later edition of the Guidance published in 2009, but effects continued to be classified according to their 'clinical importance,' apparently using the same criteria proposed in the 2004 Guidance [12]. For example, based on a standardized mean difference of 0.34, the 2009 updated NICE guidance concluded that the difference between SSRIs and placebos is "unlikely to be of clinical importance" (p. 317).

Subsequently, an empirical method of establishing the clinical relevance of change scores has been reported in a number of studies [13–16]. The method links scores on various scales used in psychiatric outcome trials to scores on the commonly used Clinical Global Impressions-Improvement (CGI-I) scale, a scale which rates improvement on a scale of 1 (very much improved) through 4 (no change) to 7 (very much worse) [17]. The CGI-I is said to be 'intuitively understood by clinicians' ([16], p 243) and has good inter-rater reliability, between 0.65 and 0.92 [18]. It has been judged to be a useful measure in clinical trials [19] and shown to have concurrent validity with other measures, including CGI severity ratings [20–22]. Spearman correlations ranging

* Corresponding author.
E-mail addresses: j.moncrieff@ucl.ac.uk (J. Moncrieff), irvkirsch@gmail.com (I. Kirsch).

http://dx.doi.org/10.1016/j.cct.2015.05.005

between .70 and .80 have been reported between CGI-I and HAM-D [17]. Thus, this method allows one to align the degree of change on a symptom scale to clinician perception of improvement, and provides a means of establishing an empirically derived criterion for clinical significance. The method has been applied to scales measuring symptoms of schizophrenia [14,15], and more recently to depression scales, specifically the HAM-D. We suggest that a CGI-I rating of 3, which indicates that the patient has "minimally improved" provides the most liberal criterion possible, as the next step on the scale is "no change."

2. The clinical relevance of antidepressant effects

Leucht et al. [16] used the raw data on the antidepressant mirtazapine gathered from 43 trials in more than 7000 people diagnosed with 'major depressive disorder'. The data were derived from placebo-controlled, comparative and open label trials that had been sponsored by the drug company, Organon. The linking analysis of absolute change in Hamilton scores to CGI-improvement scores at four time points is presented in Fig. 1.

Leucht and colleagues described these data as follows: 'The results were consistent for all assessment points examined. A CGI-I score of 4 ("no change") corresponds with a slight reduction on the HAM-D-17 of up to 3 points' ([16], p 245–246). In other words, clinicians could not detect a difference of 3 points on the Hamilton when asked to rate a patient's overall improvement. Examination of the figure reveals that a CGI-I score of 3 ('minimally improved') corresponded to changes in Hamilton score of around 7 points after two to four weeks of treatment. To attain a CGI score of 2 ('much improved'), required a change in Hamilton score of 14 points at the four week assessment.

To date, this method has been used to establish the clinical relevance of pre–post treatment differences. We propose that it can also serve as an empirically validated method of evaluating the clinical significance of drug-placebo differences, since these are also frequently calibrated in terms of differences on the Hamilton scale. Applying this to placebo-controlled antidepressant trials, Leucht et al.'s [16] data reveal that the 3-point difference in HAM-D scores proposed by NICE is overly lenient. It results in classifying a difference that cannot be detected by clinicians as clinically important. These data suggest that a difference of 7-points on the HAM-D might be a more reasonable cut-off, as it corresponds to a clinician rating of minimal improvement.

Leucht and colleagues also reported that the correspondence of HAM-D change scores to clinical ratings varied somewhat as a function of baseline severity. For less severely depressed patients, a clinician rating of minimal improvement corresponded to a 6-point HAM-D difference, whereas for very severely depressed patients, it corresponded to an 8-point change.

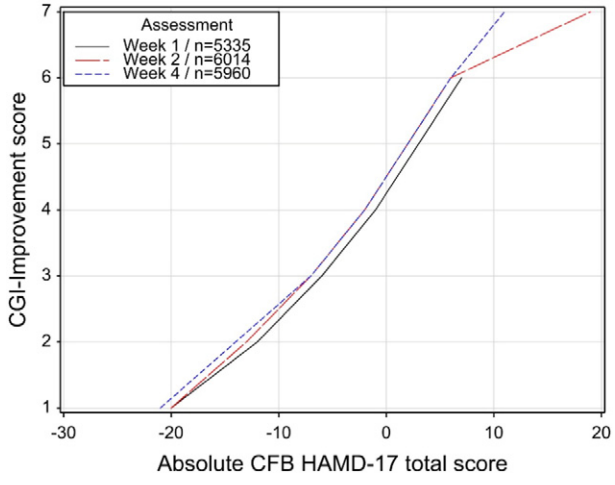

Fig. 1. *Reprinted from J Affect Disord, 148 (2,3), Leucht S, Fennema H, Engel R, Kaspers-Janssen M, Lepping P, Szegedi A. What does the HAMD mean? 243–8, (2013), with permission from Elsevier.

3. Interpreting effect sizes (*d*)

One problem with the cut-offs proposed by NICE (2004) is that a 3 point difference in HAM-D change scores does not correspond well to the effect size of $d = 0.50$ that was proposed to indicate clinical significance. The pooled SD of change scores in the Kirsch et al. meta-analysis (N = 5133) was 8.0 (7.9 for the investigational drug and 8.2 for placebo) [6]. However, that meta-analysis did not include the medication assessed in the Leucht et al. analysis (i.e., mirtazapine). More important, it did not include comparator studies without placebo arms, which were included in the Leucht et al. paper. Thus, it seemed important to assess the reliability of our SD estimate using other data.

A meta-analysis of 5 placebo-controlled mirtazapine trials yielded change score SDs of 7.7 for mirtazapine and 8.3 for placebo [23]. Reported in the same paper, a meta-analysis of 5 trials comparing mirtazapine to amitriptyline yielded SDs of 7.9 and 7.8, respectively. A later comparator trial [24] reported SDs of 7.5 for mirtazapine and 7.7 for paroxetine. These data reveal substantial consistency in the variance of HAM-D change scores across different trial designs, antidepressants, and placebos.

Using an SD of 8.0, the effect size (d) corresponding to a difference score of 7-points (i.e., a clinician rating of minimally improved) is 0.875. For very severely depressed patients, the effect size corresponding to a minimal difference would be 1.00, and for less severely depressed patients it would be a 0.75. These are the effect sizes that are required to indicate a 'minimal' difference as rated by clinicians. They are more than twice the magnitude of the effect sizes derived from meta-analyses, including those examining separately people with the most severe levels of depression [6,7,9].

Conventionally, an effect size of 0.50 is considered 'medium' and 0.80 is considered 'large.' However, Cohen proposed these cut-offs with "invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition" [25] (p 534). The data considered here suggest that with respect to changes on the HAM-D, effect sizes as large as 1.00 may be required to indicate 'minimal' differences as rated by clinicians.

4. Discussion

Over the last few decades antidepressants have become some of the most widely used and profitable drugs in history. Rates of prescriptions have risen throughout the developed world [26], leading to debates about the inappropriate medicalization of misery [27]. The more fundamental question, however, is whether antidepressants achieve worthwhile effects in depression in general. Guidelines have attempted to consider the issue of clinical relevance of antidepressant effects, but have not constructed empirically validated criteria.

The commonly used method of estimating the 'response' to drug treatment in clinical trials of antidepressants (arbitrarily set at a 50% reduction in symptoms), involves the categorisation of continuous data from symptom scales, and therefore does not provide an independent arbiter of clinical significance. Moreover, this method can exaggerate small differences between interventions such as antidepressants and placebo [28], and statisticians note that it can distort data and should be avoided [29,30]. Response rates in double-blind antidepressant trials are typically about 50% in the drug groups and 35% in the placebo groups (e.g., [31,32]). This 15% difference is often defended as clinically significant on the grounds that 15% of depressed people who get better on antidepressants would not have gotten better on placebo. However, a 50% reduction in symptoms is close to the mean and median of drug improvement rates in placebo-controlled antidepressant trials [31–33] and thus near the apex of the distribution curve. Thus, with an SD of 8 in change scores, a 15% difference in response rates is about (an odds

ratio of 1.86, a relative risk of 0.77, and an NNT of 7) is exactly what one would expect from a mean 3-point difference in HAM-D scores [28]. Lack of response does not mean that the patient has not improved; it means that the improvement has been less, by as little as one point, than the arbitrary criterion chosen for defining a therapeutic response.

The small differences detected between antidepressants and placebo may represent drug-induced mental alterations (such as sedation or emotional blunting) or amplified placebo effects rather than specific 'antidepressant' effects [34]. At a minimum, therefore, it is important to ascertain whether differences correlate with clinically detectable and meaningful levels of improvement. The CGI has been criticised for not reflecting the patient's perspective [35], and other data such as functioning and quality of life measures are also required to fully assess the value of antidepressant treatment. Cuijpers et al. [36] have proposed a different method of establishing a 'minimal important difference' (MID) based on 'utility' measures derived from quality of life scales. However, the study from which the MID was estimated did not include samples of depressed individuals, and the values obtained were found to be unstable. As a result, the authors were only able to provide a "very rough estimate of the cutoff for clinical relevance" (p. 376). Use of a patient-rated version of the CGI might allow for a more reliable and valid complement to the clinician-rated data used here to assess the clinical relevance of HAM-D scores. In its absence, CGI improvement scores provide the first empirically validated method for establishing the clinical relevance of antidepressant effects. Based on the Leucht et al. data [16], empirically derived criteria for minimal clinically relevant drug-placebo differences would be, a 7-point difference in HAM-D change scores (8 points for very severely depressed patients), and a drug-placebo effect size (d) of 0.875 (1.00 for very severely depressed patients). Currently, drug effects associated with antidepressants fall far short of these criteria.

This leaves the problem of how to treat depressed patients, given data indicating little if any difference in clinically relevant effects between one treatment and another [33]. Patients and healthcare funders need to be aware that all treatments, including placebo, produce at least a minimal average response to treatment on symptom scales (i.e., improvement of at least 7 points on the HAM-D), while none outperforms a pill placebo to a meaningful degree. We suggest that decisions about treatment should involve the balancing of criteria including patient preference, safety, and cost. Given the choice, most depressed patients prefer psychotherapy over medication [37], and with respect to safety, antidepressant medication would be the last choice between empirically assessed treatment alternatives [38].

Acknowledgements

The authors would like to thank Stefan Leucht for discussion of the original analysis of the mirtazapine data.

There was no funding for the current analysis.

The authors have no competing interests.

References

[1] P. Gotzsche, Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare, Radcliffe Publishing Ltd., London, 2013.
[2] D.O. Antonuccio, W.G. Danton, G.Y. DeNelsky, R.P. Greenberg, J.S. Gordon, Raising questions about antidepressants, Psychother. Psychosom. 68 (1) (1999) 3–14.
[3] J. Moncrieff, I. Kirsch, Efficacy of antidepressants in adults, BMJ 331 (7509) (Jul 16 2005) 155–157.
[4] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. Psychiatry 23 (Feb 1960) 56–62.
[5] N.A. Khin, Y.F. Chen, Y. Yang, P. Yang, T.P. Laughren, Exploratory analyses of efficacy data from major depressive disorder trials submitted to the US Food and Drug Administration in support of new drug applications, J. Clin. Psychiatry 72 (4) (Apr 2011) 464–472.
[6] I. Kirsch, B.J. Deacon, T.B. Huedo-Medina, A. Scoboria, T.J. Moore, B.T. Johnson, Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration, PLoS Med. 5 (2) (Feb 2008) e45.
[7] E.H. Turner, A.M. Matthews, E. Linardatos, R.A. Tell, R. Rosenthal, Selective publication of antidepressant trials and its influence on apparent efficacy, N. Engl. J. Med. 358 (3) (Jan 17 2008) 252–260.
[8] R.D. Gibbons, K. Hur, C.H. Brown, J.M. Davis, J.J. Mann, Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine, Arch. Gen. Psychiatry 69 (6) (Jun 2012) 572–579.
[9] J.C. Fournier, R.J. DeRubeis, S.D. Hollon, S. Dimidjian, J.D. Amsterdam, R.C. Shelton, et al., Antidepressant drug effects and depression severity: a patient-level meta-analysis, JAMA 303 (1) (Jan 6 2010) 47–53.
[10] National Institute for Health and Clinical Excellence, Depression: Management of Depression in Primary and Secondary Care. Clinical Practice Guideline Number 23, National Institute for Clinical Excellence, London, 2004.
[11] E.H. Turner, R. Rosenthal, Efficacy of antidepressants, BMJ 336 (7643) (Mar 8 2008) 516–517.
[12] National Institute for Health and Clinical Excellence, Depression: Management of Depression in Primary and Secondary Care. Revised Clinical Practice Guideline Number 23, National Institute for Health and Clinical Excellence, London, 2009.
[13] S. Leucht, J.M. Kane, W. Kissling, J. Hamann, E. Etschel, R.R. Engel, What does the PANSS mean? Schizophr. Res. 79 (2–3) (Nov 15 2005) 231–238.
[14] S. Leucht, J.M. Kane, E. Etschel, W. Kissling, J. Hamann, R.R. Engel, Linking the PANSS, BPRS, and CGI: clinical implications, Neuropsychopharmacology 31 (10) (Oct 2006) 2318–2325.
[15] P. Lepping, R.S. Sambhi, R. Whittington, S. Lane, R. Poole, Clinical relevance of findings in trials of antipsychotics: systematic review, Br. J. Psychiatry 198 (5) (May 2011) 341–345.
[16] S. Leucht, H. Fennema, R. Engel, M. Kaspers-Janssen, P. Lepping, A. Szegedi, What does the HAMD mean? J. Affect. Disord. 148 (2–3) (Jun 2013) 243–248.
[17] W. Guy, The Clinical Global Impression Scale, ECDEU Assessment Manual for Psychopharmacology — Revised, US Department of Education, Health and Welfare, Rockville, MD, 1976. 218–222.
[18] A. Kadouri, E. Corruble, B. Falissard, The improved Clinical Global Impression Scale (iCGI): development and validation in depression, BMC Psychiatry 7 (2007) 7.
[19] A.C. Leon, M.K. Shear, G.L. Klerman, L. Portera, J.F. Rosenbaum, I. Goldenberg, A comparison of symptom determinants of patient and clinician global ratings in patients with panic disorder and depression, J. Clin. Psychopharmacol. 13 (5) (Oct 1993) 327–331.
[20] D.W. Hedges, B.L. Brown, D.A. Shwalb, A direct comparison of effect sizes from the clinical global impression-improvement scale to effect sizes from other rating scales in controlled trials of adult social anxiety disorder, Hum. Psychopharmacol. 24 (1) (Jan 2009) 35–40.
[21] A. Khan, R.M. Leventhal, S.R. Khan, W.A. Brown, Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database, J. Clin. Psychopharmacol. 22 (1) (Feb 2002) 40–45.
[22] M. Berk, F. Ng, S. Dodd, T. Callaly, S. Campbell, M. Bernardo, et al., The validity of the CGI severity and improvement scales as measures of clinical effectiveness suitable for routine clinical use, J. Eval. Clin. Pract. 14 (6) (Dec 2008) 979–983.
[23] S. Kasper, Clinical efficacy of mirtazapine: a review of meta-analyses of pooled data, Int. Clin. Psychopharmacol. 10 (Suppl. 4) (Dec 1995) 25–35.
[24] A. Wade, G.M. Crawford, M. Angus, R. Wilson, L. Hamilton, A randomized, double-blind, 24-week study comparing the efficacy and tolerability of mirtazapine and paroxetine in depressed patients in primary care, Int. Clin. Psychopharmacol. 18 (3) (May 2003) 133–141.
[25] J. Cohen, Statistical Power Analysis for the Behavioural Sciences, second edition Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.
[26] OECD, Health at a Glance 2013: OECD indicators, OECD Publishing, Paris, 2013. ([updated 10-10-2014]; Available from: http://www.oecd.org/els/health-systems/Health-at-a-Glance-2013.pdf?_ga=1.159836794.650866776.1405853144).
[27] P. Gotzsche, Psychiatric drugs are doing us more harm than good, The Guardian2014.
[28] I. Kirsch, J. Moncrieff, Clinical trials and the response rate illusion, Contemp. Clin. Trials 28 (2007) 348–351.
[29] P. Royston, D.G. Altman, W. Sauerbrei, Dichotomizing continuous predictors in multiple regression: a bad idea, Stat. Med. 25 (1) (Jan 15 2006) 127–141.
[30] R.C. MacCallum, S. Zhang, K.J. Preacher, D.D. Rucker, On the practice of dichotomization of quantitative variables, Psychol. Methods 7 (1) (Mar 2002) 19–40.
[31] B.R. Rutherford, J.R. Sneed, S.P. Roose, Does study design influence outcome? The effects of placebo control and treatment duration in antidepressant trials, Psychother. Psychosom. 78 (3) (2009) 172–181.
[32] M. Sinyor, A.J. Levitt, A.H. Cheung, A. Schaffer, A. Kiss, Y. Dowlati, et al., Does inclusion of a placebo arm influence response to active antidepressant treatment in randomized controlled trials? Results from pooled and meta-analyses, J. Clin. Psychiatry 71 (3) (Mar 2010) 270–279.
[33] A. Khan, J. Faucett, P. Lichtenberg, I. Kirsch, W.A. Brown, A systematic review of comparative efficacy of treatments and controls for depression, PLoS One 7 (7) (2012) e41778.
[34] J. Moncrieff, D. Cohen, Rethinking models of psychotropic drug action, Psychother. Psychosom. 74 (3) (2005) 145–153.
[35] T. Forkmann, A. Scherer, M. Boecker, M. Pawelzik, R. Jostes, S. Gauggel, The Clinical Global Impression Scale and the influence of patient or staff perspective on outcome, BMC Psychiatry 11 (2011) 83.
[36] P. Cuijpers, E.H. Turner, S.L. Koole, A. van Dijke, F. Smit, What is the threshold for a clinically relevant effect? The case of major depressive disorders, Depress. Anxiety 31 (5) (May 2014) 374–378.
[37] R.K. McHugh, S.W. Whitton, A.D. Peckham, J.A. Welge, M.W. Otto, Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review, J. Clin. Psychiatry 74 (6) (Jun 2013) 595–602.
[38] P.W. Andrews, J.A. Thomson Jr., A. Amstadter, M.C. Neale, Primum non nocere: an evolutionary analysis of whether antidepressants do more harm than good, Front. Psychol. 3 (2012) 117.