Published in final edited form as:
J Psychiatr Res. 2011 January ; 45(1): 96–103. doi:10.1016/j.jpsychires.2010.04.032.

Concordance between clinician and patient ratings as predictors of response, remission, and recurrence in major depressive disorder

Boadie W. Dunlop[a,*], **Thomas Li**[b], **Susan G. Kornstein**[d], **Edward S. Friedman**[e], **Anthony J. Rothschild**[f], **Ron Pedersen**[b], **Philip Ninan**[c], **Martin Keller**[g], and **Madhukar H. Trivedi**[h]

[a]Department of Psychiatry, Emory University School of Medicine, 1256 Briarcliff Road, Building A, 3rd Floor, Atlanta, GA 30306, USA

[b]Global Biostatistics and Programming, Wyeth Research, Collegeville, Pennsylvania, USA

[c]Global Medical Affairs, Wyeth Research, Collegeville, Pennsylvania, USA

[d]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA

[e]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

[f]Department of Psychiatry, University of Massachusetts Medical School and UMass Memorial Health Care, Worcester, Massachusetts, USA

[*]Corresponding author. Tel.: +1 404 727 8969; fax: +1 404 727 3700. bdunlop@emory.edu (B.W. Dunlop).

Conflicts of interest

Dr. Rothschild received grants or funding from the National Institute of Mental Health, Cyberonics, Takeda, and Wyeth. He is a consultant for Pfizer, GlaxoSmithKline, Forest Laboratories, and Eli Lilly & Company. Royalties include the Rothschild Scale for Anti-depressant Tachyphylaxis (RSAT), Clinical Manual for the Diagnosis and Treatment of Psychotic Depression, American Psychiatric Press, 2009.

Dr. Dunlop has served as a consultant for Wyeth and Bristol-Myers Squibb and has served on the Speaker's Bureau for Bristol-Myers Squibb. He has received research support from GlaxoSmithKline, Novartis, Takeda, the National Institute of Mental Health, Ono Pharmaceuticals, and Wyeth.

Dr. Friedman is a consultant for Pfizer. He has received grant/research support from Northstar, Sanofi Aventis, Novartis, Cyberonics, Medtronics, and the National Institute of Mental Health.

Dr. Kornstein has received grants/research from the National Institute of Mental Health, Departments of Health and Human Services, Pfizer, Bristol-Myers Squibb, Eli Lilly & Company, Forest Laboratories, Wyeth, Novartis, Sepracor, Boehringer-Ingelheim, Sanofi-Synthelabo, AstraZeneca, and Takeda. She has served on advisory boards for Wyeth, Pfizer, Eli Lilly & Company, Bristol-Myers Squibb, Endo, Forest Laboratories, Sepracor, Neurocrine, and Takeda, and has received book royalties from Guilford Press.

Dr. Trivedi has received grants/research from the Agency for Healthcare Research and Quality, Corcept Therapeutics, Cyberonics, Meade Johnson, National Alliance for Research in Schizophrenia and Depression, the National Institute of Mental Health, NIDA, Novartis, Pharmacia & Upjohn, Solvay, and Targacept. He is a consultant for Abbott, Abdi Brahim, Akso, AstraZeneca, Bristol-Myers Squibb, Cephalon, Fabre-Kramer, Forest Laboratories, GlaxoSmithKline, Janssen, Johnson & Johnson, Eli Lilly & Company, Meade Johnson, Neuronetics, Parke-Davis, Pfizer, Sepracor, and Vantage Point.

Dr. Ninan is a Wyeth employee and stockholder.

Dr. Li is a Wyeth employee and stockholder.

Dr. Pedersen is a Wyeth employee and stockholder.

Dr. Keller is a consultant for Abbott, CENEREX, Cephalon, Cupress Bioscience, Cyberonics, Forest Laboratories, Janssen, JDS, Medtronic, Organon, Novartis, Pfizer, Roche, Solvay, and Wyeth. He has received grants/research from Pfizer and is part of the advisory board for Abbott, Bristol-Myers Squibb, CENEREX, Cyberonics, Cypress, Forest Laboratories, Janssen, Neurotics, Novartis, Organon, and Pfizer.

Contributors: Dr. Dunlop developed the first draft of the manuscript, contributed to the statistical analytic plan and performed literature searches. Dr. Li and Mr. Pederson performed the statistical analyses and contributed to the manuscript. Drs. Kornstein, Friedman, Ninan, Rothschild, and Trivedi contributed to the conduct of the study and the development of the analytic plan, and provided revisions to the first draft of the manuscript. Dr. Keller led the design of the protocol and contributed to the development of the statistical analytic plan and provided revisions to the manuscript. All authors contributed to approved the final manuscript.

[g]Department of Psychiatry and Human Behavior, Brown University, Providence, Rhode Island, USA

[h]University of Texas Southwestern Medical School, Dallas, Texas, USA

Abstract

We conducted a secondary analysis of data from the Prevention of Recurrent Episodes of Depression With Venlafaxine Extended Release (ER) for Two Years (PREVENT) trial to evaluate whether discrepancies between clinician and patient ratings of depression severity were predictive of response, remission, and recurrence during treatment for a depressive episode. Patients who self-rated depression severity in concordance with the clinician ("concordant patients") were defined as having a standardized patient-rated Inventory of Depressive Symptoms-Self Report (IDS-SR30) score minus standardized clinician-rated Hamilton Rating Scale for Depression ($HAM\text{-}D_{17}$) score <1 SD from mean. Non-concordant patients ("underrating patients" [−1 SD], "overrating patients" [+1 SD]) were identified. Cohorts were compared for remission and response on the $HAM\text{-}D_{17}$, Clinician Global Impression–Severity (CGI-S), and $IDS\text{-}SR_{30}$ during acute and continuation therapy and time to recurrence during maintenance therapy. During acute treatment female patients were more likely to overrate their depression severity compared to the clinician; older age predicted overrating during continuation treatment. Overrating patients had a slower onset of response on the $HAM\text{-}D_{17}$ during acute treatment ($P = 0.004$). There were no differences between cohorts for remission or response on the $HAM\text{-}D_{17}$ or CGI-S. Overrating patients at week 10 had lower remission and response rates on the IDS-SR30 during continuation therapy (32% and 50%, respectively; $P \leq 0.001$) compared with underrating patients (76%, 77%) or concordant patients (64%, 78%). Patient concordance at the end of continuation therapy did not predict recurrence during maintenance therapy, indicating that patient rating scales may be useful in tracking recurrence during maintenance therapy. Poor agreement between patient- and clinician-ratings of depression severity is primarily a state phenomenon, although it is trait-like for some patients.

Keywords

Depression; Psychiatric status rating scales; Reliability and validity; Outcome assessment; Treatment outcome; Anxiety

1. Introduction

Clinical research is a complex, time-consuming, and costly endeavor (Collier, 2009). Any method that increases the efficiency of the psychiatric research process while maintaining the veracity of study results is of interest to clinical researchers and clinicians alike (Rush et al., 2006a). Patient factors, namely early clinical improvement, are one feature of antidepressant trial design that has been scrutinized in this regard. Recently published reports of early improvement during antidepressant treatment predicting later clinical outcome are germane to both clinical trial design and patient care (Katz et al., 2009; Szegedi et al., 2009).

Another feature of interest in antidepressant trial design is the method used to rate depression severity. Clinician-administered symptom rating scales, such as the Hamilton Rating Scale for Depression (HAM-D) (Hamilton, 1960) and the Montgomery Åsberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979), are well-established instruments used to document depression severity at baseline and track changes throughout the trial. However, these methods necessitate training and certification of experienced psychiatric clinicians (Rush et al., 2006a, 2003). The use of validated patient self-rating

instruments in place of clinician rating scales is an attractive time- and cost-saving option that has gained momentum in the psychiatric research community (Bernstein et al., 2007; Rush et al., 2006a). The landmark Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial emphasized the patient-rated 16-item version of the Quick Inventory of Depressive Symptomatology–Self-Report ($QIDS\text{-}SR_{16}$) as a primary reported outcome measure of remission and response (Trivedi et al., 2006). Despite the attractiveness of self-reports, it has not been demonstrated conclusively that patient-rating instruments reliably provide assessments that are consistent with those of clinician-administered rating scales for measuring the severity of depressive symptoms at baseline and throughout a clinical trial (Corruble et al., 1999; Dorz et al., 2004).

We compared clinician-rated and patient-rated measures of depression severity and treatment response in a secondary analysis of data from the acute and continuation phases of the Prevention of Recurrent Episodes of Depression with Venlafaxine for Two Years (PREVENT) trial, a multiphase, multicenter, randomized, double-blind study that evaluated the efficacy of venlafaxine extended release (ER) as long-term therapy for recurrent depression (Keller et al., 2007b). The analysis demonstrated that patient-rated measures of depression severity did not correspond well with clinician ratings. After 6 months of continuation treatment, correlations between patient-rated scores on the 30-item Inventory of Depressive Symptomatology–Self-Rated ($IDS\text{-}SR_{30}$) and clinician ratings on the 17-item HAM-D ($HAM\text{-}D_{17}$) for remission and response were poor (κ values of 0.45 and 0.32, respectively) (Dunlop et al., 2010). These previous findings suggest that patient-rated scales cannot wholly substitute for clinician ratings when evaluating antidepressant treatment effects.

To better understand the factors contributing to the observed inconsistencies in clinician-rated and patient-rated measures of depressive symptom severity, we conducted further analyses of data from the PREVENT trial. The objectives of the analyses reported here were to identify factors contributing to differences between clinician- and patient-rated measures of remission and response during acute and continuation therapy and to determine if outcomes varied as a function of clinician-patient rating discrepancies.

2. Methods

2.1. The PREVENT trial

Details about the methods and results from the PREVENT trial are reported elsewhere (Keller et al., 2007a,b). In brief, eligible patients (n = 1,096) entered the trial and were randomly assigned to receive a 10-week, acute treatment course with double-blind venlafaxine ER or fluoxetine (Keller et al., 2007a). Patients achieving response or remission during the acute phase entered a 6-month, double-blind continuation phase on the same treatment (n = 715) (Keller et al., 2007a). The continuation phase was followed by 2 consecutive 12-month maintenance phases. At that start of each maintenance phase, responding or remitting patients in the venlafaxine ER group were randomly assigned to treatment with venlafaxine ER or placebo, and patients in the fluoxetine group continued taking fluoxetine (Keller et al., 2007a).

Patients were adults with at least a 1-month history of *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) major depressive disorder (MDD) (American Psychiatric Association, 1994) and recurrent depression (≥3 major depressive episodes, with ≥2 episodes including the current episode occurring in the past 5 years, with ≥2 months between end of the previous episode and the beginning of the current episode). All patients had a HAM-D17 total score of ≥20 at screening and ≥18 at randomization (Keller et al., 2007b). Certified raters administered the $HAM\text{-}D_{17}$, the Hamilton Anxiety Rating Scale

(HARS) (Hamilton, 1959), the Clinical Global Impressions–Severity (CGI-S), and the Clinical Global Impressions–Change (CGI-C) rating scales (Keller et al., 2007b). Patients rated the severity of their depression using the IDS-SR30 (Keller et al., 2007b). Data for the QIDS-SR_{16} was derived from the IDS-SR_{30}.

There were no statistically significant differences at the end of the acute phase in rates of response (79% for both groups; $P = 0.719$) or remission (venlafaxine ER 49%; fluoxetine 50%; $P = 0.719$). After completion of the 6-month continuation phase, rates of response (venlafaxine ER 90%; fluoxetine 92%) and remission (venlafaxine ER 72%; fluoxetine 69%) were similar for both groups ($P = 0.696$) (Keller et al., 2007b). The probability of recurrence after completing the first 12-month maintenance phase was 23.1% for venlafaxine ER and 42.0% for placebo ($P = 0.005$). At the conclusion of the second 12-month maintenance phase, the probability of recurrence was 8.0% for venlafaxine ER compared with 44.8% for placebo ($P < 0.001$) (Keller et al., 2007a).

2.2. Patient categories by discrepancies between patient- and clinician-rated scales

Differences between clinician rating and patient rating scores were used to identify patients who rated their depression as less severe (underrating patients) or more severe (overrating patients) than their clinician rating. A discrepancy (D) score based on the HAM-D_{17} and IDS-SR_{30} scores for each patient was calculated according to the method reported by Dorz and colleagues (Dorz et al., 2004), in which the standardized score on the clinician-rated HAM-D17 was subtracted from the standardized score on the patient-rated IDS-SR_{30}. Negative D-scores indicated underrating, positive D-scores indicated overrating. Patients whose D-scores were ±1 standard deviation (SD) from the mean were considered to be nonconcordant patients (underrating patients [−1 SD] and overrating patients [+1 SD]), and all other patients were concordant patients. Fig. 1 displays the overall design of the PREVENT trial and the number of patients in each category through the study.

2.3. Predictors of nonconcordance

Patient characteristics at baseline were evaluated to assess for predictors of discrepancy between patient- and clinician-ratings. The effect of anxiety on the concordance between patient and clinician ratings of depression severity was assessed by dividing the sample into high anxiety (HARS ≥18) or low anxiety (HARS < 18) at the baseline and again at week 10 visits.

2.4. Acute phase

Three cohorts of patients (concordant, underrating, and overrating) were identified at baseline, and outcomes for each cohort were compared at the conclusion of the acute phase. The acute-phase outcomes were rates of response and remission at week 10 (or last visit), time to response and remission, and within-patient effect sizes for HAM-D_{17}, IDS-SR_{30}, and QIDS-SR_{16} at week 10 (or last visit). Discontinuation rates were identified for each of the patient categories. In addition, discontinuation rates were compared for nonconcordant raters (i.e., over- and underrating patients grouped together) and concordant raters.

2.5. Continuation phase

New cohorts of concordant, underrating, and overrating patients were identified based on clinician and patient ratings after completing acute-phase treatment at week 10. Outcomes for grouped nonconcordant vs concordant patients and the 3 cohorts were compared for response and remission rates and discontinuation rates.

2.6. Change in patient category over time

In order to evaluate whether the categorical assignment of patients to underrating, concordant, or overrating categories was stable over time, the patients' individual rating category was tracked through from baseline to the 10-week and 6-month end points. The rating status of the patients who withdrew from the study prior to the 10-week or 6-month end points was not tracked.

2.7. Maintenance phase

After completing the 6-month continuation phase, new cohorts were identified for concordant, underrating, and overrating patients. The probability of and time to protocol-defined recurrence and rates of discontinuation were compared for the 3 groups.

2.8. Statistical analysis

Statistical analyses were performed on the intent-to-treat (ITT) population for the acute (N = 1047) and continuation (N = 715) phases (Keller et al., 2007b). The ITT population included all patients who had at least one dose of study medication and at least one postbaseline HAM-D_{17} evaluation (Keller et al., 2007b). In this secondary analysis, remission was defined as a HAM-D_{17} total score ≤7, CGI-S score ≤2, and IDS-SR_{30} total score ≤14. Response was defined as a ≥50% reduction from baseline on HAM-D_{17}, IDS-SR_{30}, and QIDS-SR_{16} total scores and a Clinical Global Impressions–Change (CGI-C) score ≤2. Recurrence during the maintenance phases was defined as a HAM-D_{17} total score ≥12 and a reduction of ≤50% on the HAM-D_{17} total score from baseline at 2 consecutive visits or at the last valid visit prior to study discontinuation.

Multiple logistic regression modeling identified baseline patient characteristics that were significant predictors of nonconcordance. The Mantel-Haenszel correlation chi-square was used to compare rates of response and remission and rates of discontinuation among the 3 cohorts during the acute and continuation phases. The time to response and remission during the acute phase and the time to recurrence in the maintenance phase were calculated and described with Kaplan–Meier methods. Differences between cohorts were compared using log-rank tests.

Descriptive statistics were used to report within-patient effect sizes using standardized response means (SRM) for each outcome measure during the acute phase. The SRM was calculated as follows: (baseline score – end point score) / (standard deviation of baseline – end point difference). SRMs were calculated for concordant underrating, and overrating patients. The distribution of the 3 cohorts in the venlafaxine ER, fluoxetine, and placebo arms at the beginning of the maintenance phase was compared using chi-square tests. All statistical analyses were conducted using the SAS analytic software (SAS Institute Inc, Cary, NC).

3. Results

3.1. Predictors of nonconcordance

Several baseline features predicted differences between clinician and patient ratings (Table 1). During the acute phase of treatment, men were significantly less likely than women to view their depression as more severe than the clinician (OR: 0.53,95% CI: 0.29–0.98; P<0.05 at week 10), but this difference was not apparent during the continuation phase. Younger age and greater body mass index (BMI) also predicted more severe rating scores by patients compared with clinicians during the early weeks of the acute phase ($P \leq 0.01$ for both), but the magnitude of the odds ratios for these predictors was very small, indicating they are of little clinical importance. During the continuation phase, older age was

associated with more severe patient- versus clinician-ratings at months 1 ($P < 0.05$), 5 ($P < 0.05$), and 6 ($P \leq 0.001$). The number of lifetime depressive episodes, sex, and BMI did not predict differences between patient- and clinician-rated depression severity during continuation treatment.

At baseline, 66.9% of high anxiety patients ($N = 486$) were concordant with the clinician rating of depression severity, versus 72.0% of the low anxiety patients ($N = 542$) ($p = .08$). Level of anxiety did not predict whether the patient would be classified into the under- over- or concordant category ($p = .50$).

3.2. Acute phase

After the first 10 weeks of treatment, there were no statistically significant differences in rates of remission or response between the 3 cohorts on the HAM-D_{17}, CGI-S, or IDS-SR_{30} scales (Table 2). A trend for a lower remission rate on the IDS-SR_{30} score was apparent for overrating patients (35%) compared with underrating patients (45%) and concordant patients (41%) at week 10, but these differences did not quite achieve statistical significance ($P = 0.058$). Response rates on the HAM-D_{17}, CGI-C, or IDS-SR_{30} during acute treatment were similar for the 3 cohorts (Table 2). There were no differences in time to onset of remission during the acute phase among the 3 cohorts ($P = 0.949$) (Fig. 2), but there was a significant difference on time to response ($P = 0.004$), with overrating patients achieving response more slowly on the HAM-D_{17} than underraters (Fig. 3).

Within group SRMs for change on the HAM-D_{17} (baseline to endpoint) during acute treatment were larger (2.36) for underrating patients compared with concordant patients (1.99; $P < 0.001$), but SRMs for overrating patients (1.72) were smaller than for concordant patients ($P = 0.002$). Across the 3 categories of patients, the standardized scores for the IDS-SR_{30} and QIDS-SR_{16} were essentially identical, but within group SRMs for both patient rating scales were higher for overrating patients compared with concordant patients ($P < 0.001$) and lower for underrating patients compared with concordant patients ($P \leq 0.002$).

At week 10, HARS scores identified 298 high anxiety and 522 low anxiety patients. In contrast to the finding of no difference in rating category by anxiety level at baseline, at week 10 the HAM-D_{17} versus IDS-SR_{30} correlations did find significant differences between patients with and without high levels of anxiety. Among the high anxiety patients, 65 (21.8%) were classed as underrating, 47 (15.8%) as over-rating, and 62.4% as concordant. Among the 522 low anxiety patients, under-, over- and concordant rating occurred in 49 (9.4%), 70 (13.4%) and 403 (77.2%), respectively. These categorical differences were statistically significant ($p < .0001$).

3.3. Continuation phase

New cohorts were identified at the beginning of the continuation phase based on clinician-patient discrepancies at the end of the 10-week acute treatment phase. At the last visit of the 6-month continuation treatment phase, week 10 nonconcordant patients had significantly lower rates of remission and response on the HAM-D_{17}, CGI-S, and IDS-SR_{30} compared with concordant patients (Table 3). These findings were consistent in both the completer (Month 6) and intent-to-treat (Final Visit) samples, with the exception that the difference in response rates among the completer sample on the HAM-D_{17} outcome was not significant. The lower rate of response among nonconcordant patients was significant ($p < .05$) on the HAM-D_{17} and IDS-SR_{30} outcomes at every monthly rating visit during the continuation phase, with the exception of month 3. Remission rates were significantly lower ($p < .05$) among nonconcordant patients at every monthly visit on the IDS-SR_{30} (data not shown).

When concordant, underrating, and overrating patients were assessed in a 3-way comparison (Table 4), final remission and response rates on the HAM-D_{17} and CGI-S for week 10 overrating patients were lower than the 2 other groups (except for response rates on the HAM-D_{17}, which were lowest for underrating patients), but these differences did not reach statistical significance. The overrating patients at week 10 continued to overrate throughout the continuation phase, based on the lower rates of remission and response on the IDS-SR at every visit of continuation compared with concordant patients (overall $P \leq 0.001$).

3.4. Change in patient category over time

Changes in patient category for each of the 3 original cohorts were tracked from baseline to week 10 of the acute phase and again from week 10 to the conclusion of the 6-month continuation phase (Table 5). Of the patients who completed the 10-week acute treatment phase, approximately 70% were categorized as concordant patients at week 10, regardless of baseline discrepancy rating (71.8% underrating, 73.6% concordant, 68.0% overrating). Only 8% of baseline nonconcordant raters converted to the opposing non-concordant rating (e.g., underrating patient converting to overrating patinet) at week 10. Similarly, 60% of week 10 underrating patients, 76.5% of concordant patients, and 51.7% of overrating patients were classified as concordant patients at the conclusion of the continuation phase. However, a sizable minority of week 10 underrating patients (37.8%) and overrating patients (43.1%) remained in their original category (Table 5). Only a small minority of nonconcordant patients (≤5%) switched to the opposite rating category at 6 months.

3.5. Maintenance phase

New cohorts were identified at the beginning of the first maintenance phase based on physician-patient concordance at the conclusion of the continuation phase. There were no significant differences in the distribution of patient categories to the 3 treatment groups (data not shown). There also were no statistically significant differences in time to recurrence for the combined treatment groups of concordant vs nonconcordant patients ($P = 0.40$; data not shown) or concordant patients vs underrating or overrating patients ($P = 0.68$; data not shown).

3.6. Discontinuation

Discontinuation rates during the acute treatment phase for underrating patients (34.1%) and overrating patients (36.5%) were not significantly different than for concordant patients (30.1%; $P = 0.24$) or for concordant (30.1%) vs nonconcordant patients (35.3%; $P = 0.10$). A 2-way comparison demonstrated that week 10 nonconcordant patients had significantly higher dropout rates during the continuation phase (43.2%) compared with concordant raters (32.2%; $P = 0.009$). Week-10 underrating patients (40.5%) and overrating patients (45.1%) had higher rates of discontinuation during the continuation phase compared with concordant patients (32.2%), but these differences were not statistically significant. There were no statistically significant differences among the 3 cohorts in rates of discontinuation during the maintenance phases.

4. Discussion

In this analysis of data from the PREVENT trial, we found several baseline features that predicted which patients may rate their depression as more severe than clinicians. Women were significantly more likely than men to perceive their symptoms as more severe than the clinician during acute treatment. Younger patients and those with greater BMI were also slightly more likely to self-rate their symptoms as more severe than the clinician, but the magnitude of these differences was very small. These predictors of nonconcordance, particularly age, are consistent with some reports (Rush et al., 2006b; Paykel et al., 1973).

However, other studies found that demographic or clinical features did not predict non-concordance (Corruble et al., 1999; Domken et al., 1994; Rush et al., 1987). Nevertheless, our findings suggest that reliance on patient self-rating instruments to screen patients for eligibility may result in selection bias, particularly favoring women.

We found no difference among the 3 cohorts in rates of remission or response or the time to reach remission on the HAM-D_{17} during acute therapy. In contrast, overrating patients at baseline fulfilled clinician-rated response criteria significantly more slowly than underrating or concordant patients. During the continuation phase, rates of clinician-rated remission and response for overrating patients at week 10 vs the other 2 cohorts were numerically lower, but did not achieve statistical significance. However, self-rated measures of remission and response on the IDS-SR30 among the week 10 overrating patients were significantly lower (32%; 50%) compared with underrating (76%, 77%) or concordant patients (64%, 78%) ($P < 0.001$, $P < 0.001$, respectively) during the 6 months of continuation therapy.

These findings can be interpreted to mean that nonremitting overrating patients emphasize their symptoms to such a degree as to impact clinician ratings on the HAM-D_{17} resulting in slower achievement of remission and response during acute treatment. The slower improvement among overrating patients corresponded with a nonsignificant, numerical trend toward lower remission and response rates on the HAM-D_{17} and CGI-S scores during continuation treatment. Use of self-rating instruments by such “high-distress expression” patients to track changes in symptom severity during treatment may result in less favorable study results. Interestingly, patients with high anxiety scores (HARS ≥ 18) after 10 weeks of treatment were more than twice as likely to under-rate their depression severity as compared to low anxiety patients. This finding suggests that high distress expression does not simply reflect current anxiety, but rather stems from other factors.

The within group effect size of the clinician- and patient-rated scales varied between the cohorts, which may have relevance for power calculations used in study design, and for making comparisons of benefits across studies, such as in meta-analyses. When depression severity was measured by clinicians using the HAM-D_{17}, the effect size was larger for underrating patients (2.36) than for concordant patients (1.99) or overrating patients (1.72). Therefore, the clinician’s ability to detect changes in depression severity over the course of treatment may be greatest for patients who downplay their symptoms during an episode of depression. The opposite is true for self-reports. Patient ratings of symptom severity on the IDS-SR_{30} resulted in larger effect sizes for overrating patients (2.02) compared with concordant patients (1.62) or underrating patients (1.31). These findings also raise questions about the value of self-report depression severity measures as criteria for entry into clinical trials. When a study’s primary outcome is a clinician-rated scale, but a score above a certain threshold on a self-rated depression measure is required for study inclusion, the study may be at risk for over-selecting for subjects least likely to show improvement on the primary outcome measure. This concern is even more relevant for trials examining speed of response, given the significantly slower rate of response observed among overrating patients in this analysis.

We followed the original cohorts from baseline through the completion of the 6-month continuation phase and tracked the proportion of patients whose baseline discrepancy rating changed. Patients who were nonconcordant at baseline were of particular interest because changes in their discrepancy ratings over time, which presumably corresponded with improvement in depression severity, might suggest that nonconcordant ratings represent a state, rather than trait characteristic. For example, finding that patients who downplayed pretreatment depression severity compared with the clinician (i.e., an underrating patient), but who agreed with the clinician after 6 months of treatment (i.e., switch from underrating

to concordant patient) might be interpreted to mean that baseline underrating reflects a state-like phenomenon that changes with the resolution of the depressive episode. Our findings suggest that the majority of under- and overrating patients become concordant during 10 weeks of treatment, implying that patient-clinician rating discrepancies are largely a state phenomenon. Increased concordance during acute treatment may result from: 1) patients developing a greater shared understanding with the clinician regarding the meaning of the descriptors used on rating scales to describe symptoms; 2) enhanced concentration and cognitive function due to improvement over time; 3) reduced patient mood bias in recall of the previous week's symptom levels; or 4) increased disclosure to the clinician due to improving rapport over time. During continuation, however, smaller proportions of nonconcordant patients became concordant between week 10 and month 6, suggesting that for some patients, divergent self- and clinician ratings represents a trait-like characteristic.

Our finding that time to onset of a recurrent depressive episode during the maintenance phase was not different between non-concordant and concordant patients suggests that using patient-rating scales during the maintenance phase of antidepressant trials may permit some efficiencies in time and expense. This hypothesis could be replicated through secondary analyses of existing maintenance study data sets.

The finding of higher study discontinuation rates during the continuation phase has relevance for both clinical trials and clinical care of depressed patients. Nonconcordance may represent a disjunction in understanding between the clinician and the patient. Underrating patients may perceive their clinician as making too much of their symptoms and consequently "over-treating" them. Overrating patients may perceive their clinician as not fully understanding their distress and drop out of treatment to seek care elsewhere. To reduce treatment drop out, highly discordant scores between clinicians and patients may warrant an explicit conversation with the patient about these differences and the patient's perception about the appropriateness of the care they are receiving.

Clinician-administered rating scales remain the standard by which depressive symptoms are evaluated in antidepressant drug trials. However, these scales are time-consuming and increase the cost of trials, leading some clinical trial researchers to suggest that substituting patient ratings for clinician ratings of depression severity may be rational (Rush et al., 2006a, 2003). Our findings advise a more cautious approach. In an earlier analysis of data from the PREVENT trial, we found that correlations between clinician and patient ratings were relatively poor (Dunlop et al., 2010). The data reported herein expand on these findings, demonstrating that discrepancies between patient and clinician ratings lead to meaningful differences in outcome measurements. Certainly, our findings require replication by other studies.

The value of the patient perspective on outcomes through using self-rated scales is significant and should not be abandoned. Although the clinician's rating historically has carried the most significance for entry into depression trials, there is in fact no "gold standard" to assess depression severity and thereby determine whether the clinician or the patient is more accurately capturing the level of depression. There was no evidence that the patient-clinician discrepancies arose from clinician-rater "inflation" of the baseline HAM-D_{17} scores to increase enrolment. The percentage decline in the HAM-D_{17} and IDS-SR_{30} scores was identical at the first post-baseline assessment (week 1) (Dunlop et al., 2010).

More research is required to further understand why some patients report differing levels of depression severity on a questionnaire than are detected during a clinical interview. Possible contributors to these discrepancies include unmeasured factors such as physical pain or loneliness, the patient's inability on a questionnaire to make distinctions between anxiety

and depression, or a low literacy level or misinterpretation of the question's focus. Another contributor that might be relevant for this sample is that patients with recurrent depression may be comparing their current symptom levels versus those of past episodes, and imparting a sense that the current episode is much worse (or not as bad) as the patient's prior experiences of depression, whereas the clinician is (ideally) rating symptoms without reference to past episode severity. It would also be of interest to conduct factor analyses to identify the specific items on the HAM-D_{17} and IDS-SR_{30} that most contribute to nonconcordance.

Differences between the findings of this study and other published data may be due in part to differences in characteristics of the study participants. In particular, the PREVENT study enrolled patients with a history of recurrent depression, whereas other studies (e.g., STAR*D) enrolled patients with more diverse psychiatric histories (Trivedi et al., 2006). Patients with recurrent depression may tend to view their symptoms within the context of symptoms experienced during previous episodes.

Future studies comparing concordance ratings between patients with recurrent depression and patients with first-episode depression would be of interest. It may be possible to design and validate a mathematical correction factor to account for differences between clinician- and patient-ratings and thereby produce a more consistent measure of outcome.

Acknowledgments

This study was sponsored by Wyeth which was acquired by Pfizer Inc in October 2009. Medical writing support for this manuscript was provided by Sally K. Laden, MS, of MSE Communications LLC, and medical editing support was provided by Nicole Hilberth, MS, of Advogent, and this support was funded by Pfizer, formerly Wyeth Research, Collegeville, PA.

Role of funding source: Research supported by Pfizer Research.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed.. Washington, DC: American Psychiatric Association;

Bernstein IH, Rush AJ, Carmody TJ, Woo A, Trivedi MH. Clinical vs. self-report versions of the quick inventory of depressive symptomatology in a public sector sample. Journal of Psychiatric Research. 2007; 41:23–46.

Collier R. Drug development cost estimates hard to swallow. Canadian Medical Association Journal. 2009; 180:279–280. [PubMed: 19188620]

Corruble E, Legrand JM, Zvenigorowski H, Duret C, Guelfi JD. Concordance between self-report and clinician's assessment of depression. Journal of Psychiatric Research. 1999; 33:457–465. [PubMed: 10504014]

Domken M, Scott J, Kelly P. What factors predict discrepancies between self and observer ratings of depression? Journal of Affective Disorders. 1994; 31:253–259. [PubMed: 7989640]

Dorz S, Borgherini G, Conforti D, Scarso C, Magni G. Comparison of self-rated and clinician-rated measures of depressive symptoms: a naturalistic study. Psychology and Psychotherapy: Theory, Research and Practice. 2004; 77:353–361.

Dunlop BW, Li T, Kornstein SG, et al. Correlation between patient and clinician assessments of depression severity in the PREVENT study. Psychiatry Research. 2010; 177:177–183. [PubMed: 20304503]

Hamilton M. The assessment of anxiety states by rating. British Journal of Medical Psychology. 1959; 32:50–55. [PubMed: 13638508]

Hamilton M. A rating scale for depression. Journal of Neurology Neurosurgery and Psychiatry. 1960; 23:56–62.

Katz MM, Meyers AL, Prakash A, Gaynor PJ, Houston JP. Early symptom change prediction of remission in depression treatment. Psychopharmacology Bulletin. 2009; 42:94–107. [PubMed: 19204654]

Keller M, Trivedi MH, Thase ME, et al. The Prevention of Recurrent Episodes of Depression with Venlafaxine for Two Years (PREVENT) study: outcomes from the two-year and combined maintenance phases. Journal of Clinical Psychiatry. 2007a; 68:1246–1256. [PubMed: 17854250]

Keller MB, Trivedi MH, Thase ME, et al. The Prevention of Recurrent Episodes of Depression with Venlafaxine for Two Years (PREVENT) study: outcomes from the acute and continuation phases. Biological Psychiatry. 2007b; 62:1371–1379. [PubMed: 17825800]

Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. British Journal of Psychiatry. 1979; 134:382–389. [PubMed: 444788]

Paykel ES, Prusoff BA, Klerman GL, DiMascio A. Self-report and clinical interview ratings in depression. Journal of Nervous and Mental Disease. 1973; 156:166–182. [PubMed: 4698664]

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. Biological Psychiatry. 2006a; 59:493–501. [PubMed: 16199008]

Rush AJ, Carmody TJ, Ibrahim HM, Trivedi MH, Biggs MM, Shores-Wilson K, Crismon ML, Toprac MG, Kashner TM. Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatric Services. 2006b; 57:829–837. [PubMed: 16754760]

Rush AJ, Hiser W, Giles DE. A comparison of self-reported versus clinician-related symptoms in depression. Journal of Clinical Psychiatry. 1987; 48:246–248. [PubMed: 3584081]

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. Biological Psychiatry. 2003; 54:573–583. [PubMed: 12946886]

Szegedi A, Jansen WT, van Willigenburg AP, van der ME, Stassen HH, Thase ME. Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. Journal of Clinical Psychiatry. 2009; 70:344–353. [PubMed: 19254516]

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. American Journal of Psychiatry. 2006; 163:28–40. [PubMed: 16390886]

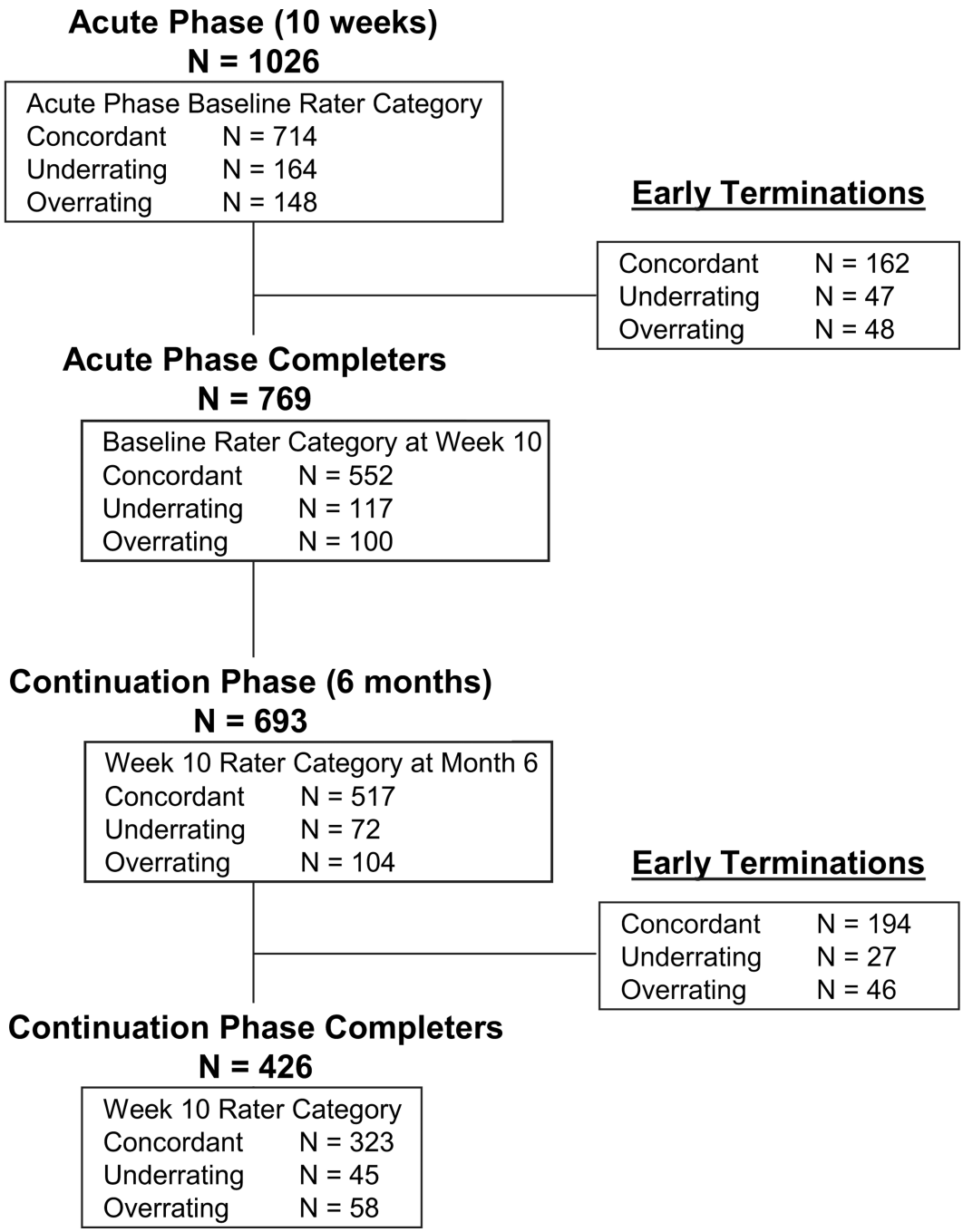

Fig. 1.
Flow chart of the PREVENT study and patient categories at each stage.

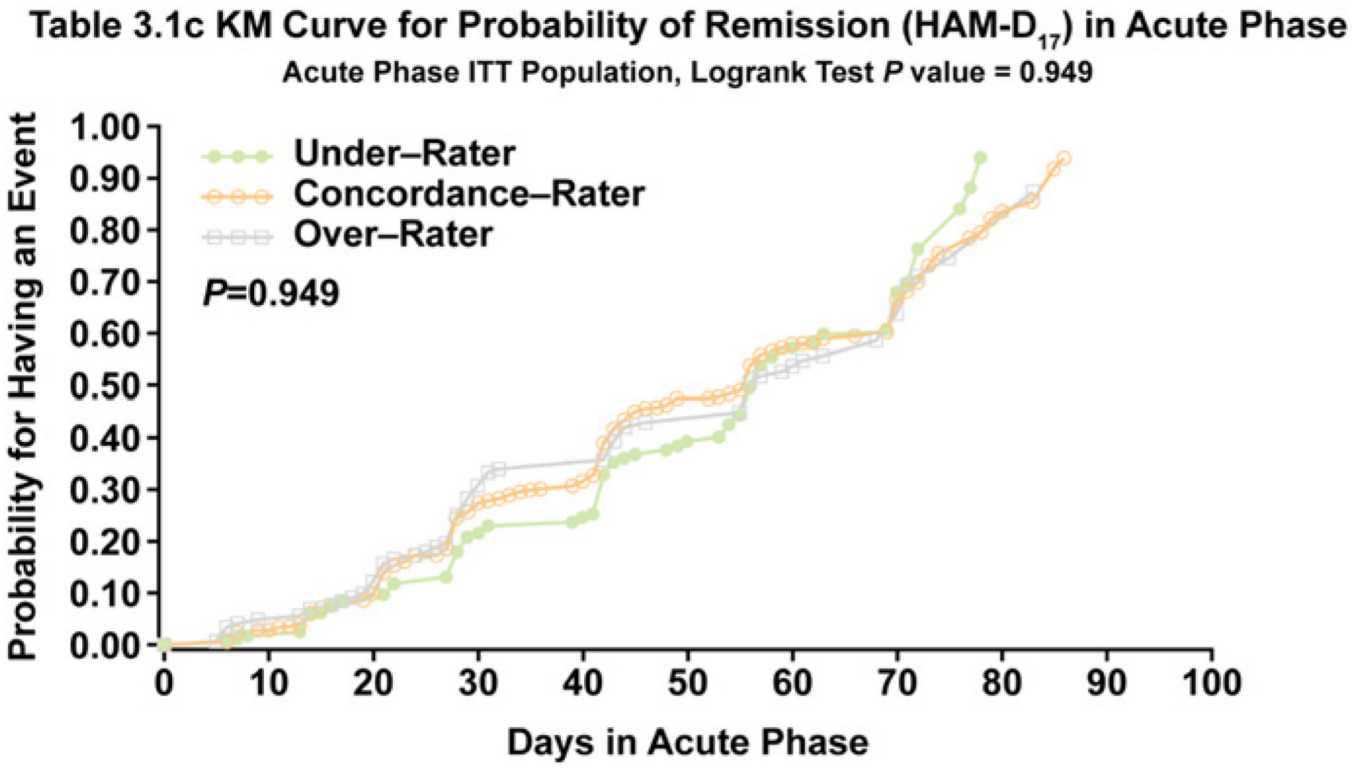

Fig. 2.
Time to HAM-D_{17} Remission During Acute Treatment for Concordance, Underrating, and Overrating Patients.

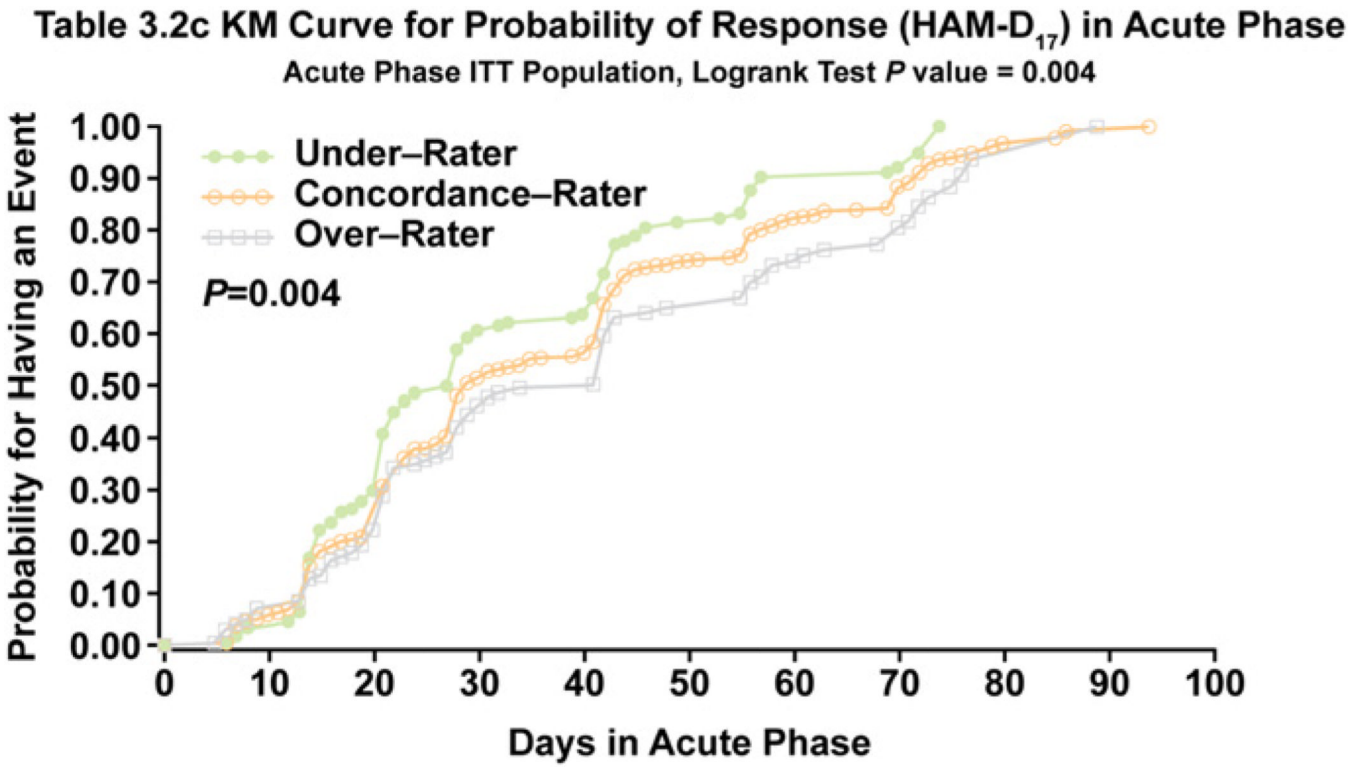

Fig. 3.

Time to HAM-D_{17} Response During Acute Treatment for Concordance, Underrating, and Overrating Patients.

Table 1

Baseline predictors of more severe patient- vs clinician-rated depression severity during acute and continuation treatment.

	Odds Ratio (95% CI)			
	Age	**Body mass index**	**Male vs female**	**Lifetime MDD episodes**
Acute Phase				
Baseline	0.96 (0.94–0.98) ‡	1.09 (1.05–1.13)§	0.54 (0.32–0.90) *	1.01 (1.00–1.02) *
Week 1	0.97 (0.95–0.99) †	1.06 (1.02–1.10) †	0.45 (0.27–0.76) †	1.00 (0.99–1.01)
Week 2	0.98 (0.96–1.00)	1.05 (1.02–1.09) †	0.47 (0.28–0.79) †	1.01 (1.00–1.02) *
Week 3	1.00 (0.98–1.02)	1.05 (1.01–1.09) †	0.76 (0.44–1.29)	1.00 (0.99–1.01)
Week 4	1.00 (0.98–1.03)	1.07 (1.03–1.11) ‡	0.42 (0.24–0.74) †	1.00 (1.00–1.01)
Week 6	1.01 (0.99–1.03)	1.06 (1.02–1.10) †	0.46 (0.25–0.82) †	1.00 (0.99–1.01)
Week 8	1.02 (1.00–1.04)	1.03 (0.99–1.07)	0.67 (0.38–1.17)	1.00 (0.99–1.01)
Week 10	1.02 (1.00–1.05)	1.02 (0.98–1.06)	0.53 (0.29–0.98) *	1.00 (0.99–1.01)
Continuation Phase				
Month 1	1.03 (1.00–1.06) *	1.02 (0.98–1.07)	0.59 (0.29–1.20)	1.01 (1.00–1.02)
Month 2	1.02 (0.99–1.04)	0.99 (0.95–1.04)	0.67 (0.33–1.38)	1.01 (1.00–1.02)
Month 3	1.03 (1.00–1.06)	1.00 (0.95–1.05)	1.52 (0.71–3.29)	1.01 (0.99–1.02)
Month 4	1.02 (0.99–1.06)	1.02 (0.97–1.08)	1.16 (0.53–2.52)	1.01 (0.99–1.02)
Month 5	1.04 (1.00–1.07) *	1.00 (0.95–1.05)	0.98 (0.43–2.26)	1.00 (0.98–1.01)
Month 6	1.06 (1.02–1.09) ‡	1.02 (0.97–1.08)	0.99 (0.43–2.28)	1.01 (0.99–1.02)

CI = confidence interval; MDD = major depressive disorder.

* $P<0.05$;

† $P\leq 0.01$;

‡ $P< 0.001$;

§ $P< 0.0001$.

Table 2

Summary of remission or response by baseline patient-clinician assessment discrepancy for acute phase ITT population.

Patient Category at Acute Phase Baseline	Remission						Response					
	HAM-D_{17}		CGI-S		IDS-SR		HAM-D_{17}		CGI-C		IDS-SR	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Underrating	75 (46)	89 (54)	97 (59)	67 (41)	74 (45)	90 (55)	121 (74)	43 (26)	125 (76)	39 (24)	89 (54)	75 (46)
Concordant	361 (51)	353 (49)	425 (60)	289 (40)	294 (41)	419 (59)	528 (74)	186 (26)	545 (76)	169 (24)	418 (59)	295 (41)
Overrating	67 (45)	81 (55)	81 (55)	67 (45)	51 (35)	97 (66)	102 (69)	46 (31)	108 (73)	40 (27)	85 (57)	63 (42)

CGI-C = Clinical Global Impressions–Change; CGI-S = Clinical Global Impressions–Severity; HAM-D_{17} = 17-item Hamilton Rating Scale for Depression; IDS-SR = Inventory of Depressive Symptomatology–Self-Report; ITT = intent to treat.

Remission is defined as: HAM-$D_{17} \leq 7$; CGI-S ≤2; IDS-SR ≤ 14.

Response is defined as: HAM-D_{17} decrease ≥ 50% from acute phase baseline; CGI-C ≤ 2; IDS-SR decrease ≥ 50% or more from acute phase baseline.

All Mantel-Haenszel chi-square test *P*-values > .05 (not significant).

Table 3

Remission or response in continuation phase by week 10 clinician-patient assessment discrepancy.

Visit	Patient Category at Week 10	Remission						Response					
		HAM-D_{17}		CGI-S		IDS-SR		HAM-D_{17}		CGI-C		IDS-SR	
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
		n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Month 6	Concordant	275 (83)	56† (17)	299 (91)	31† (9)	234 (71)	97† (29)	314 (95)	17 (5)	299 (91)	31† (9)	277 (84)	54‡ (16)
	Nonconcordant	71 (68)	33 (32)	82 (79)	22 (21)	59 (57)	45 (43)	97 (93)	7 (7)	82 (79)	22 (21)	71 (68)	33 (32)
Final Visit	Concordant	379 (76)	123‡ (24)	415 (83)	87‡ (17)	321 (64)	180† (36)	451 (90)	51* (10)	415 (83)	87† (17)	389 (78)	112‡ (22)
	Nonconcordant	104 (59)	72 (41)	124 (70)	52 (30)	88 (50)	87 (50)	146 (83)	30 (17)	124 (70)	52 (30)	107 (61)	68 (39)

CGI-S = Clinical Global Impression—Severity; HAM-D_{17}= 17-item Hamilton Rating Scale for Depression; IDS-SR = Inventory of Depressive Symptomatology–Self-Report; ITT = intent to treat.

Remission is defined as: HAM-D_{17}≤ 7; CGI-S ≤2; IDS-SR ≤ 14.

Response is defined as: HAM-D_{17} decrease ≥50% from acute phase baseline; CGI-C ≤ 2; IDS-SR decrease ≥50% or more from acute phase baseline.

**P*≤ 0.05 from Mantel-Haenszel chi-square test concordance-rater vs nonconcordance.

†*P*≤ 0.01 from Mantel-Haenszel chi-square test concordance-rater vs nonconcordance.

‡*P*< 0.001 from Mantel-Haenszel chi-square test concordance-rater vs nonconcordance.

Table 4

Remission or response in continuation phase by Week 10 clinician-patient assessment discrepancy.

Visit	Patient Category at Week 10	Remission						Response					
		HAM-D_{17}		CGI-S		IDS-SR		HAM-D_{17}		CGI-C		IDS-SR	
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
		n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Month 6	Underrating	34 (74)	12 (26)	38 (83)	8 (17)	39 (85)	7‡ (15)	40 (87)	6* (13)	38 (83)	8 (17)	39 (85)	7‡ (15)
	Concordant	275 (83)	56 (17)	299 (91)	31 (9)	234 (71)	97 (29)	314 (95)	17 (5)	299 (91)	31 (9)	277 (84)	54 (16)
	Overrating	37 (64)	21 (36)	44 (76)	14 (24)	20 (34)	38 (66)	57 (98)	1 (2)	44 (76)	14 (24)	32 (55)	26 (45)
Final Visit	Underrating	44 (60)	30 (40)	55 (74)	19 (26)	56 (76)	18‡ (24)	58 (78)	16 (22)	55 (74)	19 (26)	57 (77)	17‡ (23)
	Concordant	379 (76)	123 (24)	415 (83)	87 (17)	321 (64)	180 (36)	451 (90)	51 (10)	415 (83)	87 (17)	389 (78)	112 (22)
	Overrating	60 (59)	42 (41)	69 (68)	33 (32)	32 (32)	69 (68)	88 (86)	14 (14)	69 (68)	33 (32)	50 (50)	51 (50)

CGI-S = Clinical Global Impressions–Severity; HAM-D_{17}= 17-item Hamilton Rating Scale for Depression; IDS-SR = Inventory of Depressive Symptomatology – Self-Report.

Remission is defined as: HAM-D_{17}≤ 7; CGI-S ≤2; IDS-SR ≤ 14.

Response is defined as: HAM-D_{17} decrease >50% from acute phase baseline; CGI-C ≤2; IDS-SR decrease ≥50% or more from acute phase baseline.

**P*≤ 0.05 from 3-way Mantel-Haenszel chi-square test.

†*P*≤ 0.01 from 3-way Mantel-Haenszel chi-square test.

‡*P*≤ 0.001 from 3-way Mantel-Haenszel chi-square test.

Table 5

Changes in patient-clinician assessment discrepancy from baseline through the completion of the 6-Month continuation phase.

Time Point	Baseline Patient Category		
	Underrating	**Concordant**	**Overrating**
Acute Phase (n, % of acute phase population)			
Baseline	164 (100.0)	714 (100.0)	148 (100.0)
Discontinued during acute phase	47 (28.7)	162 (22.7)	48 (32.4)
Week 10 completers	117 (71.3)	552 (77.3)	100 (67.6)
Rater Category at Week 10 (n, % of week 10 completers)			
Underraters	24 (20.5)	72 (13.0)	8 (8.0)
Concordance raters	84 (71.8)	406 (73.6)	68 (68.0)
Overraters	9 (7.7)	74 (13.4)	24 (24.0)
Continuation Phase (n, % of continuation phase population)			
Entered continuation phase	72 (100.0)	517 (100.0)	104 (100.0)
Discontinued during continuation phase	27 (37.5)	194 (37.5)	46 (44.2)
Month 6 completers	45 (62.5)	323 (62.5)	58 (55.8)
Rater Category at Month 6 (n, % of month 6 completers)			
Underraters	17 (37.8)	41 (12.7)	3 (5.2)
Concordance raters	27 (60.0)	247 (76.5)	30 (51.7)
Overraters	1 (2.2)	35 (10.8)	25 (43.1)