Journal of Affective Disorders 130 (2011) 171–179

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research report

Clinical value of early partial symptomatic improvement in the prediction of response and remission during short-term treatment trials in 3369 subjects with bipolar I or II depression ☆

David E. Kemp [a,*], Stephen J. Ganocy [a], Martin Brecher [b], Berit X. Carlson [c], Suzanne Edwards [d], James M. Eudicone [c], Gary Evoniuk [d], Wim Jansen [e], Andrew C. Leon [f], Margaret Minkwitz [g], Andrei Pikalov [h], Hans H. Stassen [i], Armin Szegedi [j], Mauricio Tohen [k], Arjen P.P. Van Willigenburg [l], Joseph R. Calabrese [a]

[a] Case Western Reserve University, University Hospitals Case Medical Center, Cleveland, OH, USA
[b] SK Life Science, Fair Lawn, NJ, USA
[c] Bristol-Myers Squibb Co., Plainsboro, NJ, USA
[d] GlaxoSmithKline, NC, USA
[e] Schering-Plough, The Netherlands
[f] Weill Cornell Medical College, New York, NY, USA
[g] AstraZeneca Wilmington, DE, USA
[h] Dainippon Sumitomo Pharma America, Inc., Fort Lee, NJ, USA
[i] Psychiatric University Hospital Zurich, Switzerland
[j] Merck Research Laboratories, Summit, NJ, USA
[k] University of Texas Health Science Center, San Antonio, Texas, USA
[l] Complete Healthcare Communications Europe, The Netherlands

ARTICLE INFO

Article history:
Received 26 September 2010
Accepted 12 October 2010
Available online 10 November 2010

Keywords:
Bipolar disorder
Early improvement
Bipolar depression
Onset of action
Operating characteristics
Aripiprazole
Lamotrigine
Olanzapine-fluoxetine combination
Quetiapine

ABSTRACT

Objective: To evaluate the clinical value of early partial symptomatic improvement in predicting the probability of response during the short-term treatment of bipolar depression.
Methods: Blinded data from 10 multicenter, randomized, double-blind, placebo-controlled trials in bipolar I or II depression were used to determine if early improvement (≥20% reduction in depression symptom severity after 14 days of treatment) predicted later short-term response or remission. Sensitivity, specificity, efficiency, and positive and negative predictive values (PPV, NPV) were calculated using an intent to treat analysis of individual and pooled study data.
Results: 1913 patients were randomized to active compounds (aripiprazole, lamotrigine, olanzapine/olanzapine–fluoxetine, and quetiapine), and 1456 to placebo. In the pooled positive studies, early improvement predicted response and remission with high sensitivity (86% and 88%, respectively), but rates of false positives were high (53% and 59%, respectively). Pooled negative predictive values for response/remission (i.e. confidence in knowing the drug will not result in response or remission) were 74% and 82%, respectively, with low rates of false negatives (14% and 12%, respectively).
Conclusion: Early improvement in an individual patient does not appear to be a reliable predictor of eventual response or remission due to an unacceptably high false positive rate. However, the absence of early improvement appears to be a highly reliable predictor of eventual non-response,

☆ Previous presentation: Part of these data has been presented previously at the 161st Annual Meeting of the American Psychiatric Association, May 3–8, 2008, Washington, DC.

* Corresponding author. 10524 Euclid Ave., 12th Floor, Cleveland, OH 44106, USA. Tel.: +1 216 844 2865; fax: +1 216 844 2875.
E-mail address: kemp.david@gmail.com (D.E. Kemp).

0165-0327/$ – see front matter
doi:10.1016/j.jad.2010.10.026

suggesting that clinicians can have confidence in knowing when a drug is not going to work during short-term treatment. Patients who fail to demonstrate early improvement within the first two weeks of treatment may benefit from a change in therapy.

1. Introduction

Depressive symptoms dominate the lifetime symptom course for most individuals with bipolar disorder, occurring 3–4 times more frequently than manic symptoms (Kupka et al., 2007; Post et al., 2003; Judd et al., 2002). Persistent depressive symptoms are associated with high rates of disability, functional and occupational deficits (Kessler et al., 2006; Judd et al., 2005; Calabrese et al., 2004; Bauer et al., 2001), greater risk for recurrence (Perlis et al., 2006), and an increased rate of suicide (Calabrese et al., 2004).

When using traditional antidepressant medications, the conventional belief is that true antidepressant effects (in contrast to nonspecific effects of treatment or a placebo response) do not occur until after several weeks of treatment (Gelenberg and Chesen, 2000). This belief originally stemmed from the use of pattern analysis in the 1980s, in which more patients receiving active medication than placebo displayed treatment response patterns characterized by a delay of 3 weeks or more in the onset of initial improvement (Quitkin et al., 1987, 1984). These findings led to the belief that trials of antidepressants required 4–6 weeks of exposure, a practice that may contribute to treatment non-adherence for individuals experiencing little to no benefit shortly after antidepressant initiation. However, the conclusions derived from use of pattern analysis methodology have been challenged by other investigators who found evidence for a gradual alleviation of depressive symptoms in a step-wise fashion that begins as early as 1–2 weeks after starting antidepressant treatment (Katz et al., 1997, 1996–1997). Survival analytic techniques have also been applied to short-term randomized controlled trials of major depressive disorder by Stassen and colleagues (Stassen et al., 1999, 1996, 1993) and suggest that the conditional probability of achieving response or remission in patients experiencing early improvement is high.

More recently, a published meta-analysis of 47 double-blind, placebo-controlled antidepressant trials in major depressive disorder concluded that benefit does occur within the first 2 weeks of treatment (Posternak and Zimmerman, 2005). A separate meta-analysis of selective serotonin reuptake inhibitor (SSRI) antidepressants confirmed that an early response within the first week of SSRI administration is not necessarily a placebo response (Taylor et al., 2006).

With the establishment that true onset of antidepressant action occurs early in the treatment course, it was hypothesized that clinicians may be able to make rational treatment decisions after as little as two weeks of treatment (Szegedi et al., 2003; Nierenberg et al., 1995). In a randomized, controlled trial comparing mirtazapine and paroxetine in patients with major depression, Szegedi et al. (2003) reported that an improvement of only 20% in depression severity within the first 2 weeks of treatment is a clinically useful predictor of later outcome. Early improvement has also been shown to predict stable response and stable remission (i.e. response or remission criteria met at Week 4 of treatment and at all subsequent assessments) with high sensitivity. Even with cognitive behavioral therapy, early improvement has been found to be a highly sensitive predictor of later response in patients with mild major, minor and subsyndromal depression (Tadić et al., 2010). Perhaps the most clinically meaningful finding to arise from these studies is the observation that patients who fail to achieve early improvement within the first 2 weeks will have little chance of achieving response or remission with continued treatment (Szegedi et al., 2009, 2003).

To our knowledge, the predictive value of early improvement has never been studied in bipolar depression. It is problematic to extrapolate findings from unipolar depression to bipolar depression, given the differences in phenomenology, illness course, and response to treatment (Muzina et al., 2007; Akiskal, 2005). For instance, though standard antidepressants have proven to be efficacious in treating unipolar depression, their efficacy in bipolar depression is inconclusive, with one large, randomized, placebo-controlled effectiveness study finding antidepressants to confer no additional benefit when added to a mood stabilizer compared with a mood stabilizer alone (Sachs et al., 2007). Given that depression is the leading cause of morbidity and mortality in patients with bipolar disorder, if the premise were true that early improvement is a sensitive indicator of later outcome, there would be great clinical value in being able to predict treatment non-response as early as possible so that alternative treatments could be instituted. A factor complicating the exploration of early improvement in bipolar depression is the wide array of medications with diverse mechanisms of action that are prescribed for the treatment of acute depressive episodes. The current study therefore assessed whether early improvement is predictive of short-term treatment outcome in bipolar depression across a variety of different pharmacologic mechanisms and examined whether the predictive accuracy differs by the individual drug treatment. We hypothesized that (1) early improvement would be a sensitive predictor of later short-term response and remission across all antidepressant treatments and that (2) patients who did not exhibit early improvement would be unlikely to respond or remit to treatment 7–10 weeks later.

2. Method

2.1. Inclusion of studies

A pooled analysis was conducted on the individual drug data from 10, multi-center, randomized, double-blind, placebo-controlled trials in 3369 patients with bipolar I or II disorder experiencing a major depressive episode. At the time this analysis was planned, these 10 studies represented the extent of available large-scale clinical trials undertaken in bipolar depression. The manufacturers of each respective compound were contacted for participation and each agreed to release the requested data for an independent analysis. The raw data for each trial were individually submitted by the study sponsors to a data management infrastructure within the bipolar disorders research

center at Case Western Reserve University for analysis. Each of the studies was approved by the institutional review board of the participating site. Table 1 describes the study designs of all trials included in the analysis.

In summary, the following compounds were evaluated:

1. Aripiprazole; 2 studies of bipolar I depression conducted over 8 weeks (Thase et al., 2008).
2. Lamotrigine; 1 study of bipolar I depression conducted over 7 weeks (Calabrese et al., 1999); 2 studies of bipolar I depression conducted over 8 weeks (Calabrese et al., 2008); 1 study of bipolar I and II depression conducted over 10 weeks (Calabrese et al., 2008); 1 study of bipolar II depression conducted over 8 weeks (Calabrese et al., 2008).
3. Olanzapine and olanzapine–fluoxetine combination (OFC); 1 study of bipolar I depression conducted over 8 weeks (Tohen et al., 2003) [olanzapine and OFC data were pooled];
4. Quetiapine; 2 studies of bipolar I and II depression conducted over 8 weeks (Thase et al., 2006; Calabrese et al., 2005).

2.2. Statistical analyses

The mean change from baseline to study endpoint (last observation carried forward; LOCF) in the Montgomery-Asberg Depression Rating Scale (MADRS) total score was the primary or key secondary endpoint and available for analysis in all studies (Montgomery and Asberg, 1979). The LOCF principle was applied in determining outcome status in order to account for the negative effects of treatment discontinuation due to lack of efficacy or loss of previous symptom improvement. An individual and combined analysis of the predictive effect of early improvement on endpoint response and remission was performed on all active agents and their corresponding placebo arms. However, active drug did not separate from placebo in all of the studies. Therefore, additional comparisons were made across the following four groups: agents in trials that separated from placebo on the MADRS (4 positive studies), agents in trials that failed to separate from placebo (6 negative/failed studies) and the corresponding placebo arms for the positive and negative/failed studies.

The primary aim of these post-hoc analyses was to determine whether early improvement to a treatment for depressive episodes in bipolar disorder predicts later response or remission at acute study endpoint (7–10 weeks). Early improvement was defined by ≥20% reduction from baseline in the MADRS Total Score at Week 2. An improvement of ≥20% after the first 2 weeks of treatment has been identified as the most accurate cut-off for predicting later response or remission in major depression and schizophrenia, as it balances sensitivity and specificity while optimizing the negative predictive value for later response/remission (Kemp et al., 2010; Kinon et al., 2008; Szegedi et al., 2003). Response at study endpoint was defined as ≥50% reduction in baseline MADRS Total Score, and remission was defined as MADRS Total Score ≤10 at endpoint.

2.3. Operating characteristics

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and efficiency were calculated. For those instances in which studies with different study drugs were pooled, a weighted analysis using LOCF was conducted. In this case, weighting was performed using the number of patients contributed per study drug. For example, if Drug A

Table 1
Summary of randomized, double-blind, placebo-controlled trials of bipolar I or II depression.

Drug and author	Study duration	Bipolar subtype	Symptom threshold entrance criteria*	Dose (daily)	Number randomized	Percentage response/remission
Studies separating from placebo						
Lamotrigine Calabrese et al. (1999)	7 weeks	Type I	Minimum score ≥18 on 17-item HAM-D	50 mg 200 mg	LTG: 129 PBO: 66	LTG 50 mg: 48.0/37.5 LTG 200 mg: 54.0/46.0 PBO: 29.0/21.5
Olanzapine–fluoxetine Tohen et al. (2003)	8 weeks	Type 1	Minimum score ≥20 on MADRS	OLZ: 5–20 mg OFC: 6/25–12/50 mg	OLZ:370 OFC: 86 PBO:377	OLZ: 39.0/32.8 OFC: 56.1/48.8 PBO: 34.0/24.5
Quetiapine Thase et al. (2006)	8 weeks	Type I and II	Minimum score ≥20 on 17-item HAM-D	300 mg 600 mg	QUE: 341 PBO: 168	QUE 300 mg: 60.0/51.6 QUE: 600 mg: 58.3/52.3 PBO: 44.7/37.3
Quetiapine Calabrese et al. (2005)	8 weeks	Type I and II	Minimum score ≥20 on 17-item HAM-D	300 mg 600 mg	QUE: 361 PBO: 181	QUE: 300 mg: 58.0/52.9 QUE: 600 mg: 58.0/52.9 PBO: 36.1/28.4
Studies failing to separate from placebo						
Aripiprazole Thase et al. (2008)	8 weeks	Type I	Minimum score ≥18 on 17-item HAM-D	5-30 mg	ARI: 186 PBO: 188	ARI: 43.2/30.2 PBO: 39.0/27.8
Aripiprazole Thase et al. (2008)	8 weeks	Type I	Minimum score ≥18 on 17-item HAM-D	5–30 mg	ARI: 187 PBO: 188	ARI: 44.6/25.7 PBO: 44.3/29.0
Lamotrigine Calabrese et al. (2008)	10 weeks	Type I and II	Minimum score ≥18 on 17-item HAM-D	100–400 mg	LTG: 103 PBO: 103	LTG: 45.1/NA PBO: 49.0/NA
Lamotrigine Calabrese et al. (2008)	8 weeks	Type I	Minimum score ≥18 on 17-item HAM-D	200 mg	LTG: 133 PBO: 124	LTG: 46.0/NA PBO: 39.3/NA
Lamotrigine Calabrese et al. (2008)	8 weeks	Type II	Minimum score ≥18 on 17-item HAM-D	200 mg	LTG: 111 PBO: 110	LTG: 54.1/NA PBO: 45.7/NA
Lamotrigine Calabrese et al. (2008)	8 weeks	Type I	Minimum score ≥18 on 17-item HAM-D	200 mg	LTG:131 PBO: 128	LTG: 45.5/NA PBO: 44.0/NA

contributed 100 individuals, Drug B contributed 200 individuals, and Drug C contributed 500 individuals to the calculation of sensitivity, then the final value for sensitivity was a combination of 12.5% of Drug A's sensitivity plus 25% of Drug B's sensitivity plus 62.5% of Drug C's sensitivity. Weighting by individual study drug and unweighted analyses yielded similar findings.

Sensitivity is calculated by dividing the number of endpoint responders/remitters who demonstrated early improvement by the total number of endpoint responders/remitters. Specificity is calculated by dividing the number of endpoint non-responders/non-remitters who did not demonstrate early improvement by the total number of non-responders/non-remitters at study endpoint.

The PPV represents the probability that patients will achieve a response if they show early improvement. It may be thought of as "knowing that the drug is working". It is calculated as the number of true positives (patients who showed early improvement *and* responded at endpoint) divided by the number of patients categorized as positive (all patients who showed early improvement). The NPV represents the probability that a patient will not achieve a response if they do not show early improvement. It may be thought of as "knowing that the drug is not working". It is calculated as the number of true negatives (number of patients who did not show an early improvement *and* did not respond at endpoint) divided by the total number of patients categorized as negative (all patients who did not show early improvement). Efficiency represents the proportion of patients for which the early improvement metric correctly identified the response or remission status at study endpoint.

It should be noted that the operating characteristics just described do not adjust for chance agreement. In order to adjust for chance, including the varying rates of response and remission among the individual drug trials, three separate chance-corrected operating characteristics were calculated (Kraemer, 1992). These included a chance-corrected measure of sensitivity, $\kappa_{(1,0)}$, a chance-corrected measure of specificity, $\kappa_{(0,0)}$, and Cohen's κ (Cohen, 1960), which is a chance-corrected measure of efficiency. The interpretation of chance-corrected operating characteristics has been outlined by Landis and Koch (1977), where it is suggested that 0.6 to 0.8 is "substantial", 0.4 to 0.6 is "moderate", and 0.2 to 0.4 is "fair".

3. Results

Across the 10 studies, 1913 patients were randomized to active drug and 1456 patients were randomized to placebo. The results of individual trials with regards to efficacy, reasons for dropout, and tolerability have been published in detail elsewhere (Calabrese et al., 2008, 2005, 1999; Thase et al., 2008, 2006; Tohen et al., 2003).

3.1. Rates of early improvement

With early improvement defined as ≥20% reduction from baseline in MADRS Total Score at Week 2, 60% of patients on active compounds (n = 1155) fulfilled this criterion compared to 47% of patients receiving placebo (n = 683). Pooled positive studies had broadly similar rates of early improvement (lamotrigine, 54%; olanzapine and OFC, 69%; and quetiapine, 75%;) as pooled negative studies (aripiprazole, 68%; and lamotrigine 44%) and placebo from pooled positive (54%) and negative (44%) studies (see Fig. 1).

3.2. Predictive value of early improvement

Table 2 presents sensitivity, specificity, PPV and NPV for prediction of response and remission. Early improvement predicted later response (≥50% reduction in MADRS Total Score at study endpoint) with moderately high sensitivity across all studies (75%), indicating that patients who achieve response or remission will typically demonstrate ≥20% improvement within the first 2 weeks of treatment. This finding was similar in examination of the pooled positive studies (77–88%), and negative/failed studies (63–81%), including the corresponding placebo arms (69–80%). The sensitivity was lowest among the four pooled negative studies of lamotrigine (63%). Similarly, early improvement predicted later remission (MADRS Total Score ≤10) with a moderately high sensitivity for all combinations of pooled study arms (63–89%), with the sensitivity lowest in the pooled negative studies of lamotrigine (63%). Additionally, the chance-corrected sensitivity characteristics for response and remission were moderate for pooled positive trials (52%/58%) but only fair for pooled negative trials (35%/37%). Similar to the unadjusted values, the chance corrected sensitivity was lowest in the pooled negative studies of lamotrigine.

The rate of false negatives was variably low across pooled positive and negative studies. The rate of false positives for response and remission was high across all active study arms (28–64%), but was numerically highest in the pooled positive studies of olanzapine/OFC, pooled positive studies of quetiapine, and the pooled negative studies of aripiprazole. A false positive occurs when a patient demonstrates early improvement but does not go on to respond or remit to treatment. The pooled analysis of all active agents demonstrated a nominal specificity of 57% for prediction of response and 53% for prediction of remission. Specificity values were similar for analysis of pooled positive and negative studies. Across all trials, early improvement was more sensitive than specific for predicting later response or remission as indicated by the smaller chance-corrected specificity (28%/16%) values in comparison with the chance-corrected measurements of sensitivity (38%/32%). Chance-corrected specificity values consistently demonstrated that the specificity for early improvement to predict later response was greater than the specificity to predict later remission. Chance-corrected specificity did not differ appreciably between pooled positive (25%/17%) and pooled negative (26%/18%) studies.

Approximately 68% (range 54–72% across drugs) of individuals who exhibited early improvement to an active drug went on to have a positive response at study endpoint. The probability of achieving remission in those with early improvement was 56% across all drugs studied (ranging from 44 to 62%). A high NPV for response and remission was consistently observed for pooled positive and negative trials across each compound studied (68–86%). Pooled positive studies demonstrated similar NPVs for response/remission (lamotrigine, 74%/83%; olanzapine and OFC, 77%/86%; quetiapine, 71%/79%) as pooled negative studies (aripiprazole, 74%/81%; lamotrigine 68%/76%) and the placebo arm from pooled positive (83%/90%)

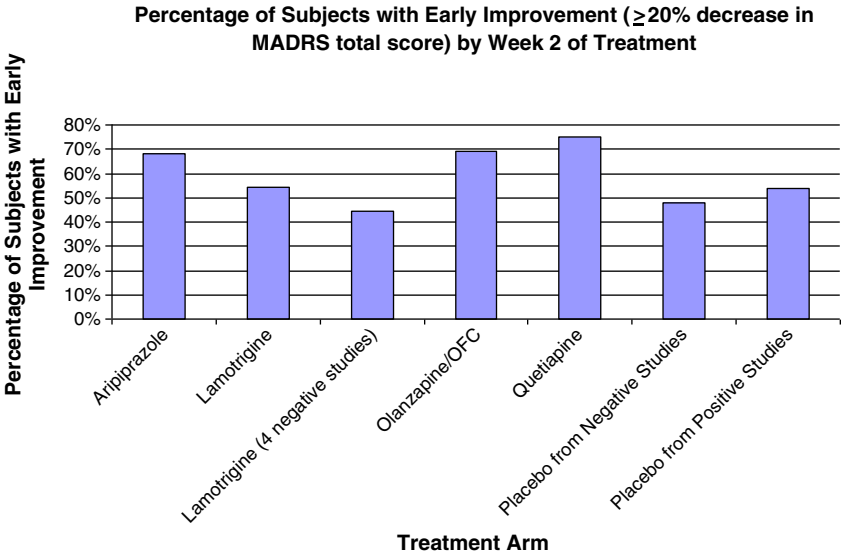

Fig. 1. MADRS = Montgomery-Asberg Depression Rating Scale; OFC = olanzapine–fluoxetine combination.

and negative (74%/80%) studies. The NPV's indicate that patients who fail to demonstrate at least a 20% reduction in depressive severity after 2 weeks of treatment were unlikely to meet response or remission criteria later in the course of treatment.

As reflected by the Cohen's κ ((.5,0)), early improvement predicted later response with greater chance-corrected efficiency (23–46%) than it predicted later remission (20–44%). The chance-corrected efficiency for all trials was fair, with the exception of the positive lamotrigine study, for which it reached a moderate level.

Fig. 2 illustrates rates of response at study endpoint by early improvement status for the pooled studies of aripiprazole, lamotrigine, olanzapine/olanzapine–fluoxetine, and quetiapine.

4. Discussion

This article represents a comprehensive analysis of the clinical utility of early partial symptomatic improvement to predict later outcomes in the acute treatment of bipolar depression. Building upon prior observations of early improvement in major depressive disorder and schizophrenia, this study provides for the first time an estimate of the predictive value of early improvement in bipolar depression by pooling data from 10 placebo-controlled trials of four different agents, representing a total of 3369 bipolar patients who received acute treatment for a major depressive episode.

Improvement was observed in at least half of patients within the first 2 weeks of treatment with all four compounds

Table 2
Operating characteristics of early improvement* in MADRS total score to predict later response or remission.

Arm	Outcome	Prevalence	Sensitivity	Specificity	PPV	NPV	False Positives	False Negatives	Efficiency	κ(1,0)	κ(0,0)	κ(.5,0)
All Studies Active	Resp/Rem	.52/.39	.75/.70	.57/.53	.68/.56	.72/.80	.43/.47	.25/.30	.66/.60	.38/.32	.28/.16	.32/.21
All Studies Placebo	Resp/Rem	.45/.36	.67/.62	.70/.65	.64/.50	.72/.75	.30/.35	.33/.38	.69/.64	.38/.31	.36/.22	.37/.26
Pooled Positive Studies	Resp/Rem	.55/.42	.86/.88	.47/.41	.66/.52	.74/.82	.53/.59	.14/.12	.68/.61	.52/.58	.25/.17	.34/.26
Pooled Negative Studies	Resp/Rem	.47/.36	.70/.71	.60/.56	.61/.47	.70/.77	.40/.44	.30/.29	.65/.61	.35/.37	.26/.18	.30/.24
Positive Studies												
Lamotrigine (1)	Resp/Rem	.51/.42	.77/.81	.69/.65	.72/.62	.74/.83	.31/.35	.23/.19	.73/.72	.49/.59	.43/.35	.46/.44
Olanzapine/OFC (1)	Resp/Rem	.50/.34	.86/.88	.48/.41	.62/.44	.77/.86	.52/.59	.14/.13	.67/.57	.55/.61	.25/.15	.34/.23
Quetiapine (2)	Resp/Rem	.58/.46	.88/.89	.42/.36	.68/.54	.71/.79	.58/.64	.32/.46	.69/.60	.51/.55	.23/.15	.32/.24
Negative Studies												
Aripiprazole (2)	Resp/Rem	.54/.36	.81/.83	.43/.41	.54/.44	.74/.81	.57/.60	.19/.17	.60/.56	.41/.47	.16/.13	.23/.20
Lamotrigine (4)	Resp/Rem	.48/.35	.63/.63	.72/.66	.67/.50	.68/.76	.28/.34	.37/.37	.68/.65	.33/.34	.37/.23	.35/.27
Placebo Arms Pooled												
Placebo Pooled Positive (4)	Resp/Rem	.38/.25	.80/.81	.62/.56	.56/.38	.83/.90	.38/.44	.20/.19	.69/.62	.57/.59	.29/.18	.39/.27
Placebo Pooled Negative (6)	Resp/Rem	.44/.34	.69/.70	.68/.64	.63/.50	.74/.80	.32/.36	.31/.30	.68/.66	.40/.43	.34/.24	.37/.31

*Improvement of ≥20% reduction in MADRS Total Score at Week 2.
MADRS = Montgomery-Asberg Depression Rating Scale; NPV = negative predictive value; OFC = olanzapine–fluoxetine combination; PPV = positive predictive value. Rem = Remission (MADRS Total score ≤ 10); Resp = Response (≥50% reduction in MADRS Total socre); κ(1,0) = chance-corrected sensitivity; κ(0,0) = chance-corrected specificity; κ(.5,0) = chance corrected efficiency.

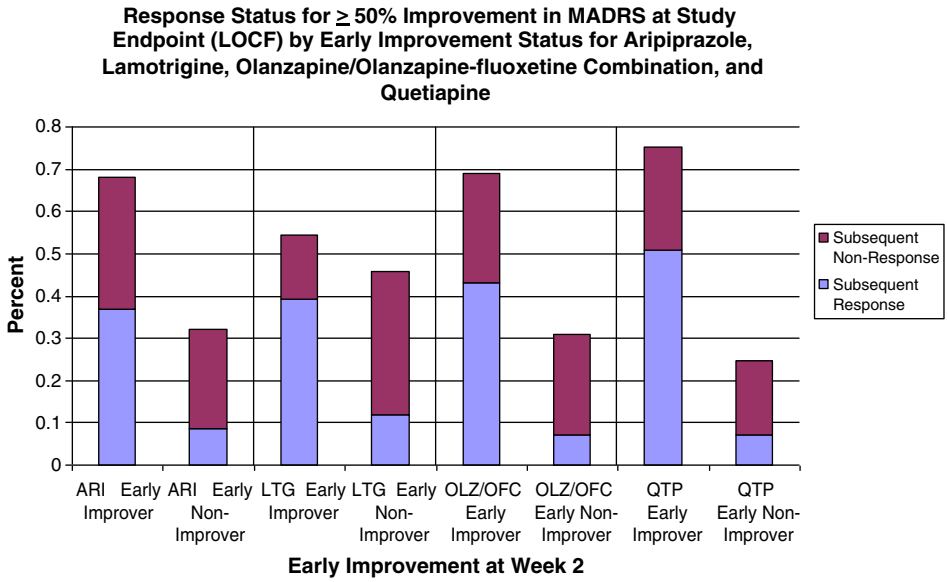

Fig. 2. Early improvement defined as ≥20% reduction in MADRS Total Score at Week 2. ARI = aripiprazole; LOCF = last observation carried forward; LTG = lamotrigine; MADRS = Montgomery-Asberg Depression Rating Scale; OFC = olanzapine–fluoxetine combination; OLZ = olanzapine; QTP = Quetiapine.

studied in bipolar depression. Overall, four distinct findings emerged in our analysis, each of which was consistent with a large meta-analysis of early improvement involving unimodal antidepressants in major depressive disorder (Szegedi et al., 2009). First, more than half of patients demonstrated at least 20% improvement in MADRS Total Score by Week 2. Second, the NPVs for response and remission were high, indicating that patients who fail to demonstrate at least a 20% reduction in depressive severity after 2 weeks of treatment are unlikely to meet response or remission criteria later in the course of treatment. Third, early improvement predicted later response or remission with high sensitivity, indicating that the large majority of patients who ultimately achieve response or remission will display measurable improvement within the first 2 weeks of treatment. Fourth, a high rate of false positives was observed across all studies, indicating that early improvement in bipolar depression may not be stable or enduring, and cannot be reliably used to predict later response or remission to acute treatment.

Of the above findings, the most clinically useful predictor to emerge from this analysis may be the reliably high NPVs. This indicates that the absence of early improvement during the first 2 weeks of treatment is a highly reliable predictor of subsequent lack of response and remission at study endpoint. Applied to usual practice, these data suggest that clinicians could confidently initiate a medication change after 2 weeks of treatment in the absence of early improvement. Such a strategy diverges from conventional practice in which trials are prolonged for as many as 6–12 weeks to optimize any potential for response. The ability to predict outcome early in the course of treatment may shorten the length of an unsuccessful treatment trial, and has the potential to decrease morbidity and risk of suicide from protracted depressive symptoms (Szegedi et al., 2003). It should be pointed out that prospective comparisons are still needed to determine if patients without early improvement will benefit from altering the treatment regimen compared with those who continue the same treatment.

The predictor of at least 20% improvement within the first 2 weeks was chosen because it has been reported to be a clinically meaningful change that is easily recognizable by clinicians. It was also found to be a reliable and sensitive indicator for sustained improvement in studies of unipolar depression. Overall, the predictive value of early improvement appears less consistent and less robust in trials of bipolar depression compared to unipolar major depression. In this bipolar cohort, early improvement predicted later response and remission with high sensitivity (75%/70%), a value lower than in unipolar depression where the sensitivity was found to be greater than 90% (Szegedi et al., 2009, 2003).

We assessed the predictive potential of early improvement for each compound individually, in all studies combined, and then in subsequent secondary analysis of pooled positive and negative trials. Data from the pooled positive studies appear to better predict endpoint response/non-response, suggesting that heterogeneity in the design and conduct of studies—which may ultimately lead to the failure of an active compound to separate from placebo—may also affect the predictive accuracy of the early improvement analyses.

The predictive accuracy was notably reduced in the pooled negative/failed studies of lamotrigine (n = 4), for which the rates of early improvement and sensitivity values were lowest. The dosing schedule for lamotrigine is slower than the other drugs studied in order to minimize the risk of serious rash, thus patients did not show improvement until Week 3 (Calabrese et al., 2002). This may have inhibited our ability to predict eventual outcomes by Week 2. For these reasons, a longer initial trial may be necessary to thoroughly assess the potential for response to lamotrigine. Interestingly, the positive and negative pooled trials for lamotrigine exhibited the lowest rate of false positives for all agents assessed in these studies. On the other hand, the highest

false positive rates occurred with aripiprazole, olanzapine/olanzapine–fluoxetine, and quetiapine. These agents have a wider variety of potential side effects that may have contributed to an artificial perception of early response as measured by MADRS Total score improvement at Week 2. For instance, the early soporific effects of quetiapine and olanzapine/olanzapine–fluoxetine and the early activating effects of aripiprazole may have contributed to a higher rate of false positives if these effects were not later accompanied by an improvement in other core symptoms of depression.

Confirming the work of other authors in major depressive disorder, the predictive values generated from placebo-treated patients appear remarkably similar to patients receiving an active drug treatment (Szegedi et al., 2009; Stassen and Angst, 1998). The absence of early improvement remained a highly reliable predictor of non-response and non-remission, regardless of whether the patient was receiving an active drug or placebo. Moreover, the probability of falsely predicting response/remission was higher among patients receiving active compounds (43%/47%) than among those receiving a placebo (30%/35%).

We believe a distinguishing feature of the present analysis as compared with previously published studies is the use of chance-corrected operating characteristics to adjust for chance agreement in sensitivity and specificity that occurs with random testing. Failure to assess the chance-corrected quantities may lead to overstating the operating characteristics of early improvement as a predictor of later response or remission. Consistent with this premise, each of the chance-corrected values were lower than the unadjusted sensitivity and specificity values. However, the pooled positive studies continued to demonstrate better sensitivity in comparison with the pooled negative trials, particularly for the olanzapine/OFC, quetiapine, and positive lamotrigine trials.

The majority of chance-corrected sensitivity measurements fell within the moderate range of 0.4 to 0.6 as described by Landis and Koch (1977), even despite the heterogeneity of medications studied in this analysis. One exception was the pooled negative trials of lamotrigine, for which the chance-corrected sensitivity ranged from 0.33 to 0.34. Although the chance-corrected sensitivity values identified that early improvement for the prediction of remission was most sensitive in the olanzapine/OFC trials, chance-corrected specificity values for each of the trials were similar, with the largest values occurring in the positive lamotrigine trial (43%/35%). It should be noted that when utilizing the chance-corrected calibrations, a test that optimizes chance-corrected sensitivity also optimizes the negative predictive value of a test. Likewise, a test that optimizes chance-corrected specificity also optimizes the positive predictive value of a test (Kraemer, 1992). Again, the better chance-corrected sensitivity and negative predictive values occurred in the pooled positive trials of all agents studied.

Limitations of the present analysis include not investigating the predictive value of early improvement for a range of timepoints (i.e. ≥20% improvement at week 1, 3, or 4) or different MADRS cutpoints (i.e. ≥15, 25, or 30% reduction in MADRS Total Score at Week 2). In trials of schizophrenia, there is some evidence that early response thresholds and predictive accuracy increase over time (i.e. from Week 1 to Week 4) (Chen et al., 2009). In support of this finding, a post hoc analysis of the aripiprazole bipolar depression studies showed that early improvement at Week 3 was a slightly stronger predictor of later response or remission than Week 2, although the area under the receiver operating characteristics curves for both time points was similar and clinically meaningful (Kemp et al., 2010). Also, our study was not designed to examine whether a symptom-specific or global evaluation is the most appropriate metric for evaluating early improvement as these data were not available. Future investigations of early improvement may be strengthened by employing an individual item-analysis, as one study recently identified that non-early improvers with major depressive disorder have at least a 3-fold higher risk of developing later treatment-emergent suicidal ideation (Seemüller et al., 2010).

The data are limited by the acute nature of the trials; thus we do not know if improvement is maintained or accelerated beyond 8 weeks, or whether early improvement may predict outcomes experienced at a later point in time. As this analysis was performed solely on trials of bipolar depression, it does not provide any information on the predictive effect of early improvement in states of bipolar mania, for which limited data exist (Ketter et al., 2010; Kemp et al., in press).

Although early improvement analyses of unipolar depression originating from naturalistic samples appear to replicate those derived from randomized controlled trials (Henkel et al., 2009), it should be noted that each of the trials included in this report were conducted for the purposes of regulatory approval, and employed strict inclusion and exclusion criteria. It is unknown whether these results can be generalized to naturalistic samples, such as patients receiving these agents when used in combination with other psychotropic medications, a common occurrence in the treatment of bipolar disorder (Goldberg et al., 2009; Baldessarini et al., 2008). Additionally, dosing regimens in clinical practice may differ from those used in clinical trials. If the dose of a drug is titrated at a much slower pace, it may affect the rapidity of improvement and alter the positive and negative predictive value of the early improvement metric.

4.1. Conclusion

Patients with bipolar disorder experience tremendous burden from chronic depressive symptoms and recurrent depressive episodes. As a field, we need more knowledge about what treatments are most effective for bipolar depression and how to better tailor treatment approaches for individual patients. This includes improved identification of reliable predictors of treatment efficacy. The current study suggests that while early improvement is not a reliable predictor of ultimate response or remission, the lack of early improvement is an important prognostic indicator, and should be utilized to avoid sustained trials of ineffective treatments.

Role of Funding Source

The individual trials included in this report were funded by AstraZeneca, Bristol-Myers Squibb, Eli Lilly, and GlaxoSmithKline Dr. Kemp's salary is supported in part by 1KL2RR024990. The NIH had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Conflict of Interest

Disclosures:

Dr. Brecher is an employee of SK Life Science.

Dr. Calabrese has received grant support, lecture honoraria, or has participated in advisory boards with Abbott, AstraZeneca, Bristol-Myers Squibb/Otsuka, Cephalon, Dainippon Sumitomo, Forest, France Foundation, GlaxoSmithKline, Janssen, Johnson and Johnson, Lilly, Lundbeck, Merck, Neurosearch, OrthoMcNeil, Pfizer, Repligen, Sanofi, Schering-Plough, Servier, Solvay, Synosia, Supernus Pharmaceuticals, Takeda and Wyeth.

Dr. Carlson is an employee of Bristol-Myers Squibb.

Dr. Edwards is an employee of GlaxoSmithKline.

Mr. Eudicone is an employee of Bristol-Myers Squibb.

Dr. Evoniuk is an employee of GlaxoSmithKline.

Dr. Ganocy receives grant support from AstraZeneca and Eli Lilly.

Mr. Jansen is an employee of Schering-Plough.

Dr. Kemp has acted as a consultant to Bristol-Myers Squibb and is on the speaker's bureau for AstraZeneca and Pfizer.

Dr. Leon serves on independent Data and Safety Monitoring Boards for AstraZeneca, Dainippon Sumitomo Pharma America, and Pfizer; serves as a Consultant/Advisor to NIMH, MedAvante and Roche; has equity in MedAvante.

Dr. Minkwitz is an employee of AstraZeneca.

Dr. Pikalov is an employee of Dainippon Sumitomo Pharma America, Inc.

Dr. Stassen reports no financial relationships relevant to the subject of this article.

Dr. Szegedi is an employee of Merck.

Dr. Tohen is a former Eli Lilly employee (2008). He has received honoraria from AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Forest, GlaxoSmithKline, Merck, Otsuka, Sepracor, and Wyeth. His spouse is an Eli Lilly employee and minor stock holder.

Mr. Van Willigenburg is an employee of Complete Healthcare Communications Europe.

Acknowledgements

The authors thank Eli Lilly and Company for providing data which were used in our analyses and Ellen Dennehy, PhD for editorial assistance with this manuscript.

References

Akiskal, H.S., 2005. The dark side of bipolarity: detecting bipolar depression in its pleomorphic expressions. J. Affect. Disord. 84, 107–115.

Baldessarini, R., Henk, H., Skla, A., Chang, J., Leahy, L., 2008. Psychotropic medications for patients with bipolar disorder in the United States: polytherapy and adherence. Psychiatr. Serv. 59, 1175–1183.

Bauer, M.S., Kirk, G.F., Gavin, C., Williford, W.O., 2001. Determinants of functional outcome and healthcare costs in bipolar disorder: a high-intensity follow-up study. J. Affect. Disord. 65, 231–241.

Calabrese, J.R., Bowden, C.L., Sachs, G.S., Ascher, J.A., Monaghan, E., Rudd, G.D., 1999. A double-blind placebo-controlled study of lamotrigine monotherapy in outpatients with bipolar I depression. Lamictal 602 Study Group. J. Clin. Psychiatry 60, 79–88.

Calabrese, J.R., Sullivan, J.R., Bowden, C.L., Suppes, T., Goldberg, J.F., Sachs, G.S., Shelton, M.D., Goodwin, F.K., Frye, M.A., Kusumakar, V., 2002. Rash in multicenter trials of lamotrigine in mood disorders: clinical relevance and management. J. Clin. Psychiatry 63, 1012–1019.

Calabrese, J.R., Hirschfeld, R.M., Frye, M.A., Reed, M.L., 2004. Impact of depressive symptoms compared with manic symptoms in bipolar disorder: results of a U.S. community-based sample. J. Clin. Psychiatry 65, 1499–1504.

Calabrese, J.R., Keck Jr., P.E., Macfadden, W., Minkwitz, M., Ketter, T.A., Weisler, R.H., Cutler, A.J., McCoy, R., Wilson, E., Mullen, J., 2005. A randomized, double-blind, placebo-controlled trial of quetiapine in the treatment of bipolar I or II depression. Am. J. Psychiatry 162, 1351–1360.

Calabrese, J.R., Huffman, R.F., White, R.L., Edwards, S., Thompson, T.R., Ascher, J.A., Monaghan, E.T., Leadbetter, R.A., 2008. Lamotrigine in the acute treatment of bipolar depression: results of five double-blind, placebo-controlled clinical trials. Bipolar Disord. 10, 323–333.

Chen, L., Ascher-Svanum, H., Stauffer, V., Kinon, B.J., Kollack-Walker, S., Ruberg, S., 2009. Optimal thresholds of early response to atypical antipsychotics: application of signal detection methods. Schizophr. Res. 113, 34–40.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Gelenberg, A.J., Chesen, C.L., 2000. How fast are antidepressants? J. Clin. Psychiatry 61, 712–721.

Goldberg, J.F., Brooks, J.O., Kurita, K., Hoblyn, J.C., Ghaemi, S.N., Perlis, R.H., Miklowitz, D.J., Ketter, T.A., Sachs, G.S., Thase, M.E., 2009. Depressive illness burden associated with complex polypharmacy in patients with bipolar disorder: findings from STEP-BD. J. Clin. Psychiatry 70, 155–162.

Henkel, V., Seemüller, F., Obermeier, M., Adli, M., Bauer, M., Mundt, C., Brieger, P., Laux, G., Bender, W., Heuser, I., Zeiler, J., Gaebel, W., Mayr, A., Möller, H.J., Riedel, M., 2009. Does early improvement triggered by antidepressants predict response/remission? Analysis of data from a naturalistic study on a large sample of inpatients with major depression. J. Affect. Disord. 115, 439–449.

Judd, L.L., Akiskal, H.S., Schettler, P.J., Endicott, J., Maser, J., Solomon, D.A., Leon, A.C., Rice, J.A., Keller, M.B., 2002. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch. Gen. Psychiatry 59, 530–537.

Judd, L.L., Akiskal, H.S., Schettler, P.J., Endicott, J., Leon, A.C., Solomon, D.A., Coryell, W., Maser, J.D., Keller, M.B., 2005. Psychosocial disability in the course of bipolar I and II disorders: a prospective, comparative, longitudinal study. Arch. Gen. Psychiatry 62, 1322–1330.

Katz, M.M., Koslow, S.H., Frazer, A., 1996–1997. Onset of antidepressant activity: reexamining the structure of depression and multiple actions of drugs. Depress Anxiety 4, 257–267.

Katz, M.M., Bowden, C., Stokes, P., Casper, R., Frazer, A., Koslow, S.H., Kocsis, J., Secunda, S., Swann, A., Berman, N., 1997. Can the effects of antidepressants be observed in the first two weeks of treatment? Neuropsychopharmacology 17, 110–115.

Kemp, D.E., Calabrese, J.R., Eudicone, J., Ganocy, S., Tran, Q.V., Pikalov, A., Marcus, R., Vester-Blokland, E., Owen, R., Carlson, B.X., 2010. Predictive Value of Early Improvement in Bipolar Depression Trials: A Post-hoc Pooled Analysis of Two 8-week Aripiprazole Studies. Psychopharmacol. Bull. 43, 5–27.

Kemp, D.E., Johnson, E., Wang, W.V., Tohen, M., Calabrese, J.R., in press. Clinical utility of early improvement to predict response or remission in acute mania: focus on olanzapine and risperidone. J. Clin. Psychiatry.

Kessler, R.C., Akiskal, H.S., Ames, M., Birnbaum, H., Greenberg, P., Hirschfeld, R.M.A., Jin, R., Merikangas, K.R., Simon, G.E., Wang, P.S., 2006. Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers. Am. J. Psychiatry 163, 1561–1568.

Ketter, T.A., Agid, O., Kapur, S., Loebel, A., Siu, C.O., Romano, S.J., 2010. Rapid antipsychotic response with ziprasidone predicts subsequent acute manic/mixed episode remission. J. Psychiatr. Res. 44, 8–14.

Kinon, B.J., Chen, L., Ascher-Svanum, H., Stauffer, V.L., Kollack-Walker, S., Sniadecki, J.L., Kane, J.M., 2008. Predicting response to atypical antipsychotics based on early response in the treatment of schizophrenia. Schizophr. Res. 102, 230–240.

Kraemer, H.C., 1992. Evaluating Medical Tests: Objective and Quantitative Guidelines. Sage Publication, Newbury Park, CA.

Kupka, R.W., Altshuler, L.L., Nolen, W.A., Suppes, T., Luckenbaugh, D.A., Leverich, G.S., Frye, M.A., Keck Jr., P.E., McElroy, S.L., Grunze, H., Post, R.M., 2007. Three times more days depressed than manic or hypomanic in both bipolar I and bipolar II disorder. Bipolar Disord. 9, 531–535.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389.

Muzina, D.J., Kemp, D.E., McIntyre, R.S., 2007. Differentiating bipolar disorders from major depressive disorders: treatment implications. Ann. Clin. Psychiatry 19, 305–312.

Nierenberg, A.A., McLean, N.E., Alpert, J.E., Worthington, J.J., Rosenbaum, J.F., Fava, M., 1995. Early nonresponse to fluoxetine as a predictor of poor 8-week outcome. Am. J. Psychiatry 152, 1500–1503.

Perlis, R.H., Ostacher, M.J., Patel, J., Marangell, L.B., Zhang, H., Wisniewski, S.R., Ketter, T.A., Miklowitz, D.J., Otto, M.W., Gyulai, L., Reilly-Harrington, N., Nierenberg, A.A., Sachs, G.S., Thase, M.E., 2006. Predictors of recurrence in bipolar disorder: primary outcomes from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). Am. J. Psychiatry 163, 217–224.

Post, R.M., Leverich, G.S., Nolen, W.A., Kupka, R.W., Altshuler, L.L., Frye, M.A., Suppes, T., McElroy, S., Keck, P., Grunze, H., Walden, J., 2003. Stanley Foundation Bipolar Network: a re-evaluation of the role of antidepressants in the treatment of bipolar depression: data from the Stanley Foundation Bipolar Network. Bipolar Disord. 5, 396–406.

Posternak, M.A., Zimmerman, M., 2005. Is there a delay in the antidepressant effect? A meta-analysis. J. Clin. Psychiatry 66, 148–158.

Quitkin, F.M., Rabkin, J.G., Ross, D., Stewart, J.W., 1984. Identification of true drug response to antidepressants. Arch. Gen. Psychiatry 41, 782–786.

Quitkin, F.M., Rabkin, J.D., Markowitz, J.M., Stewart, J.W., McGrath, P.J., Harrison, W., 1987. Use of pattern analysis to identify true drug response: a replication. Arch. Gen. Psychiatry 44, 259–264.

Sachs, G.S., Nierenberg, A.A., Calabrese, J.R., Marangell, L.B., Wisniewski, S.R., Gyulai, L., Friedman, E.S., Bowden, C.L., Fossey, M.D., Ostacher, M.J., Ketter, T.A., Patel, J., Hauser, P., Rapport, D., Martinez, J.M., Allen, M.H., Miklowitz, D.J., Otto, M.W., Dennehy, E.B., Thase, M.E., 2007. Effectiveness of adjunctive antidepressant treatment for bipolar depression. N Engl J. Med. 356, 1711–1722.

Seemüller, F., Wolff, R.S., Obermeier, M., Henkel, V., Möller, H.J., Riedel, M., 2010. Does early improvement in major depression protect against treatment emergent suicidal ideation? J. Affect. Disord. 124, 183–186.
Stassen, H., Angst, J., 1998. Delayed onset of action of antidepressants: fact or fiction? CNS Drugs 9, 177–184.
Stassen, H.H., Delini-Stula, A., Angst, J., 1993. Time course of improvement under antidepressant treatment: a survival–analytical approach. Eur. Neuropsychopharmacol. 3, 127–135.
Stassen, H.H., Angst, J., Delini-Stula, A., 1996. Delayed onset of action of antidepressant drugs? Survey of results of Zurich meta-analyses. Pharmacopsychiatry 29, 87–96.
Stassen, H.H., Angst, J., Delini-Stula, A., 1999. Fluoxetine versus moclobemide: cross-comparison between the time courses of improvement. Pharmacopsychiatry 32, 56–60.
Szegedi, A., Müller, M.J., Anghelescu, I., Klawe, C., Kohnen, R., Benkert, O., 2003. Early improvement under mirtazapine and paroxetine predicts later stable response and remission with high sensitivity in patients with major depression. J. Clin. Psychiatry 64, 413–420.
Szegedi, A., Jansen, W.T., van Willigenburg, A.P., van der Meulen, E., Stassen, H.H., Thase, M.E., 2009. 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. J. Clin. Psychiatry 70, 344–353.
Tadić, A., Helmreich, I., Mergl, R., Hautzinger, M., Kohnen, R., Henkel, V., Hegerl, U., 2010. Early improvement is a predictor of treatment outcome in patients with mild major, minor or subsyndromal depression. J. Affect. Disord. 120, 86–93.
Taylor, M.J., Freemantle, N., Geddes, J.R., Bhagwagar, Z., 2006. Early onset of selective serotonin reuptake inhibitor antidepressant action: systematic review and meta-analysis. Arch. Gen. Psychiatry 63, 1217–1223.
Thase, M.E., Macfadden, W., Weisler, R.H., Chang, W., Paulsson, B., Khan, A., Calabrese, J.R., 2006. BOLDER II Study Group: efficacy of quetiapine monotherapy in bipolar I and II depression: a double-blind, placebo-controlled study (the BOLDER II study). J. Clin. Psychopharmacol. 26, 600–609 (Erratum in: J Clin Psychopharmacol. 27, 51).
Thase, M.E., Jonas, A., Khan, A., Bowden, C.L., Wu, X., McQuade, R.D., Carson, W.H., Marcus, R.N., Owen, R., 2008. Aripiprazole monotherapy in nonpsychotic bipolar I depression: results of 2 randomized, placebo-controlled studies. J. Clin. Psychopharmacol. 28, 13–20.
Tohen, M., Vieta, E., Calabrese, J., Ketter, T.A., Sachs, G., Bowden, C., Mitchell, P. B., Centorrino, F., Risser, R., Baker, R.W., Evans, A.R., Beymer, K., Dube, S., Tollefson, G.D., Breier, A., 2003. Efficacy of olanzapine and olanzapine–fluoxetine combination in the treatment of bipolar I depression. Arch. Gen. Psychiatry 60, 1079–1088.