Bipolar Disorders 2005: 7: 507–517

BIPOLAR DISORDERS

Original Article

Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings

Youngstrom E, Meyers O, Demeter C, Youngstrom J, Morello L, Piiparinen R, Feeny N, Calabrese JR, Findling RL. Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings.
Bipolar Disord 2005: 7: 507–517. © Blackwell Munksgaard, 2005

Objectives: To compare six promising mania measures, the Parent Mood Disorder Questionnaire (P-MDQ), the Adolescent self-report MDQ, the 10-item short form of the Parent General Behavior Inventory (PGBI-SF10), the 28-item Adolescent General Behavior Inventory (AGBI), the Parent Young Mania Rating Scale (P-YMRS), and the adolescent YMRS, in a demographically diverse outpatient sample.

Methods: Participants were 262 outpatients (including 164 males and 131 African-Americans) presenting to either an academic medical center (n = 153) or a community mental health center (n = 109). Diagnoses were based on semi-structured interviews with the parent and then youth sequentially.

Results: Ninety youths (34%) met criteria for a bipolar spectrum disorder. Parent measures yielded Areas Under the Receiver Operating Curve (AUROC) values of 0.81 for the PGBI-SF10 to 0.66 for the P-YMRS. Adolescent report measures performed significantly less well, with AUROCs ranging from 0.65 to 0.50. There were no significant differences in the diagnostic performance of the measures across the sites or by racial groups, although the reliability of measures tended to be lower in the urban community mental health site. The PGBI-SF10 made a significant contribution to logistic regression models examining all combinations of the instruments. The P-MDQ added information in the younger age group, and no measure improved classification of bipolar cases after controlling for the PGBI-SF10 in the older age group.

Discussion: Results replicate previous findings that, in decreasing order of efficiency, the PGBI-SF10, P-MDQ, and P-YMRS significantly discriminate bipolar from non-bipolar cases in youths aged 5–18; and they appear robust in a demographically diverse community setting. Adolescent self-report measures are significantly less efficient, sometimes performing no better than chance at detecting bipolar cases.

Eric Youngstrom[a], Oren Meyers[b], Christine Demeter[b], Jen Youngstrom[c], Laura Morello[b], Richard Piiparinen[b], Norah Feeny[b], Joseph R Calabrese[b] and Robert L Findling[b]

[a]Department of Psychology, Case Western Reserve University, [b]Department of Psychiatry, Case Western Reserve University/University Hospitals of Cleveland, [c]Applewood Centers Inc., Cleveland, OH, USA

Key words: bipolar disorder – children and adolescents – sensitivity and specificity

Received 7 December 2004, revised and accepted for publication 6 June 2005

Corresponding author: Eric A Youngstrom, PhD, Department of Psychology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7123. Fax: 216 368 4891;
e-mail: eric.youngstrom@case.edu

Attention to bipolar disorder in children and adolescents has increased rapidly in the last decade. Research in this area has been somewhat hampered by the lack of good instruments to screen samples to identify youths at risk of bipolar disorder for participation in prevention or treatment studies as well as investigations of genetics. There also would be great value in having tools that could be used in a standardized manner across research groups. Such measures would help determine whether apparent differences in phenomenology across sites

The authors of this paper do not have any commercial associations that might pose a conflict of interest in connection with this manuscript.

are due to actual differences in presentation, versus being attributable to differences in interview procedures or the conceptual definition of 'bipolar disorder' used by different groups (1). Also, bipolar disorder has proven challenging to recognize clinically, even in adults, with patients often going more than 10 years between the onset of mood symptoms and the formal diagnosis of a bipolar spectrum disorder (2). Well-validated rating scales could help increase the rate of identification, helping address unmet need for treatment of bipolar disorder (3).

To date, all published investigations of diagnostic efficiency in pediatric bipolar disorder have been based on outpatient samples at academic medical research centers. More than 80% of all participants have been European-American and middle or upper socioeconomic status (SES) families. It would be highly valuable to establish whether these measures perform comparably well in other clinical settings, such as community mental health or justice settings, where much of the unmet need for treatment of mood disorder is likely to occur. A community mental health setting is likely to involve challenging conditions, including a low base rate of bipolar disorder, high rates of attention-deficit/hyperactivity disorder (ADHD) and disruptive behavior disorders, increased demographic diversity, a larger percentage of patients coming from lower SES families, and potentially influential differences in reading ability and attitudes toward mental health issues.

It also would be useful to look specifically at whether measures performed similarly in African-American and European-American subsamples. Almost no work has been published on the presentation or treatment of pediatric bipolar disorder in minority populations. However, bipolar disorder is more likely to be diagnosed in European-Americans, and schizophrenia more likely in African-Americans presenting with similar psychotic symptoms (4–7), even though there do not appear to be significant racial differences in the symptom presentation or family history of cases diagnosed with bipolar I disorder (8–10). Similar patterns have been found in clinical diagnoses of psychiatrically hospitalized adolescents (11), and differences in rate of bipolar diagnosis persist even when using semi-structured research interviews (5). These findings suggest that rating scales could be particularly helpful in detecting bipolar disorder in minority populations, inasmuch as they de-emphasize clinical judgment in making ratings as well as differences in interview content (7), and thus may offer a relatively objective indicator of risk of mood disorder.

The primary goal of the present study was to compare the diagnostic efficiency of six promising measures to assess bipolar disorder in youths, investigating the parent- and self-report versions of the General Behavior Inventory (GBI) (12–19), Mood Disorder Questionnaire (MDQ) (20–22), and Young Mania Rating Scale (YMRS). All these measures possess potential advantages as aids to differential diagnosis in a wide range of clinical settings: they are inexpensive, readily available, and require minimal training to consistently administer and interpret (23). The most widely used instrument across research and clinical groups, the Achenbach Child Behavior Checklist (24, 25), has been helpful as a method of comparing phenotypes across sites (19, 26–28). However, the Child Behavior Checklist (CBCL) has several shortcomings when applied to bipolar disorder, the paramount being that it does not have a mania scale, and does not include many of the core DSM symptoms of mania (25, 29). Data indicate that the less familiar rating scales investigated in this study outperform the CBCL at detecting bipolar disorder, presumably because they focus more specifically on manic content (19).

Based on prior findings (16, 19, 21, 30), we hypothesized that parent report would do significantly better than self-report on the same instrument, e.g., Parent GBI or Parent MDQ would outperform adolescent report on the Adolescent GBI or Adolescent MDQ. Because all the measures studied here concentrate on symptoms of mania, we anticipated much smaller differences in diagnostic efficiency within informant (i.e., all three parent measures might perform similarly).

A second major goal was to compare the performance of the measures across two different clinical settings, comparing cases drawn from an urban community mental health setting to cases presenting at an outpatient academic research center. We hypothesized that performance would be significantly lower in the community mental health setting, because of differences in demographics and rates of other disruptive behavior disorders.

A third goal was to re-analyze the data comparing diagnostic efficiency in European-American versus African-American sub-groups, formally testing whether the measures' performance differed significantly across racial groups. We anticipated that if there were statistically significant differences in performance, then the measures would perform worse in the African-American subsample.

A fourth goal was to examine whether any combination of measures could significantly improve the detection of bipolar disorder.

Although prior studies have found that combinations of measures do not significantly improve classification of bipolar disorder after controlling for the best measure (16, 19), the current study includes more measures that focus specifically on mania and mixed states, and the items range broadly from the DSM-IV symptoms (comprising the MDQ) to associated features of mania (the YMRS versions), or a mix of both (as found on the GBI forms). There is some evidence that data from multiple informants could help identify mania, or at least indicate more impaired cases (31, 32).

Methods

Participants

The Institutional Review Boards of University Hospitals of Cleveland, Case Western Reserve University, and Applewood Centers, Incorporated approved all procedures used here. Participants were recruited from two distinct clinical infrastructures. One was a community mental health center (CMHC) with four urban sites (28). A random subsample of families presenting for outpatient treatment were invited to participate. The only exclusionary criteria were that the patient needed to be between the age of 5 and 18, and the patient and caregiver needed to be conversant in spoken English in order to complete the interviews.

The other infrastructure was an outpatient academic medical center with more than a dozen different pharmacotherapy studies, depending upon currently open protocols (described more fully in Findling et al., 33). Target diagnoses for study recruitment included bipolar disorder [bipolar I, bipolar II, cyclothymia or bipolar not otherwise specified (NOS)], unipolar depression, ADHD, conduct disorder, and aggressive behavior regardless of diagnosis. Recruitment was based on presenting symptoms and willingness to participate in treatment protocols. Advertisements and referrals described treatment studies, and those families interested in various treatment studies completed the diagnostic assessment as a screening or baseline evaluation. The sample was enriched by referrals of children whose parents had a diagnosed bipolar disorder and were participating in treatment or research at an affiliated adult mood disorders clinic. In addition, youths (including normal controls) were recruited by flyers and word of mouth to complete these descriptive psychometric instruments under the auspices of a Child/Adolescent Psychiatric Clinical Research Center.

Inclusion criteria were (i) youths between the ages of 5 years 0 months and 17 years 11 months of age, (ii) of either gender, (iii) of any ethnicity, (iv) presenting for an outpatient evaluation for which the youth provided written assent and the guardian provided written consent for participation, and (v) both the youth and the primary caregiver presented for the assessment. In addition, at both sites, both the youth and the parent needed to be able to communicate orally at a conversational level in English for inclusion in this study.

Subjects were excluded from enrollment into this study at the academic site if a pervasive developmental disorder, as determined by psychiatric history, psychiatric interview, or having an Autism Screening Questionnaire score of 15 or higher (34), was present. In addition, patients with suspected moderate, severe or profound mental retardation – documented via educational history, standardized cognitive ability test scores < 70, or a Peabody Picture Vocabulary Test-Third Edition (35) score < 70 if there was suspicion that global ability might fall below 70 – was exclusionary, as the family would not meet eligibility criteria to participate in the clinical trials for which this protocol acted as a screening battery. All participants completed the same assessment procedures, including the index tests and reference standard diagnostic interview, regardless of presenting symptoms or treatment study eligibility. The design was 'prospective' in the sense that data collection and analyses were planned before the index test and reference standard were performed (36), as opposed to *post hoc* examination of a variety of measures collected for a different purpose.

Measures

Reference standard: semi-structured diagnostic interview using the Schedule of Affective Disorders and Schizophrenia for Children. All participants and their families completed a semi-structured diagnostic interview by a highly trained research assistant, using the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime (KSADS-PL) (37) combined with the mood disorders module from the Washington University KSADS (WASH-U-KSADS) (38). The mood disorders module of the WASH-U-KSADS includes additional symptoms and associated features of depression and mania not captured by other structured or semi-structured instruments. Diagnoses of bipolar I, bipolar II, cyclothymia, and bipolar NOS were made in strict accordance with diagnostic criteria published in the DSM-IV (American Psychiatric Association, 1994). Failure to meet strict durational criteria was the most

common reason for diagnosing bipolar NOS instead of one of the other bipolar diagnoses (39).

Research assistants (n = 4 predoctoral interns, 3 PhDs, 1 MA, and 2 psychology BA raters), were trained to criterion by having them rate along while observing five KSADS interviews by an experienced rater. New raters then led five KSADS interviews with an experienced rater and achieved an overall $\kappa > 0.85$ at the symptom severity level and 1.0 agreement about the presence or absence of diagnoses on each in order to graduate from training. Acceptable inter-rater reliability ($\kappa > 0.85$ about symptom severity) was maintained by having monthly joint rating sessions after training was completed. The same interviewer worked with both informants, resolving discrepancies using best clinical judgment. All cases were reviewed by an expert consensus team, with the review always involving a licensed clinical psychologist (EAY, JKY, or NCF) and the rater conducting the KSADS. The consensus meeting could use the KSADS, family history, and prior treatment history to assign the consensus diagnosis. All individuals involved in the consensus meeting remained blind to the scores on the index tests. Kappa was 0.95 about bipolar diagnoses and 0.91 about all diagnoses when comparing the expert consensus to KSADS diagnoses. Two cases that met criteria for bipolar disorder on the KSADS were assigned other diagnoses by the expert consensus team. One case met criteria for bipolar NOS on the KSADS, but this was changed to depression NOS with comorbid post-traumatic stress disorder upon review. The other case met criteria for cyclothymia on the KSADS, and this was revised to ADHD-combined type and comorbid mood disorder NOS upon review by the expert consensus team. Both cases represent modest disagreement between the KSADS and the expert consensus, and reflect a conservative approach to diagnosing bipolar spectrum disorders.

Index tests

Mood Disorder Questionnaire-Adolescent self-report (A-MDQ) (20). The MDQ was designed specifically as a screening instrument for bipolar disorder. It includes an item for each of the DSM-IV symptoms of mania, along with an item asking if many of the symptoms co-occurred at the same time, and another item asking if there was impairment associated with the symptoms. Items are scored as being present or absent. The MDQ was validated in an adult population. The present study represents one of the first efforts to examine the validity of self-report in an adolescent population (21).

Mood Disorder Questionnaire-Parent report about youth (P-MDQ) (21). Parents were also asked to complete a slightly modified version of the MDQ, where they reported about potential manic symptoms in their child. The MDQ-P is promising as a diagnostic aid, because it uses parent report, it is brief, and it is focused specifically on symptoms of mania. Preliminary data from another sample suggest that it would out-perform self report on the MDQ (21).

Adolescent self-report Young Mania Rating Scale (A-YMRS). The A-YMRS is a 11-item questionnaire adapted from the YMRS (40) for the present study. To our knowledge, this is the first published exploration of whether adolescents could use the YMRS as a questionnaire and provide reliable and clinically valid information. Adolescents rate their own manic symptoms on five explicitly defined grades of severity, with item scores ranging from 0 to 4 (and three items ranging from 0 to 8). The A-YMRS yields a total score that can range from 0 to 56, with higher scores representing greater pathology. Ratings were based on the reported presence of symptoms over the past 2 weeks.

Parent Young Mania Rating Scale (P-YMRS) (41). The P-YMRS is a 11-item questionnaire adapted from the YMRS. Parents rated their child's manic symptoms on five explicitly defined grades of severity, using the same anchors as the Adolescent-rated and clinician-rated versions of the YMRS. Ratings were based on the reported presence of symptoms over the past 2 weeks. Internal consistency has been adequate in previous samples (e.g., $\alpha = 0.80$ in the age 5–10 sample, and 0.69 in the older sample) (19, 41, 42).

Adolescent self-report on the General Behavior Inventory (A-GBI) (12). The GBI is a 73-item self-report questionnaire measuring depressive, hypomanic, manic, and mixed ('biphasic') mood symptoms used with adolescents as young as age 11 (15). Respondents rate each symptom on a 0 (*never or hardly ever*) to 3 (*very often or almost constantly*) Likert-type scale, with higher scores indicating greater severity. The GBI yields two scales scores, a depressive ($\alpha = 0.96$) and a hypomanic/biphasic score ($\alpha = 0.94$) (15). Present analyses use the hypomanic/biphasic score, as preliminary findings indicate that this is the scale that best discriminates bipolar spectrum disorders from other diagnoses (15, 16, 19).

Parent General Behavior Inventory (P-GBI) (17). The P-GBI is an adaptation of the GBI, modified so that parents complete it to rate the depressive, hypomanic, manic, and biphasic mood symptoms of their children aged 5–17. The two scales of depressive and hypomanic/biphasic symptoms have strong construct validity and exceptionally high internal consistency (e.g., alphas of 0.97 for depression and 0.94 for hypomanic/biphasic in both age groups) (17). A 10-item mania short form that was developed on an independent sample appears quite promising as a diagnostic aid (PGBI-SF10) (18). The 10 items were extracted from the full-length version at the CMHC site to replicate the performance of the short form and investigate its performance in a more demographically heterogeneous sample. At the academic site, the PGBI-SF10 was administered as a short form (i.e., only the 10 items were collected), to begin to evaluate its psychometric properties when completed in the short format (43).

Procedure

The parent or guardian provided written consent for the participation of their child, and all youths provided written assent to participation. All participants and their families completed the KSADS diagnostic interview. The interviewer met the adolescent and parent separately. While the youth was being interviewed, parents completed the P-GBI, P-MDQ, and P-YMRS questionnaires. When the parent was completing the KSADS interview, then youths aged 11–17 were given the A-MDQ, A-YMRS, and A-GBI to complete. Youths younger than 11 years did not complete any of the self-report instruments. Widely used, nationally standardized instruments such as the Achenbach CBCL begin using youth self-report information at age 11; so our research protocol also began gathering other self-report measures at age 11. Youths and parents did not have access to each other's responses on the rating scales. The KSADS diagnoses were blind to the content of the rating scales, which were scored after the completion of the interview.

Statistical methods

The primary criterion measure for all analyses grouped youths into two categories: (i) those with no diagnosis of a bipolar spectrum disorder, although multiple other Axis I diagnoses might be present, and (ii) those with any bipolar spectrum disorder (i.e., bipolar I, bipolar II, cyclothymia, or bipolar NOS) present, regardless of comorbidity. The overall diagnostic efficiency of each test was quantified using nonparametric estimates of the Area Under the Receiver Operating Curve (AUROC) analyses. AUROC simultaneously considers two aspects of test performance: sensitivity (or the percentage of bipolar cases correctly identified) versus specificity (or the percentage of non-bipolar cases correctly identified). We compared the diagnostic efficiency of the different index tests (e.g., parent versus youth report) within each sample using the *z*-test of dependent AUROCs (44). *Z*-tests of independent samples compared the performance of the measures across sites, and across race. Logistic regression analyses determined whether combinations of the index tests provided any incremental value after interpreting an individual index test (45).

Results

Participants

Across both sites, 261 youths aged 5–17 participated, comprising a consecutive case series over the period from December 2002 to October 2004. More than half of the caregivers earned < $20,000 per year, and another third earned in the range of $20,000 to $40,000. Table 1 presents demographic characteristics and KSADS primary diagnoses separately by site. The CMHC sample (n = 109) was 55% male, 86% black, and 80% qualifying for Medicaid; whereas the academic research sample (n = 153) was 68% male, 24% black, and < 20% of families qualifying for Medicaid (all demographic differences significant at $p < 0.01$). No adverse events were reported as a result of completing the index tests or KSADS.

Diagnostic efficiency statistics

Table 2 presents descriptive statistics for the index tests and global functioning separately for the criterion groups (KSADS bipolar diagnosis present or absent). Table 3 presents the internal consistency estimates of reliability separately by site. The reliability statistics are somewhat lower in the CMHC than in the academic center, consistent with concerns about differences in the reading level of participants. The differences in reliability are unlikely to be the result of differences in the severity of presentation of bipolar illness across sites: there was no site-by-diagnosis interaction in levels of global functioning ($F(1,214) = 0.00$, $p = 0.984$) and there was a significant tendency for cases at the CMHC to be more impaired regardless

Table 1. Demographic and diagnostic characteristics presented separately by site

Characteristic	Academic center (n = 153)	Community mental health center (n = 109)
Age in years (SD)	10.5 (3.5)	11.1 (3.1)
Gender (male) (%)	104 (68)*	60 (55)
Ethnicity (%)		
African-American	37 (24)	94 (86)***
Hispanic	7 (5)*	0 (0)
White	106 (69)***	9 (8)
Other	3 (2)	6 (6)
Reference standard positive (%)		
Bipolar I	42 (27)**	5 (5)
Bipolar II, NOS, cyclothymia	39 (25)*	5 (5)
Reference standard negative (%)		
Unipolar depression (MDD, dysthymia, adjustment d/o + depressed mood)	20 (13)	29 (27)*
ADHD or disruptive behavior w/o mood disorder	39 (25)	51 (47)*
Residual (anxiety, post-traumatic stress disorder, psychotic disorders, or no Axis I)	14 (9)	18 (17)
Any ADHD (%)	103 (67)	62 (57)
Number of Axis I diagnoses (SD)	2.3 (1.1)	2.3 (1.3)

For present purposes, any mood diagnosis was considered 'primary.' Those with primary bipolar diagnoses also met criteria for 0–6 (median = 1) other DSM-IV Axis I diagnoses. The most common comorbidity was bipolar and ADHD, occurring in 72% of the cases diagnosed with bipolar disorder. ***p < 0.0005, **p < 0.005, *p < 0.05, two-tailed; NOS = not otherwise specified; MDD = major depressive disorder; ADHD = attention-deficit/hyperactivity disorder.

Table 2. Index test scores and global assessment of functioning (GAF) for youths with and without bipolar diagnoses (combined across both sites)

	Non-bipolar		Bipolar		
Index Test	Mean	SD	Mean	SD	Cohen's *d*
Ages 5–10 (n = 141)	n = 82		n = 59		
P-YMRS	10.64	6.72	15.40	8.64	0.62***
P-MDQ	5.30	3.20	7.70	2.54	0.81***
P-GBI-SF-10	11.19	8.07	20.60	8.59	1.13***
GAF	56.38	9.23	53.78	7.87	0.33 ns
Ages 11–17 (n = 124)	n = 94		n = 30		
P-YMRS	9.83	6.47	14.43	6.67	0.62**
P-MDQ	4.99	3.26	8.17	3.42	0.84***
P-GBI-SF-10	11.49	8.55	21.95	8.20	1.08***
A-YMRS	10.34	6.82	10.24	6.66	0.00 ns
A-MDQ	5.14	3.27	6.69	3.06	0.41*
A-GBI-BH	22.90	16.48	31.09	15.79	0.43*
GAF	54.73	7.33	54.90	5.72	0.00 ns

Cohen's *d* of 0.2 constitutes a small effect size, 0.5 a medium, and 0.8 a large effect for the social sciences. Bipolar versus non-bipolar differences ***p < 0.0005, **p < 0.005, *p < 0.05, two-tailed; ns = not significant. PGBI-SF10 = Parent report on the 10-item short form of the General Behavior Inventory; P-MDQ = Parent report on the Mood Disorder Questionnaire; P-YMRS = Parent report on the Young Mania Rating Scale questionnaire; AGBI-BH = Adolescent report on the Biphasic/Hypomanic scale of the General Behavior Inventory; A-MDQ = Adolescent report on the Mood Disorder Questionnaire; A-YMRS = Adolescent report on the Young Mania Rating Scale questionnaire.

Table 3. Internal consistency reliability of the index tests of manic and hypomanic symptoms presented separately by site

		Cronbach's alpha		
Index test	Items	CMHC	Academic center	Combined
A-MDQ	13	0.76	0.84	0.80
P-MDQ	13	0.83	0.82	0.83
A-YMRS	11	0.66	0.75	0.71
P-YMRS	11	0.67	0.74	0.73
A-PGBI HB	28	0.91	0.96	0.94
P-GBI HB	28	0.92	0.94	0.92
PGBI-SF10	10	0.86	0.92	0.93

CMHC = Community Mental Health Center; PGBI-SF10 = Parent report on the 10 item short form of the General Behavior Inventory; P-MDQ = Parent report on the Mood Disorder Questionnaire; P-YMRS = Parent report on the Young Mania Rating Scale questionnaire; AGBI-BH = Adolescent report on the Biphasic/Hypomanic scale of the General Behavior Inventory; A-MDQ = Adolescent report on the Mood Disorder Questionnaire; A-YMRS = Adolescent report on the Young Mania Rating Scale questionnaire.

of diagnosis, with an average GAF of 53.0 versus 56.5 (F (1,214) = 14.38, p < 0.0005, partial eta-squared = 0.06). Bipolar cases tended to show poorer functioning regardless of site (F(1,214) = 7.56, p = 0.006, partial eta-squared = 0.03). The differences in reliability were more pronounced in self-report than in parent report.

Table 4 presents correlations among the potential screening variables as well as the AUROC. The PGBI-SF10, P-MDQ, and P-YMRS earned AUROCs larger than for the A-MDQ, A-GBI, or A-YMRS. The PGBI-SF10 significantly outperformed all three adolescent report measures as well as the P-YMRS in the older cohort (ages 11–17)

Table 4. Index test correlations and global measures of diagnostic efficiency, combined across sites

Correlations	PGBI-SF10	P-MDQ[a]	P-YMRS	AGBI-BH	A-MDQ[a]	A-YMRS
Ages 5–10 (n = 141)						
PGBI-SF10	1.00					
P-MDQ	0.58***	1.00				
P-YMRS	0.54***	0.55***	1.00			
Area under curve	0.79***	0.72***	0.66**			
95% CI	0.71–0.87	0.63–0.81	0.57–0.76			
Ages 11–17 (n = 124)						
PGBI-SF10	1.00					
P-MDQ	0.56***	1.00				
P-YMRS	0.57***	0.59***	1.00			
AGBI-BH	0.28**	0.35***	0.35***	1.00		
A-MDQ	0.12	0.22*	0.27**	0.63***	1.00	
A-YMRS	0.14	0.31**	0.31**	0.46***	0.42***	1.00
Area under curve	0.81***	0.75***	0.70**	0.65*	0.63*	0.50
95% CI	0.72–0.90	0.64–0.86	0.59–0.81	0.53–0.76	0.51–0.75	0.37–0.63

***p < 0.0005, **p < 0.005, *p < 0.05, two-tailed. Results are combined across sites because there were no statistically significant differences in Areas Under Curve. PGBI-SF10 = Parent report on the 10-item short form of the General Behavior Inventory; P-MDQ = Parent report on the Mood Disorder Questionnaire; P-YMRS = Parent report on the Young Mania Rating Scale questionnaire; AGBI-BH = Adolescent report on the Biphasic/Hypomanic scale of the General Behavior Inventory; A-MDQ = Adolescent report on the Mood Disorder Questionnaire; A-YMRS = Adolescent report on the Young Mania Rating Scale questionnaire. Confidence intervals based on nonparametric estimation.

[a]Logistic regressions demonstrated that the items asking about co-occurrence (no. 14) and impairment (no. 15) added nothing to the prediction of bipolar disorder in either age group after controlling for the MDQ total score. Requiring co-occurrence and at least moderate self-reported impairment (as per the MDQ instructions) increased specificity slightly, but at the cost of marked reductions in sensitivity.

(all $p < 0.05$). The PGBI-SF10 also outperformed the P-YMRS in the younger cohort ($z = 2.92$, $p < 0.01$). The PGBI-SF10 and the P-MDQ were not statistically reliably different in their performance in either age group.

The full-length, 28-item PGBI Hypomanic/Biphasic scale was available in a subset of 152 cases, as noted above. The PGBI-SF10 performed equally as well as the full-length scale, generating slightly higher AUROCs in both age cohorts than did the full-length scale.

Comparing parent versus youth report

As a set, the parent-report measures all produced larger AUROCs than did the adolescent-report scales. Based on Hanley & McNeil's test of dependent AUROCs, the PGBI-SF10 outperformed all adolescent-report measures ($p < 0.01$ in all cases). Unexpectedly, the PGBI-SF10 also performed significantly better than parent report on the P-YMRS ($p < 0.05$). When comparing parent to adolescent report on the same instrument, parent report was significantly better on both the GBI and the YMRS. The MDQ showed a trend in favor of parent report, with AUROC values of 0.75 versus 0.63 ($p = 0.13$). The PGBI-SF10 tended to show larger AUROC values than the P-MDQ in both age groups, but the differences were not statistically significant.

Comparing performance in academic versus community settings

There were no statistically significant differences in the performance of the measures between the CMHC and academic sites. The largest z-value was 0.98, comparing the P-MDQ AUROCs of 0.82 at the CMHC versus 0.72 at the academic site ($p = 0.325$).

Comparing performance between races

There were no statistically significant differences between the performance of the measures in the African-American versus European-American patients on any of the measures. The largest discrepancy was on the PGBI-SF10, where the AUROC was 0.75 in European-American participants and 0.82 in African-American participants. Both groups had a standard error of the AUROC of 0.05, and the z-value of 1.01 was not significant ($p = 0.315$). There was no trend evident when comparing the performance of the measures across site or racial groups, suggesting that the similarity of performance was not due to a lack of statistical power.

Evaluating performance of combinations of index tests

The PGBI-SF10 produced the largest AUROC values of any of the tests. Logistic regressions

evaluated whether any of the other index tests provided statistically significant information after controlling for the PGBI-SF10 score. In the 5–10 year-old group, the P-MDQ provided a statistically significant improvement in the prediction of bipolar spectrum disorder, with an associated Wald value of 3.89 ($p = 0.049$). The addition of the P-MDQ resulted in the correct identification of two bipolar cases that would have been misclassified on the basis of the PGBI-SF10 alone. This finding is interesting, because the P-MDQ conforms directly to the DSM-IV symptoms of mania, and the PGBI-SF10 relies on a statistically driven set of associated features that maximally discriminate bipolar cases (18). At the same time, it will require cross-validation in larger samples, given the general negative findings about the value of combining scales and the relatively modest improvement shown here.

In the older sample (ages 11–17), no scale made a significant contribution after controlling for the PGBI-SF10. None of the adolescent self-report measures came close to making a statistically significant contribution after controlling for the PGBI-SF10, contradicting the conventional emphasis on gathering data from multiple informants.

An additional set of logistic regressions tested whether the PGBI-SF10 discriminated bipolar cases because of manic symptomatology, or because it was a marker for general impairment. Neither GAF score nor number of comorbid Axis I diagnoses (both indicators of severity) predicted a bipolar diagnosis, and PGBI-SF10 continued to make a statistically significant ($p < 0.0005$) contribution even after controlling for both GAF and comorbid diagnoses.

Discussion

This study compared the diagnostic efficiency of six different rating scales as tools to facilitate the accurate diagnosis of bipolar disorders in youths aged 11–17. It also compared the performance of three parent measures in a younger sample aged 5–10. Results replicate previous findings that the Parent versions of the PGBI Short Form, MDQ, and YMRS significantly discriminate bipolar from non-bipolar cases in youths aged 5–17 (19, 21, 41). The PGBI Short Form performed significantly better than the P-YMRS. Consistent with previous findings, adolescent self-report is significantly less efficient on each measure, with adolescent report on the YMRS not performing better than chance in these data. The A-MDQ and A-GBI discriminated bipolar disorder at rates comparable to what had been reported in the initial validation samples (15, 21). These findings underscore the value of involving a parent or other familiar adult in the assessment process when evaluating potential bipolar disorder. The relatively good performance of parent report makes some sense given the substantial time that parents typically spend observing child behaviors, as well as the cognitive developmental constraints on the reliability and validity of younger children's self-report (46). Manic behaviors in youths appear to be more externalizing than internalizing in nature, and collateral informants are often more valid as reporters of externalizing symptoms. It is also likely that manic symptoms, which are often associated with a lack of insight into one's own behavior, may have further undermined the validity of adolescent self-report (47).

The present results are also encouraging because measures that have performed well in academic research samples performed approximately equally well in an urban community mental health setting. The academic and community health settings were significantly different on a variety of demographic measures, including gender, ethnicity, SES, and education and insurance status. The two samples presented with similar rates of Axis I diagnoses, but with the community setting having a high rate of externalizing disorders and a statistically significantly worse level of global functioning. In spite of these pronounced differences, there was no reliable difference in the diagnostic efficiency of the measures between sites or when comparing the African-American to European-American subsets. There was a slight decrease in the reliability of the adolescent self-report measures in the community mental health site, but this was not associated with any meaningful change in diagnostic performance. These results suggest that the measures are likely to perform well as diagnostic aids for the detection of pediatric bipolar disorder across a range of ages, gender, and demographic variables. This is encouraging, because although it is widely believed that sensitivity and specificity are intrinsic properties of a test, they actually can vary dramatically from sample to sample (48, 49). For the sake of comparison, the AUROC of ~0.80 for the PGBI-SF10 is roughly comparable to a Cohen's *d* effect size of 1.19 (where 0.80 would be considered a 'large' effect based on Cohen's guidelines), or a test yielding sensitivity of 0.75 and specificity of 0.85 (or vice versa) if the test responses were normally distributed in both the bipolar and non-bipolar groups.

The results also demonstrated that there was no advantage to combining parent and youth report

measures. This contradicts the conventional recommendation to gather information from multiple sources, but it is consistent with previous evaluations of batteries to identify pediatric bipolar disorder (16, 19). Based on these replicated findings, it seems prudent to involve parents in assessment of potential pediatric bipolar disorder, and to have them complete the most valid instrument available. Although low scores on a variety of different parent measures appear similarly good at ruling out bipolar disorder, the PGBI-SF10 appears to be one of the best measures available for helping increase the likelihood of an accurate positive diagnosis (18). When both parent and youth measures are available, then the parent measure should take precedence, and there does not appear to be value added by having the teenager complete any of the measures investigated here, nor the Achenbach Youth Self Report based on other findings (19, 30). These results should not be interpreted as indicating that the youth should not be involved in the assessment process (32, 50), but rather as a rebuttal of the value of having them complete these specific instruments as a rating scale to aid in the diagnostic process.

The findings suggest that the PGBI-SF10 and the P-MDQ both could be used as screening devices to identify bipolar disorder in youths. The P-YMRS performed significantly less well than the PGBI-SF10, replicating previous results in a larger sample (18). The PGBI-SF10 showed a statistical advantage as a screener in logistic regression analyses. If these results prove robust, then the PGBI-SF10 and P-MDQ could be combined in a multiple gating strategy for use with prepubertal children, or either one could be used as a screen and the other could be used as a symptom measure for quantifying outcomes in the same sample. The PGBI-SF10 and P-MDQ are intriguing to use in tandem, because they are quite different in terms of format and symptom content. The A-MDQ and A-YMRS are not recommended for use in adolescents as a screener based on weak performance in the present samples as well as the data reported by Wagner et al. (21). If only one measure could be used, then the PGBI-SF10 appears most promising based on its performance compared with the other measures in this and prior samples (18).

A major strength of the present study is that analyses are based on heterogeneous clinical cohorts. The community sample was randomly drawn from all families presenting to the community clinics, in order to be maximally generalizable to those settings. Diagnostic efficiency can change a lot, depending on whether highly purified or comparatively unfiltered and comorbid samples are compared. The present research emphasizes comparisons in clinically complex samples, with few exclusionary criteria and high rates of comorbidity. Thus, results of this study are more likely to generalize to clinical practice, where there also are few exclusionary criteria. Other strengths include adherence to the recommendations of the Standardized Reporting of Diagnostic studies (STARD) guidelines for reporting diagnostic test results (36). To our knowledge, this is also the first paper to compare these measures simultaneously in the same samples, and it also is the first investigation of adapting the YMRS for use as an adolescent questionnaire. Another strength is the simultaneous inclusion of multiple index tests, affording comparisons both between measures and different sources of information (parent versus youth). The analyses also relied on multiple methods for evaluating diagnostic efficiency, including both ROC and logistic regression analyses. Finally, results provide important information about the robustness of these measures' performance when moved from research settings into urban community mental health and demographically diverse families. To our knowledge, this sample is at present the most ethnically and economically diverse to be evaluated in the area of pediatric bipolar disorder.

Limitations

Limitations of the study include that the present sample contained few Hispanic youths; and it will be important for future research to establish whether these measures perform similarly in Hispanic and other diverse populations. Analyses were limited to comparisons of total scale scores, and did not examine potential differences in presentation at the symptom level across demographic groups (7). Another limitation is that specific cut scores or likelihood ratios are not presented (51). The sample size is not yet large enough to justify the presentation of cut scores or likelihood ratios, particularly when broken down by age or ethnicity (48). Those who are interested in interpreting scores on these measures clinically can use the likelihood ratios published in larger samples (18, 19), although these need to be treated with some caution if the tests are being used with families that differ in clinically relevant respects from the modal participants in the validation samples (51). It is also crucial to note that none of the measures assessed in this study are sufficient for determining a bipolar diagnosis in isolation. These questionnaires were not originally intended to be diagnostic instruments, they do not comprehensively evaluate

mood cycling, duration, or course of illness. In the present study, a licensed psychologist or psychiatrist reviewed the KSADS protocols and notes to assign a diagnosis for all cases.

Research and clinical implications

The present findings indicate that several inexpensive and convenient parent-completed rating scales could facilitate accurate diagnosis, particularly by reducing the number of false-positive diagnoses in children and adolescents seen at outpatient and community settings. These tests can contribute to the assessment process by raising 'red flags' when high scores occur during an initial assessment or screening, indicating when more specialized evaluation is warranted. Low scores on parent measures are also more decisive in helping 'rule out' bipolar disorder, even in fairly ambiguous situations. These instruments are also promising for distilling research samples, creating a smaller sample that is enriched for bipolar disorder that can then more cost-effectively be evaluated using the KSADS and other research tools. Conversely, for genetic or treatment studies needing to enroll a non-bipolar comparison arm, then low scores on these measures would be an efficient filter to exclude bipolar cases. Finally, it is encouraging that these instruments appear to perform comparably in low-income and racially diverse samples. This finding suggests that these tools can facilitate research and clinical work in underserved populations.

Acknowledgements

This research was supported by NIMH R01 MH-066647, a Clinical Research Center Grant from the Stanley Medical Research Institute, and NIMH P20 MH-066054. Special thanks to Diana Hays, Denise Delporto Bedoya, and Jodie Beaver for their assistance with data collection.

References

1. Nottelmann E, Biederman J, Birmaher B et al. National Institute of Mental Health research roundtable on prepubertal bipolar disorder. J Am Acad Child Adolesc Psychiatry 2001; 40: 871–878.
2. Lish JD, Dime-Meenan S, Whybrow PC et al. The National Depressive and Manic-Depressive Association (DMDA) survey of bipolar members. J Affect Disord 1994; 31: 281–294.
3. Hirschfeld RM. Bipolar spectrum disorder: improving its recognition and diagnosis. J Clin Psychiatry 2001; 62 (Suppl. 14): 5–9.
4. Kilbourne AM, Haas GL, Mulsant BH, Bauer MS, Pincus HA. Concurrent psychiatric diagnoses by age and race among persons with bipolar disorder. Psychiatr Serv 2004; 55: 931–933.
5. Neighbors HW, Trierweiler SJ, Ford BC, Muroff JR. Racial differences in DSM diagnosis using a semi-structured instrument: the importance of clinical judgment in the diagnosis of African Americans. J Health Soc Behav 2003; 44: 237–256.
6. Neighbors HW, Trierweiler SJ, Munday C et al. Psychiatric diagnosis of African Americans: diagnostic divergence in clinician-structured and semistructured interviewing conditions. J Natl Med Assoc 1999; 91: 601–612.
7. Strakowski SM, Hawkins JM, Keck PE Jr et al. The effects of race and information variance on disagreement between psychiatric emergency service and research diagnoses in first-episode psychosis. J Clin Psychiatry 1997; 58: 457–463.
8. Goodwin FK, Jamison KR. Manic-Depressive Illness. New York: Oxford University Press, 1990.
9. Helzer JE. Bipolar affective disorder in Black and White men: a comparison of symptoms and familial illness. Arch Gen Psychiatry 1975; 32: 1140–1143.
10. Strakowski SM, Flaum M, Amador X et al. Racial differences in the diagnosis of psychosis. Schizophr Res 1996; 21: 117–124.
11. DelBello MP, Lopez-Larson MP, Soutullo CA, Strakowski SM. Effects of race on psychiatric diagnosis of hospitalized adolescents: a retrospective chart review. J Child Adolesc Psychopharmacol 2001; 11: 95–103.
12. Depue RA, Slater JF, Wolfstetter-Kausch H, Klein DN, Goplerud E, Farr DA. A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. J Abnorm Psychol 1981; 90: 381–437.
13. Klein DN, Depue RA, Slater JF. Inventory identification of cyclothymia. IX. Validation in offspring of bipolar I patients. Arch Gen Psychiatry 1986; 43: 441–445.
14. Lewinsohn PM, Klein DN, Seeley JR. Bipolar disorders in a community sample of older adolescents: prevalence, phenomenology, comorbidity, and course. J Am Acad Child Adolesc Psychiatry 1995; 34: 454–463.
15. Danielson CK, Youngstrom EA, Findling RL, Calabrese JR. Discriminative validity of the General Behavior Inventory using youth report. J Abnorm Child Psychol 2003; 31: 29–39.
16. Findling RL, Youngstrom EA, Danielson CK et al. Clinical decision-making using the General Behavior Inventory in juvenile bipolarity. Bipolar Disord 2002; 4: 34–42.
17. 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 2001; 13: 267–276.
18. Youngstrom EA, Frazier TW, Findling RL, Calabrese JR. A ten item brief screen for manic-depression in youths age 5–17 years. Paper presented at: Annual meeting of the American Psychiatric Association; May, 2004; New York, NY.
19. Youngstrom EA, Findling RL, Calabrese JR et al. Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. J Am Acad Child Adolesc Psychiatry 2004; 43: 847–858.
20. Hirschfeld RMA, Williams JBW, Spitzer RL et al. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. Am J Psychiatry 2000; 157: 1873–1875.

21. Wagner KD, Emslie GJ, Findling RL, Gracious B, Reed ML. Clinic Validation of the Adolescent Mood Disorder Questionnaire (A-MDQ). Paper Presented at Annual Meeting of the American Psychiatric Association. New York: APA, 2004.
22. Isometsa E, Suominen K, Mantere O et al. The mood disorder questionnaire improves recognition of bipolar disorder in psychiatric care. BMC Psychiatry 2003; 3: 8.
23. Drotar D, Stein REK, Perrin EC. Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. Special Issue: Methodological issues in clinical child psychology research. J Clin Child Psychol 1995; 24: 184–192.
24. Achenbach TM. Manual for the Child Behavior Checklist/4-18 and 1991 Profile. Burlington, VT: University of Vermont, 1991.
25. Achenbach TM, Rescorla LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: University of Vermont, 2001.
26. Mick E, Biederman J, Pandina G, Faraone SV. A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. Biol Psychiatry 2003; 53: 1021–1027.
27. Kahana SY, Youngstrom EA, Findling RL, Calabrese JR. Employing parent, teacher, and youth self-report checklists in identifying pediatric bipolar spectrum disorders: an examination of diagnostic accuracy and clinical utility. J Child Adolesc Psychopharmacol 2003; 13: 471–488.
28. Youngstrom EA, Youngstrom JK, Starr M. Bipolar diagnoses in community mental health: Achenbach CBCL profiles and patterns of comorbidity. Biological Psychiatry in press.
29. Lengua LJ, Sadowski CA, Friedrich WN, Fisher J. Rationally and empirically derived dimensions of children's symptomatology: expert ratings and confirmatory factor analyses of the CBCL. J Consult Clin Psychol 2001; 69: 683–698.
30. Hazell PL, Lewin TJ, Carr VJ. Confirmation that Child Behavior Checklist clinical scales discriminate juvenile mania from attention deficit hyperactivity disorder. J Paediatr Child Health 1999; 35: 199–203.
31. Thuppal M, Carlson GA, Sprafkin J, Gadow KD. Correspondence between adolescent report, parent report, and teacher report of manic symptoms. J Child Adolesc Psychopharmacol 2002; 12: 27–35.
32. Carlson GA, Youngstrom EA. Clinical implications of pervasive manic symptoms in children. Biol Psychiatry 2003; 53: 1050–1058.
33. Findling RL, Gracious BL, McNamara NK, Youngstrom EA, Demeter C, Calabrese JR. Rapid, continuous cycling and psychiatric co-morbidity in pediatric bipolar I disorder. Bipolar Disord 2001; 3: 202–210.
34. Berument SK, Rutter M, Lord C, Pickles A, Bailey A. Autism screening questionnaire: diagnostic validity. Br J Psychiatry 1999; 175: 444–451.
35. Dunn LM, Dunn LM. Examiner's Manual for the Peabody Picture Vocabulary Test, 3rd edn. Circle Pines, MN: American Guidance Service, 1997.
36. Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138: 40–44.
37. Kaufman J, Birmaher B, Brent D et al. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 1997; 36: 980–988.
38. Geller B, Zimerman B, Williams M et al. Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. J Am Acad Child Adolesc Psychiatry 2001; 40: 450–455.
39. Leibenluft E, Charney DS, Towbin KE, Bhangoo RK, Pine DS. Defining clinical phenotypes of juvenile mania. Am J Psychiatry 2003; 160: 430–437.
40. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity, and sensitivity. Br J Psychiatry 1978; 133: 429–435.
41. Gracious BL, Youngstrom EA, Findling RL, Calabrese JR. Discriminative validity of a parent version of the Young Mania Rating Scale. J Am Acad Child Adolesc Psychiatry 2002; 41: 1350–1359.
42. Youngstrom EA, Gracious BL, Danielson CK, Findling RL, Calabrese JR. Toward an integration of parent and clinician report on the Young Mania Rating Scale. J Affect Disord 2003; 77: 179–190.
43. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess 2000; 12: 102–111.
44. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839–843.
45. Hosmer DW, Lemeshow S. Applied Logistic Regression, 2nd edn. New York: Wiley, 2000.
46. Anastasi A, Urbina S. Psychological Testing, 7th edn. New York: MacMillan, 1997.
47. Youngstrom EA, Findling RL, Calabrese JR et al. Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. J Am Acad Child Adolesc Psychiatry 2004; 43: 847–858.
48. Kraemer HC. Evaluating Medical Tests: Objective and Quantitative Guidelines. Newbury Park, CA: Sage Publications, 1992.
49. Zhou X-H, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: Wiley, 2002.
50. Tillman R, Geller B, Craney JL, Bolhofner K, Williams M, Zimerman B. Relationship of parent and child informants to prevalence of mania symptoms in children with a prepubertal and early adolescent bipolar disorder phenotype. Am J Psychiatry 2004; 161: 1278–1284.
51. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM, 2nd edn. New York: Churchill Livingstone, 2000.