JOURNAL OF CHILD AND ADOLESCENT PSYCHOPHARMACOLOGY
Volume 23, Number 2, 2013

Pp. 72–79
DOI: 10.1089/cap.2012.0024

Original Articles

Clinical Significance of Treatment Effects with Aripiprazole versus Placebo in a Study of Manic or Mixed Episodes Associated with Pediatric Bipolar I Disorder

Eric Youngstrom, PhD,[1] Joan Zhao, PhD,[2] Raymond Mankoski, MD,[3] Robert A. Forbes, PhD,[2] Ronald M. Marcus, MD,[4] William Carson, MD,[2] Robert McQuade, PhD,[2] and Robert L. Findling, MD, MBA[5]

Abstract

Objective: Published studies in adult and pediatric bipolar disorder have used different definitions of treatment response. This analysis aimed to compare different definitions of response in a large sample of children and adolescents.

Methods: Anexploratory analysis of a 4-week, multicenter, placebo-controlled study assessed patients ($n = 296$; ages, 10–17 years) with an acute manic/mixed episode associated with bipolar I disorder who were randomized to aripiprazole (10 or 30 mg/day) or placebo. The primary efficacy endpoint was mean change from baseline to week 4 in Young Mania Rating Scale (YMRS) total score. Additional assessments included: Clinical Global Impressions–Bipolar Disorder (CGI-BP) Overall and Mania scales, Child Global Assessment Scale (CGAS), and parent and subject General Behavior Inventory. Response was compared across seven operational definitions. Cohen's κ and Spearman's correlation tested relationships between various response definitions or changes in outcome measures and clinically meaningful improvement (defined as a CGI-BP Overall Improvement score of 1 or 2).

Results: Response rates varied depending upon the operational definition, but were highest for 95% reliable change (statistical method used to determine individual change from previous assessment) and $\geq 33\%$ reduction in YMRS total score. Response rate definitions with the highest validity in terms of predicting clinically meaningful improvement were: $\geq 50\%$ reduction on YMRS ($\kappa = 0.64$), a composite definition of response (YMRS < 12.5, Children's Depression Rating Scale-Revised (CDRS-R) ≤ 40, and CGAS ≥ 51; $\kappa = 0.59$), and 95% reliable change on the CGAS or 33% reduction on YMRS ($\kappa = 0.56$). Parent ratings of symptoms were generally better at detecting symptom improvement than were subject ratings ($\kappa = \sim 0.4$–0.5 vs. ~ 0.2 when compared with CGI-BP Overall Improvement score).

Conclusions: Clinically meaningful definitions of response in acute treatment of a manic/mixed episode in pediatric subjects include a 50% change in YMRS and a composite measure of response. Parent-reported measures of symptom improvement appear reliable for assessing symptom change.

Introduction

Pediatric bipolar disorder is a serious mental health illness that disrupts the lives of patients and their families. Although the exact prevalence of pediatric bipolar disorder is not known, a recent meta-analysis of community epidemiological studies indicates that bipolar spectrum conditions may affect 2% of children and adolescents around the world (Youngstrom et al. 2010; Van Meter et al. 2011).

Validated measures of symptom severity, such as the Young Mania Rating Scale (YMRS) (Young et al. 1978), are commonly used to evaluate the efficacy of pharmacologic treatment in clinical trials of bipolar disorder in both children and adults. However, published studies in adult and pediatric populations have used different definitions of response, which can increase the difficulty in comparing results across studies. Proposed definitions vary from a 33% or 50% reduction in symptoms (measured using the YMRS) to more complex definitions that combine scale scores with measures of overall functional outcomes (Kowatch et al. 2000; Kafantaris et al. 2001; Findling et al. 2003).

In addition to the more standard approaches mentioned, an interesting model that is widely used in studies of psychosocial interventions has been developed to evaluate clinically significant

[1]Department of Psychology, University of North Carolina, Chapel Hill, North Carolina.
[2]Otsuka Pharmaceutical Development & Commercialization, Inc., Princeton, New Jersey. [3]Bristol-Myers Squibb, Plainsboro, New Jersey.
[4]Bristol-Myers Squibb, Wallingford, Connecticut.
[5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins Medicine and the Kennedy Krieger Institute, Baltimore, Maryland.
Funding: This study was supported by Bristol-Myers Squibb (Princeton, NJ) and Otsuka Pharmaceutical Co., Ltd. (Tokyo, Japan).
Clinical trial registration: ID number: NCT00110461; registry: www.clinicaltrials.gov

change (Jacobson and Truax 1991; Atkins et al. 2005). This method, proposed by Jacobson and Truax, involves a two-step approach that focuses on whether there is a *reliable change* between an individual's pre- and post-treatment functioning on a given outcome measure, plus comparison against normative benchmarks (three cutoff criteria, defined as A [move the patient *Away* from the clinically impaired range of functioning]; B [move the patient *Back* into the normal range of functioning]; and C [move the patient *Closer* to the normal than the clinically impaired range of functioning]). Reliable change is a difference in the individual's scores considered large enough that it is unlikely to be caused by imprecision in the measure. Jacobson and colleagues have suggested using a 95% confidence interval (CI) constructed using the standard error of the difference score for the measure as a way of operationally defining reliable change. They suggested setting the normative benchmarks at two standard deviations (SD) below average for the clinical population of interest (the *Away* benchmark, assuming that high scores reflect greater pathology) and two SDs around the nonclinical mean on the same measure. The third benchmark is usually operationally defined as the weighted average of the clinical and nonclinical means from the normative groups, taking into account the possibility that the two groups have different variations in scores. The choice of a 95% CI for the reliable change index, and the use of a two SD threshold for the benchmarks, has become an established convention to align with the α level of 0.05 norm in statistical significance testing.

The Jacobson and Truax approach permits classification of individuals into one of four categories: *recovered* – reliable change moving past one of the normative benchmarks; *improved* – reliable change, but not yet sufficiently great to pass a normative benchmark; *unchanged* – when score movement is too small to surpass the reliable change threshold imposed by the precision of the instrument; or *deteriorated* – when there is reliable change for the worse (Atkins et al. 2005; Youngstrom et al. 2008). A major difference between the Jacobson and Truax method and conventions focusing on effect size is that, with the former, clinically significant individual outcomes, rather than an average effect across cases, are evaluated. All of the benchmarks were constructed for each potential outcome measure, as one of the major aims of this article was to compare the alternate definitions of change and treatment response across measures. We believe that improved understanding of clinically meaningful improvement in symptoms is of value to researchers, as well as to clinicians.

Since its introduction, a variety of refinements and alternatives have been proposed to the basic Jacobson and Truax model. Some are modifications to further enhance the psychometric properties; for example, by adjusting the standard errors to reflect changes in precision across different score ranges (Speer 1992). However, these methods are more complex and do not always improve performance (Ogles et al. 2001). A second alternative is the use of mixed regression models or growth curve models to estimate individual trajectories of change (Speer et al. 1995). Although mixed regression or growth curve models are conceptually attractive in a variety of ways, they are challenging for practicing clinicians to estimate, and difficult or impossible to apply to individual cases. They also do not map onto metrics such as the number needed to treat (NNT), which evidence-based medicine uses to evaluate trials and apply results to individual patients (Guyatt and Rennie 2002; Straus et al. 2011). A third approach is to use the equivalence testing framework (Westlake 1976) to evaluate whether the post treatment distribution is negligibly different from nonclinical distributions (Kendall et al. 1999). This method has conceptual appeal as a complement to group effect size estimation, but it remains a group-oriented summary itself that cannot be applied by clinicians to individual cases (Follette and Callaghan 2001). Experts have pointed to the need to evaluate consumer satisfaction and quality of life as important components of clinically significant outcomes, in addition to measures of symptom reduction (Kazdin 1999). Of the different definitions of clinically significant change in symptoms, the Jacobson model is the one to accrue the most evidence for corresponding with satisfaction and improved functioning (Ankuta and Abeles 1993; Lunnen and Ogles 1998). For all of these reasons, the ''classic'' Jacobson and Truax method has remained the dominant model in psychotherapy trials, and, therefore, was our choice for comparison with the operational definitions used in prior pharmacological trials in pediatric bipolar disorder.

Currently, no consensus has been established regarding any particular definition of response as a measure of clinically significant change in the treatment of manic or mixed episodes associated with bipolar I disorder in either the adult or pediatric patient population. The purpose of this secondary analysis was to compare different definitions of response in a large sample of children and adolescents (ages 10–17 years) who had participated in a 4-week trial of aripiprazole (10 or 30 mg/day) for the treatment of an acute manic or mixed episode associated with bipolar I disorder (Findling et al. 2009). In these analyses, simultaneous comparisons were made across multiple outcome measures using a range of definitions of response in order to evaluate which definitions of response may be most relevant to clinical practice. One goal was to explore whether any definitions were consistently more liberal or conservative across the various outcome measures. A second goal was to examine whether any measures appeared particularly sensitive or unresponsive to treatment effects. Because different studies have selected varying definitions of response, comparing these operational definitions in a single sample provides an important sense of how the definitions are calibrated against each other.

Methods

Study design and patients

This was a secondary analysis of a 4-week, multicenter, placebo-controlled, randomized study of the efficacy and safety of aripiprazole 10 or 30 mg/day in children and adolescents ($n = 296$; ages, 10–17 years) with an acute manic or mixed episode associated with bipolar I disorder. The study enrolled patients across 59 sites in the United States between March 2005 and February 2007. Study procedures were adherent to the Declaration of Helsinki and International Conference on Harmonization/Good Clinical Practice Guidelines, and were approved by the institutional review boards at each site. All parents/guardians provided written informed consent to participate, and subjects provided written, informed assent when possible.

Detailed information on the study design and patient population has been published previously (Findling et al. 2009). Briefly, the study population consisted of children and adolescents with a confirmed *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. (DSM-IV) diagnosis of bipolar I disorder with current manic or mixed episodes (American Psychiatric Association 1994), with or without psychotic features, and a YMRS total score ≥ 20 at baseline. Subjects were diagnosed by a board-certified or board-eligible child and adolescent psychiatrist and the diagnosis was confirmed using the Kiddie Schedule for Affective Disorders and Schizophrenia (Kim et al. 2004). Subjects with comorbid attention-deficit/hyperactivity disorder (ADHD),

conduct disorder, oppositional defiant disorder, or anxiety disorders (except posttraumatic stress disorder or obsessive-compulsive disorder) were also eligible.

Assessments

The primary efficacy outcome measure was the mean change from baseline to endpoint (week 4) in the YMRS total score. During the study, patients, parents, and clinicians completed additional rating scales at baseline and at each study visit up to week 4, including: the Clinical Global Impressions–Bipolar Disorder (CGI-BP) Overall, – Depression, and – Mania Improvement and Severity scores (Spearing et al. 1997); Children's Depression Rating Scale-Revised (CDRS-R) (Poznanski and Mokros 1995); Children's Global Assessment Scale (CGAS) (Shaffer et al. 1983); and parent- and subject-rated 10-item versions of the General Behavior Inventory Mania (parent/subject-GBI-M10) and Depression (parent/subject-GBI-D10) (Youngstrom et al. 2001; Danielson et al. 2003) scales. GBI items were scaled from 0 to 3; therefore, scores could range from 0 to 30 points. Further details on efficacy and safety assessments undertaken in the study have been published elsewhere (Findling et al. 2009).

Definitions of response

Response, as measured by change from baseline in YMRS total score, was defined according to one of seven different operational criteria: 1) 33% reduction from baseline in YMRS total score; 2) ≥50% reduction from baseline in YMRS total score; 3) 95% reliable change (Jacobson and Truax 1991); 4) reliable change plus moving more than two SDs away from the clinically impaired range of functioning for bipolar disorder cases (Jacobson and Truax definition A, *Away* from the clinical distribution; Fig. 1; [Jacobson and Truax 1991]); 5) reliable change plus moving back within two SDs of the nonclinically impaired, that is, normal functioning mean (Jacobson and Truax definition B, *Back* within the nonclinical range; Fig. 1; [Jacobson and Truax 1991]); and 6) reliable change plus moving closer to the nonclinically impaired than the clinically impaired range of functioning (Jacobson and Truax definition C, *Closer* to nonclinical than clinical levels; Fig. 1; [Jacobsonand Truax 1991]).

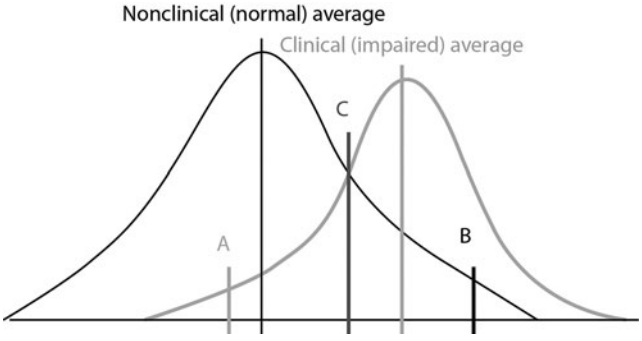

FIG. 1. Jacobsen and Truax definitions of response (Jacobson and Truax 1991). **A** (Jacobson and Truax definition A): reliable change plus moving more than two standard deviations (SDs) *Away* from the impaired range of functioning; **B** (Jacobson and Truax definition B): reliable change plus moving *Back* within two SDs of the impaired, that is, in the normal functioning mean; **C** (Jacobson and Truax definition C): reliable change plus moving *Closer* to the normal functioning than to the impaired range of functioning.

The seventh response definition composite (COMP) comprised a composite of scores achieved on the YMRS (< 12.5), CDRS-R (≤ 40), and CGAS (≥ 51),which has been used previously to define clinical response in a study of lithium and divalproex sodium for the treatment of bipolar disorder in pediatric patients (Findling et al. 2003).

Statistical analysis

Secondary analyses were conducted on the efficacy sample, which consisted of patients who had received at least one dose of study treatment and had had at least one post baseline efficacy assessment. Rates of response were derived from the number of individual cases meeting each operational definition of change over the total number of cases evaluated for that change. Rates were estimated separately for the placebo and aripiprazole 10 mg and 30 mg treatment arms. The NNT (Straus et al. 2005) compared response rates with aripiprazole 10 mg/day versus placebo, and aripiprazole 30 mg/day versus placebo separately. The absolute rates indicated which definitions produced more liberal or conservative rates of response, whereas the NNT focused on the extent to which active treatment differentiated from placebo response. Two sets of analyses quantified the criterion validity of each operational definition of individual improvement compared with the clinical global impressions of change; a CGI-BP Overall Improvement score of 1 or 2 (''very much improved'' or ''much improved,'' respectively) was specified *a priori* as a clinically relevant benchmark of symptom improvement. Kappa-measured agreement above chance with a dichotomized ''improved, yes or no'' standard (Cohen 1960) and Spearman's correlation quantified agreement with an ordinal scale of degree of improvement (Well and Myers 2003). All analyses were conducted using the last observation carried forward approach and were descriptive in nature; formal statistical comparisons were not conducted.

Results

Patient population

A total of 296 subjects were randomized and included in the efficacy sample: aripiprazole 10 mg/day, $n = 98$; aripiprazole 30 mg/day, $n = 99$; placebo, $n = 99$. Of these, 237 (80.1%) completed the 4-week study (Findling et al. 2009).

Baseline demographic characteristics of subjects in each treatment group are shown in Table 1, and have been published previously (Findling et al. 2009). In brief, the overall study population had a mean age of 13.4 years and was predominantly white (65.2%) and male (53.7%) (Findling et al. 2009). Acute treatment yielded moderate-to-large group effect sizes (Cohen's d values of 0.60 and 0.93 for aripiprazole 10 and 30 mg/day, respectively) for improvement in the CGI-BP Overall Improvement score, the main criterion for examining the validity of differing operational definitions of clinically significant change in the present analyses.

Response rates by differing levels of response

The proportions of subjects achieving response to aripiprazole 10 or 30 mg/day or placebo according to predefined thresholds of symptom improvement are presented in Figure 2.

Response rates varied substantially across operational definitions. Utilizing a response definition of ≥ 33% reduction in YMRS total score, rates were > 70% for both aripiprazole 10 (73%, NNT = 3) and 30 (77%, NNT = 3) mg/day and 38% for placebo, whereas the magnitude of response to treatment with aripiprazole or placebo was generally lower when the more stringent definition

Table 1. Baseline Demographics

Characteristic	*Aripiprazole 10 mg/day* (n = *98*)	*Aripiprazole 30 mg/day* (n = *99*)	*Placebo* (n = *99*)
Mean age, years (SD)	13.7 (2.2)	13.3 (2.3)	13.3 (2.1)
Males/females, (%)	53.1/46.9	51.5/48.5	56.6/43.4
Mean weight, kg (SD)	63.8 (20.1)	60.5 (21.5)	60.5 (17.3)
Race, *n* (%)			
Caucasian	65 (66.3)	68 (68.7)	60 (60.6)
Black	24 (24.5)	18 (18.2)	23 (23.2)
Pacific Islander	2 (2.0)	0 (0.0)	0 (0.0)
Other	7 (7.1)	13 (13.1)	16 (16.2)
Symptom rating scales, mean[a]			
YMRS total score	29.8	29.5	31.1
CGAS	46.9	47.5	45.5
CDRS-R	35.2	34.1	33.8
CGI-BP mania severity score	4.7	4.6	4.9
CGI-BP depression severity score	2.9	2.9	2.8
CGI-BP overall severity score	4.7	4.6	4.8
GBI-M (parent)	17.7	17.4	19.1
GBI-M (subject)	15.1	14.8	14.8
GBI-D (parent)	13.4	12.4	13.4
GBI-D (subject)	12.1	11.3	10.5
Psychotic symptoms in current episode, *n* (%)[b]			
Yes	7 (7.1)	4 (4.0)	3 (3.0)
No	58 (59.2)	58 (58.6)	64 (64.7)
Unknown	33 (33.7)	37 (37.4)	32 (32.3)
History of rapid cycling, *n* (%)[b]			
Yes	17 (17.4)	13 (13.1)	15 (15.2)
No	49 (50.0)	46 (46.5)	51 (51.5)
Unknown	32 (32.7)	40 (40.4)	33 (33.3)

[a]Patient populations in the aripiprazole (10 mg/day), aripiprazole (30 mg/day), and placebo groups were: 96, 99, and 94, respectively for YMRS, CGAS, CGI-BP; 91, 94, and 86, respectively, for CDRS-R; 95, 96, and 93, respectively, for parent versions of GBI-M and GBI-D; and 96, 96, and 93, respectively, for subject versions of GBI-M and GBI-D.

[b]These data were collected *post hoc*; as capture of data on comorbid diagnoses was not required by investigators, there is a high percentage of missing data.

CDRS-R, Children's Depression Rating Scale-Revised; CGAS, Children's Global Assessment Scale; CGI-BP, Clinical Global Impressions–Bipolar Disorder; GBI-D, General Behavior Inventory Depression; GBI-M, General Behavior Inventory Mania; YMRS, Young Mania Rating Scale.

of ≥50% reduction in YMRS total score was applied (45% [NNT = 5] or 64% [NNT = 3] for aripiprazole 10 or 30 mg/day, respectively; 25.5% for placebo).

The Jacobson and Truax criterion of 95% reliable change resulted in the highest response rates; 88% (NNT = 6) and 91% (NNT = 5) being observed in the aripiprazole 10 and 30 mg/day treatment groups, respectively, and 70% in the placebo group. Conversely, rates of response to aripiprazole did not exceed 30% using Jacobson and Truax definition A (15% [NNT = 9]; 27% [NNT = 4]), B (5% [NNT = 19]; 10% [NNT = 10]) or C (5% [NNT = 19]; 10% [NNT = 10]) for aripiprazole 10 mg/day and 30 mg/day, respectively, whereas placebo response rates were negligible (< 3.2%).

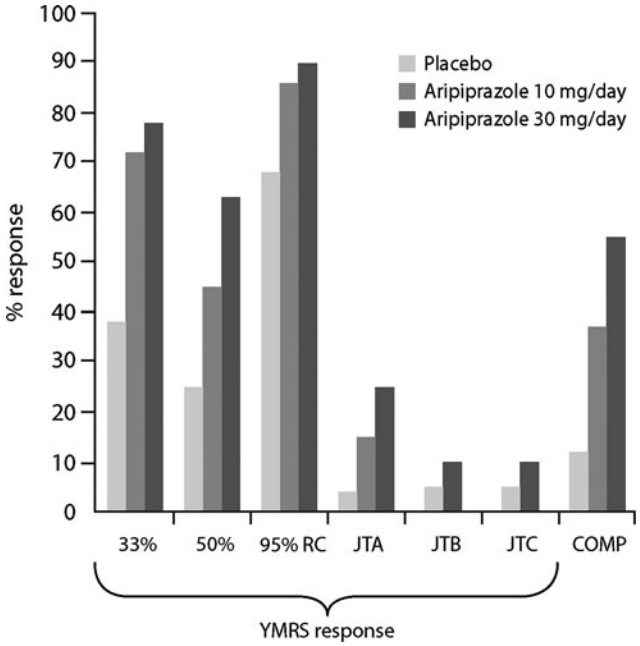

FIG. 2. YMRS response rates using different definitions of response and composite response (COMP*). *A composite of scores achieved on the YMRS (< 12.5), CDRS-R (≤ 40), and CGAS (≥ 51) (Findling et al. 2003). COMP, composite response; CDRS-R, Children's Depression Rating Scale-Revised; CGAS, Children's Global Assessment Scale; JTA, Jacobson and Truax efinition A; JTB, Jacobson and Truax definition B; JTC, Jacobson and Truax definition C; RC, reliable change; YMRS, Young Mania Rating Scale.

Calculation of response derived from COMP of scores achieved on the YMRS, CDRS-R, and CGAS resulted in response rates of 35% (NNT = 4) and 55% (NNT = 2) in the aripiprazole 10 and 30 mg/day groups, respectively, compared with 13% for placebo.

Criterion validity/agreement

Table 2 presents the results of κ tests of agreement between symptom improvement, defined according to different thresholds of response, and a CGI-BP Overall Improvement score of 1 or 2. Overall, the degree of agreement ranged from 0.15 to 0.64 among the various definitions of symptom improvement applied.

The measure with the highest validity (highest agreement) in terms of predicting clinician-rated global symptom improvement was ≥50% reduction on the YMRS ($\kappa = 0.64$) followed by the composite definition of response (COMP; $\kappa = 0.59$) and 95% reliable change on the CGAS or 33% reduction on the YMRS ($\kappa = 0.56$; Table 2). Notably, 50% and 33% reduction on the parent version of the GBI-M10 were associated with a magnitude of agreement equivalent to κ values of 0.48 and 0.43, respectively (Table 2).

Agreement was lowest for measures based on subject self-report, with κ values ranging from 0.20 to 0.25 for S-GBI-M10 and from 0.15 to 0.21 for S-GBI-D10. Additional analyses using CGI-BP Mania Improvement score as a benchmark (as opposed to CGI-BP Overall Improvement score) revealed κ values that ranged between 0.11 and 0.63. As generally observed for CGI-BP Overall Improvement score, highest agreement in terms of predicting mania symptom improvement was observed for response definitions comprising at least 50% reduction in YMRS score ($\kappa = 0.63$), 33%

Table 2. Kappa Values Using Various Outcome Measures (with Different Definitions of Response) and CGI-BP Overall Improvement Score = 1 or 2 as the Benchmark

Response definition	*YMRS*	*P-GBI-M10*	*S-GBI-M10*	*P-GBI-D10*	*S-GBI-D10*	*CDRS-R*	*CGAS*	*COMP*
33% reduction	**0.56**	0.43	0.20	0.35	0.21	0.27	0.52	–
≥50% reduction	**0.64**	0.48	0.22	0.33	0.21	0.10	0.37	–
95% reliable change	0.26	0.36	0.24	0.27	0.18	0.24	**0.56**	–
JTA	0.30	0.24	–	–	–	0.24	0.52	–
JTB	0.11	0.46	0.20	0.30	0.21	0.27	0.37	–
JTC	0.11	0.43	0.25	0.27	0.15	0.27	0.47	–
COMP[a]	–	–	–	–	–	–	–	**0.59**

Bold values represent the highest kappa values.

[a]A composite of scores achieved on the YMRS (<12.5), CDRS-R (≤40), and CGAS (≥51) (Findling et al. 2003).

CDRS-R, Children's Depression Rating Scale-Revised; CGAS, Children's Global Assessment Scale; COMP, composite response; JTA, Jacobson and Truax definition A; JTB, Jacobson and Truax definition B; JTC, Jacobson and Truax definition C; P-GBI-M10/S-GBI-M10, parent/subject 10-item version of General Behavior Inventory Mania; P-GBI-D10/S-GBI-D10, parent/subject 10-item version of General Behavior Inventory Depression; YMRS, Young Mania Rating Scale.

reduction on the YMRS ($\kappa = 0.58$) and COMP, or 95% reliable change on the CGAS ($\kappa = 0.53$).

Correlations between estimates of change

As shown in Figure 3, the relationship, as determined by correlation coefficients, between change in score on measures of depression and mania symptom severity, and change in CGI-BP Overall Improvement score, varied across measures. Among these, Spearman's rank correlation coefficients with change in CGI-BP Overall Improvement score were highest for CGI-BP Mania severity score ($r = 0.94$), CGAS ($r = 0.73$) and YMRS ($r = 0.69$), all of which were ratings made by the clinician based on the same interviews; thus sharing substantial method variance (Campbell and Fiske 1959). The magnitude of the correlation between measures of depression severity and CGI-BP Overall Improvement score was generally lower than for measures of mania (Fig. 3).

Of note, parent-reported versions of GBI-M10 and GBI-D10 demonstrated higher correlations with the CGI-BP Overall Improvement score than the corresponding subject-reported versions (GBI-M10: $r = 0.48$ vs. 0.29; GBI-D10: $r = 0.30$ vs. 0.22). Observed correlations between change in symptom severity scales and change in mania (as measured using CGI-BP Mania Improvement score) were highest for CGI-BP Overall Improvement score ($r = 0.94$), YMRS ($r = 0.72$) and CGAS ($r = 0.71$). Similar to CGI-BP Overall Improvement score, correlations between change in measures of depression severity and change in CGI-BP Mania Improvement score were of a lower order of magnitude ($r = 0.21$ to 0.45) than measures of mania severity ($r = 0.30$ to 0.72).

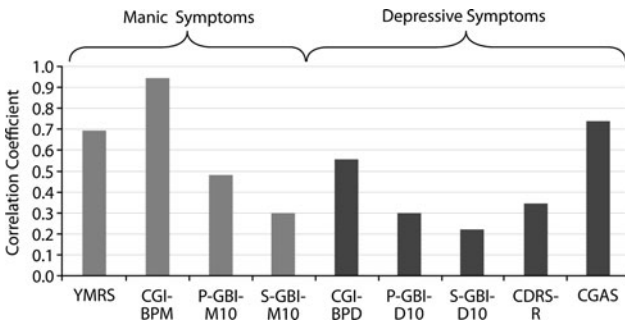

FIG. 3. Spearman correlation coefficients (absolute values) for changes in outcome measures compared with changes in CGI-BP Overall Improvement score (aripiprazole [10 and 30 mg/day] and placebo combined). CDRS-R, Children's Depression Rating Scale-Revised; CGAS, Children's Global Assessment Scale; CGI-BPD, Clinical Global Impressions–Bipolar Disorder Depression severity score; CGI-BPM, Clinical Global Impressions–Bipolar Disorder Mania severity score; P-GBI-M10/S-GBI-M10, parent/subject 10-item version of General Behavior Inventory Mania; P-GBI-D10/S-GBI-D10, parent/subject 10-item version of General Behavior Inventory Depression, YMRS, Young Mania Rating Scale.

Additional analyses of the correlations between change in parent- and clinician-rated measures on the GBI revealed statistically significant ($p < 0.05$) relationships between the change scores (parent-GBI-M10: YMRS, $r = 0.44$; CGI-BP Mania Improvement score, $r = 0.45$; CGAS, $r = 0.43$; and parent-GBI-D10: CGAS, $r = 0.28$; CDRS-R, $r = 0.37$; CGI-BP Depression Improvement score, $r = 0.33$).

Discussion

There is a lack of consensus regarding which definitions of response for assessment of treatment effects in both adult and pediatric clinical trials of bipolar disorder are the most clinically relevant. Available data from a study of nearly 300 children and adolescents participating in a randomized, placebo-controlled trial of aripiprazole for the treatment of an acute manic or mixed episode associated with bipolar I disorder provided an opportunity to conduct simultaneous comparisons of different definitions of response and clinically relevant change using multiple outcome measures in order to explore the significance of observed treatment effects. The comparison of multiple operational definitions within the same sample provides a sort of ''Rosetta Stone'' to help calibrate results across multiple studies. Comparing different trials is much more feasible when they use a shared comparator (e.g., placebo) and consistent operational definitions of outcome (e.g., Cipriani et al. 2011).

Not unexpectedly, our findings demonstrated considerable variability in response rates based on operational definitions. The composite algorithm or clinician ratings on the YMRS (a reduction in total score of ≥50%) and CGAS showed the best validity in terms of predicting a CGI-BP Overall Improvement score of 1 or 2 at endpoint ($\kappa = 0.6$ for each), which was chosen as the ''benchmark'' for meaningful clinical improvement. As the study population was enriched for patients with mania, analyses were also undertaken using CGI-BP Mania Improvement score as a benchmark, and provided comparable results. These observations support the use of a reduction of ≥50% in YMRS total score or a composite

measure combining response on mania, depression, and global functioning (YMRS < 12.5, CDRS-R ≤ 40, and CGAS ≥ 51) as a clinically relevant criterion for response in clinical trials for acute treatment of manic or mixed presentations of pediatric bipolar I disorder. The Jacobson and Truax definitions were much more conservative, basically eliminating placebo response. The rates of clinically significant change were low for active treatment as well, reflecting that normalization of functioning is a challenging goal for treatment of bipolar disorder. The NNT for active drug versus placebo ranged from 2 to 19 depending upon the definition of response used. The most liberal definition of response (e.g., 95% reliable change) had moderate NNT values because of high placebo response rates, whereas the most stringent Jacobson and Truax definitions had poorer NNT values because of lower response rates for active treatment, as well as placebo. The composite definition produced the most favorable NNT, but the Jacobson and Truax definitions highlight the difficulty of achieving normalization even with substantial treatment response.

Although the YMRS and CGAS scales and the CGI-BP Mania Improvement scores were among the three measures that demonstrated the highest correlation with change in CGI-BP Overall Improvement scores, the magnitude of these correlations (0.7–0.8) suggests that there is incomplete overlap between the symptoms measured using the various scales, and the symptom improvements measured by the CGI-BP Overall Improvement score. Conversely, the CGI-BP, YMRS, and CGAS all share similar methodology (i.e., clinician-rated measures based on a combination of interview and observation). This shared ''source variance'' ultimately boosts their correlation with each other, because each is asking the same clinician to make judgments about related aspects of the patient's functioning (Campbell and Fiske 1959).

Measurements based on parental responses were generally better indicators of changes in symptom severity compared with measurements based on subject response. For example, κ values denoting agreement between definitions of symptom improvement and CGI-BP Overall Improvement were lower when rated by subjects than when rated by their parents (~0.2 vs. ~0.4–0.5); although it should be noted that these differences were not tested statistically. This may be largely because of a relative lack of insight on the part of an adolescent experiencing an acute manic or mixed episode (Pini et al. 2001; Dell'Osso et al. 2002; Youngstrom et al. 2004). However, the baseline score of the patient versus parent reports should also be considered; the lower baseline average and smaller SD in patient-reported symptoms at baseline reduces the amount of change that could be observed in scores, as well as attenuating the observed correlations with other ratings (Cohen et al. 2003).

Overall, the correlation between symptom severity change and overall improvement in bipolar illness appeared to be stronger for symptoms of mania than for those of depression. Moreover, correlations between parent- and clinician-rated outcomes tended to be greater for mania than for depressive symptoms. Two aspects may partly account for the latter observations. First, the study population was enriched for mania (YMRS total score ≥ 20 required for enrollment), leading to mania showing a stronger contribution to overall clinical presentation. Second, despite the established antimanic effects of antipsychotics, fewer published data exist to substantiate their efficacy in the treatment of depressive symptoms (Fountoulakis 2010). The smaller amount of improvement in depression symptoms would contribute to correlations being of smaller magnitude.

The demonstrated relationship between the parent-reported GBI and change in other measures of mania and depression symptom severity suggests that the parent-GBI may provide a useful adjunct to evaluating symptom response in clinical research, particularly in situations in which clinicians have variable training (Garb 1998; Mackin et al. 2006; Dubicka et al. 2008; Jenkins et al. 2011) or when there are constraints on time or expense for assessment (Camara et al. 1998). Although agreement about most aspects of child behavior is often low among parent, youth, and teacher (Achenbach et al. 1987; Youngstrom et al. 2003; Carlson and Youngstrom 2011), consensus is that obtaining cross-informant report about youth behavior adds incremental value about severity (Carlson and Youngstrom 2003), situational specificity of problems and functioning (De Los Reyes and Kazdin 2005), and treatment response in clinical research. A National Institute of Mental Health expert consensus paper advocates using parent-reporting measures as a way of building a cross-talk between research groups and studies, providing an information source not subject to differences in clinical training or conceptualization (Nottelmann 2001). Present findings suggest potential value for adding parent-reporting measures in studies of bipolar disorder, not just for investigations of diagnosis and phenomenology, but also for studies of treatment outcome (West et al. 2011). A further argument for the inclusion of parent-reporting measures as an adjunct to clinician ratings is that parents are less likely than clinicians to ''inflate'' baseline scores in order to meet study inclusion criteria. In addition to the value of parent ratings in clinical research, the relationship between parent ratings and other measures of bipolar symptom severity also suggest that parent-reported measures may have value in general clinical practice.

Another implication is that the Jacobson and Truax definitions have been used to date in psychotherapy trials rather than in pharmacological treatment studies (Jacobson and Truax 1991). Although the definitions used more commonly in pharmacological trials produce higher response rates, they also result in higher placebo response rates. Observed results suggest that it is essential to examine in greater depth rather than simply comparing trials based on ''response rate,'' as choosing different definitions of response within this single study yielded response rates ranging from 0% to 91%. Although psychotherapy studies have adopted more stringent standards, it has not been established whether these more stringent definitions have greater validity when judged against criteria such as consumer satisfaction, quality of life, or academic progress versus simply being more conservative.

Limitations

Limitations of these analyses include the use of the YMRS scale, which does not cover all of the DSM-IV defined symptoms of mania; nor was the YMRS originally designed to be developmentally appropriate for children or adolescents (Fristad et al. 1992, 1995; Youngstrom et al. 2002). Results still support the use of the YMRS as an outcome measure, although it is possible that other instruments could be even more sensitive to treatment effects in this patient population. Second, this was a brief study of acute treatment response; longer study duration might provide a different picture regarding continued response and potential response to depressive symptoms. It should also be considered that differences between treatment groups at baseline may have influenced the results; this was not specifically evaluated. Finally, it would also be interesting for future studies to examine if the pattern of clinically significant change was associated with premature discontinuation from the treatment protocol.

Conclusions

In conclusion, this exploratory analysis of a large sample of adolescents and children with bipolar I disorder demonstrates that

clinically meaningful definitions of response in acute treatment of a manic or mixed episode include a 50% change in YMRS and a composite measure of response, as well as others. Results also provide evidence for the feasibility of implementing parent-reported measures of symptom improvement to complement clinician assessments.

The significant variability among definitions of response highlights a clear need for greater consensus among pediatric studies in order to enable more effective assessment of treatment efficacy, especially when the goal is not merely improvement on a scale but an overall clinically meaningful response.

Clinical Significance

No consensus has been established regarding the agreed definitions of response and clinically significant change in treatment of manic or mixed episodes associated with bipolar I disorder. Using a range of definitions of response, this analysis showed considerable variability in response rates, depending upon the choice of operational definition. Measurements based on parental responses were generally better at detecting changes in symptom severity than were measurements based on subject response.

Disclosures

Raymond Mankoski was an employee of Bristol-Myers Squibb at the time of this study. Ron Marcus is an employee of Bristol-Myers Squibb. Qiong Zhao, William Carson, and Robert McQuade are employees of Otsuka Pharmaceutical Development & Commercialization, Inc. Eric Youngstrom has received travel support from Bristol-Myers Squibb and consulted with Lundbeck about assessment. Robert L. Findling receives or has received research support, acted as a consultant, received royalties from, and/or served on a speaker's bureau for Abbott, Addrenex, Alexza, American Psychiatric Press, AstraZeneca, Biovail, Bristol-Myers Squibb, Dainippon Sumitomo Pharma, Forest, GlaxoSmithKline, Guilford Press, Johns Hopkins University Press, Johnson & Johnson, Kem-Pharm Lilly, Lundbeck, Merck, National Institutes of Health, Neuropharm, Novartis, Noven, Organon, Otsuka, Pfizer, Physicians' Post-Graduate Press, Rhodes Pharmaceuticals, Roche, Sage, Sanofi-Aventis, Schering-Plough, Seaside Therapeutics, Sepracore, Shionogi, Shire, Solvay, Stanley Medical Research Institute, Sunovion, Supernus Pharmaceuticals, Transcept Pharmaceuticals, Validus, WebMD, and Wyeth.

Acknowledgments

Editorial support for the preparation of this manuscript was provided by Ogilvy Healthworld Medical Education. We thank the patients and their parents for participating in this study.

References

Achenbach TM, McConaughy SH, Howell CT: Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. Psychol Bull 101:213–232, 1987.

American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV). Washington, DC: American Psychiatric Association; 1994.

Ankuta GY, Abeles N: Client satisfaction, clinical significance, and meaningful change in psychotherapy. Prof Psychol Res Pract 24:70–74, 1993.

Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP: Assessing clinical significance: Does it matter which method we use? J Consult Clin Psychol 73:982–989, 2005.

Camara W, Nathan J, Puente A. Psychological test usage in professional psychology: Report of the APA practice and science directorates. Washington, DC: American Pyschological Association Press; 1998.

Campbell DT, Fiske DW: Convergent and discriminant validation by multitrait-multimethod matrix. Psychol Bull 56:81–105, 1959.

Carlson GA ,Youngstrom EA: Clinical implications of pervasive manic symptoms in children. Biol Psychiatry 53:1050–1058, 2003.

Carlson GA, Youngstrom EA: Two Opinions About One Child–What's the Clinician To Do? J Child Adolesc Psychopharmacol 21:385–387, 2011.

Cipriani A, Barbui C, Salanti G, Rendell J, Brown R, Stockton S, Purgato M, Spineli LM, Goodwin GM, Geddes JR: Comparative efficacy and acceptability of antimanic drugs in acute mania: A multiple-treatments meta-analysis. Lancet 378:1306–1315, 2011.

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

Cohen J, Cohen P, West SG, Aiken LS: Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed. Hillsdale, NJ: Lawrence Erlbaum; 2003.

Danielson CK, Youngstrom EA, Findling RL, Calabrese JR: Discriminative validity of the general behavior inventory using youth report. J Abnorm Child Psychol 31:29–39, 2003.

De Los Reyes A, Kazdin AE: Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychol Bull 131:483–509, 2005.

Dell'Osso L, Pini S, Cassano GB, Mastrocinque C, Seckinger RA, Saettoni M, Papasogli A, Yale SA, Amador XF: Insight into illness in patients with mania, mixed mania, bipolar depression and major depression with psychotic features. Bipolar Disord 4:315–322, 2002.

Dubicka B, Carlson GA, Vail A, Harrington R: Prepubertal mania: Diagnostic differences between US and UK clinicians. Eur Child Adolesc Psychiatry17:153–161, 2008.

Findling RL, McNamara NK, Gracious BL, Youngstrom EA, Stansbrey RJ, Reed MD, Demeter CA, Branicky LA, Fisher KE, Calabrese JR: Combination lithium and divalproex sodium in pediatric bipolarity. J Am Acad Child Adolesc Psychiatry 42:895–901, 2003.

Findling RL, Nyilas M, Forbes RA, McQuade RD, Jin N, Iwamoto T, Ivanova S, Carson WH, Chang K: Acute treatment of pediatric bipolar I disorder, manic or mixed episode, with aripiprazole: A randomized, double-blind, placebo-controlled study. J Clin Psychiatry 70:1441–1451, 2009.

Follette WC, Callaghan GM: The evolution of clinical significance. Clin Psychol 8:431–435, 2001.

Fountoulakis KN: An update of evidence-based treatment of bipolar depression: Where do we stand? Curr Opin Psychiatry 23:19–24, 2010.

Fristad MA, Weller EB, Weller RA: The mania rating scale: Can it be used in children? A preliminary report. J Am Acad Child Adolesc Psychiatry 31:252–257, 1992.

Fristad MA, Weller RA, Weller EB: The mania rating scale (MRS): Further reliability and validity studies with children. Ann Clin Psychiatry 7:127–132, 1995.

Garb HN: Studying the Clinician: Judgment Research and Psychological Assessment. Washington, DC: American Psychological Association; 1998.

Guyatt GH, Rennie D: Users' Guides to the Medical Literature. Chicago: AMA Press; 2002.

Jacobson NS, Truax P: Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 59:12–19, 1991.

Jenkins MM, Youngstrom EA, Washburn JJ, Youngstrom JK: Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. Prof Psychol Res Pr 42:121–129, 2011.

Kafantaris V, Coletti DJ, Dicker R, Padula G, Kane JM: Adjunctive antipsychotic treatment of adolescents with bipolar psychosis. J Am Acad Child Adolesc Psychiatry 40:1448–1456, 2001.

Kazdin AE: The meanings and measurement of clinical significance. J Consult Clin Psychol 67:332–339, 1999.

Kendall PC, Marrs–Garcia A, Nath SR, Sheldrick RC: Normative comparisons for the evaluation of clinical significance. J Consult Clin Psychol 67:285–299, 1999.

Kim YS, Cheon KA, Kim BN, Chang SA, Yoo HJ, Kim JW, Cho SC, Seo DH, Bae MO, So YK, Noh JS, Koh YJ, McBurnett K, Leventhal B: The reliability and validity of Kiddie-Schedule for Affective Disorders and Schizophrenia–Present and Lifetime Version– Korean version (K-SADS-PL-K). Yonsei Med J 45:81–89, 2004.

Kowatch RA, Suppes T, Carmody TJ, Bucci JP, Hume JH, Kromelis M, Emslie GJ, Weinberg WA, Rush AJ: Effect size of lithium, divalproex sodium, and carbamazepine in children and adolescents with bipolar disorder. J Am Acad Child Adolesc Psychiatry 39:713–720, 2000.

Lunnen KM, Ogles BM: A multiperspective, multivariable evaluation of reliable change. J Consult Clin Psychol 66:400–410, 1998.

Mackin P, Targum SD, Kalali A, Rom D, Young AH: Culture and assessment of manic symptoms. Br J Psychiatry189:379–380, 2006.

Nottelmann ED: National Institute of Mental Health research roundtable on prepubertal bipolar disorder. J Am Acad Child Adolesc Psychiatry 40:871–878, 2001.

Ogles BM, Lunnen KM, Bonesteel K: Clinical significance: History, application, and current practice. Clin Psychol Rev 21:421–446, 2001.

Pini S, Cassano GB, Dell'Osso L, Amador XF: Insight into illness in schizophrenia, schizoaffective disorder, and mood disorders with psychotic features. Am J Psychiatry158:122–125, 2001.

Poznanski EO, Mokros HB. Children's Depression Rating Scale, Revised (CDRS-R). Los Angeles: Western Psychological Services; 1995.

Shaffer D, Gould MS, Brasic J, Ambrosini P, Fisher P, Bird H, Aluwahlia S: A children's global assessment scale (CGAS). Arch Gen Psychiatry 40:1228–1231., 1983.

Spearing MK, Post RM, Leverich GS, Brandt D, Nolen W: Modification of the Clinical Global Impressions (CGI) Scale for use in bipolar illness (BP): The CGI-BP. Psychiatry Res 73:159–171, 1997.

Speer DC: Clinically significant change: Jacobson and Truax (1991) revisited. J Consult Clin Psychol 60:402–408, 1992.

Speer DC, Greenbaum PE: Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. J Consult Clin Psychol 63:1044–1048, 1995.

Straus SE, Glasziou P, Richardson WS, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM, 4th ed. New York: Churchill Livingstone; 2011.

Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine: How to practice and teach EBM, 3rd ed. New York: Churchill Livingstone; 2005.

Van Meter AR, Moreira AL, Youngstrom EA: Meta-analysis of epidemiological studies of pediatric bipolar disorder. J Clinical Psychiatry72:1250–1256, 2011.

Well AD, Myers JL. Research Design and Statistical Analysis. 2nd ed.Hillsdale, NJ: Lawrence Erlbaum; 2003.

West AE, Celio CI, Henry DB, Pavuluri MN: Child Mania Rating Scale–Parent Version: A valid measure of symptom change due to pharmacotherapy. J Affect Disord 128:112–119, 2011.

Westlake WJ: Symmetrical confidence intervals for bioequivalence trials. Biometrics 32:741–744, 1976.

Young RC, Biggs JT, Ziegler VE, Meyer DA: A rating scale for mania: Reliability, validity and sensitivity. Br J Psychiatry 133:429–435, 1978.

Youngstrom E, Van Meter A, Algorta GP: The bipolar spectrum: myth or reality? Curr Psychiatry Rep12:479–489, 2010.

Youngstrom EA, Danielson CK, Findling RL, Gracious BL, Calabrese JR: Factor structure of the Young Mania Rating Scale for use with youths ages 5 to 17 years. J Clin Child Adolesc Psychol 31:567–572, 2002.

Youngstrom EA, Findling RL, Calabrese JR: Effects of adolescent manic symptoms on agreement between youth, parent, and teacher ratings of behavior problems. J Affect Disord 82 Suppl 1:S5–S16, 2004.

Youngstrom EA, Findling RL, Calabrese JR: Who are the comorbid adolescents? Agreement between psychiatric diagnosis, parent, teacher, and youth report. J Abnorm Child Psychol 31:231–245, 2003.

Youngstrom EA, Findling RL, Danielson CK, Calabrese JR: Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. Psychol Assess 13:267–276, 2001.

Youngstrom EA, Frazier TW, Demeter C, Calabrese JR, Findling RL: Developing a 10-item mania scale from the Parent General Behavior Inventory for children and adolescents. J Clin Psychiatry 69:831–839, 2008.

Address correspondence to:
Eric Youngstrom, PhD
Department of Psychology
University of North Carolina
Davie Hall, CB3270
Chapel Hill, NC 27599

E-mail: eay@unc.edu