



Published in final edited form as:

Psychol Assess. 2012 June ; 24(2): 269–281. doi:10.1037/a0025775.

Generalizability of Evidence-Based Assessment Recommendations for Pediatric Bipolar Disorder

Melissa M. Jenkins, MA,
University of North Carolina at Chapel Hill

Eric A. Youngstrom, PhD,
University of North Carolina at Chapel Hill

Jennifer Kogos Youngstrom, PhD,
University of North Carolina at Chapel Hill

Norah C. Feeny, PhD, and
Case Western Reserve University

Robert L. Findling, MD
Case Western Reserve University

Abstract

Bipolar disorder is frequently clinically diagnosed in youths who do not actually satisfy DSM-IV criteria, yet cases that would satisfy full DSM-IV criteria are often undetected clinically. Evidence-based assessment methods that incorporate Bayesian reasoning have demonstrated improved diagnostic accuracy, and consistency; however, their clinical utility is largely unexplored. The present study examines the effectiveness of promising evidence-based decision-making compared to the clinical gold standard. Participants were 562 youth, ages 5-17 and predominantly African American, drawn from a community mental health clinic. Research diagnoses combined semi-structured interview with youths' psychiatric, developmental, and family mental health histories. Independent Bayesian estimates relied on published risk estimates from other samples discriminated bipolar diagnoses, Area Under Curve=.75, $p<.00005$. The Bayes and confidence ratings correlated $r_s = .30$. Agreement about an evidence-based assessment intervention "threshold model" (wait/assess/treat) had $K=.24$, $p<.05$. No potential moderators of agreement between the Bayesian estimates and confidence ratings, including type of bipolar illness, were significant. Bayesian risk estimates were highly correlated with logistic regression estimates using optimal sample weights, $r=.81$, $p<.0005$. Clinical and Bayesian approaches agree in terms of overall concordance and deciding next clinical action, even when Bayesian predictions are based on published estimates from clinically and demographically different samples. Evidence-based assessment methods may be useful in settings that cannot routinely employ gold standard assessments, and they may help decrease rates of overdiagnosis while promoting earlier identification of true cases.

Correspondence concerning this manuscript may be sent to Dr. Eric A. Youngstrom, Professor of Psychology and Psychiatry, Associate Director, Center for Excellence in the Research & Treatment of Bipolar Disorder University of North Carolina, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270. eay@unc.edu. 919-962-3997 phone, 919-962-2537 fax.

Author Note Melissa M. Jenkins, Department of Psychology, University of North Carolina at Chapel Hill; Eric A. Youngstrom, Departments of Psychology and Psychiatry, University of North Carolina at Chapel Hill; Jennifer Kogos Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill; Norah C. Feeny, Department of Psychology, Case Western Reserve University; Robert L. Findling, Department of Psychiatry, Case Western Reserve University.

The other authors have no conflicts of interest to disclose.

Keywords

pediatric bipolar disorder; evidence-based assessment; decision-making; diagnosis; clinical judgment

Bipolar disorder is a high stakes diagnosis. Untreated, it often follows a recurrent and progressively worsening course, with increased risk of substance use, incarceration, and death by suicide or other causes (Goodwin & Jamison, 2007). If bipolar is misdiagnosed as depression or attention-deficit/hyperactivity disorder, there are delays initiating effective mood stabilizing treatments (Kowatch, Fristad, et al., 2005). There also are concerns that the options selected instead could worsen the course of the bipolar disorder (Akiskal, et al., 2003; Altshuler, et al., 1995; cf. Carlson, 2003; Hirschfeld, et al., 2002; cf. Joseph, Youngstrom, & Soares, 2009; Scheffer, Kowatch, Carmody, & Rush, 2005). Earlier identification may lead to better outcomes (DelBello, Adler, Whitsel, Stanford, & Strakowski, 2007; Findling, et al., 2003; Fristad, Verducci, Walters, & Young, 2009), delay or prevent relapse (Findling, et al., 2007), and afford more titrated or benign interventions (Miklowitz & Chang, 2008). The median lag exceeds five years between when mood symptoms begin to cause problems versus when a clinician arrives at a bipolar diagnosis in youths (Marchand, Wirth, & Simon, 2006) as well as adults (Hirschfeld, Lewis, & Vornik, 2003; Lish, Dime-Meenan, Whybrow, Price, & Hirschfeld, 1994). To close this gap, early identification has been promoted both in scholarly research (Miklowitz & Chang, 2008; Youngstrom, Findling, Youngstrom, & Calabrese, 2005) and in the popular media (Kluger & Song, 2002; Papolos & Papolos, 2002).

There is concern that the pendulum may have swung too far in favor of early diagnosis, though (Healy, 2006; Parens, Johnston, & Carlson, 2010). Clinical diagnoses of pediatric bipolar disorder (PBD) have risen steeply in the U.S. (Blader & Carlson, 2007), with some estimates showing a 40-fold increase in diagnoses over the last decade (Moreno, et al., 2007). Although evidence supports the validity of research diagnoses of bipolar disorder in terms of phenotypic, genetic, structural, functional, neurocognitive, and treatment response similarities to bipolar disorder in adults (Geller & Luby, 1997; Youngstrom, Birmaher, & Findling, 2008); and longitudinal studies are demonstrating considerable developmental continuity (Birmaher, et al., 2009; Geller, Tillman, Bolhofner, & Zimmerman, 2008), the crucial issue from a public health and consumer point of view is whether the research diagnoses and clinical diagnoses identify the same cases.

Clinical diagnosis itself is a complex enterprise, fraught with potential biases and sources of error (Garb, 1998). Clinicians tend to rely on unstructured interviews and observations that have proven problematic (Neisworth & Bagnato, 2004). Clinicians also are prone to cognitive biases that often result in suboptimal diagnostic decisions (Croskerry, 2002; Galanter & Patel, 2005). More than 100 studies over the last 50 years have consistently demonstrated that clinical judgment is no better than simple actuarial algorithms, and often significantly worse (Ægisdóttir, et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954).

The situation appears even worse with regard to clinical diagnoses of bipolar disorder. A recent meta-analysis comparing clinical diagnoses to those based on structured interviews found that bipolar disorder was an outlier even compared to the general trend of mediocrity in diagnostic reliability, with a kappa of less than .10 (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). Unfortunately, it is also very difficult to assess because of the phenomenology of the illness (Bowring & Kovacs, 1992; Youngstrom, Findling, et al., 2005) as there is substantial overlap in symptomatology (Leibenluft, Charney, Towbin,

Bhangoo, & Pine, 2003; Weller, Danielyan, & Weller, 2004) making it hard to tease apart bipolar symptoms from symptoms of more prevalent diagnoses, such as Attention Deficit Hyperactivity Disorder (ADHD). In addition, comorbidity complicates diagnosis, as youths frequently meet criteria for multiple psychiatric disorders (Axelson, et al., 2006; Kowatch, Youngstrom, Danielyan, & Findling, 2005). Consequently, a “typical” presentation of bipolar in isolation is rare, and clinicians may focus on the comorbid condition and neglect PBD (Youngstrom, Findling, et al., 2005). Further, varied presentations of PBD threaten the reliability of diagnostic impressions. For example, classic bipolar I disorder can present as florid mania, severe depression, a mixed state, or as normal functioning, depending on the mood state.

In addition to PBD being arguably one of the most difficult diagnoses to make correctly, broader practice issues of training, burden, and reimbursement can make research diagnostic instruments, such as structured and semi-structured interviews, impractical for use in many clinical settings. As a result, many common assessment methods are not evidence-based (EB). Thus, the difficulty diagnosing PBD is compounded by complex presentation of illness, contextual constraints common to real-world clinical practice, and clinicians’ cognitive vulnerabilities.

Clinical diagnoses of bipolar disorder in adults also have waxed and waned in popularity, without any evidence of a corresponding change in the base rate of the condition (Zuckerman, 1999). There is debate about whether the contemporary rise in clinical diagnoses is a correction from previous underdiagnosis—similar to historical changes in patterns of diagnosis for depression (Kovacs, 1989), versus being evidence of overdiagnosis, perhaps due to nonspecific use of the “bipolar” label to include a broad group of youths who have marked irritability and aggression, but may not otherwise meet criteria for mania (Leibenluft, et al., 2003). Overall, it appears many youth diagnosed with bipolar do not actually have the disorder while many true cases of bipolar go undiagnosed (Ghaemi, Sachs, Chiou, Pandurangi, & Goodwin, 1999).

Pediatric Bipolar Disorder and Evidence-Based Assessment

Clinical decision-making could benefit from evidence based (EB) assessment and decision-making strategies in diagnosing high-stakes conditions such as PBD. An alternative to traditional clinical assessment is taking an actuarial approach that quantifies the risk of illness using Bayesian reasoning. Moreover, Jaeschke, Guyatt, & Sackett (1994) recommend using a “nomogram” as a simple, practical method for combining information about risk with the “diagnostic likelihood ratios” associated with test results or other clinical findings. Nomograms are charts scaled in a way that eliminates the need for multiplication or division while working with probabilities. The nomogram can correctly combine information (i.e., base rate, familial risk, and test score) into consistent (less spread in opinion), unbiased (neither systematically over- or under-estimating risk), and efficient (using a parsimonious amount of information to arrive at the posterior probability) estimates (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011). This estimate, the Bayesian posterior probability, can be used in the assessment of PBD to determine the likelihood that a youth has PBD, and to guide next steps in clinical care (Youngstrom, Freeman, & Jenkins, 2009). An example of what a nomogram looks like and information about appropriate times for using this tool are provided in the Appendix (Jenkins, et al., 2011; Youngstrom & Duax, 2005, provide worked examples).

Although the nomogram is central in EB medicine (Guyatt & Rennie, 1993; Straus, Richardson, Glasziou, & Haynes, 2005), it is not widely utilized in mental health. In two recent special issues devoted to EB assessment, the guest editors (Mash & Hunsley, 2005)

and one of the invited commentaries (McFall, 2005) advocated for the adoption of this type of framework, but only three papers mentioned related approaches (such as Receiver Operating Curves and diagnostic likelihood ratios) and only one described the nomogram approach. Recent evidence suggests that taking a Bayesian approach can significantly increase diagnostic accuracy and consensus among community clinicians as well as significantly decrease overdiagnosis of PBD (Jenkins, et al., 2011). Specifically, clinicians who learned how to use a nomogram (i.e., training took less than 30 minutes) and then participated in a clinical vignette exercise (i.e., testing proficiency in skill post-training) were less likely to overestimate the risk of PBD and improved in consistency and accuracy (Jenkins, et al., 2011). Despite the fact that the actuarial strategies are not widely used in mental health settings, they have started gaining attention.

A second contribution of EB medicine is providing a framework for mapping the assessment results onto a decision-making “threshold model.” The threshold model maps the probability estimate onto a continuum with two major thresholds: the Wait-Test Threshold, and the Test-Treat Threshold. Probabilities lower than the Wait-Test Threshold are sufficiently small that the diagnosis is considered “ruled out” and diagnostic assessment is discontinued. The clinician adopts a “wait and see if anything changes” approach. If the probability rises above the Test-Treat Threshold, then the diagnosis is treated as if present, and treatment begins. In between the two thresholds is the zone where intensive assessment is indicated, until the additional information pushes the probability below the Wait-Test or above the Test-Treat Threshold. Using actuarial methods, like the nomogram, in conjunction with an EB intervention “threshold model” allows clinicians to incorporate patient preferences into the decision-making process, by negotiating adjusted threshold points (Straus, et al., 2005). The nomogram can also help anchor the likelihood of a bipolar diagnosis and inform next steps in clinical care (Youngstrom, et al., 2009, provide a worked example). It is possible that taking this approach can help avoid unnecessary assessment batteries (i.e., cost-benefit analysis) as well as protect against prematurely ruling out bipolar disorder.

Evidence-based approaches to assessment could help considerably. Improving the accuracy of diagnoses can simultaneously reduce the number of false positives, averting the overdiagnosis of bipolar disorder, while also diminishing the delay between onset and correct identification. At present, several different measures had demonstrated diagnostic validity in research samples, but it is unknown how well the performance of these tools would generalize to new settings. Psychometric performance tends to “shrink” when cross-validated in new samples, and the amount of shrinkage can change depending on new sample characteristics. In addition, the Bayesian approach recommended by EB medicine involves some simplifications, such as using scoring thresholds, which may also change performance across samples.

Aims of the Study

Although the current research on actuarial strategies is encouraging (Jenkins, et al., 2011; Youngstrom, et al., 2009), to date no research has examined the degree to which the published actuarial estimates from one sample generalize to new samples. A crucial next step is to examine how well the EB assessment recommendations in the literature might extrapolate to new clinical settings. In the present study, we tested four main hypotheses to examine the effectiveness of actuarial methods for assessing real-world cases with potential PBD. First, we predicted a positive relationship between Bayesian estimates (actuarial estimates of having PBD based on the combination of family history and test score on the Parent General Behavior Inventory) and research diagnoses as well as best estimate clinical probabilities. Using the Bayesian estimates to predict diagnoses provides one means of estimating “shrinkage” or reduced validity when applying the weights in a new sample.

Second, we predicted a positive relationship between Bayesian estimates and empirical estimates based on logistic regressions using the same predictors as the Bayesian model. The clinical ratings provide a second criterion against which to measure the extent to which a simple algorithm maps onto expert evaluation using additional sources of information, and the regression estimates provide a criterion for quantifying the “shrinkage” when using published weights compared to weights that are statistically optimal for the new sample.

Third, we anticipated that applying an EB assessment intervention “threshold model” (wait/assess/treat) to clinical ratings and Bayesian estimates would show clinically significant agreement between the two methods (e.g., Cicchetti, et al., 2006). Fourth, we examined potential moderators of agreement between the clinical and actuarial approaches. We predicted that type of bipolar would statistically moderate agreement between the two approaches.

Method

Participants

Youth participants ($N=562$) were a consecutive case series recruited from an urban community mental health center with four urban sites (Youngstrom, Youngstrom, & Starr, 2005) as part of a larger project (R01 MH066647, PI: E. Youngstrom). The majority (80%) of youth were enrolled by their mothers to participate in the study. See Table 1 for youth participant demographics and diagnostic characteristics.

Youths were included if they were between 5 years 0 months and 17 years 11 months of age and both the youth and the primary caregiver were available for the assessment. Youth were excluded if they (or their respective caregivers) could not communicate orally at a conversational level in English to complete the interview, had a pervasive developmental disorder, or there was suspected moderate, severe, or profound mental retardation. The same assessment procedures were administered to all eligible participants.

Measures

Reference Standard: Semistructured Diagnostic Interview Using the Schedule of Affective Disorders and Schizophrenia for Children (KSADS)—The Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime (KSADS-PL) (Kaufman, et al., 1997) combined with the mood disorders module from the Washington University KSADS (WASH-U-KSADS; Geller, Zimmerman, et al., 2001) was administered to all participants and their families. The KSADS is the most widely used semi-structured diagnostic procedure for investigations of PBD (Nottelmann, et al., 2001). Bipolar I (BP-I), Bipolar II (BP-II), cyclothymic disorder, and Bipolar Not Otherwise Specified (BP-NOS) diagnoses were made in accordance with *DSM-IV* diagnostic criteria (American Psychiatric Association, 2001) including a strong emphasis that mood symptoms needed to represent a clear change in functioning, and follow an episodic presentation. The most frequent reason for diagnosing BP NOS was failure to meet strict DSM duration criteria, requiring four days for a hypomanic episode and seven days or hospitalization for mania or mixed episodes (American Psychiatric Association, 2001; Leibenluft, et al., 2003), consistent with emerging data about the duration of mood episodes in clinical and epidemiological samples in youths and adults (see Youngstrom, 2009, for review). Research assistants received extensive training prior to administering KSADS ($K>.85$ at item level on ten cases) (Youngstrom, Meyers, et al., 2005).

Mini International Neuropsychiatric Interview—(MINI) (Sheehan, et al., 1998). The MINI is a brief (about 20 minutes) structured, reliable diagnostic interview covering

common DSM-IV and International Classification of Diseases (ICD-10) psychiatric disorders. The MINI has been validated against the Structured Clinical Interview for DSM diagnoses (SCID-P) (Lecrubier, et al., 1997; Otsubo, et al., 2005; Sheehan, et al., 1998) and the Composite International Diagnostic Interview (CIDI; Kadri, Agoub, Gnaoui, & Alami, 2005; Lecrubier, et al., 1997) in several languages. It was consistently faster than both other interviews.

The primary caregivers completed the MINI first about themselves, and then repeated the interview focusing on first degree relatives of the youth proband; using a modified family History – Research Diagnostic Criteria format (Andreasen, Endicott, Spitzer, & Winokur, 1977).

Parent General Behavior Inventory—(PGBI; Youngstrom et al., 2001). The PGBI has parents of youths 5 to 18 years old rate the severity of depressive, manic, and mixed mood symptoms using a 0 to 3 Likert scale. The Hypomanic/Biphasic Scale has 28 items, alpha of .93 in this sample. Higher scores on the PGBI indicate a higher probability that a given youth has a bipolar disorder. The present study uses risk estimates (Straus, et al., 2005) associated with PGBI scores based on a completely independent sample with higher average SES and more mood disorders in the referral stream, published in Youngstrom et al. (2004).

Procedure

Clinical Assessment—All study procedures were approved by the Institutional Review Boards of University Hospital of Cleveland, Case Western Reserve University, and Applewood Centers. Parents or guardians provided written consent and all youths gave written assent. Highly trained research assistants administered the Kiddie Schedule for Schizophrenia and Affective Disorders (KSADS; Kaufman, et al., 1997) diagnostic interview to all participant families.

Parents completed a series of questionnaires while youths were interviewed, including the PGBI. Similarly, youths completed questionnaires while the parents completed the KSADS and family history interviews. KSADS interviewers were blind to rating scales.

After the client and his/her family completed the psychiatric evaluation, research diagnoses were reached through a Longitudinal Expert Evaluation of All Data (LEAD; Spitzer, 1983) conference. Specifically, all cases were reviewed by an expert consensus team, which always consisted of at least one licensed psychologist in addition to the rest of the members of the interview team for the given family. LEAD diagnoses were based on: (1) results from the KSADS; (2) developmental history; (3) family history of mental illness; and, (4) psychiatric history, including any current diagnoses. Clinical chart reviews often provided information regarding youths' developmental and psychiatric histories. For each diagnosis, the expert clinician assigned a confidence rating based on how likely he/she viewed the diagnosis given all of the available information. These diagnoses and corresponding confidence ratings (likelihood of illness from 0-100%) represent the closest criterion to a gold standard in clinical assessment that field currently has; these ratings, the LEAD confidence ratings, were compared to the Bayesian estimates.

Actuarial Assessment—Procedures for actuarial assessments involved four steps which are described in more detail below: (1) Determine the prevalence or starting base rate for the study population; (2) Obtain family history of bipolar illness and test scores on the PGBI; (3) Translate family history of bipolar illness from the MINI and test scores on the PGBI for these cases into diagnostic likelihood ratios (DLRs) using published estimates from prior samples and research (Youngstrom, Frazier, Demeter, Calabrese, & Findling, 2008; Youngstrom, Meyers, et al., 2005); (4) Use Bayesian methods to combine the base rate of

PBD and the DLRs for family history and PGBI scores to provide the probability of a bipolar diagnosis (i.e., the Bayesian estimate). The Bayesian estimates were probabilities that theoretically could range from 0 to 100, and in the present sample ranged from 0.004 to .746. These were the Bayesian estimates of the likelihood of a bipolar diagnoses based on combining the likelihood ratios for prevalence rate (6%), family history of mood disorder, and test score on the Parent General Behavior Inventory (PGBI) questionnaire.

Prevalence—The present study estimated prevalence rates for PBD on the basis of a literature review, as recommended in EB medicine books (Guyatt & Rennie, 2002; Straus, et al., 2005), giving precedent to well-conducted research over local estimates. A recent meta-analysis of 12 epidemiological studies found an average rate of 2% for bipolar disorder in youths under age 19 (Van Meter, Moreira, & Youngstrom, 2011). In outpatient clinical populations, evidence suggests prevalence estimates between 0.6 to 15%, depending on the diagnostic instrument, clinic specialization, and referral source (Geller, Craney, et al., 2001; Lewinsohn, Klein, & Seeley, 1995; Strober, Schmidt-Lackner, Freeman, Bower, & al, 1995). Because prevalence of bipolar disorder varies substantially by type of setting, it is important to consider the starting base rate in light of clinical context. To determine the starting base rate for the present study, we used benchmarks from the literature that approximated local conditions (e.g., Youngstrom, 2007; Youngstrom, et al., 2009, Table 2) --similar to how clinicians could access base rate information if they were to take an actuarial approach in their clinical practice. Two published estimates indicated a base rate of 6% for bipolar spectrum disorders in an outpatient clinic, which by some standards is conservative (Pavuluri, Birmaher, & Naylor, 2005).

Diagnostic Likelihood Ratios (DLRs)—A DLR quantifies the extent to which new information, such as a test result or risk factor, changes the odds of a diagnosis. A DLR divides the percentage of cases with bipolar showing that result by the percentage of cases without bipolar that would also have similar test results. This is identical to dividing the diagnostic sensitivity by the false alarm rate (Pepe, 2003). Family history and PGBI test scores can be translated into DLRs and plotted on the middle column of a probability nomogram.

For the present study, we conducted a search using the terms “pediatric bipolar disorder” and “sensitivity and specificity” and “diagnostic likelihood ratios” to find the DLRs for PGBI test scores, and “bipolar disorder” and “risk estimate” and “systematic review” to find DLRs for family history of bipolar disorder. See Youngstrom et al. (2004) for the DLRs associated with test scores on the PGBI.

The present investigation concentrated on direct interview information about first degree relatives. For a biological parent (or other first degree relative) with bipolar disorder, the EB familial risk estimate = 5 (Hodgins, Faucher, Zarc, & Ellenbogen, 2002; Tsuchiya, Byrne, & Mortensen, 2003; Youngstrom & Duax, 2005).

Results

Less than 2% of data points were missing. Given the small amount, missing data were excluded listwise. This approach provides less bias than pairwise deletion and is adequately suited for small amounts of missing data (Allison, 2002).

Agreement between LEAD Diagnoses, Confidence and the Bayesian Estimates

Each case received a Bayesian estimate based on DLRs for their age group on the PGBI as well as family history. Receiver Operating Characteristic (ROC) analyses comparing Bayesian estimates to the LEAD diagnoses of bipolar disorder found an Area Under the Curve (AUC)

of .75, $p < .00005$. This indicated a marginal degree of shrinkage from the diagnostic efficiency of the PGBI, with AUCs of .83 in the original report (Youngstrom, et al., 2004) and .78 in the present sample. Bayesian estimates based on the combination of family history and test score on the PGBI using a base rate of 6% correlated $r_s = .30$ with the LEAD confidence ratings, $p < .00005$, indicating a medium positive relationship (Cohen, 1988). This suggests that clinicians agree with the Bayesian estimate more often than is typically the case comparing clinical versus structured diagnoses of bipolar disorder (Rettew, et al., 2009). On the other hand, these are not redundant or identical pieces of information, as the LEAD confidence rating also integrated other information not included in the Bayesian estimate.

Generalizability of Actuarial Approach

Pearson's correlation quantified how well Bayesian estimates using independent, published DLRs generalized to logistic regression estimates using optimal weights for the sample. The two estimates were highly correlated, $r = .81$, $p < .0005$, indicating a strong positive relationship (Cohen, 1988); see Figure 1.

Agreement about Next Clinical Action

Cohen's kappa coefficient tested if applying an EB assessment intervention "threshold model" (wait/assess/treat) to LEAD confidence ratings and Bayesian estimates showed clinically significant agreement between the two assessment methodologies. Results indicate a $\kappa = .24$, $p < .0005$, when test-wait threshold and test-treat thresholds were set at 20% and 90%, respectively, indicating "fair" agreement (per Cicchetti, 2006).

Figure 2 superimposes the Wait-Test and Test-Treat Thresholds on the scatter plot displaying Bayesian estimates and LEAD confidence ratings. The Bayesian estimates were generally more conservative, almost never crossing the treatment threshold, and often falling in the low or indeterminate ranges for cases where the LEAD confidence ratings were high. This pattern suggests: (a) the Bayesian approach by itself would not be sufficient to initiate treatment, at least not with the current set of inputs; (b) additional information captured in the LEAD process may be necessary to cross the treatment threshold; and, (c) the Bayesian approach is unlikely to generate "false positives" that would be exposed to unwarranted treatments.

Potential Moderators of Agreement between Nomogram and Clinical Confidence

Clinician confidence was significantly higher for cases with bipolar I or II ($M = 84.4$) versus the softer spectrum ($M = 76.7$), $t(60.60) = 3.12$, $p = .003$. An OLS regression approach tested whether type of bipolar moderated agreement between LEAD confidence ratings and Bayesian estimates. Specifically, the Bayesian predicted the LEAD confidence rating, along with dummy codes for bipolar type, and interaction terms for bipolar type with Bayesian estimates. None of the interaction terms were significant, indicating that type of bipolar did not statistically moderate agreement between LEAD confidence ratings and Bayesian estimates; see Table 2.

Discussion

The overarching goal of the present study was to compare the current gold standard for clinical assessment of pediatric bipolar disorder -- a LEAD diagnosis integrating a KSADS interview with collateral information and treatment history -- to an innovative actuarial approach. We also examined agreement about next clinical action using the threshold model developed in EB medicine (Straus, et al., 2005). Additional analyses examined potential moderators of agreement between the Bayesian and clinical approaches, including whether

agreement was higher for fully syndromal cases (i.e., bipolar I) versus other bipolar spectrum presentations.

Consistent with hypotheses, Bayesian estimates derived from published risk estimates showed clinically meaningful diagnostic efficiency even when generalized to a new sample with different clinical and demographic characteristics. The AUC for the Bayesian estimates shrank compared to the original published estimates, but still remained large (AUC .75) and highly significant. Also as hypothesized, LEAD confidence ratings and Bayesian estimates—based on a much more circumscribed set of variables--showed medium-sized correlation. Clinician confidence integrated substantially more information via a LEAD process: clinicians' LEAD confidence ratings reflect findings from the KSADS interview, detailed family history, and clinical chart information. The relationship between LEAD confidence ratings and Bayesian estimates would likely differ if clinicians did not have this additional information (see Jenkins et al., 2011). Given that LEAD confidence ratings tended to be higher than Bayesian estimates, clinicians may have been more confident in their bipolar diagnoses because of the additional “supporting” information from the KSADS.

As hypothesized, Bayesian estimates using published estimates were highly correlated with logistic regression estimates optimized for the present sample, indicating a high degree of generalizability. These findings are enhanced by the fact that the present sample is substantially different in terms of demography and SES as well as clinical referral patterns from the prior research on the assessment of PBD (cf. Youngstrom, et al., 2004). Whereas most prior work has relied on middle class, predominantly white participants with high rates of mood disorder seeking services in specialty clinics at academic centers (Hodgins, et al., 2002), the present sample was low income, predominantly underserved ethnic minority families seeking services at a community mental health center, mostly for attention problems and disruptive behavior disorders. The high correlation between new regression estimates and Bayesian estimates using published weights provides strong confirmation that the EB assessment recommendations generalize across a wide range of demographic and clinical facets.

When a clinical action threshold model was applied to LEAD confidence ratings and Bayesian estimates, these different approaches evidenced fair agreement (Cicchetti, et al., 2006; Landis & Koch, 1977). When they disagreed, the actuarial approach consistently recommended a more conservative approach--waiting or active assessment--rather than indicating active treatment. This finding has important clinical implications. Gold standard clinical assessments are often not feasible in real-world practice (Garb, 1998). For example, issues related to insurance reimbursement and staff training can make it difficult to implement a comprehensive assessment such as the KSADS (Eisman, et al., 1998). Moreover, many clinical settings lack the training and supervision resources to ensure acceptable administration of semi-structured diagnostic interviews, which can result in unreliable diagnostic impressions and inappropriate treatment plans. Results suggest that the actuarial approach could be implemented using published likelihood ratios, rapidly and inexpensively identifying cases for further assessment without overdiagnosing or overtreating bipolar disorder. There also is precedent for third party payors such as Medicaid reimbursing for KSADS or other follow-up procedures when they are clinically indicated rather than being administered to all comers.

Contrary to study predictions, the type of bipolar did not significantly moderate the level of agreement between LEAD confidence ratings and Bayesian estimates (i.e., the interaction terms were not significant). This finding is surprising given that clinician confidence was significantly higher about “fully syndromal” bipolar I or II versus “softer spectrum” cyclothymic or bipolar illness cases, consistent with findings from vignette studies

(Dubicka, Carlson, Vail, & Harrington, 2008). Clinicians are less confident when diagnosing bipolar that does not fit into current categorical definitions of the disorder, consistent with current scientific debate about the diagnostic status of the “NOS” cases (e.g., Axelson, et al., 2006; Dubicka, et al., 2008; Findling, 2005; Leibenluft, et al., 2003; Parens, et al., 2010). Apparently, type of bipolar may affect clinicians’ diagnostic confidence; but it does not influence the degree to which actuarial and clinical assessment methodologies agree.

It is noteworthy that the present sample was predominantly African American. Given the well-documented difficulties assessing bipolar using other methods in minority samples -- especially African Americans -- (DelBello, Lopez-Larson, Soutullo, & Strakowski, 2001; Neighbors, Trierweiler, Ford, & Muroff, 2003), the effectiveness of the actuarial approach is all the more impressive. Actuarial assessment methods may help decrease the rate at which clinicians misdiagnose African Americans and other minority groups by encouraging more systematic and thorough assessment, as well as preventing or adjusting faulty clinical judgment or differences in initial description of the presenting problem (Jenkins, et al., 2011). For example, if an individual presents with a family history of mental illness and/or receives a positive test score on a bipolar screening instrument, the actuarial approach generates the same probability of bipolar disorder regardless of race/ethnicity. If there were evidence that demographic characteristics moderated risk or test performance, then the actuarial approach could incorporate optimal weights to make the appropriate adjustments.

Limitations

A large portion of the bipolar spectrum cases had diagnoses of cyclothymic disorder and bipolar NOS. By definition, these cases did not present with distinct episodes of week-long mania or sufficient severity to require hospitalization. However, the research diagnoses required an episodic presentation and that symptoms be a clear change in functioning if they were counted towards a mood diagnosis. The majority of cases with NOS or cyclothymic diagnoses failed to meet criteria for bipolar I or II due to insufficient duration of the activated mood states, but virtually all had multiple episodes, and a mean episode duration well in excess of other research definitions (Birmaher, et al., 2006). The emphasis on episodicity and differentiating mood from other possible sources of symptoms increases the specificity of the bipolar diagnoses. It also is important to understand how findings generalize across the bipolar spectrum, as the cyclothymic and NOS cases are proving to be more frequent in epidemiological studies (Merikangas, et al., 2010; Van Meter, et al., 2011) as well as clinical settings (Axelson, et al., 2006), and continue to be associated with high degrees of impairment (Birmaher, et al., 2009; Findling, et al., 2010; Findling, et al., 2005).

It is possible that the conservative approach used in defining risk due to family history of bipolar illness may have lowered estimates of agreement. Some research suggests that individuals with familial probands of bipolar disorder may be as much as ten times more likely to manifest the illness (Smoller & Finn, 2003). Using a smaller likelihood ratio for family risk contributed to the Bayesian approach yielding lower risk estimates. This is not a general limitation of the EB assessment approach: In fact, a strength is that the EB assessment methods can flexibly integrate new information. Also, clinicians have the option of performing “sensitivity analyses,” where they examine the effects of changing the inputs on the range of resulting probabilities. These “what if” scenarios inform the clinician about how changes to assumptions, or the addition of new evidence, would change diagnostic probabilities.

Another caveat is that family history information sometimes was inaccurate or incomplete. Most family history information was gathered from participants’ mothers’ using a semi-structured interview. Collateral report, or a family history method, is less accurate than direct interview of all relatives. The rate of bipolar disorder reported in the other family

members was lower than the rates found in epidemiological studies in the general U.S. population (Merikangas & Pato, 2009), suggesting that the interview did not demonstrate high diagnostic sensitivity to bipolar conditions in fathers or second degree relatives. However, these same factors of low awareness of bipolar history in other family members, along with potential bias, are likely to apply to most clinical evaluations as well.

Some of the results are also dependent on starting values, such as the base rate of bipolar disorder or the choice of values for the Wait-Test and Test-Treat thresholds. This is intrinsic to using a Bayesian approach, but it is likely to be disconcerting at first because it seems both unfamiliar and potentially subjective (Gigerenzer & Goldstein, 1996). In fact, the Bayesian methods are often more consistent, efficient, and unbiased than clinical judgment in synthesizing information (Jenkins, et al., 2011), and these advantages remain even when using inaccurate starting values. Additionally, EB assessment methods have will be self-correcting over time: Using the techniques will change diagnostic patterns to converge on the true base rate even when starting from inaccurate values (Meehl, 1954). The potential for subjectivity in choice of thresholds also is a potential strength, as it makes it possible to have meaningful discussions about costs and benefits with the patient and family (Kraemer, 1992; Straus, et al., 2005; Swets, Dawes, & Monahan, 2000). Finally, clinician confidence is not completely equivalent to clinical judgment as operationalized in some other studies (cf. Meehl, 1954).

Future Directions and Clinical Implications

Dissemination efforts are more successful when procedures are acceptable, feasible, and likely to permit adherence (Arndorfer, Allen, & Aljazireh, 1999). Better understanding of clinician perspectives about actuarial assessment along with identifying potential practice barriers can enhance acceptance, feasibility, and adherence. Before attempting to create large scale change in current decision-making practice, researchers should first pilot EB tools with providers. It will be important to understand if clinicians on the front line are willing to adopt actuarial approaches in real-world practice. Recent research suggests that a majority of MH professionals who received a brief nomogram training (< 30 minutes) endorsed using it in practice (Jenkins, et al., 2011). Qualitative methodologies may shed light on clinicians' attitudes and impressions that can inform future education and training of the nomogram.

More work needs to evaluate barriers related to more general policy considerations (e.g., supervision, clinical training models) as well as consumer preferences (Suppiger, et al., 2009). Documenting and addressing these obstacles could accelerate uptake of EB assessment methods. Changes in "packaging" and interpretation software may also accelerate uptake. Innovative technology combined with EB decision-making tools might greatly appeal to audiences that find statistics intimidating. Using technology to expedite the delivery of EB mental health services is a rapidly growing niche (Bucholz, et al., 1991; Erdman, et al., 1992; Finfgeld, 1999; Kobak, et al., 1997).

Another area of work will be studying the effects of varying the Wait-Test and Test-Treat Thresholds, as well as incorporating patient preferences about risks and benefits. Some of the controversy around the bipolar diagnosis has focused on the consequences of erroneously prescribed pharmacological treatment (DSM5.org, accessed January 19, 2011), rather than the underlying validity of the diagnosis, which has accumulated substantial evidence (Geller & Tillman, 2005; Youngstrom, Birmaher, et al., 2008). Psychopharmacological agents used to treat bipolar disorder carry a range of potential side effects extending from the unpleasant to the life-threatening. The perceived adversity of any treatment's costs and side effects will naturally vary on an individual basis. More work is needed to understand how the interaction of clinicians' recommendations and patient

preferences map on to the threshold model. Adjusting the Test-Treat Threshold provides a mechanism for clinicians to directly address the potential for harm by literally setting the bar higher, requiring greater confirmatory evidence before initiating the riskiest or most expensive treatments (Straus, et al., 2005).

EB assessment offers a fast and frugal approach to gathering information about bipolar disorder in youngsters. Combined with a two-threshold (Wait-Test, Test-Treat) model, the current generation of EB assessment tools can increase detection of bipolar without increasing the rate of false positives who are then exposed to unnecessary treatments. Using brief, free/public domain tools can identify a limited number of cases for more intensive assessment and follow up. Medicaid and other payors have been found to reimburse this additional evaluation when medical necessity has been demonstrated. Upgrades in the tools and data available will further enhance the accuracy of this approach without requiring a change to the interpretive framework. Present findings also indicated that risk estimates derived from the published literature would generalize well, even to new samples with markedly different demographic and clinical characteristics. The EB approach could likely yield immediate improvements in many clinical settings, which would be further refined as new tools, risk factors, and moderators of assessment and treatment performance are identified.

Acknowledgments

This work was supported in part by NIH R01 MH066647 (PI: E. Youngstrom) Dr. Findling receives or has received research support, acted as a consultant and/or served on a speaker's bureau for Abbott, Adrenex, Alexza, AstraZeneca, Biovail, Bristol-Myers Squibb, Forest, GlaxoSmithKline, Johnson & Johnson, KemPharm Lilly, Lundbeck, Merck, Neuropharm, Novartis, Noven, Organon, Otsuka, Pfizer, Rhodes Pharmaceuticals, Sanofi-Aventis, Schering-Plough, Seaside Therapeutics, Sepracore, Shire, Solvay, Sunovion, Supernus Pharmaceuticals, Transept Pharmaceuticals, Validus, and Wyeth.

Dr. E. Youngstrom has received travel support from Bristol-Myers Squibb.

Appendix

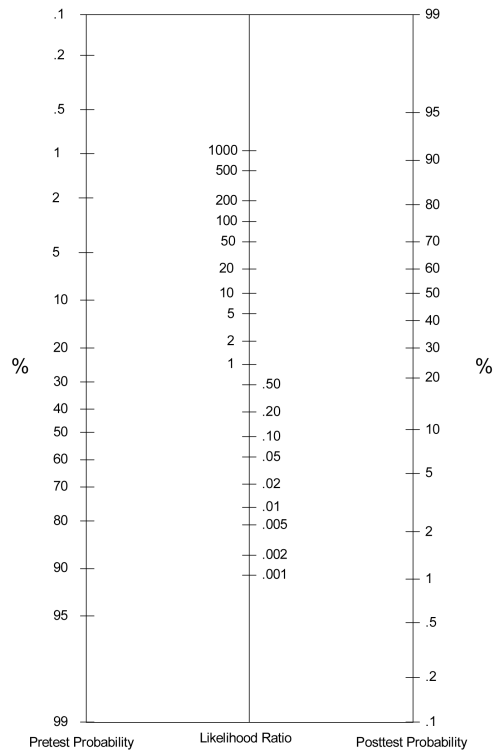


Figure A1.

Nomogram for Combining Probability with Diagnostic Likelihood Ratios

Note. A nomogram is a particularly helpful tool for quantifying the risk of bipolar disorder when certain warning signs are present (Youngstrom et al., 2009). Warning signs can include family history of bipolar disorder, a high score on a parent report questionnaire that is sensitive to manic symptoms, and/or a youth's clinical presentation of decreased need for sleep, elevated, expansive mood, and grandiosity, or possible psychotic features.

References

- Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*. 2006; 34(3):341–382. doi: 10.1177/0011000005285875.
- Akiskal HS, Hantouche EG, Allilaire JF, Sechter D, Bourgeois ML, Azorin JM, et al. Validating antidepressant-associated hypomania (bipolar III): a systematic comparison with spontaneous hypomania (bipolar II). *Journal of Affective Disorders*. 2003; 73(1-2):65–74. doi:10.1016/S0165-0327(02)00325-7. [PubMed: 12507739]
- Allison, PD. *Missing data*. Sage; Thousand Oaks: 2002.
- Altshuler LL, Post RM, Leverich GS, Mikalauskas K, Rosoff A, Ackerman L. Antidepressant-induced mania and cycle acceleration: a controversy revisited. *American Journal of Psychiatry*. 1995; 152(8):1130–1138. doi: unknown. [PubMed: 7625459]
- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. 4th--Text Revision ed. Author; Washington, DC: 2001.

- Andreasen NC, Endicott J, Spitzer RL, Winokur G. The family history method using diagnostic criteria. Reliability and validity. *Archives of General Psychiatry*. 1977; 34(10):1229–1235. doi: unknown. [PubMed: 911222]
- Arndorfer RE, Allen KD, Aljazireh L. Behavioral health needs in pediatric medicine and the acceptability of behavioral solutions: Implications for behavioral psychologists. *Behavior therapy*. 1999; 30(1):137–148. doi: 10.1016/S0005-7894(99)80050-1.
- Axelson DA, Birmaher B, Strober M, Gill MK, Valeri S, Chiappetta L, et al. Phenomenology of children and adolescents with bipolar spectrum disorders. *Archives of General Psychiatry*. 2006; 63(10):1139–1148. doi: 10.1001/archpsyc.63.10.1139. [PubMed: 17015816]
- Birmaher B, Axelson D, Goldstein B, Strober M, Gill MK, Hunt J, et al. Four-Year Longitudinal Course of Children and Adolescents With Bipolar Spectrum Disorders: The Course and Outcome of Bipolar Youth (COBY) Study. *American Journal of Psychiatry*. 2009; 166:795–804. doi: 10.1176/appi.ajp.2009.08101569. [PubMed: 19448190]
- Blader JC, Carlson GA. Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996–2004. *Biological Psychiatry*. 2007; 62(2):107–114. doi: 10.1016/j.biopsych.2006.11.006. [PubMed: 17306773]
- Bowring MA, Kovacs M. Difficulties in diagnosing manic disorders among children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1992; 31(4):611–614. doi: 10.1097/00004583-199207000-00006. [PubMed: 1644722]
- Bucholz KK, Robins LN, Shayka JJ, Przybeck TR, Helzer JE, Goldring E, et al. Performance of two forms of a computer psychiatric screening interview: version I of the DISSI. *Journal of Psychiatric Research*. 1991; 25(3):117–129. doi: 10.1016/0022-3956(91)90005-U. [PubMed: 1941708]
- Carlson GA. The bottom line. *Journal of Child and Adolescent Psychopharmacology*. 2003; 13(2): 115–118. doi:10.1089/104454603322163826. [PubMed: 12886908]
- Cicchetti D, Bronen R, Spencer S, Haut S, Berg A, Oliver P, et al. Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. *Journal of Nervous and Mental Disease*. 2006; 194(8):557–564. doi: 10.1097/01.nmd.0000230392.83607.c5. [PubMed: 16909062]
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum; Hillsdale: 1988.
- Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine*. 2002; 9(11):1184–1204. doi: 10.1197/aemj.9.11.1184. [PubMed: 12414468]
- DelBello MP, Adler CM, Whitsel RM, Stanford KE, Strakowski SM. A 12-week single-blind trial of quetiapine for the treatment of mood symptoms in adolescents at high risk for developing bipolar I disorder. *The Journal of Clinical Psychiatry*. 2007; 68(5):789–795. doi:10.4088/JCP.v68n0520. [PubMed: 17503991]
- DelBello MP, Lopez-Larson MP, Soutullo CA, Strakowski SM. Effects of race on psychiatric diagnosis of hospitalized adolescents: A retrospective chart review. *Journal of Child and Adolescent Psychopharmacology*. 2001; 11(1):95–103. doi:10.1089/104454601750143528. [PubMed: 11322750]
- Dubicka B, Carlson GA, Vail A, Harrington R. Prepubertal mania: diagnostic differences between US and UK clinicians. *European Child & Adolescent Psychiatry*. 2008; 17(3):153–161. doi: 10.1007/s00787-007-0649-5. [PubMed: 17876503]
- Eisman, EJ.; Dies, RR.; Finn, SE.; Eyde, LD.; Kay, GG.; Kubiszyn, TW., et al. Problems and limitations in the use of psychological assessment in contemporary health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part II (pp. 22). American Psychological Association; Washington, DC: 1998.
- Erdman HP, Klein MH, Greist JH, Skare SS, Husted JJ, Robins LN, et al. A comparison of two computer-administered versions of the NIMH Diagnostic Interview Schedule. *Journal of Psychiatric Research*. 1992; 26(1):85–95. doi: 10.1016/0022-3956(92)90019-K. [PubMed: 1560412]

- Findling RL. Update on the treatment of bipolar disorder in children and adolescents. *European Psychiatry: the Journal of the Association of European Psychiatrists*. 2005; 20(2):87–91. doi: unknown. [PubMed: 15797690]
- Findling RL, Frazier TW, Youngstrom EA, McNamara NK, Stansbrey RJ, Gracious BL, et al. Double-blind, placebo-controlled trial of divalproex monotherapy in the treatment of symptomatic youth at high risk for developing bipolar disorder. *Journal of Clinical Psychiatry*. 2007; 68(5):781–788. doi:10.4088/JCP.v68n0519. [PubMed: 17503990]
- Findling RL, Youngstrom EA, Fristad MA, Birmaher B, Kowatch RA, Arnold LE, et al. Characteristics of children with elevated symptoms of mania: the Longitudinal Assessment of Manic Symptoms (LAMS) study. *Journal of Clinical Psychiatry*. 2010; 71:1664–1672. doi: 10.4088/JCP.09m05859yel. [PubMed: 21034685]
- Findling RL, McNamara NK, Gracious BL, Youngstrom EA, Stansbrey RJ, Reed MD, et al. Combination lithium and divalproex in pediatric bipolarity. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2003; 42(8):895–901. doi:10.1097/01.CHI.0000046893.27264.53. [PubMed: 12874490]
- Findling RL, Youngstrom EA, McNamara NK, Stansbrey RJ, Demeter CA, Bedoya D, et al. Early symptoms of mania and the role of parental risk. *Bipolar Disorders*. 2005; 7(6):623–634. doi: 10.1111/j.1399-5618.2005.00260.x. [PubMed: 16403188]
- Fingfeld DL. Computer-based mental health assessments. Panaceas, pariahs, or partners in research and practice? *Computers in Nursing*. 1999; 17(5):215. doi: unknown. [PubMed: 10502881]
- Fristad MA, Verducci JS, Walters K, Young ME. Impact of multifamily psychoeducational psychotherapy in treating children aged 8 to 12 years with mood disorders. *Archives of General Psychiatry*. 2009; 66(9):1013–1021. doi:10.1001/archgenpsychiatry.2009.112. [PubMed: 19736358]
- Galanter C, Patel V. Medical Decision making: a selective review for child psychiatrists and psychologists. *Journal of Child Psychology & Psychiatry*. 2005; 46(7):675–689. doi: 10.1111/j.1469-7610.2005.01452.x. [PubMed: 15972065]
- Garb, HN. *Studying the clinician: Judgment research and psychological assessment*. American Psychological Association; Washington, DC: 1998.
- Geller B, Craney JL, Bolhofner K, DelBello MP, Williams M, Zimmerman B. One-year recovery and relapse rates of children with a prepubertal and early adolescent bipolar disorder phenotype. *The American Journal of Psychiatry*. 2001; 158(2):303–305. doi:10.1176/appi.ajp.158.2.303. [PubMed: 11156815]
- Geller B, Luby J. Child and adolescent bipolar disorder: A review of the past 10 years. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1997; 36(9):1168–1176. doi: 10.1097/00004583-199709000-00008. [PubMed: 9291717]
- Geller B, Tillman R. Prepubertal and early adolescent bipolar I disorder: review of diagnostic validation by Robins and Guze criteria. *The Journal of Clinical Psychiatry*. 2005; 66(Suppl 7):21–28. doi: unknown. [PubMed: 16124838]
- Geller B, Tillman R, Bolhofner K, Zimmerman B. Child bipolar I disorder: prospective continuity with adult bipolar I disorder; characteristics of second and third episodes; predictors of 8-year outcome. *Archives of general psychiatry*. 2008; 65(10):1125–1133. doi:10.1001/archpsyc.65.10.1125. [PubMed: 18838629]
- Geller B, Zimmerman B, Williams M, Bolhofner K, Craney JL, DelBello MP, 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. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2001; 40(4):450–455. doi:10.1097/00004583-200104000-00014. [PubMed: 11314571]
- Ghaemi SN, Sachs GS, Chiou AM, Pandurangi AK, Goodwin FK. Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? *Journal of affective disorders*. 1999; 52(1-3): 135–144. doi: 10.1016/S0165-0327(98)00076-7. [PubMed: 10357026]
- Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*. 1996; 103(4):650–669. doi: 10.1037/0033-295X.103.4.650. [PubMed: 8888650]

- Goodwin, FK.; Jamison, KR. Manic-depressive illness. 2nd ed. Oxford University Press; New York: 2007.
- Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*. 2000; 12(1):19–30. doi:10.1037/1040-3590.12.1.19. [PubMed: 10752360]
- Guyatt GH, Rennie D. Users' guides to the medical literature. *Journal of the American Medical Association*. 1993; 270(17):2096–2097. doi: 10.1001/jama.1993.03510170083036. [PubMed: 8411578]
- Guyatt, GH.; Rennie, D., editors. Users' guides to the medical literature. AMA Press; Chicago: 2002.
- Healy D. The latest mania: selling bipolar disorder. *PLoS Medicine*. 2006; 3(4):185. doi: 10.1371/journal.pmed.0030185.
- Hirschfeld RM, Lewis L, Vornik LA. Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry*. 2003; 64(2):161–174. doi: 10.4088/JCP.v64n0209. [PubMed: 12633125]
- Hirschfeld RMA, Bowden CL, Gitlin MJ, Keck PE, Suppes T, Thase ME, et al. American Psychiatric Association: Practice Guidelines for the treatment of patients with bipolar disorder. Washington DC: American Psychiatric Association. 2002:2.
- Hodgins S, Faucher B, Zarac A, Ellenbogen M. Children of parents with bipolar disorder. A population at high risk for major affective disorders. *Child & Adolescent Psychiatric Clinics of North America*. 2002; 11(3):533–553. doi: 10.1016/S1056-4993%2802%2900002-0.
- Jaesckhe R, Guyatt GH, Sackett DL. Users' guides to the medical literature: III. How to use an article about a diagnostic test: A. Are the results of the study valid? *Journal of the American Medical Association*. 1994; 271(5):389–391. doi:10.1001/jama.271.5.389. [PubMed: 8283589]
- Jenkins MM, Youngstrom EA, Washburn JJ, Youngstrom JK. Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice*. 2011; 42(2):121–129. doi: 10.1037/a0022506.
- Joseph M, Youngstrom EA, Soares JC. Antidepressant-Coincident Mania in Children and Adolescents Treated with Selective Serotonin Reuptake Inhibitors. *Future Neurology*. 2009; 4(1):87–102. doi: 10.2217/14796708.4.1.87. [PubMed: 19884978]
- Kadri N, Agoub M, Gnaoui SE, Alami KH. Moroccan colloquial Arabic version of the Mini International Neuropsychiatric Interview (MINI): qualitative and quantitative validation. *European psychiatry*. 2005; 20(2):193–195. doi:10.1016/j.eurpsy.2004.11.007. [PubMed: 15797707]
- Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1997; 36(7):980–988. doi: 10.1097/00004583-199707000-00021. [PubMed: 9204677]
- Kluger J, Song S. Young and bipolar. *Aug 19.2002* :39–47. 51. Time.
- Kobak KA, Taylor LH, Dotti SL, Greist JH, Jefferson JW, Burroughs D, et al. Computerized screening for psychiatric disorders in an outpatient community mental health clinic. *Psychiatric Services*. 1997; 48(8):1048. doi: unknown. [PubMed: 9255838]
- Kovacs M. Affective disorders in children and adolescents. Special Issue: Children and their development: Knowledge base, research agenda, and social policy application. *American Psychologist*. 1989; 44(2):209–215. doi:10.1037/0003-066X.44.2.209. [PubMed: 2653132]
- Kowatch RA, Fristad MA, Birmaher B, Wagner KD, Findling RL, Hellander M. Treatment guidelines for children and adolescents with bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2005; 44(3):213–235. doi:10.1097/00004583-200503000-00006. [PubMed: 15725966]
- Kowatch RA, Youngstrom EA, Danielyan A, Findling RL. Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders*. 2005; 7(6):483–496. doi: 10.1111/j.1399-5618.2005.00261.x. [PubMed: 16403174]
- Kraemer, HC. Evaluating medical tests: Objective and quantitative guidelines. Sage Publications; Newbury Park, CA: 1992.

- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33(1):159. doi: 10.2307/2529310. [PubMed: 843571]
- Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K. et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry*. 1997; 12(5):224–231. doi: 10.1016/S0924-9338(97)83296-8.
- Leibenluft E, Charney DS, Towbin KE, Bhangoo RK, Pine DS. Defining clinical phenotypes of juvenile mania. *The American Journal of Psychiatry*. 2003; 160:430–437. doi: 10.1176/appi.ajp.160.3.430. [PubMed: 12611821]
- Lewinsohn PM, Klein DN, Seeley JR. Bipolar disorders in a community sample of older adolescents: Prevalence, phenomenology, comorbidity, and course. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1995; 34(4):454–463. doi:10.1097/00004583-199504000-00012. [PubMed: 7751259]
- Lish JD, Dime-Meenan S, Whybrow PC, Price RA, Hirschfeld RM. The National Depressive and Manic-Depressive Association (DMDA) survey of bipolar members. *Journal of Affective Disorders*. 1994; 31(4):281–294. doi: 10.1016/0165-0327(94)90104-X. [PubMed: 7989643]
- Marchand WR, Wirth L, Simon C. Delayed diagnosis of pediatric bipolar disorder in a community mental health setting. *Journal of Psychiatric Practice*. 2006; 12(2):128–133. doi: 10.1097/00131746-200603000-00011. [PubMed: 16728912]
- Mash EJ, Hunsley J. Evidence-based assessment of child and adolescent disorders: issues and challenges. *Journal of clinical child and adolescent psychology*. 2005; 34(3):362–379. doi: 10.1207/s15374424jccp3403_1. [PubMed: 16026210]
- McFall RM. Theory and utility--key themes in evidence-based assessment: comment on the special section. *Psychological assessment*. 2005; 17:312–323. doi: 10.1037/1040-3590.17.3.312. [PubMed: 16262457]
- Meehl, PE. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press; Minneapolis: 1954.
- Merikangas KR, Pato M. The epidemiology of bipolar disorder in adults and children: Magnitude, correlates and future directions. *Clinical Psychology: Science & Practice*. 2009; 16(2):121–133. doi: 10.1111/j.1468-2850.2009.01152.x.
- Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, et al. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*. 2010; 49:980–989. doi: S0890-8567(10)00476-4 10.1016/j.jaac.2010.05.017. [PubMed: 20855043]
- Miklowitz DJ, Chang KD. Prevention of bipolar disorder in at-risk children: theoretical assumptions and empirical foundations. *Development and Psychopathology*. 2008; 20(3):881–897. doi: 10.1017/S0954579408000424. [PubMed: 18606036]
- Moreno C, Laje G, Blanco C, Jiang H, Schmidt AB, Olfson M. National Trends in the Outpatient Diagnosis and Treatment of Bipolar Disorder in Youth. *Archives of General Psychiatry*. 2007; 64(9):1032–1039. doi: 10.1001/archpsyc.64.9.1032. [PubMed: 17768268]
- 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. *Journal of Health and Social Behavior*. 2003; 44(3):237–256. doi: 10.2307/1519777. [PubMed: 14582306]
- Neisworth JT, Bagnato SJ. The Mismeasure of Young Children: The Authentic Assessment Alternative. *Infants and Young Children*. 2004; 17(3):198–212. doi: 10.1097/00001163-200407000-00002.
- Nottelmann E, Biederman J, Birmaher B, Carlson GA, Chang KD, Fenton WS, et al. National Institute of Mental Health research roundtable on prepubertal bipolar disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2001; 40(8):871–878. doi: 10.1097/00004583-200108000-00007. [PubMed: 11501685]
- Otsubo T, Tanaka K, Koda R, Shinoda J, Sano N, Tanaka S, et al. Reliability and validity of Japanese version of the Mini International Neuropsychiatric Interview. *Psychiatry and Clinical*

- Neurosciences. 2005; 59(5):517–526. doi: 10.1111/j.1440-1819.2005.01408.x. [PubMed: 16194252]
- Papoulos, DF.; Papoulos, J. The bipolar child: The definitive and reassuring guide to childhood's most misunderstood disorder. 2nd ed. Broadway Books; New York: 2002.
- Parens E, Johnston J, Carlson G. Pediatric Mental Health Care Dysfunction Disorder? *New England Journal of Medicine*. 2010; 362:1853–1855. doi: 10.1056/NEJMp1003175. [PubMed: 20484395]
- Pavuluri MN, Birmaher B, Naylor MW. Pediatric bipolar disorder: a review of the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2005; 44(9):846–871. doi: 10.1097/01.chi.0000170554.23422.c1. [PubMed: 16113615]
- Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Wiley; New York: 2003.
- Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*. 2009; 18(3):169–184. doi: 10.1002/mpr.289. [PubMed: 19701924]
- Scheffer RE, Kowatch RA, Carmody T, Rush AJ. Randomized, placebo-controlled trial of mixed amphetamine salts for symptoms of comorbid ADHD in pediatric bipolar disorder after mood stabilization with divalproex sodium. *The American Journal of Psychiatry*. 2005; 162(1):58–64. doi:10.1176/appi.ajp.162.1.58. [PubMed: 15625202]
- Sheehan DV, Lecrubier Y, Sheehan KH, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry*. 1997; 12(5):232–241. doi: 10.1016/S0924-9338(97)83297-X.
- Smoller JW, Finn CT. Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics. Part C, Seminars in medical genetics*. 2003; 123C(1):48–58. doi: 10.1002/ajmg.c.20013.
- Spitzer RL. Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*. 1983; 24(5):399–411. doi: 10.1016/0010-440X(83)90032-9. [PubMed: 6354575]
- Straus, SE.; Richardson, WS.; Glasziou, P.; Haynes, RB. Evidence-based medicine: How to practice and teach EBM. 3rd ed. Churchill Livingstone; New York: 2005.
- Strober M, Schmidt-Lackner S, Freeman R, Bower S. Recovery and relapse in adolescents with bipolar affective illness: A five-year naturalistic, prospective follow-up. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1995; 34(6):724–731. al, e. doi: 10.1097/00004583-199506000-00012. [PubMed: 7608045]
- Suppiger A, In-Albon T, Hendriksen S, Hermann E, Margraf J, Schneider S. Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy*. 2009; 40(3):272–279. doi:10.1016/j.beth.2008.07.002. [PubMed: 19647528]
- Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*. 2000; 1(1):1–26. doi: doi:10.1111/1529-1006.001.
- Tsuchiya KJ, Byrne M, Mortensen PB. Risk factors in relation to an emergence of bipolar disorder: A systematic review. *Bipolar Disorders*. 2003; 5(4):231–242. doi:10.1034/j.1399-5618.2003.00038.x. [PubMed: 12895201]
- Van Meter A, Moreira AL, Youngstrom EA. Meta-analysis of epidemiological studies of pediatric bipolar disorder. *Journal of Clinical Psychiatry*, in press. 2011 doi:10.4088/JCP.10m06290.
- Weller EB, Danielyan AK, Weller RA. Somatic treatment of bipolar disorder in children and adolescents. *Psychiatric Clinics of North America*. 2004; 27(1):155–178. doi: 10.1016/S0193-953X(03)00116-3. [PubMed: 15062636]
- Youngstrom, EA. Pediatric bipolar disorder. In: Mash, EJ.; Barkley, RA., editors. *Assessment of Childhood Disorders*. 4th ed. Guilford Press; New York: 2007. p. 253-304.
- Youngstrom EA. Definitional issues in bipolar disorder across the life cycle. *Clinical Psychology: Science & Practice*. 2009; 16:140–160.
- Youngstrom EA, Birmaher B, Findling RL. Pediatric bipolar disorder: Validity, phenomenology, and recommendations for diagnosis. *Bipolar Disorders*. 2008; 10(Supplement 1):194–214. doi: 10.1111/j.1399-5618.2007.00563.x. [PubMed: 18199237]

- Youngstrom EA, Duax J. Evidence based assessment of pediatric bipolar disorder, part 1: Base rate and family history. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2005; 44:712–717. doi: 10.1097/01.chi.0000162581.87710.bd. [PubMed: 15968241]
- Youngstrom EA, Findling RL, Calabrese JR, Gracious BL, Demeter C, DelPorto Bedoya D. et al. Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2004; 43:847–858. doi: 10.1097/01.chi.0000125091.35109.1e. [PubMed: 15213586]
- Youngstrom EA, Findling RL, Danielson CK, Calabrese JR. Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment*. 2001; 13(2):267–276. doi: 10.1037/1040-3590.13.2.267. [PubMed: 11433802]
- Youngstrom EA, Findling RL, Youngstrom JK, Calabrese JR. Toward an evidence-based assessment of pediatric bipolar disorder. *Journal of Clinical Child and Adolescent Psychology*. 2005; 34(3): 433–448. doi: 10.1207/s15374424jccp3403_4. [PubMed: 16026213]
- 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. *Journal of Clinical Psychiatry*. 2008; 69(5):831–839. doi: 10.4088/JCP.v69n0517. [PubMed: 18452343]
- Youngstrom EA, Freeman AJ, Jenkins MM. The assessment of children and adolescents with bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America*. 2009; 18(2):353–390. doi: 10.1016/j.chc.2008.12.002. [PubMed: 19264268]
- Youngstrom EA, Meyers OI, Demeter C, Kogos Youngstrom J, Morello L, Piiparinen R, et al. Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders*. 2005; 7(Special Issue: Pediatric Bipolar Disorder):507–517. doi: 10.1111/j.1399-5618.2005.00269. [PubMed: 16403176]
- Youngstrom EA, Youngstrom JK, Starr M. Bipolar Diagnoses in Community Mental Health: Achenbach CBCL Profiles and Patterns of Comorbidity. *Biological Psychiatry*. 2005; 58:569–575. doi: 10.1016/j.biopsych.2005.04.004. [PubMed: 15950197]
- Zuckerman, M. *Vulnerability to psychopathology: A biosocial model*. American Psychological Association; Washington, DC: 1999.

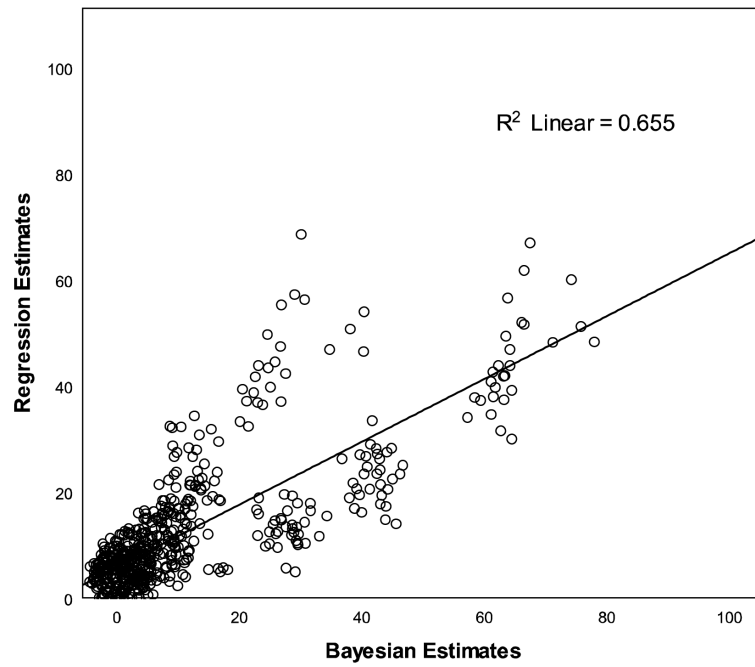


Figure 1. Generalizability of the nomogram estimates based on published likelihood ratios compared to logistic regression estimates optimized for new sample ($N=562$).

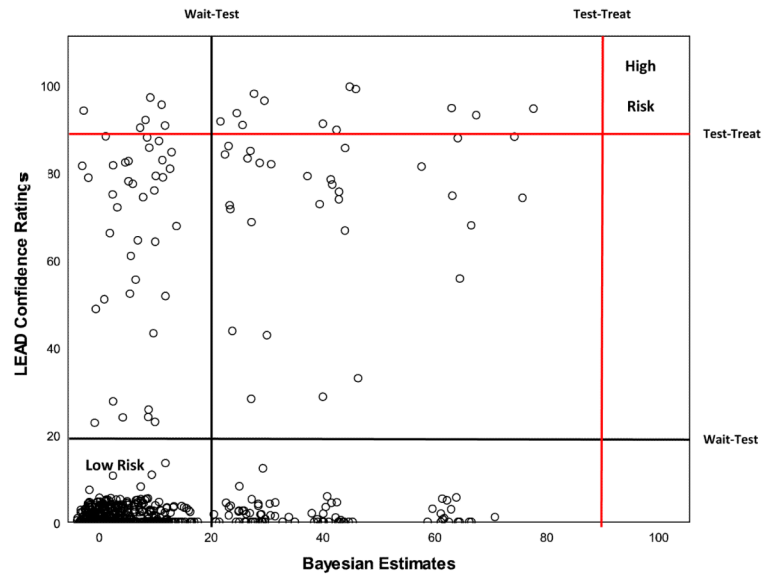


Figure 2. Agreement between clinician confidence ratings and Bayesian nomogram estimates based on published likelihood ratios.
Note: The data are skewed for the clinician confidence ratings. The distribution is zero-inflated, with 0 being the modal response because the clinicians assigned a low probability of bipolar disorder to the majority of the cases.

Table 1

Participant Demographic and Diagnostic Characteristics (N = 562)

Characteristic	<i>n</i> (% of full sample)	% Within Bipolar Subset
Caregiver		
Marital Status		
Single Parent (Never Married)	326 (58%)	59%
Married	112 (20%)	25%
Separated/Divorced	96 (17%)	4%
Widowed	28 (5%)	9%
Education		
Partial High School	163 (29%)	30%
High School Graduate or GED	197 (35%)	30%
1-3 Years College or Trade School	169 (30%)	36%
4 Year Degree	28 (5%)	3%
Some Graduate School	17 (2%)	1%
Reported Annual Income		
< \$5,000	174 (31%)	25%
\$5,000 to \$14,999	214 (38%)	41%
\$15,000 to \$29,999	124 (22%)	23%
\$30,000+	51 (9%)	11%
Youth		
Gender: Male	341 (61%)	59%
Average Age	10.6 (<i>SD</i> = 3.4)	10.5 (<i>SD</i> = 3.7)
Race/Ethnicity		
Black	495 (88%)	80%
White	37 (7%)	9%
Hispanic	8 (1%)	4%
Other	22 (4%)	7%
Average # of Axis I Diagnoses	3.8 (<i>SD</i> = 1.7)	4.7 (<i>SD</i> = 1.6)***
LEAD Diagnoses on Bipolar Spectrum		
Bipolar I	72 (13%)	--
Bipolar II	16 (3%)	--
Cyclothymic Disorder	7 (1%)	--
Bipolar NOS	22 (4%)	--
LEAD Diagnoses All Not Bipolar Spectrum		
Unipolar Depression (MDD, dysthymia, adjustment-depressed mood)	168 (30%)	--
ADHD or disruptive behavior without mood disorder	272 (48%)	--
Residual (anxiety, PTSD, psychotic disorders, no Axis I)	49 (9%)	--
Any ADHD (no Bipolar diagnosis)	315 (64%)	74%
Global Assessment of Functioning Score	52.7 (<i>SD</i> = 8.4)	47.0 (<i>SD</i> = 6.2)***

*** Bipolar group different from rest of sample $p < .0005$, two-tailed; all other comparisons between bipolar and non-bipolar cases not significant, uncorrected p value $> .05$.

Table 2

Regression model examining whether bipolar subtype moderates agreement between Bayesian estimates and LEAD ratings (N=562)

Predictor	<i>B</i>	<i>SE</i>	<i>p</i>	<i>Part r</i>
Bayesian Estimate	25.4	4.8	.000	.14
LEAD Diagnostic Status				
BP-I Yes or No	78.9	7.3	.000	.29
BP-II Yes or No	82.6	10.7	.000	.21
Cyclothymic Disorder Yes or No	78.1	6.0	.000	.34
Interaction Terms (Diagnosis * Actuarial Bayesian Estimate)				
BP-I	-11.4	23.2	.622	-.01
BP-II	-27.4	32.5	.399	-.02
Cyclothymic Disorder	-34.8	18.5	.061	-.05
Constant	2.2	0.9	.015	

Note. The significance of the interaction terms was the direct test of the hypothesis; none were significant, $p > .05$. Overall model $R^2 = .61$, adjusted $R^2 = .61$, $p < .0005$.