Published in final edited form as:
Depress Anxiety. 2012 December ; 29(12): 1043–1049. doi:10.1002/da.21993.

SELF-REPORT AND CLINICIAN-RATED MEASURES OF DEPRESSION SEVERITY: CAN ONE REPLACE THE OTHER?

Rudolf Uher, M.D., Ph.D.[1,2,*], **Roy H. Perlis, M.D.**[3], **Anna Placentino, Psy.D.**[4,5], **Mojca Zvezdana Dernovšek, M.D.**[6], **Neven Henigsberg, M.D.**[7], **Ole Mors, M.D., Ph.D.**[8], **Wolfgang Maier, M.D.**[9], **Peter McGuffin, F.R.C.P., F.R.C.Psych., Ph.D.**[2], and **Anne Farmer, M.D., F.R.C.Psych.**[2]

[1]Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada

[2]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, United Kingdom

[3]Center for Experimental Drugs and Diagnostics, Department of Psychiatry and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts

[4]Psychiatric Unit (UOP 23), Department of Mental Health, Spedali Civili Hospital of Brescia, Lombardei, Italy

[5]Biological Psychiatry Unit, IRCCS-FBF, Brescia; Faculty of Psychology, University of Milano-Bicocca, Italy

[6]University Psychiatric Clinic, Ljubljana, Slovenia

[7]Croatian Institute for Brain Research, Medical School University of Zagreb, Croatia

[8]Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Denmark

[9]Department of Psychiatry, University of Bonn, Germany

Abstract

Background—It has been suggested that clinician-rated scales and self-report questionnaires may be interchangeable in the measurement of depression severity, but it has not been tested whether clinically significant information is lost when assessment is restricted to either clinician-rated or self-report instruments. The aim of this study is to test whether self-report provides information relevant to short-term treatment outcomes that is not captured by clinician-rating and vice versa.

Methods—In genome-based drugs for depression (GENDEP), 811 patients with major depressive disorder treated with escitalopram or nortriptyline were assessed with the clinician-rated Montgomery–Åsberg Depression Rating Scale (MADRS), Hamilton Rating Scale for Depression (HRSD), and the self-report Beck Depression Inventory (BDI). In sequenced treatment alternatives to relieve depression (STAR*D), 4,041 patients treated with citalopram were assessed with the clinician-rated and self-report versions of the Quick Inventory of Depressive Symptomatology (QIDS-C and QIDS-SR) in addition to HRSD.

*Correspondence to: Rudolf Uher, Department of Psychiatry, Dalhousie University, Mood Disorders Program, Abbie J. Lane Building, Room 3089, 5909 Veterans' Memorial Lane, Halifax, Nova Scotia B3H 2E2, Canada. rudolf.uher@kcl.ac.uk.

Conflict of interest. Uher consults for the World Health Organization. Perlis has received consulting fees from Proteus Biomedical, Concordant Rater Systems, and RIDventures. Henigsberg has received honoraria for participating in expert panels from pharmaceutical companies including Lundbeck. Placentino, Dernovsek, Mors, Maier; McGuffin and Farmer have no conflicts of interest.

Results—In GENDEP, baseline BDI significantly predicted outcome on MADRS/HRSD after adjusting for baseline MADRS/HRSD, explaining additional 3 to 4% of variation in the clinician-rated outcomes (both P < .001). Likewise, each clinician-rated scale significantly predicted outcome on BDI after adjusting for baseline BDI and explained additional 1% of variance in the self-reported outcome (both P < .001). The results were confirmed in STAR*D, where self-report and clinician-rated versions of the same instrument each uniquely contributed to the prediction of treatment outcome.

Conclusion—Complete assessment of depression should include both clinician-rated scales and self-reported measures.

Keywords

depression; *assessment/diagnosis*; *clinical trials*; *antidepressants*; *treatment*; *mood disorders*

INTRODUCTION

Current guidelines for major depressive disorder (MDD) recommend measurement-based care with routine assessment of depression severity guiding treatment options.[1–4] The implementation of these recommendations requires an informed debate on *how* depression severity should be measured.[5,6] Potentially, the most important distinction is between clinician rating and self-report questionnaires.[7]

In most clinical trials, especially those of pharmacotherapy, depression severity has been assessed by trained clinicians using depression-rating scales, such as the Hamilton Rating Scale for Depression (HRSD) or the Montgomery–Åsberg Depression Rating Scale (MADRS). Current treatment guidelines and antidepressants drug licenses are largely based on treatment efficacy measured with these scales. However, use of clinician-rated scales in routine clinical practice is costly and puts additional requirements on clinicians' training and consultation times. It has therefore been suggested that cheaper self-report instruments may replace clinician-rating scales in routine practice.[8–10] Self-report instruments have a long tradition, especially in psychotherapy research, with the Beck Depression Inventory (BDI) being the most widely used questionnaire.[11–13] However, the BDI differs from clinician-rated scales not just in the mode of administration, but also in terms of what symptoms are assessed. This has prompted the development of parallel self-report and clinician-rated scales with matching content to facilitate the translation of evidence between research studies and clinical practice.[9,14,15]

The moderate-to-strong correlations between clinician-rated scales and self-report questionnaires suggest that the two modes of measuring depression may indeed be interchangeable.[9] However, it has not been tested whether clinically significant information is lost when assessment is restricted to either clinician rating or self-report. The agreement between self-reported and clinician-rated measures of depression severity is far from perfect.[6,16–23] Although some of the differences may be due to variation in scale content, significant discrepancies between self-report and clinician-rated versions of the same scale suggest that other factors play a role.[9,24,25] The most important question is whether the information that is uniquely captured by self-report or by clinician rating is clinically relevant. It is possible that the discrepancy between self-report and clinician rating is due to measurement error and therefore is inconsequential. The alternative possibility is that clinician rating obtains unique information that is not accessible by self-report or vice versa. Differences in treatment effect sizes[26,27] and systematic relationship of self–clinician discrepancies with personality factors[25,28] and with treatment outcome[29] suggest that the latter may be the case.

The present study aims to test the hypotheses that clinician-rated scales provide clinically relevant information that is not captured by self-report questionnaires, and that self-report questionnaires obtain clinically important information that is not accessible through clinician-rating. Since the outcome of treatment in terms of symptom reduction and remission is the most clinically meaningful validator, clinical relevance is assessed as prediction of outcome of up to 12 weeks of treatment with antidepressants in two large clinical trials. One of these studies includes the three most widely used scales: the HRSD, MADRS, and BDI. The other study included clinician-rated and self-report versions of the same scale, thus offering an opportunity to distinguish the effect of rater from scale content.

METHODS

We tested whether self-report and clinician-rated scales uniquely predict outcomes in two large clinical trials of antidepressant treatment for MDD, in which participants were assessed with multiple self-report and clinician-rated measures of depression severity at pretreatment baseline, during and after treatment with antidepressant drugs. Jointly, the two studies allowed the exploration and comparison of the commonly used outcome scales.

CLINICIAN-RATED MEASURES OF DEPRESSION SEVERITY

The 17-item *Hamilton Rating Scale for Depression (HRSD)* is a clinician-rated scale designed to measure the severity of illness in patients diagnosed with a depressive disorder.[30] The 17 items assess mood, guilt, suicidal thoughts, early insomnia, middle insomnia, late insomnia, activity, psychomotor retardation, agitation, psychic anxiety, somatic anxiety, appetite, fatigue, libido, hypochondriasis, weight loss, and insight. Nine items are scored on a 5-point (0 to 4) ordinal scale and eight items are scored on a 3-point (0, 1, 2) scale. A total score is calculated as sum of the 17 items and can range from 0 to 52. Higher scores reflect more severe depression. The relatively large number of items assessing sleep, appetite, weight loss, libido, and fatigue mean that somatic and neurovegetative symptoms contribute disproportionately to the total score.[20,22] HRSD is the most widely used depression rating scale. It has overall acceptable reliability, although several of its items are unreliable and add little to the measurement of depression severity.[31,32]

The *Montgomery–Åsberg Depression Rating Scale (MADRS)* is a 10-item clinician-rated scale assessing symptoms of depression that were selected to be responsive to treatment.[33] Sad mood is assessed by two items that capture the observer perspective and reported subjective experience, respectively. The other eight items assess tension, sleep, appetite, concentration, lassitude (activity), inability to feel (anhedonia), pessimism, and suicidal thoughts. Each item is rated on a 7-point (0 to 6) ordinal scale. A total score is computed as the sum of the 10 items and can range from 0 to 60. Higher scores reflect more severe depression. MADRS has been found to be internally consistent and to discriminate levels of depression severity more accurately than HRSD and other rating scales.[20,34,35]

The 16-item Quick Inventory of Depressive Symptomatology (QIDS) assesses the nine symptoms diagnostic of depressive episode according to the Diagnostic and Statistical Manual, fourth edition (DSM-IV)[36] sad mood, interest/anhedonia, energy/fatigue, concentration, outlook on self, suicidal thoughts, psychomotor retardation/agitation, sleep, and appetite. The clinician-rated version (QIDS-C) is rated by a trained clinical interviewer. The 16 items are scores on 4-point (0 to 3) ordinal scale. The QIDS total score includes only the highest scored item among the four items assessing sleep, the highest scored item of the four items assessing appetite and weight change, and the highest scored item of the two items assessing psychomotor retardation and agitation. As a result, the total QIDS score is calculated as a sum of nine items, one for each DSM-IV symptom, and ranges from 0 to 27.

Higher scores describe more severe depression. QIDS-C has good psychometric properties in clinical populations and concurrent validity with established depression rating scales.[9,37]

SELF-REPORT MEASURES OF DEPRESSION SEVERITY

The 21-item *Beck Depression Inventory (BDI)* was developed by J. Erbaugh based on records of statements made by individuals with depressive disorders during psychotherapeutic sessions.[11] Its 21-items assess all DSM-IV diagnostic symptoms of depression and additional symptoms (e.g. irritability). A large proportion of BDI items focus on the cognitive symptoms of depression, such as self-esteem, guilt, feeling disappointed in oneself, feeling of being punished, and pessimism. As a result, cognitive symptoms of depression contribute disproportionately to the BDI score.[20] Each item is composed of four first-person statements graded by the degree of depression severity it typically represents and rated on a 4-point ordinal scale (0 to 3). BDI was originally designed to be read out to the patient by an interviewer, but has commonly been used as self-report questionnaire for literate patients. The total BDI score is calculated by summing the 21 items and can range from 0 to 63. BDI has good psychometric properties with acceptable internal consistency and moderate concurrent validity.[15,20,23,38]

The Quick Inventory of Depressive Symptomatology Self-Report (QIDS-SR) is the self-report version of QIDS-C with identical item content and scoring. It assesses the nine diagnostic symptoms of depression according to DSM-IV, has a total scores range from 0 to 27 (higher scores indicate more severe depression) and good psychometric properties.[9,39]

TREATMENT STUDIES AND SAMPLES

The *Genome-based Drugs for Depression (GENDEP) study* is a European multicenter open-label part-randomized clinical and pharmacogenetic study. GENDEP recruited 811 treatment-seeking men and women with MDD diagnosed in a semistructured interview.[40] They were treated with protocol-guided flexible doses of escitalopram (10 to 30 mg daily) or nortriptyline (50 to 200 mg daily) for 12 weeks. Participants were assessed at pretreatment baseline and each treatment week with two clinician rated scales and one self-report questionnaire.[20,41] MADRS and HRSD were administered by psychiatrists and psychologists trained to achieve high interrater reliability.[20] BDI was completed by the participants. Participants treated by the two antidepressants achieved a similar level of improvement on the three outcome measures.[41] The GENDEP sample is described in Table 1 and further details are available in previous publications.[41–44] GENDEP was approved by ethics boards in all centers. All participants provided a written informed consent. GENDEP is registered at EudraCT (No.2004-001723-38, http://eudract.emea.europa.eu) and ISRCTN (No. 03693000, http://www.controlled-trials.com).

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study

—The primary purpose of STAR*D was to determine which treatments work best if the first antidepressant treatment does not produce remission. STAR*D included 4,041 treatment-seeking adult outpatients with DSM-IV nonpsychotic major depression, recruited in 31 centers in the United States. The treatment included protocol-guided citalopram 20 to 60 mg daily.[45] The assessment of depression severity comes from three sources: (1) the participants completed QIDS-SR at every treatment visit (baseline and follow-up every 2 weeks); (2) the treating clinician administered the QIDS-C at every treatment visit (baseline and follow-up visits every 2 weeks) based on a face-to-face interview; (3) an independent research outcome assessor (ROA) administered HRSD in telephone interviews at baseline and at study/level exit (end of treatment). This study uses 3,637 subjects with at least one postbaseline measurement during citalopram treatment (level 1) on any of the three outcome measures from the limited access data set (version no. 2) distributed by the National

Institutes of Health (NIH). The numbers of subjects available for analyses differ depending on which outcome measure was used: valid postbaseline HRSD was available for 2,796 (69.2%), QIDS-C for 3,630 (89.8%), and QIDS-SR for 3,607 (89.3%) subjects. The STAR*D sample is described in Table 1 and further details are available elsewhere.[45,46] The study was approved by institutional ethics review boards in participating centers. All participants provided a written consent after the procedures were explained. STAR*D is registered at ClinicalTrials.gov (NCT00021528).

STATISTICAL ANALYSIS

In each study, we used three different depression severity measures as predictors and outcomes. In GENDEP, we used clinician-rated MADRS and HRSD and self-report BDI. In STAR*D, we used clinician rated QIDS-C, self-report QIDS-SR, and ROA-rated HRSD. To use all available information and account for missing data and repeated measurements within individual, we used mixed-effect linear regression for hypothesis testing.[41,47,48] To test the hypothesis that one type of measurement contributes *unique* information to the prognosis that is not contained in the other type of measurement, we always corrected for the baseline score on the measure that was used as the outcome. For example, when we tested whether baseline BDI predicted outcome as measured on the MADRS, we included baseline MADRS score as a covariate. Consequently, all reported results reflect the *unique* contribution of one scale over and above the baseline measure of depression severity on another scale (the one used as the outcome measure). In addition, all regression models corrected for age, sex, and center of recruitment. To correct for the number of independent tests (six in each study), we consider findings with a $P < .0083$ (.05/6) to be statistically significant.

To quantify the strength of each prediction, we also calculated percentage variance explained as the r^2 in a simple linear regression using the residual of the exit score (or its best unbiased linear estimate in case of missing values) adjusted for baseline score on the same depression measure, age, sex, and center of recruitment as the outcome (dependent) variable and the score on another measure of depression as the only predictor.

RESULTS

RELATIONSHIP BETWEEN RATING SCALES AT BASELINE

In GENDEP, the three scales were strongly correlated at baseline and on study exit (Table 2). At 0.77 (baseline) and 0.94 (exit), the correlation between the two clinician-rated scales was stronger than the correlation of either with the self-report BDI.

Likewise, in STAR*D, the three depression rating scales were strongly correlated at baseline and on study exit (Table 3). At 0.69 (baseline) and 0.90 (exit), the correlation between the two versions of QIDS was stronger than the correlations of either with the ROA-rated HRSD.

PREDICTION OF OUTCOMES IN GENDEP

The results of linear mixed-effect model prediction in the GENDEP sample are given in Table 4. Each model includes the baseline score on the outcome measure as a covariate, so that only the unique contribution of the predictor is tested.

Baseline BDI significantly and strongly predicted outcome on MADRS even after adjusting for baseline MADRS score and explained an additional 3% of variation in the clinician-rated outcome (Table 4). Baseline BDI also strongly predicted outcome on HRSD even after adjusting for baseline HRSD score and explained an additional 4% of variation in the

clinician-rated outcome (Table 4). MADRS and HRSD also predicted outcomes on the other clinician-rated scale, albeit with significantly smaller effect size (both estimates below the 95% confidence interval of the prediction by BDI, Table 4).

Both MADRS and HRSD significantly predicted outcome on the BDI after adjusting for the baseline BDI score and explained an additional 1% of variation in the self-reported outcome over and above the baseline BDI score (Table 4).

Overall, the results were similar for the two treatment arms and there was no significant predictor-by-drug interaction.

PREDICTION OF OUTCOMES IN STAR*D

The results of linear mixed-effect model prediction in STAR*D are given in Table 5. Each model includes the baseline score on the outcome measure as a covariate, so that only the unique contribution of the tested predictor is tested.

We first examined the relative contribution of the self-report and clinician-rated versions of QIDS. Baseline self-report QIDS-SR significantly predicted outcome on the clinician-rated QIDS-C even after adjusting for baseline QIDS-C and explained an additional 1.3% of variance in the clinician-rated outcome (Table 5). Conversely, baseline clinician-rated QIDS-C predicted outcome on QIDS-SR after adjusting for baseline QIDS-SR and explained an additional 0.3% of variation in the self-reported outcome.

We next considered the HRSD rated by the ROAs. The baseline ROA-rated HRSD significantly and strongly predicted outcomes measured by the clinician-rated and self-report scales and explained around 2% of variation in these outcomes (Table 5).

DISCUSSION

Equivalence of self-reported questionnaires and clinician-rated scales would be convenient, since it would allow efficient translation of evidence on efficacy into routine measurement-based care. The present study results demonstrate that self-reported and clinician-rated outcomes are not equivalent. Self-report and clinician rating each provide unique information that is relevant to clinical prognosis.

It has been debated whether clinician-led assessment or self-report should be primary in the assessment of treatment outcome and efficacy.[7,27] In both GEN-DEP and STAR*D, self-report scores contributed more to the prediction of outcomes on clinician-rated instruments and clinician assessment contributed relatively less to the prediction of outcome on self-reported scales. This suggests that self-report measures are preferable to clinician's rating if only one can be used. This finding supports the current trend toward using self-report instruments in clinical practice, primarily motivated by the fact that routine uptake of clinician-rated scales is impractical.[49] However, we also demonstrate that clinician-rated scales contain unique information that is not captured with self-report and is prognostically relevant. Therefore, in academic specialist settings, clinician-rated scales should be combined with self-report instrument to provide an accurate assessment since each assessment modality provides unique nonredundant information that complements the other in predicting treatment outcomes.

It has been shown that depressive symptoms can be separated into several dimensions[20] that differ in their response to treatments[41] and differentially predict outcomes.[50] Therefore, one reason why self-report instruments might contribute information that is different from clinician-rated scales is because they differ in their content and the relative

weighting of different symptom dimensions. For example, the cognitive symptom dimension that disproportionately contributes to the BDI has been shown to strongly predict outcome of antidepressant treatment[50] and this is likely to contribute to the prediction of outcome by BDI. However, the results from the STAR*D study show that even self-report and clinician-rated versions of the same instrument uniquely contribute to the prediction of treatment outcome, suggesting that content differences do not entirely account for the prediction of outcome. Therefore, in agreement with previous proposals,[7,17] we conclude that self-reported instruments and clinician-rated scales have complementary roles in the assessment of patients with MDD.

Although not directly relevant to our hypotheses, an additional unexpected finding of the present study deserves comment. In STAR*D, ratings by the independent ROAs strongly contributed to the prediction of outcomes on both self-report and clinician-rated measures. A post hoc examination of this result suggests that the structured telephone-based ROA rating represents a modality of assessment that differs from both clinician-rating based on unstructured clinical interview and self-report. Specifically, we found that a longer version of the Inventory of Depressive Symptomatology (IDS-30) rated by the ROA, was highly correlated with the HRSD, which was rated in the same telephone interview ($r = .89$), and less strongly related with the structurally more similar QIDS-C ($r = .58$) and QIDS-SR ($r = .64$) even when these were rated within a day of the ROA-interview. Surprisingly, the ROA-rated outcomes were somewhat closer to self-report than to clinician rating. This may be because they were rated in a telephone interview without access to the visual assessment or because they were performed by raters that were trained to perform a more structured type of interview. This may explain why ROA-rated HRSD contributed more to the prediction of clinician-rated than to self-reported outcomes. A similar pattern of findings was obtained for the ROA-rated IDS-30 (data available on request).

The present results should be interpreted with respect to several limitations. First, none of the studies were designed to test the hypotheses addressed in this article. The concordant results in the two samples with complementary methodological strengths and limitations present a relatively strong evidence to support the main result. Second, the present study is limited to the outcomes of pharmacological treatments for MDD. The relative contribution of self-report and clinician-rated outcomes to the naturalistic course of depression and to the specific response to antidepressants remains to be explored in routine outcome monitoring and placebo-controlled studies. The relevance of self-report and clinician-rated scales to outcomes of psychological treatment will also require a separate study.

With these limitations in mind, we conclude that clinician-rated and self-report measured of depression severity provide nonredundant information, which is relevant to clinical prognosis of antidepressant treatment. The most accurate prediction of outcome is achieved when both clinician assessment and self-rating are available.

Acknowledgments

Lundbeck provided nortriptyline and escitalopram for the GENDEP study. Glaxo-SmithKline and the UK National Institute for Health Research of the Department of Health contributed to the funding of the sample collection at the Institute of Psychiatry, London. The sponsors had no role in the design and conduct of the study, in data collection, analysis, interpretation or writing the report. Data for the replication study were obtained from the limited access datasets (version no. 2) distributed from the NIH-supported “Sequenced Treatment Alternatives to Relieve Depression” (STAR*D). STAR*D was supported by NIMH Contract no. N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is NCT00021528. This manuscript reflects the views of the authors and may not reflect the opinions or views of the STAR*D Study Investigators or the NIH. Dr. Uher is supported by the Canada Research Chair program (http://www.chairs-chaires.gc.ca/) and a grant from the Innovative Medicines Initiative of the European Commission (Grant Agreement no. 115008). Dr. Perlis is supported by NIMH MH086026. We thank A. John Rush for comments on an earlier version of the manuscript.

Contract grant sponsor: European Commission Framework 6; Contract grant number: LSHB-CT-2003–503428; Contract grant sponsor: GlaxoSmithKline; Contract grant sponsor: UK National Institute for Health Research of the Department of Health; Contract grant sponsor: NIMH; Contract grant number: N01MH90003; Contract grant sponsor: Innovative Medicines Initiative of the European Commission; Contract grant number: 115008; Contract grant sponsor: NIMH; Contract grant number: MH086026.

REFERENCES

1. American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder (revision). Am J Psychiatry. 2000; 157:1–45.
2. Gelenberg AJ. A review of the current guidelines for depression treatment. J Clin Psychiatry. 2010; 71:e15. [PubMed: 20667285]
3. National Institute for Clinical Excellence. The treatment and management of depression in adults. National Institute for Clinical Excellence (NICE), Department of Health; London, UK: 2009.
4. Nutt DJ. Highlights of the international consensus statement on major depressive disorder. J Clin Psychiatry. 2011; 72:e21. [PubMed: 21733474]
5. Adli M, Bauer M, Rush AJ. Algorithms and collaborative-care systems for depression: are they effective and why? A systematic review. Biol Psychiatry. 2006; 59:1029–1038. [PubMed: 16769294]
6. Cameron IM, Cardy A, Crawford JR, et al. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. Br J Gen Pract. 2011; 61:e419–e426. [PubMed: 21722450]
7. Moller HJ. Rating depressed patients: observer- vs self-assessment. Eur Psychiatry. 2000; 15:160–172. [PubMed: 10881213]
8. Katzelnick DJ, Duffy FF, Chung H, Regier DA, Rae DS, Trivedi MH. Depression outcomes in psychiatric clinical practice: using a self-rated measure of depression severity. Psychiatr Serv. 2011; 62:929–935. [PubMed: 21807833]
9. Rush AJ, Carmody TJ, Ibrahim HM, et al. Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatr Serv. 2006; 57:829–837. [PubMed: 16754760]
10. Trivedi MH. Tools and strategies for ongoing assessment of depression: a measurement-based approach to remission. J Clin Psychiatry. 2009; 70(Suppl 6):26–31. [PubMed: 19922741]
11. Beck AT, Ward CH, Mandelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry. 1961; 4:561–571. [PubMed: 13688369]
12. Richter P, Werner J, Bastine R, Heerlein A, Kick H, Sauer H. Measuring treatment outcome by the Beck Depression Inventory. Psychopathology. 1997; 30:234–240. [PubMed: 9239795]
13. Moller HJ, von Zerssen D. Self-rating procedures in the evaluation of antidepressants. Psychopathology. 1995; 28:291–306. [PubMed: 8838401]
14. Carroll BJ, Feinberg M, Smouse PE, Rawson SG, Greden JF. The Carroll rating scale for depression. I. Development, reliability and validation. Br J Psychiatry. 1981; 138:194–200. [PubMed: 7272609]
15. Svanborg P, Asberg M. A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Asberg Depression Rating Scale (MADRS). J Affect Disord. 2001; 64:203–216. [PubMed: 11313087]
16. Carter JD, Frampton CM, Mulder RT, Luty SE, Joyce PR. The relationship of demographic, clinical, cognitive and personality variables to the discrepancy between self and clinician rated depression. J Affect Disord. 2010; 124:202–206. [PubMed: 20004477]
17. Fava GA, Kellner R, Lisansky J, Park S, Perini GI, Zielezny M. Rating depression in normals and depressives: observer versus self-rating scales. J Affect Disord. 1986; 11:29–33. [PubMed: 2944925]
18. Feinberg M, Carroll BJ, Smouse PE, Rawson SG. The Carroll rating scale for depression. III. Comparison with other rating instruments. Br J Psychiatry. 1981; 138:205–209. [PubMed: 7272611]
19. Prusoff BA, Klerman GL, Paykel ES. Concordance between clinical assessments and patients' self-report in depression. Arch Gen Psychiatry. 1972; 26:546–552. [PubMed: 5027118]

20. Uher R, Farmer A, Maier W, et al. Measuring depression: comparison and integration of three scales in the GENDEP study. Psychol Med. 2008; 38:289–300. [PubMed: 17922940]
21. Zimmerman M, Coryell W, Wilson S, Corenthal C. Evaluation of symptoms of major depressive disorder. Self-report vs. clinician ratings. J Nerv Ment Dis. 1986; 174:150–153. [PubMed: 3950597]
22. Carroll BJ, Fielding JM, Blashki TG. Depression rating scales. A critical review. Arch Gen Psychiatry. 1973; 28:361–366. [PubMed: 4688625]
23. Schotte CK, Maes M, Cluydts R, De DD, Cosyns P. Construct validity of the Beck Depression Inventory in a depressive population. J Affect Disord. 1997; 46:115–125. [PubMed: 9479615]
24. Corruble E, Legrand JM, Zvenigorowski H, Duret C, Guelfi JD. Concordance between self-report and clinician's assessment of depression. J Psychiatr Res. 1999; 33:457–465. [PubMed: 10504014]
25. Domken M, Scott J, Kelly P. What factors predict discrepancies between self and observer ratings of depression? J Affect Disord. 1994; 31:253–259. [PubMed: 7989640]
26. Cuijpers P, Li J, Hofmann SG, Andersson G. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. Clin Psychol Rev. 2010; 30:768–778. [PubMed: 20619943]
27. Greenberg RP, Bornstein RF, Greenberg MD, Fisher S. A meta-analysis of antidepressant outcome under “blinder” conditions. J Consult Clin Psychol. 1992; 60:664–669. [PubMed: 1401382]
28. Enns MW, Larsen DK, Cox BJ. Discrepancies between self and observer ratings of depression. The relationship to demographic, clinical and personality variables. J Affect Disord. 2000; 60:33–41. [PubMed: 10940445]
29. Rane LJ, Fekadu A, Wooderson S, Poon L, Markopoulou K, Cleare AJ. Discrepancy between subjective and objective severity in treatment-resistant depression: prediction of treatment outcome. J Psychiatr Res. 2010; 44:1082–1087. [PubMed: 20471031]
30. Hamilton M. Development of a rating scale for primary depressive illness. Br J Clin Psychol. 1967; 6:278–296.
31. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry. 2004; 161:2163–2177. [PubMed: 15569884]
32. Trajkovic G, Starcevic V, Latas M, et al. Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. Psychiatry Res. 2011; 189:1–9. [PubMed: 21276619]
33. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979; 134:382–389. [PubMed: 444788]
34. Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. A Comparison of the QIDS-C16, QIDS-SR16, and the MADRS in an adult outpatient clinical sample. CNS Spectr. 2010; 15:458–468. [PubMed: 20625366]
35. Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. Eur Neuropsychopharmacol. 2006; 16:601–611. [PubMed: 16769204]
36. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed.. Washington, DC: 2000. text rev.
37. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the quick inventory of depressive symptomatology and the hamilton rating scale for depression: a sequenced treatment alternatives to relieve depression trial report. Biol Psychiatry. 2006; 59:493–501. [PubMed: 16199008]
38. Martinsen EW, Friis S, Hoffart A. Assessment of depression: comparison between Beck Depression Inventory and subscales of Comprehensive Psychopathological Rating Scale. Acta Psychiatr Scand. 1995; 92:460–463. [PubMed: 8837974]
39. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry. 2003; 54:573–583. [PubMed: 12946886]
40. Wing, JK.; Sartorius, N.; Ustin, TB. A Reference Manual for SCAN. World Health Organization; Geneva, Switzerland: 1998. Diagnosis and Clinical Measurement in Psychiatry.

41. Uher R, Maier W, Hauser J, et al. Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. Br J Psychiatry. 2009; 194:252–259. [PubMed: 19252156]

42. Uher R, Farmer A, Henigsberg N, et al. Adverse reactions to antidepressants. Br J Psychiatry. 2009; 195:202–210. [PubMed: 19721108]

43. Uher R, Muthen B, Souery D, et al. Trajectories of change in depression severity during treatment with antidepressants. Psychol Med. 2010; 40:1367–1377. [PubMed: 19863842]

44. Uher R, Dernovsek MZ, Mors O, et al. Melancholic, atypical and anxious depression subtypes and outcome of treatment with escitalopram and nortriptyline. J Affect Disord. 2011; 132:112–120. [PubMed: 21411156]

45. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006; 163:28–40. [PubMed: 16390886]

46. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006; 163:1905–1917. [PubMed: 17074942]

47. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Arch Gen Psychiatry. 2004; 61:310–317. [PubMed: 14993119]

48. Mallinckrodt CH, Sanger TM, Dube S, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry. 2003; 53:754–760. [PubMed: 12706959]

49. Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures. National survey. Br J Psychiatry. 2002; 180:101–103. [PubMed: 11823316]

50. Uher R, Perlis RH, Henigsberg N, et al. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. Psychol Med. 2012; 45:967–980. [PubMed: 21929846]

TABLE 1

Description of GENDEP and STAR*D samples

	GENDEP		STAR*D (level 1)	
	n	%	*n*	%
Number of subject recruited	811	100	4041	100
Number of subject with valid outcome	789	97	3637	90
Female	514	63	2532	63
Married/cohabiting	468	58	2372	59
Employed	422	52	2545	63
Recurrent depression	486	60	3054	76
Caucasian	811	100	3177	79
Treatment				
Citalopram			4041	100
Escitalopram	457	56		
Nortriptyline	354	44		

	Mean	SD	Mean	SD
Age	42.5	11.8	41.2	13.2
Depression severity at baseline				
Clinician-rated scales				
MADRS	28.8	6.7		
HRSD	21.7	5.3	19.8	6.5
QIDS-C			16.4	3.4
Self-rated scales				
BDI	28.0	9.7		
QIDS-SR			15.4	4.3

TABLE 2

Correlations between depression rating scales in GENDEP at study entry (baseline) and study exit (last valid score within the 12 weeks of treatment)

		Baseline scores			Exit scores		
		MADRS	**HRSD**	**BDI**	**MADRS**	**HRSD**	**BDI**
Baseline	MADRS	1.00					
	HRSD	0.77	1.00				
	BDI	0.58	0.48	1.00			
Exit	MADRS	0.39	0.39	0.36	1.00		
	HRSD	0.35	0.40	0.34	0.94	1.00	
	BDI	0.37	0.33	0.54	0.85	0.84	1.00

TABLE 3

Correlation of depression rating scales in STAR*D at study entry (baseline) and study exit (last valid score within the 12 weeks of treatment)

		Baseline scores			Exit scores		
Baseline		**QIDS-C**	**HRSD**	**QIDS-SR**	**QIDS-C**	**HRSD**	**QIDS-SR**
	QIDS-C	1.00					
	HRSD	0.53	1.00				
	QIDS-SR	0.69	0.57	1.00			
Exit	QIDS-C	0.33	0.34	0.33	1.00		
	HRSD	0.30	0.45	0.34	0.86	1.00	
	QIDS-SR	0.32	0.33	0.37	0.90	0.85	1.00

TABLE 4

Prediction of outcome by clinician-rated and self-report scales in GENDEP. *P*-vaLues significant after correction for multiple comparisons ($P < 0.0083$) are in bold

Predictor	Beta (95% CI)	*P*	r^2
Outcome: MADRS			
HRSD	0.09 (0.01, 0.15)	0.0173	0.1%
BDI	0.20 (0.14, 0.24)	**3.0×10^{-15}**	**3.1%**
Outcome: HRSD			
MADRS	0.13 (0.06, 0.19)	**0.0001**	**0.6%**
BDI	0.20 (0.14, 0.23)	**3.2×10^{-16}**	**4.1%**
Outcome: BDI			
MADRS	0.11 (0.06, 0.16)	**4.4×10^{-05}**	**1.1%**
HRSD	0.09 (0.04, 0.14)	**0.0007**	**0.9%**

TABLE 5

Prediction of outcome by clinician-rated and self-report scales in STAR*D. *P*-values significant after correction for multiple comparisons ($P < .0083$) are in bold

Predictor	Beta (95% CI)	*P*	r^2
Outcome: QIDS-C			
HRSD	0.34 (0.30, 0.38)	**2.00×10^{-57}**	**2.5%**
QIDS-SR	0.33 (0.28, 0.38)	**9.07×10^{-39}**	**1.3%**
Outcome: HRSD			
QIDS-C	0.71 (0.38, 1.04)	**2.21×10^{-05}**	**0.4%**
QIDS-SR	0.88 (0.55, 1.21)	**1.82×10^{-07}**	**0.9%**
Outcome: QIDS-SR			
QIDS-C	0.15 (0.10, 0.19)	**1.5×10^{-11}**	**0.3%**
HRSD	0.27 (0.23, 0.30)	**5.58×10^{-49}**	**1.9%**