International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 23(4): 469–473 (2014)
Published online 4 September 2013 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1393

The fallacy of thresholds used in defining response and remission in depression rating scales

FLORIAN NAUDET,[1,2,3] BRUNO MILLET,[2,4] JEAN MICHEL REYMANN[3,5] & BRUNO FALISSARD[1,6,7]

1 INSERM U669, Paris, France
2 Université de Rennes 1, EA-425 Unité Comportement et Noyaux Gris Centraux, Rennes, France
3 Centre d'Investigation Clinique CIC-P INSERM 0203, Hôpital de Pontchaillou, Centre Hospitalier Universitaire de Rennes & Université de Rennes 1, Rennes, France
4 Centre Hospitalier Guillaume Régnier, Service Hospitalo-Universitaire de Psychiatrie, Rennes, France
5 Laboratoire de Pharmacologie Expérimentale et Clinique, Faculté de Médecine, CS34317, 2 avenue du Pr Léon Bernard, 35043 Rennes, France
6 Université Paris-Sud and Université Paris Descartes, UMR-S0669, Paris, France
7 AP-HP, Hôpital Paul Brousse, Département de santé publique, Villejuif, France

Key words
response, remission, major depressive disorder, threshold, misclassification bias

Correspondence
Florian Naudet, INSERM U669, Maison de Solenn, 97 Boulevard de Port Royal, 75679 Paris cedex 14, France.
Telephone (+33) 1 58 41 28 50
Fax (+33) 1 58 41 28 43
Email: floriannaudet@gmail.com

Received 3 September 2012;
revised 5 February 2013;
accepted 28 March 2013

Abstract

Response and remission are determined by the proportion of people who fall below a threshold score on depression rating scales. This calculation implies a possibility of false positives (FP) and false negatives (FN) depending on sensitivity and specificity of the threshold used, but also on response and remission rates. A simulation illustrates the methodological consequences of this phenomenon in a comparative trial where response and remission rates differ between groups: the probability of being misclassified differs between groups, and measures of association (relative risk and odds ratio) are biased. Alternatives are proposed to cope with this misclassification bias. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

In antidepressant efficacy trials, results are usually reported as the score on a symptom rating scale. Nevertheless, the score is not always meaningful because the identification of a minimal clinically relevant difference is not straightforward (Falissard *et al.*, 2003).

Clinicians, in their day-to-day practice, are used to dealing with binary outcomes like response and remission, which have considerable prognostic value (Judd *et al.*, 1998). Although these concepts appear intuitive, they are usually determined by the proportion of people who fall below predefined threshold scores, which are rather validated by convention and tradition (Mulder *et al.*, 2003).

Since 1991 (Frank *et al.*, 1991) a consensus has emerged to define remission as a score ≤ 7 on the 17 items of the Hamilton Depression Rating Scale (HDRS-17). Response is usually defined as a reduction of 50% on the HDRS-17.

Nevertheless, imposing categories on continuous data creates the impression of clear-cut patterns, while the data does not suggest any (patients just over the cutoff score are often clinically indistinguishable from those just under the cutoff):

(1) It can inflate the differences derived from continuous outcomes between groups in clinical trials. This phenomena is interpreted as a major bias (Kirsch and Moncrieff, 2007) or as proof of antidepressant effectiveness (Gibbons *et al.*, 2012), depending on the stances of the different authors.
(2) It generates outcomes that partly differ from clinicians' representations implying the possibility of "false positives" and "false negatives".

Concerning this second point, misclassification can be described by positive predictive value (PPV, proportion of positive test results that are true positives) and negative predictive value (NPV, proportion of negative test results that are true negatives) which are two critical measures of the performance of a diagnostic method. Their values do however rely on sensitivity (Se) and specificity (Spe) of the diagnosis method, but also on the prevalence (*p*) of the outcome studied (Loong, 2003).

In a randomized controlled trial (RCT) comparing an antidepressant against placebo, one can expect higher response and remission rates in the active arm than in the placebo arm; as Se and Spe in the classification methods are constants (these intrinsic properties are stable in a given population), PPV and NPV will automatically differ between groups. This corresponds to a misclassification bias between the two groups.

The objective was here to evaluate the extent of this bias, and its impacts on measures of association [Relative Risk (RR) and Odds Ratio (OR)] using a simple numerical example.

Methods

The literature was searched for reports on Se and Spe of the HDRS-17 response and remission algorithm as compared to the response and remission classification derived from two instruments with high face validity:

(1) the Clinical Global Impression (CGI) items 1 ("very much improved") and 2 ("much improved") for response (Furukawa *et al.*, 2007);
(2) the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) definition for remission.

Eligible studies were identified from Pubmed/Medline. The search terms used were "Hamilton AND 'Depressive Disorder'[Mesh] AND ('Psychiatric Status Rating Scales'[Mesh] OR 'Psychometrics'[Mesh] OR 'Questionnaires'[Mesh]) AND ('Sensitivity and Specificity'[Mesh] OR 'Endpoint Determination'[Mesh])". Only studies using the English version of the scale were retained.

Then we considered a hypothetical RCT of antidepressant (atd) versus placebo (pbo) with an arbitrary sample size of 100 patients per arm, with the following CGI response rates (respectively 50% and 20%, $RR_{CGI} = 2.5$, $OR_{CGI} = 4.0$) and DSM-IV remission rates (respectively 35% and 15%, $RR_{DSM} = 2.4$, $OR_{DSM} = 3.1$). These values are classic in RCTs in major depressive disorder (MDD). They are coherent with a number needed to treat of three and five.

Table 1 presents the 2 × 2 table used for the calculations. PPV, NPV and measures of association (RR_{HDRS} and OR_{HDRS}) that would be obtained if a threshold was applied for these different rates of response and remission using Equations 1, 2, 3 and 4 (Bland and Altman, 2000; Loong, 2003).

$$PPV = \frac{Se \times p}{Se \times p + (1 - Sp) \times (1 - p)} \tag{1}$$

$$NPV = \frac{Sp \times (1 - p)}{(1 - Se) \times p + Sp \times (1 - p)} \tag{2}$$

$$RR_{HDRS} = \frac{(TP_{Atd} + FP_{Atd})/(TP_{Atd} + FP_{Atd} + FN_{Atd} + TN_{Atd})}{(TP_{Pbo} + FP_{Pbo})/(TP_{Pbo} + FP_{Pbo} + FN_{Pbo} + TN_{Pbo})} \tag{3}$$

$$OR_{HDRS} = \frac{(TP_{Atd} + FP_{Atd})/(FN_{Atd} + TN_{Atd})}{(TP_{Pbo} + FP_{Pbo})/(FN_{Pbo} + TN_{Pbo})} \tag{4}$$

Results

The search of Pubmed/Medline provided a total of 123 citations. Of these, 117 were discarded because, after review of the abstracts, it appeared that these papers did not meet the criteria. Four studies were excluded after careful consideration: one used a Spanish version of the HDRS, two (Dunlop *et al.*, 2011; Furukawa *et al.*, 2007) did not report values of Se and Spe. Since two papers (Zimmerman *et al.*, 2004, 2005) were reports of the "Rhode Island Methods to

Table 1. A 2 × 2 table presenting equations used for the calculations and a numerical example corresponding to the use of reduction of 50% on the HDRS-17 (threshold) to diagnose responders as detected by the Clinical Global Impression (CGI) (considered as the gold standard) in antidepressant and placebo arms of a hypothetical study

	Gold standard +	Gold standard −	Total
< Threshold	**TP = Se × p × N**	**FP = (1 − Spe) × (1 − p) × N**	**TP + FP**
Antidepressant	48	14.5	62.5
Placebo	19.2	23.2	42.4
> Threshold	**FN = (1 − Se) × p × N**	**TN = Spe × (1 − p) × N**	**FN + TN**
Antidepressant	2	35.5	37.5
Placebo	0.8	56.8	57.6
Total	**TP + FN = p × N**	**FP + TN = (1 − p) × N**	**N = TP + FP + FN + TN**
Antidepressant	50	50	100
Placebo	20	80	100

Note: TP = true positives; TN = true negatives; FN = false negatives; FP = false positives; N = number of patients in the arm; Se = sensitivity; Spe = specificity; p = prevalence of the outcome studied according to the gold standard used.
RR_{HDRS} = (62.5/100)/(42.4/100) = 1.5.
OR_{HDRS} = (62.5/37.5)/(42.4/57.6) = 2.3.

Improve Diagnostic Assessment and Services" project, only one (Zimmerman *et al.*, 2005) was retained. Another paper also met the criteria (Mulder *et al.*, 2003). Sensitivity and specificity for a reduction of 50% on the HDRS-17 to diagnose responders as detected by the CGI were estimated respectively 96% and 71% (Mulder *et al.*, 2003). Sensitivity and specificity for an HDRS-17 score ≤ 7 to diagnose remission as detected by the DSM-IV were estimated respectively 100% and 74% (Zimmerman *et al.*, 2005).

Details of calculation for response rates in the two groups of our hypothetical study are presented in Table 1.

For a response rate of 50% on the CGI, PPV and the NPV are respectively 77% and 95% whereas they are 46% and 99% for a response rate of 20% (RR_{HDRS} = 1.5, OR_{HDRS} = 2.3).

For a remission rate of 35%, the PPV and the NPV are respectively 67% and 100% whereas they are 40% and 100% for a remission rate of 15% (RR_{HDRS} = 1.3, OR_{HDRS} = 1.8).

Discussion

Main findings

This example shows that the use of thresholds on a scale generates non-trivial rates of false positives. Moreover, the probability of being misclassified differs substantially between groups in a comparative trial: the lower the remission (or response) rate, the lower the PPV and the higher the NPV. This is called a non-differential misclassification bias, since Se and Spe are assumed to be the same between groups.

It makes false positives more frequent in the less effective treatment group than in the other: the two groups appear more alike. It leads to an underestimation of the difference (i.e. RR and OR will be biased toward one) (Hofler, 2005; Mertens, 1993) as illustrated by the numerical example.

In comparative effectiveness research, this results in a loss of statistical power. In RCTs versus placebo, this bias should be taken into account when interpreting high rates of responders in the placebo group. With the same rationale, patients considered as responders (or remitters) in studies on resistant depression (where the outcome considered is rare) fit an intuitive definition of response (or remission) less well than responders (or remitters) in studies on non-resistant depression (where the outcome considered is frequent).

Limitations

In the present study, the reference instruments (CGI and DSM-IV) were chosen because of their high face validity suggesting a good congruence with clinicians' representations. Since the meaning of these instruments was clear from a semantic point of view, they are usually considered as gold standard in validation studies and are used to explore the question of the minimal clinically relevant difference in anchor-based approaches. But in theory, their definitions of response and remission also generate false positives or false negatives, where no real gold standard exists. Moreover, these instruments are far from ideal. Reliability coefficients in the 0.6 range have been reported for the DSM-IV MDD and the CGI could be ambiguous and prone to cultural misunderstanding (Kadouri *et al.*, 2007).

Nevertheless, it is possible to surmise that if a gold standard did exist, it would not strictly overlap a definition derived from depression rating scales, so that the bias would behave in the same way.

Perspectives

When reading a clinical trial, one should remember that if the terms response and remission generate intuitive representations, the reality behind these concepts can vary from one group to another. In fact, the main bias is more semantic than methodological. It concerns the interpretation of the data (i.e. representations derived from the data): the terms of remission and response are misleading; in a RCT against placebo – patients classified as responders in the placebo group may fit an intuitive definition of response less well than patients assigned to antidepressants.

Caution can be recommended towards the erroneous representations that this dichotomization can produce. Here are a few simple recommendations for clinical research:

(1) consider differences in mean scores between groups as the reference to assess effectiveness;
(2) present different ways of assessing response and remission as sensitivity analyses;
(3) develop alternative methods for assessing qualitative outcomes.

To address this last point, qualitative or mixed (qualitative/quantitative) methods should be developed enabling the measure of an “impression” related to the “essence” of the patient, two major concepts in Phenomenology.

Acknowledgements

This paper was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM). The authors thank Bertrand Prioux for his helpful comments, Claudine Naudet and Angela Swaine Verdier for revising the English.

Author’s contributions: conceived and designed the experiments: N.F.; performed the experiments: N.F.; analysed the data: N.F.; contributed reagents/materials/analysis tools: N. F.; wrote the paper: N.F.; revised the paper critically for important intellectual content: M.B., J.M.R., F.B.; final approval of the version to be published: N.F., M.B., J.M.R., F.B.

Declaration of interest statement

There are no conflicts of interest regarding this paper. All authors have completed the Unified Competing Interest form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that: (1) all authors have no support at all from any company for the submitted work; (2) N.F. has relationships (board membership or travel/accommodations expenses covered/reimbursed) with Servier, BMS, Lundbeck and Janssen who might have an interest in the work submitted in the previous three years; M.B has relationship (consultancy and travel/accommodations expenses covered/reimbursed) with Janssen, BMS, Otsuka, Lundbeck, Lilly, Servier, Astra Zeneca, Medtronics, Syneïka and has received grants for research from Medtronic, Lilly and Astra Zeneca in the previous three years; F.B has relationship (board membership or consultancy or payment for manuscript preparation or travel/accommodations expenses covered/reimbursed) with Sanofi-Aventis, Servier, Pierre-Fabre, MSD, Lilly, Janssen-Cilag, Otsuka, Lundbeck, Genzime, Roche, BMS who might have an interest in the work submitted in the previous three years; (3) N.F., F.B., J.M.R., spouses, partners, or children have no financial relationships that may be relevant to the submitted work. M.B.’s spouse is an employee of Janssen; none of the authors have any non-financial interests that may be relevant to the submitted work.

References

Arroll B., Elley C.R., Fishman T., Goodyear-Smith F.A., Kenealy T., Blashki G., Kerse N., Macgillivray S. (2009) Antidepressants versus Placebo for Depression in Primary Care, Cochrane Database System Review CD007954, Chichester, John Wiley & Sons.

Bland J.M., Altman D.G. (2000) Statistics notes. The odds ratio. *British Medical Journal*, **320**, 1468.

Dunlop B.W., Li T., Kornstein S.G., Friedman E.S., Rothschild A.J., Pedersen R., Ninan P., Keller M., Trivedi M.H. (2011) Concordance between clinician and patient ratings as predictors of response, remission, and recurrence in major depressive disorder. *Journal of Psychiatric Research*, **45**, 96–103.

Falissard B., Lukasiewicz M., Corruble E. (2003) The MDP75: a new approach in the determination of the minimal clinically meaningful difference in a scale or a questionnaire. *Journal of Clinical Epidemiology*, **56**, 618–621.

Frank E., Prien R.F., Jarrett R.B., Keller M.B., Kupfer D.J., Lavori P.W., Rush A.J., Weissman M.M. (1991) Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, **48**, 851–855.

Furukawa T.A., Akechi T., Azuma H., Okuyama T., Higuchi T. (2007) Evidence-based guidelines for interpretation of the Hamilton Rating Scale for Depression. *Journal of Clinical Psychopharmacology*, **27**, 531–534.

Gibbons R.D., Hur K., Brown C.H., Davis J.M., Mann J.J. (2012) Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Archives of General Psychiatry*, **69**(6), 572–579.

Hofler M. (2005) The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research*, **14**, 92–101.

Judd L.L., Akiskal H.S., Maser J.D., Zeller P.J., Endicott J., Coryell W., Paulus M.P., Kunovac J.L., Leon A.C., Mueller T.I., Rice J.A., Keller M.B. (1998) Major depressive disorder: a prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. *Journal of Affective Disorders*, **50**, 97–108.

Kadouri A., Corruble E., Falissard B. (2007) The improved Clinical Global Impression Scale (iCGI): development and validation in depression. *BMC Psychiatry*, **7**, 7.

Kirsch I., Moncrieff J. (2007) Clinical trials and the response rate illusion. *Contempory Clinical Trials*, **28**, 348–351.

Loong T.W. (2003) Understanding sensitivity and specificity with the right side of the brain. *British Medical Journal*, **327**, 716–719.

Mertens T.E. (1993) Estimating the effects of misclassification. *Lancet* **342**, 418–421.

Mulder R.T., Joyce P.R., Frampton C. (2003) Relationships among measures of treatment outcome in depressed patients. *Journal of Affective Disorders*, **76**, 127–135.

Zimmerman M., Posternak M.A., Chelminski I. (2004) Implications of using different cut-offs on symptom severity scales to define remission from depression. *International Clinical Psychopharmacology*, **19**, 215–220.

Zimmerman M., Posternak M.A., Chelminski I. (2005) Is the cutoff to define remission on the Hamilton Rating Scale for Depression too high? *Journal of Nervous Mental Diseases*, **193**, 170–175.