Evidence-Based Assessment of Pediatric Bipolar Disorder, Part II: Incorporating Information From Behavior Checklists

ERIC A. YOUNGSTROM, PH.D. AND JENNIFER KOGOS YOUNGSTROM, PH.D.

The previous commentary described how to gather and use information about the base rate of pediatric bipolar disorder (PBD) and the family history of bipolar illness to assess the risk of PBD for an individual case, using an evidence-based practice (EBP) framework. This month's column continues with that case and demonstrates how behavior rating scales or checklists are another inexpensive and potentially useful source of information. This column shows how to make use of likelihood ratios (how much more likely a disorder is to be present than absent for a given test score in a defined population [Guyatt and Rennie, 2002]) to assess the likelihood of PBD in an individual case. Likelihood ratios are valuable because they may enable a clinician to take a legitimate shortcut and use an easier to administer diagnostic test (here the Child Behavior Checklist [CBCL]) in place of a more time-consuming but definitive test (here the Schedule for Affective Disorders and Schizophrenia for School-Age Children [Kaufman et al., 1997]). The specific details of the case have been changed so that the person described is not recognizable.

At the time of intake, the biological mother completed the CBCL (Achenbach, 1991; Achenbach and Rescorla, 2001). Is CBCL information helpful in assessing for PBD? In *Medline*, the search terms "CBCL AND bipolar disorder" generate multiple hits, including a meta-analysis of previous studies comparing CBCL scores for cases diagnosed with PBD compared with other children (Mick et al., 2003), along with some more recent studies (Kahana et al., 2003; Youngstrom et al., 2004). The meta-analysis presents mean CBCL clinical syndrome scale *T* scores in clinical bipolar youth, with the Aggressive Behavior scale being most elevated, followed by Anxious/Depressed, Attention Problems, and Delinquent Behavior. Mick et al. (2003) propose these scales as a potential "bipolar profile" but do not suggest optimal cutoff scores to differentiate PBD. The more recent articles compared the CBCL subscales with the Externalizing problems broad band score (a composite of several syndrome scales) and found that after controlling for the Externalizing score, none of the other CBCL scales improved the identification rate of PBD (Kahana et al., 2003; Youngstrom et al., 2004). It is worth noting that all published research on PBD to date has used older versions of the CBCL. The current version in clinical use is the 2001 form, which changed some items and reorganized some scales. However, the Externalizing scale score correlates 0.99 between the 1991 and 2001 versions (Achenbach and Rescorla, 2001), suggesting that performance should be similar. Overall, based on these findings, it makes sense to focus on the Externalizing score.

Consider, therefore, the likelihood ratios associated with low, intermediate, high, and very high Externalizing CBCL scores, broken down separately by age (Youngstrom et al., 2004). This approach examines the performance of multiple segments of the test scores rather than simply dividing the test at one place (as is done when calculating the sensitivity and specificity of the test at a single threshold score) (Guyatt and Rennie,

Accepted February 11, 2005.

From the Departments of Psychology and Psychiatry, Case Western Reserve University/University Hospitals of Cleveland.

This research was supported by NIMH R01 MH-066647 as well as NIMH P20 MH 066054 and a Clinical Research Center Grant from the Stanley Medical Research Institute. Thanks also to John Hamilton, M.D., and Martha Hellander, J.D., for suggestions and comments.

Reprint requests to Dr. Eric A. Youngstrom, Department of Psychology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7123; e-mail: Eric.Youngstrom@case.edu.

DOI: 10.1097/01.chi.0000164589.10200.a4

2002). Advantages of the multilevel likelihood ratio approach include the following: (1) it preserves more information from the test, (2) even relatively mediocre tests may provide clinically useful information at extreme scores, (3) it conveys whether the test performs asymmetrically (i.e., one test may be more decisive in *ruling out* a diagnosis via low scores, and another test may be more powerful at *ruling in* a diagnosis via high scores), and (4) it compels the user to consider the base rate or previous probability to interpret the test score. Zack earned a CBCL Externalizing *T* score of 83. This is extremely high (>3 SDs above the nonclinical average score) and falls in the highest risk category for ages 5 to 10 years, with a likelihood ratio of 3.5 (from Table 4 in Youngstrom et al., 2004). Using the nomogram (included in Part I of this commentary) to combine the 24% probability with the increase in risk of 3.5 results in a revised posterior probability of roughly 54%. Using the Bayes Theorem to combine the base rate, the family history, and the test score yields a probability of 52.8%, very close to the result from the nomogram. The order in which information is combined does not matter. If we had chosen to interpret the test score first, then the nomogram would combine the 6% probability and the likelihood ratio of 3.5 to yield a probability of 18%, and then combining 18% and a likelihood ratio of 5 (for the bipolar parent) ends in an estimate of 54%.

Consider that in evaluating Zack, we have encountered two major ''red flags'' that could signal the presence of PBD: a close biological relative with a clear history of bipolar disorder and extremely high scores on an instrument that has been widely researched and documented to show elevated scores in bipolar cases. However, because PBD is rare in this clinical setting (as is true in most clinical settings), there is still a good chance that Zack currently does not have PBD. Similarly, the majority of children presenting at this agency with a positive family history and equally severe externalizing problems will also not warrant bipolar diagnoses. The nomogram approach correctly combines these three pieces of information into the most accurate estimate possible.

Human beings are prone to a variety of different ''cognitive heuristics'' (practical strategies for processing information quickly that do not otherwise have justification in theory or data) that can undermine accuracy (Dawes et al., 1989). The effects of heuristics will typically be most pronounced in situations like this scenario, in which the condition of interest is rare. There also can be tremendous differences in practitioners' estimates of risk based on just these three pieces of information. The variability is sobering for several reasons. Real-life decision making will almost always consider more information than just the three variables mentioned so far. Each additional variable adds more information but also more opportunity for differences of interpretation. Also, consider the perspective of the family: Just on the basis of the family history and a standard checklist, clinicians will often come up with contradictory conclusions about the risk of PBD (i.e., estimates of risk well above or well below 50%). This increases the chances for seemingly contradictory second opinions. Adopting an evidence-based, Bayesian (Bayes first wrote about how to calculate posttest probability from pretest probability and test characteristics) framework for combining basic variables such as base rate, familial risk, and test scores standardizes the interpretation of these factors and reduces disagreement between practitioners, helping ensure that everyone arrives in the same ''ballpark'' before entering into the more nuanced aspects of assessment.

ARE TESTS AVAILABLE THAT ARE BETTER THAN THE CBCL EXTERNALIZING SCALE?

The CBCL is used routinely at many clinics as part of an intake assessment because it provides information about a broad range of behavior problems. Is there another instrument available that would be more useful to further refine the probability of Zack's meeting criteria for PBD? There are several measures available that have better content validity for this purpose than the CBCL because they directly ask about symptoms of mania. Almost all these instruments were initially developed in adults and then have been used with teenagers and in some cases young children. Relatively few have been evaluated using receiver operating characteristic (ROC) methods to quantify their diagnostic efficiency (McFall and Treat, 1999). However, Youngstrom et al. (2004) compared six tests with each other in the same sample of adolescent and four tests with each other in a sample of children ages 5 to 10 years. ROC analyses were performed on all the tests, and then significance tests (Hanley and McNeil, 1983) examined whether any of the tests were performing significantly better than others. Multilevel likelihood ratios were also calculated

for all the tests. Results indicated that parent report on the CBCL was significantly better than teacher report (on the Teacher Report Form) or youth report in discriminating cases with semistructured interview diagnoses of PBD, and CBCL also outperformed teacher ratings in the younger cohort. Parent report on the other questionnaires overall did not significantly outperform the CBCL Externalizing score, but the likelihood ratios indicated that the CBCL was more useful for ruling out bipolar cases (with very low scores decreasing the odds of a bipolar diagnosis by a factor of 20), whereas the more specialized mania scales were more useful for increasing the odds of a bipolar diagnosis (with likelihood ratios of 6 or 9 versus 3.5 or 5 for extremely high CBCL scores). On the basis of these findings, Youngstrom et al. (in press) recommend that (1) parent report always be included in the differential diagnosis of pediatric mania; (2) teacher report on the Achenbach does not add sufficient information about bipolar diagnoses to justify its inclusion in the process of differential diagnosis of PBD; (3) if the CBCL is already available, then low scores on it will be decisive in most settings in ruling out PBD; (4) high scores on the CBCL Externalizing scale should trigger more thorough assessment, with the recognition that in most clinical settings, even cases with extremely high scores will be more likely to have other disorders in addition to mania.

It is rare to find studies that compare several tests in the same sample. More often practitioners will need to compare tests that have been validated in different samples. It is possible to use meta-analytic methods to compare the performance of tests across different samples (Hanley and McNeil, 1983). Although articles presenting ROC results will be much more useful for making decisions about individual cases, measures of effect size (such as Cohen's *d,* the difference between the means of two groups divided by the pooled SD) can be used to rank tests according to their diagnostic validity. When comparing tests, it is important to carefully consider the ''clinical validity'' of the study—the extent to which the results can appropriately be applied in clinical contexts. There are guidelines and checklists available to help rate the quality of studies (Bossuyt et al., 2003; Guyatt and Rennie, 2002). In general, it is easier for a test to appear to perform well if there are more exclusion criteria, if there is minimal diagnostic comorbidity, if there is ''criterion contamination'' (in which the diagnostician is not blind to the test results), and if there is more similarity between the testing and diagnostic procedures. Tests will appear to perform best when comparing known patients and normal controls, but this comparison is less clinically meaningful than evaluations comparing patients with one disorder with those with another disorder (e.g., PBD versus attention-deficit/hyperactivity disorder [ADHD], or PBD versus all other diagnoses of patients presenting to a clinic). Paradoxically, studies predicting diagnoses based on semistructured interviews involving multiple informants, or consensus diagnoses, often have somewhat lower reliability and may produce lower diagnostic efficiency statistics (i.e., lower areas under the ROC curve, sensitivity, specificity, or κ values), but they usually have greater clinical validity (Meyer, 2002). The higher clinical validity, which derives from their reliance on multiple information sources and the incorporation of some degree of clinical inference, makes semistructured instruments the preferred standard for evaluating tests for mental health applications.

DOES COMBINING TESTS LEAD TO BETTER RESULTS?

Because the literature search found multiple measures that might differentiate PBD, it is logical to ask whether a combination of measures might do better than any single scale. Published findings indicate that there is not much diagnostic value added by combining information from the parent, teacher, and youth reports on the Achenbach checklists beyond that gleaned from the parent report on the CBCL by itself (Kahana et al., 2003). Similarly, youth self-report on the General Behavior Inventory (Depue et al., 1981) does not improve classification after controlling for parent report on the General Behavior Inventory, the Parent Young Mania Rating Scale, or the CBCL (Youngstrom et al., 2004). Bear in mind that the results of studies such as these only speak to the value of combining the exact instruments included in the analyses. Given the growing number of instruments assessing pediatric mania, it is quite likely that new instruments will be found that add significant diagnostic information in combination with existing instruments.

In the meantime, some rules of thumb should guide the use of instruments as a component of the assessment process. First, do not gather and synthesize multiple

measures from the same person. It would be absurd to have a mother fill out the same questionnaire three times, and then revise the probability of a bipolar diagnosis based on the three scores. It is not much more valid to combine the parent's report on the CBCL with his or her impression on the Parent General Behavior Inventory; both tools are gathering the same parent's impressions about the child's behavior over similar time frames and settings. The scores are clearly correlated, and the sequential use of the nomogram assumes that the pieces of information being synthesized are independent of each other.

This is not to say that multiple questionnaires should never be administered. If the CBCL score is high, then it makes sense to ask the parent to complete a more specific measure of mania (such as the Mood Disorder Questionnaire, the parent version of the Young Mania Rating Scale, and the Parent General Behavior Inventory) and then substitute the mania scale for the CBCL when estimating the probability of a bipolar diagnosis. Conversely, it may be reasonable to combine youth and parent reports sequentially via the nomogram because these information sources are only modestly correlated. Teacher and parent reports also show relatively low correlations, but the Achenbach Teacher Report Form does not appear to provide enough information about mania even at extreme scores to warrant its inclusion in diagnostic predictions of PBD. Family history represents another situation in which it may be useful to combine information sequentially about multiple relatives with bipolar disorder. Each bipolar diagnosis is likely to increase the risk of illness in the child, although it should be remembered that it is unknown at present whether the risk is additive (as would be assumed by the nomogram framework) or more complicated.

NEXT STEPS IN CLINICAL ASSESSMENT OF PEDIATRIC BIPOLAR DISORDER

Three pieces of information, the base rate, family history of bipolar disorder, and scores on a parent checklist, can provide a considerable amount of information about the degree of risk of PBD in a specific case. In an outpatient setting with roughly a 6% base rate of PBD, family history and a high checklist score could raise the probability of a bipolar diagnosis to 75% (i.e., raw scores of ≥49 on the Hypomanic/Biphasic scale of the Parent General Behavior Inventory—a likelihood ratio of 9.2 coupled with a history of bipolar disorder in at least one biological parent) or decrease it to less than 0.3% (i.e., *T* scores <54 on the CBCL Externalizing—likelihood ratio of 0.04 and no family history of bipolar disorder). Currently available assessment tools are more helpful for *ruling out* PBD than ruling it *in*, especially in settings in which PBD is relatively uncommon. Put another way, the combination of base rate, family history, and a checklist score will often be enough to reduce the probability of a PBD diagnosis below the test threshold, indicating that no further assessment is required, but the same pieces of information will rarely be sufficient to raise the probability of PBD above the treatment threshold.

Under no circumstances should the probability based on a checklist and family history be equated with a formal diagnosis. Diagnosis is a health care decision with significant consequences in terms of treatment as well as legal and ethical ramifications. Making the formal diagnosis of PBD will require additional information, including careful assessment of the frequency, intensity, and duration of the specific symptoms of mania and depression as stipulated in the *DSM-IV*. Confidence in the accuracy of the diagnosis and responsiveness to treatment will be heightened by incorporating prospective assessment of mood and energy changes, using techniques such as life charting (Denicoff et al., 1997). There is some evidence that symptoms of elated mood, pressured speech, heightened interest in sex, and perhaps grandiosity may be more specific to mania than other associated symptoms such as aggression, irritability, distractibility, or increased motor activity (Geller et al., 2002). There are formal semistructured diagnostic interviews that are available to aid in the diagnosis of PBD, although the most detailed are currently research instruments that are not intended for use in clinical practice. Based on the recommendations in two review articles (Quinn and Fristad, 2004; Youngstrom et al., in press), your colleague decides to have the family complete prospective life charting over the next several weeks using a free format (such as available at http://www.bpkids.org/learning/6-02.pdf and http://www.bpkids.org/learning/directions.doc) to gather more information about fluctuations in mood and energy as well as planning to do "mood checkups" at follow-up visits by repeating brief mood questionnaires to monitor treatment response.

LIMITATIONS: MAKING THINGS FUZZY AGAIN

There are many factors to consider in evaluating whether a diagnostic study is valid and whether it is applicable to a particular patient. The Standards for Reporting of Diagnostic Accuracy criteria list 25 items to consider in evaluating a study (Bossuyt et al., 2003), and Guyatt and Rennie (2002) provide detailed suggestions for determining the relevance to an individual case. One issue now prominent in the PBD literature is the variability in the diagnostic gold standard used. Epidemiological studies have typically relied on structured interviews, and some have not involved a parent, despite evidence suggesting that a parent is a crucial information source, and semistructured interviews performed by appropriately trained and supervised raters produce more valid diagnoses of bipolar disorder. Research definitions of PBD also vary (Leibenluft et al., 2003), and it is unknown how these changes affect the performance of tests or the value of family history. Additionally, most published studies rely on families presenting to academic medical centers, which makes them on average better educated and less demographically diverse than would be nationally representative. Demographics and other variables may moderate the performance of diagnostic checklists or family history in unforeseen ways. Finally, the approach advocated here hinges on the stability of the likelihood ratios associated with various test scores. Because likelihood ratios are based on the sensitivity and specificity of the test, they are mathematically unrelated to the base rate of the disorder. However, it is possible for the sensitivity and specificity of a test to change in different populations, depending on factors such as the average severity of bipolar illness (e.g., lower severity would reduce sensitivity) or the rate of diagnoses that are often yield false alarms on diagnostic tools (Kraemer, 1992). For example, higher rates of ADHD or oppositional defiant disorder would lower the specificity of the test because they would increase the number of nonbipolar cases "accidentally" scoring high on screening tests. One pragmatic solution is to perform "sensitivity analyses" by changing the assumptions or

TABLE 1

General Recommendations for Starting an Evidence-Based Approach to Diagnostic Assessment of a Disorder

Step	How Implemented
1. Select target disorder	What are the most common presenting problems?
	What are the most difficult differential diagnoses?
2. Establish local base rate estimate	Identify published rate with most similar sample
	Directly estimate from own medical records (but carefully consider threats to the validity of the diagnoses)
3. Identify relevant instruments	*Medline* or *PsycINFO* search: target disorder AND "sensitivity and specificity" (MeSH term)
4. Compare relevant instruments	Compare published information to Standards for Reporting of Diagnostic Accuracy criteria (Bossuyt et al., 2003) and to clinical population of interest
	Pick instrument with largest area under the receiver operating characteristic curve or most extreme likehood ratios
5. Find likelihood ratios associated with test scores	Look for published likelihood ratios
	Convert sensitivity and specificity into likelihood ratios for positive and negative test results[a]
6. Determine whether other risk factors may be clinically useful	Medline or PsycINFO search: target disorder AND "risk factor" (MeSH) term)
7. Make nomogram convenient to use	Keep copies of nomogram at offices
	Consider "premarking" nomogram to indicate base rate estimates for different disorders in your setting
	Consider "premarking" nomogram to indicate likelihood ratios associated with different risk factor or test scores
8. Regularly review and update tool kit	Periodically repeat *Medline* searches (or *PsycINFO* for psychosocial interventions)
	Target disorder AND "sensitivity and specificity"
	Target disorder AND "risk factors"
	Adopt new tests with better norms, better criterion diagnosis, and/or better diagnostic efficiency

[a] The likelihood ratio associated with a positive test result is equal to sensitivity/(false alarm rate) = sensitivity/(1–specificity). The likelihood ratio for a negative test result is equal to (1–sensitivity)/specificity. See Guyatt and Rennie (2002) for more details.

estimates involved in the risk assessment and seeing how they affect the results (Guyatt and Rennie, 2002).

GENERAL RECOMMENDATIONS

Your colleague found the information helpful in terms of establishing an evidence-based estimate of the risk of PBD in a specific case and uncovering resources and a framework to guide continued assessment. The discussion also identified several general principles that could readily be applied to other disorders, with similar potential for improvement in the efficiency and accuracy with which clinical information is "triaged" (Table 1). Many diagnoses will not be as difficult to differentiate as PBD, but they often will be more prevalent. Developing a similar tool kit for the assessment of ADHD, anxiety disorders, and unipolar depression would provide coverage for the majority of new cases at many clinical settings. Once you have identified that a patient is likely to have bipolar disorder, then look for practice guidelines and evidence-based approaches to suggest first-line treatment strategies (Carlson et al., 2003; McClellan and Werry, 1997).

Disclosure: Dr. Youngstrom is co-investigator on investigator-initiated research grants sponsored by Abbott and AstraZeneca Pharmaceuticals and is the statistical expert for both protocols. He also consults with Otsuka Pharmaceuticals about assessment of pediatric bipolar disorder. The other author has no financial relationships to disclose.

REFERENCES

Achenbach TM (1991), *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington: University of Vermont, Department of Psychiatry

Achenbach TM, Rescorla LA (2001), *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: University of Vermont, Department of Psychiatry

Bossuyt PM, Reitsma JB, Bruns DE et al. (2003), Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Ann Intern Med* 138:40–44

Carlson GA, Jensen PS, Findling RL et al. (2003), Methodological issues and controversies in clinical trials with child and adolescent patients with bipolar disorder: report of a consensus conference. *J Child Adolesc Psychopharmacol* 13:1–15

Dawes RM, Faust D, Meehl PE (1989), Clinical versus actuarial judgment. *Science* 243:1668–1674

Denicoff KD, Smith-Jackson EE, Disney ER, Suddath RL, Leverich GS, Post RM (1997), Preliminary evidence of the reliability and validity of the prospective life-chart methodology (LCM-p). *J Psychiatr Res* 31:593–603

Depue RA, Slater JF, Wolfstetter-Kausch H, Klein DN, Goplerud E, Farr DA (1981), A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. *J Abnorm Psychol* 90:381–437

Geller B, Zimerman B, Williams M et al. (2002), DSM-IV mania symptoms in a prepubertal and early adolescent bipolar disorder phenotype compared to attention-deficit hyperactive and normal controls. *J Child Adolesc Psychopharmacol* 12:11–25

Guyatt GH, Rennie D, eds. (2002), *Users' Guides to the Medical Literature*. Chicago: AMA Press

Hanley JA, McNeil BJ (1983), A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148:839–843

Kahana SY, Youngstrom EA, Findling RL, Calabrese JR (2003), 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* 13:471–488

Kaufman J, Birmaher B, Brent D et al. (1997), 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* 36:980–988

Kraemer HC (1992), *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Newbury Park, CA: Sage

Leibenluft E, Charney DS, Towbin KE, Bhangoo RK, Pine DS (2003), Defining clinical phenotypes of juvenile mania. *Am J Psychiatry* 160: 430–437

McClellan J, Werry J (1997), Practice parameters for the assessment and treatment of children and adolescents with bipolar disorder. American Academy of Child and Adolescent Psychiatry. *J Am Acad Child Adolesc Psychiatry* 36:157S–176S

McFall RM, Treat TA (1999), Quantifying the information value of clinical assessment with signal detection theory. *Annu Rev Psychol* 50:215–241

Meyer GJ (2002), Implications of information-gathering methods for a refined taxonomy of psychopathology. In: *Rethinking the DSM: Psychological perspectives*, Beutler LE, Malik M, eds. Washington, DC: American Psychological Association, pp 69–106

Mick E, Biederman J, Pandina G, Faraone SV (2003), A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. *Biol Psychiatry* 53:1021–1027

Quinn CA, Fristad MA (2004), Defining and identifying early onset bipolar spectrum disorder. *Curr Psychiatry Rep* 6:101–107

Youngstrom EA, Findling RL, Calabrese JR et al. (2004), 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* 43:847–858

Youngstrom EA, Findling RL, Youngstrom JK, Calabrese JR (in press), Towards an evidence-based assessment of pediatric bipolar disorder. *J Clin Child Adolesc Psychol*