Available online at www.sciencedirect.com

ScienceDirect

Cognitive and Behavioral Practice xx (2014) xxx-xxx

Cognitive and Behavioral Practice

www.elsevier.com/locate/cabp

Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment

Eric A. Youngstrom, Sophia Choukas-Bradley and Casey D. Calhoun, *University of North Carolina at Chapel Hill*
Amanda Jensen-Doss, *University of Miami*

Assessment plays an essential role in diagnosis, treatment planning, and progress monitoring, but assessment data are often used in ways that are impressionistic and prone to biases. Evidence-based medicine (EBM) principles, underutilized in psychology, can be used to streamline the assessment process and increase the accuracy of conclusions. Using a case example to illustrate the application of each step, this paper outlines a 12-step approach for applying EBM assessment strategies in clinical practice. The initial steps utilize information about clinical base rates, psychopathology risk factors, rating scale scores, and selected in-depth assessment to conduct an iterative, efficient approach to estimating the probability of a given diagnosis until that probability falls into a range suggesting the diagnosis is unlikely to be present, or likely enough to warrant treatment. Once the practitioner and client agree on the treatment plan, subsequent steps monitor progress and outcomes and use that information to make decisions about termination, and then continued monitoring guards against relapse.

A huge amount of research has been conducted since we, as practitioners, completed our training. Tens of thousands of articles are published annually, and even more things compete for our attention if we consider blogs, advertisements, and the news. The problem is that many of the claims are not scientifically valid, and much of the science is not clinically relevant. Perhaps less than 0.25% of the research in most areas of health care will combine scientific validity and clinical relevance (Glasziou, 2006). Who has the time to skim 400 articles to find 1 gem, which may or may not be helpful for the clients we will see this week?

Evidence-Based Medicine (EBM) developed as a philosophy and a set of skills to help manage information overload, so that clinicians can continue to update practices with information to improve client care. EBM is relentlessly pragmatic, using search strategies and critical appraisal tools to find evidence quickly and slash away "hits" that are based on weak designs or will not matter for the client. It is client centered, with the clinician forming answerable questions and looking for evidence to guide decisions about key client issues. The methods have been honed so that updates and searches fit in between seeing clients, or during brief periods as would occur naturally with cancellations and no shows, or perhaps during 30 minutes of regularly scheduled weekly self-improvement (Straus, Glasziou, Richardson, & Haynes, 2011).

Unfortunately, EBM also has developed almost entirely independently from clinical psychology. The original proponents specialized in internal medicine (Sackett, Straus, Richardson, & Rosenberg, 1998), and most of the writings on EBM are oriented towards medicine and nursing (Straus et al., 2011). This is a shame, because EBM has much to offer psychological practice, and psychology also has much to add to EBM (Norcross, Hogan, & Koocher, 2008; Spring, 2007). Adopting these strategies enables clinicians to work more efficiently by streamlining the assessment process. There is an up-front investment of some time to reorganize the assessment process. The reorganization involves identifying reasonable estimates for local base rates, comparing the different assessments available for specific clinical problems, selecting one as the primary measure, and finding or calculating psychometric details that facilitate clinical application of the tools. Many of the most clinically helpful psychometric characteristics are not yet routinely reported in technical manuals or articles, although sufficient information is available to calculate them. The installation process for evidence-based assessment thus involves some focused searches and some one-time calculations to derive the estimates that plug into the assessment process. Once these details are in place, the cost increase and amount of time added per client are

Keywords: evidence-based medicine; evidence-based assessment; diagnosis; sensitivity and specificity; clinical decision-making

1077-7229/12/xxx-xxx$1.00/0

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

negligible (and may actually yield either net savings, or an increase in the reimbursable time). Prior articles have described the evolution of our thinking about the complementary strengths of psychological assessment and EBM, as well as a research agenda (Youngstrom, 2013a; Youngstrom, Jenkins, Jensen-Doss, & Youngstrom, 2012). The goal of this article is to describe 12 steps that integrate EBM ideas with traditional assessment into an evidence-based assessment (EBA) model, walk through the processes of installing the model in an existing clinical practice, and applying its steps to an individual client (see Figure 1). These 12 steps are grounded in EBM's probability-based approach to the assessment process. Before defining the steps, we will first describe the underlying theory.

Base Rates and Probabilities: Foundations of the EBM Diagnostic Approach

The EBM approach to diagnosis focuses on determining the probability of a client's having each diagnosis. In the absence of other information, Meehl (1954) advised "betting the base rate." In other words, if 20% of all of our clients have anxiety, prior to learning anything about a new client, there is a 20% chance that the next client has anxiety. EBM provides strategies for integrating information from risk factors and test results to revise the probability of each diagnosis. Bayes' Theorem lays out the mathematics underpinning this approach. The base rate provides an estimate of the prior probability of a diagnosis (in other words, a "best guess" before gathering additional assessment data), and then combine it with the change in risk attached to a particular assessment finding, estimating the updated posterior probability.

Although Bayesian methods are a bit complicated mathematically, there are now websites and smartphone apps that will do the number crunching (e.g., http://www.ebm.med.ualberta.ca/DiagnosisCalc.html; http://ktclearinghouse.ca/cebm/practise/ca/calculators). EBM also uses a probability nomogram (Figure 2) as a graphical method for synthesizing probabilities and changes in risk. We will use the nomogram as we work through our case example. Interested readers can refer to the "diagnosis" and "risk" chapters in Straus et al. (2011), or a series of commentaries illustrating the methods with psychiatric evaluations (Frazier & Youngstrom, 2006; Youngstrom & Duax, 2005; Youngstrom & Kogos Youngstrom, 2005).

What does the posterior probability represent? One way of thinking about it is as the average probability of a diagnosis for a large number of cases with identical scores

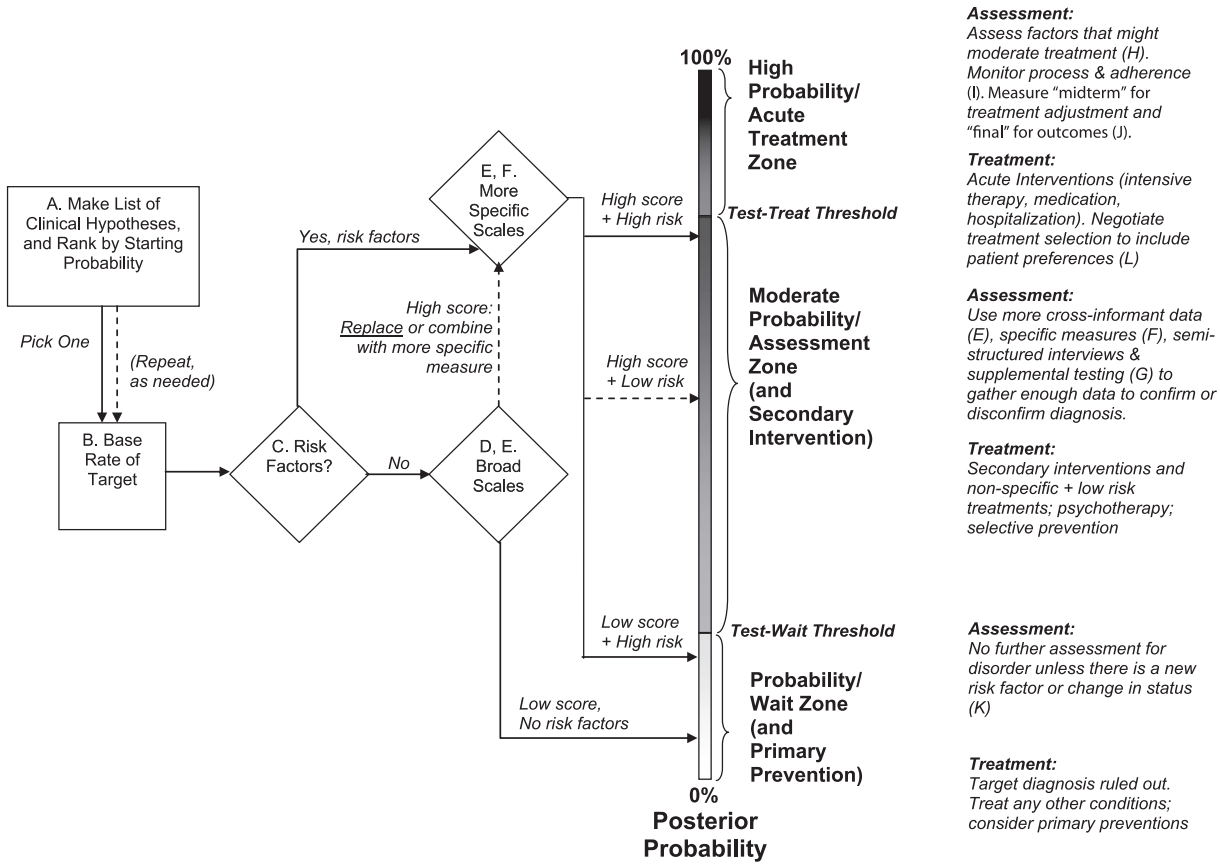

Figure 1. Mapping Assessment Results Onto Clinical Decision Making. *Note.* Letters refer to assessment step in Table 1.

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

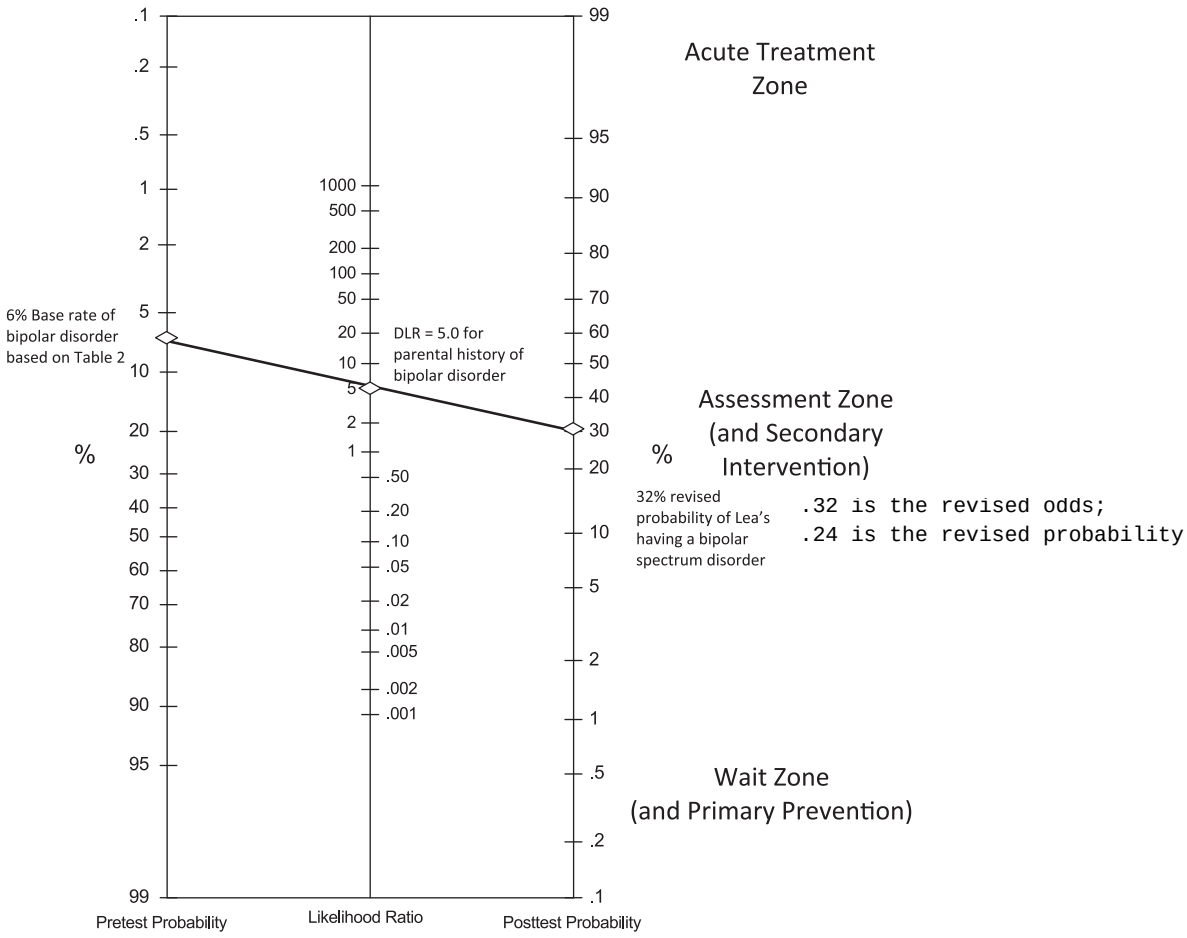

Figure 2. Probability nomogram used to combine prior probability with likelihood ratios to estimate revised, posterior probability. Straus et al. (2011) provide the rationale and medical examples; Jenkins et al. (2011) illustrate applying the nomogram to a case with possible pediatric bipolar disorder, and Frazier & Youngstrom (2006) with possible ADHD; all three sources include nomograms without marking up for example of family history and bipolar disorder.

on risk factors and tests. A posterior probability of 20% means that out of 100 cases presenting to the clinic with a similar set of risk factors and assessment results, 20 would be expected to have the disorder. An alternate way of thinking about the posterior probability is that it indicates the probability that an individual has the diagnosis (Kruschke, 2011). This is akin to reporting the probability of rain in weather reports. It is rare for the weather report to offer such extremes as 0% or 100% chance of rain, and some weather reporting now updates the probability on an hourly basis to incorporate recent data. In the case of a posterior probability of 20%, until we learn more, we can assume that a client has a 20% likelihood of having that disorder.

As with the weather report, clinicians must make decisions about next steps with no guarantees about the assessment results. If the chance of rain is 90%, it is likely enough that we probably will change our plans; and if the probability is less than 10%, many will decide not to bring the umbrella. In EBM, there are two major thresholds along the probability continuum ranging from 0% (definitely does not have the condition) to 100% (definitely does have it). The lower bar is called the Wait-Test Threshold. When the posterior probability falls below the Wait-Test Threshold, then the condition is unlikely enough that we consider it "ruled out," and no further testing is needed unless new information emerges that forces us to revisit diagnoses. In the weather analogy, the 10% probability of rain was below the Wait-Test Threshold, so we consider rain "ruled out" and do not worry about checking the weather report again for that day.

The higher bar is the Test-Treat Threshold. When the posterior probability crosses this threshold, then the probability is high enough that the condition is "ruled in" and becomes a major focus of treatment planning. A 90% chance of rain does not guarantee precipitation, but it is likely enough that it would be foolish not to plan

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

accordingly. Note that there are not fixed locations for the two thresholds. Where to set the bar depends on the costs and benefits associated with the different interventions and outcomes, as well as the client's values and preferences. A 20% chance of rain may not be enough to cancel a jog, but could be a big issue for an outdoor wedding; an 80% chance of ADHD may be sufficient for some cases to start treatment, but others may want more assessment first. Similarly, suicide attempts are so serious that even low risks might change the treatment plan, and the side effects associated with atypical antipsychotics compared to stimulants require different thresholds before initiating treatment.

The middle zone between the two thresholds is the "assessment zone," where more assessment is needed to gather data that revise the probability of the diagnosis either upwards until it crosses the Test-Treat Threshold, or downwards below the Wait-Test Threshold. EBM texts present a simplified scenario where clinicians have three modes of action: treating, testing, and waiting (Straus et al., 2011). We prefer a slightly fuzzier model, where there is always some sort of assessment and treatment available. Above the Test-Treat Threshold, the "Red Zone," assessment switches from a focus on diagnosis to measuring treatment process and outcome variables. In the "Yellow Zone" of midrange probabilities, it often makes sense to start with low risk, broad spectrum treatments (such as psychotherapy) in conjunction with increased assessment to clarify diagnostic questions. In the "Green Zone" of low probability, it could make sense to use prevention programs and low-intensity monitoring for cues of risk.

Steps A and B: Setting the Stage for EBA

Table 1 lays out a dozen steps to reorganize assessment practices to integrate EBM ideas. The first two steps lay a foundation to support a system of evaluation for all subsequent clients. They will take a few hours to fully implement, requiring a start-up investment of time and effort, but then adding no additional time or cost to individual assessments.

A. Identify the Most Common Diagnoses in Our Setting

The first step is to generate a list of the most common presenting problems, referral questions, and clinical diagnoses in our practice. For clinics or providers who use electronic records, it may be possible to query the database to create a summary list. Otherwise, reviewing case notes and assessment reports and manually building a tally would work. For clinics or providers with large caseloads, it may be more feasible to review a randomly selected subset of cases. Note that at this stage we want to track all diagnoses that might influence treatment, not just the primary diagnoses.

After ranking the list, compare it to the clinic's current assessment tool kit. Does the clinic already have good assessment instruments to address the top clinical issues? A clinic does not need a different instrument for every issue. Some measures address a variety of dimensions of emotion and behavior problems (e.g., Achenbach & Rescorla, 2001; Derogatis, 1977). Other issues are rare occurrences, often too infrequent to justify buying an instrument for the occasional case for which it would be helpful. Pareto's 80/20 rule of thumb is a good first approximation: 80% of cases are likely to have 20% of all possible diagnoses (Burr, 1990). Designing the assessment toolkit to do a good job for the most common diagnoses will typically cover the needs of most clients. If comparing the list of clinical targets to the catalog of instruments reveals gaps, then finding an appropriate EBA strategy will help improve evaluations. Looking for challengers to incumbent measures helps refine the assessment battery. If the clinic already uses a broad measure of functioning with multiple scales (e.g., Achenbach & Rescorla), do we need to add another parent or self-report rating scale to assess ADHD? Or depression? Are there tools that show significantly better diagnostic efficiency under clinically realistic conditions? Or greater sensitivity to treatment effects? Alternately, if there is a less expensive or less burdensome product that yields equivalent results, that would also be an evidence-based upgrade. Some measures may fill multiple roles, making it possible to streamline batteries.

B. Benchmark Our Base Rates

Dividing the tally for each diagnosis or clinical issue by the total number of cases reviewed provides a base rate of the clinical target in the local practice. The base rates create context, identifying the common versus rare clinical issues in each clinical setting. Practitioners can leverage even more information from base rates, though, in two ways. First, as discussed above, base rates provide the prior probability estimates for the diagnoses to be assessed. Second, benchmarking them against other data helps calibrate our diagnostic practices. If we work in an outpatient practice, but never diagnose anyone with an eating disorder or autism, we should ask why. Is there something about the local referral pattern that siphons those issues away? In our training clinic at UNC, for example, we see almost no cases with either of these issues because there are specialty clinics that families seek out instead. Often, though, when the local rate of a particular diagnosis is lower than the rates from epidemiological studies, it may indicate a gap in local assessment practices. If epidemiologists would find more ADHD by doing structured interviews with a random sample of people from the phonebook than we would find in our clinic,

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

Table 1
Twelve Steps in Implementing Evidence-Based Assessment and Applying It to Individual Cases

Assessment Step	Rationale	Steps to Put in Practice
A. Identify most common diagnoses in our setting	Planning for the typical issues helps ensure that appropriate assessment tools are available and routinely used	Review practice database, notes, reports; generate "short list" of most common diagnoses and clinical issues
B. Benchmark base rates	Base rate is an important starting point to anchor evaluations and prioritize order of investigation	Select a sample of cases (six months, random draw from past year) and tally local base rate; compare to benchmarks from other practices and published rates; identify any potential mismatches
C. Evaluate risks and moderators	Risk factors raise "index of suspicion," and he combination of multiple risk factors elevate probability into "assessment" or possibly "treatment" zones	Make short checklist of key risk factors; make second list of factors that might change treatment selection or moderate outcome; develop plan for how to routinely assess them
D. Synthesize intake instruments into revised probabilities	Probably already using in practice; upgrading the value for formulation and decision-making by clarifying what the scores mean vis changing probability for common conditions	Make a table crossing assessment instruments with common presenting problems. Identify gaps in coverage. Make cheat sheet with key information about assessment for each application.
E. Interpret cross-informant data patterns	High scores across settings or informants often mean worse pathology; do not over-interpret common patterns.	Gather collateral information to revise case formulation; consider parent, spouse, roommate; also behavioral traces such as Facebook postings. Anticipate typical level of agreement.
F. Add narrow and incremental assessments to clarify diagnoses	Often more specific measures will show better validity, or incremental value supplementing broad measures	Have follow-up tests available and criteria for when they should be used. Organize so that key information is easy to integrate
G. Add necessary intensive methods to finalize diagnoses and formulation	If screening and risk factors put revised probability in the "assessment zone," what are the evidence-based methods to confirm or rule out the diagnosis in question?	Do (semi-)structured interview or review checklist with client to confirm sufficient criteria; supplement with other methods as needed to cross treatment threshold.
H. Finish assessment for treatment planning and goal setting	Rule out general medical conditions, other medications; family functioning, quality of life, personality, school adjustment, comorbidities also must be considered	Develop systematic ways of screening for medical conditions and medication use. Assess family functioning, personality, comorbidity, SES and other potential treatment moderators.
I. Measure processes ("dashboards, quizzes and homework")	Check learning of therapy skills, evidence of early response or need for change in intervention	Track homework, session attendance, life charts, mood check-ins at each visit, medication monitoring, therapy assignments, daily report cards (Weisz et al., 2011).
J. Chart progress and outcome ("midterm and final exams")	Repeat assessment with main severity measures – interview and/or parent report most sensitive to treatment effects; if poor response, revisit diagnoses.	Make cheat sheet with Jacobson & Truax (1991) benchmarks for measures routinely used; track homework, progress on skills; Youth Top Problems (Weisz et al., 2011).
K. Monitor maintenance; relapse warnings	Consolidating treatment gains and planning for maintenance are core features of excellent termination planning, and crucial to long term management of many problems	Develop list of key predictors, recommendations about next action if starting to worsen.
L. Seek and use client preferences	Client beliefs and attitudes influence treatment seeking and engagement, and are vital for balancing risks and benefits.	Assess client concordance with treatment plan; ask about cultural factors that might affect treatment plan and engagement

then we should revisit our assessment methods to make sure that they are sensitive to the diagnosis.

Table 2 pulls together benchmarks from two large recent epidemiological studies, as well as a SAMHSA summary of 2003 Medicaid claims data diagnoses from 13 states (Substance Abuse and Mental Health Services Administration, 2012). The epidemiological benchmarks have the virtue of being based on a consistent interview method and a strong sampling design, enrolling people regardless of whether they were seeking help. The downside of these estimates is that they are likely to be low compared to rates at a clinical setting. For example,

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

Table 2
Benchmarks From Epidemiological Studies and Medicaid Surveillance

Diagnosis or Target Condition	NCS-R All Ages	NCS-R 18-29 Years*	NCS-R 30-44 Years	NCS-R 45-49 Years	NCS-R 60+ Years	NCS-A	SAMHSA Medicaid Data	Rettew et al. (2009) SDI	Rettew clinical
Any Disorder	46%	52%	55%	47%	26%		>99%	–	–
Any Anxiety	**29%**	**30%**	**35%**	**31%**	**15%**	**32%**	–	–	–
Specific Phobia	12%	13%	14%	14%	7%	19%	–	15%	6%
PTSD	7%	6%	8%	9%	3%	5%	–	9%	3%
Generalized Anxiety Disorder	6%	4%	7%	8%	4%	2%	–	10%	5%
Panic Disorder	5%	4%	6%	6%	2%	2%	–	11%	12%
Social Phobia	5%	14%	14%	12%	7%	9%	–	20%	6%
Separation Anxiety	5%	2%	2%	1%	1%	8%	–	18%	8%
Any Impulse Control Disorder	**25%**	**27%**	**23%**	**–**	**–**	**20%**	**–**	**–**	**–**
ODD	9%	10%	8%	–	–	13%	–	38%	37%
Conduct Disorder	9%	11%	8%	–	–	7%	5%	25%	17%
ADHD	8%	8%	8%	–	–	9%	18%	38%	23%
Intermittent Explosive Disorder	5%	7%	6%	5%	2%	–	–	–	–
Any Mood Disorder	**21%**	**21%**	**25%**	**23%**	**12%**	**14%**	**20%**	**–**	**–**
MDD	17%	15%	20%	19%	11%	12%	–	26%	17%
Bipolar I & II	4%	6%	5%	4%	1%	3%	–	–	–
Dysthymia	3%	2%	3%	4%	1%	(included above)	–	8%	10%
Any Substance Abuse Disorder	**15%**	**17%**	**18%**	**15%**	**6%**	**11%**	**53%**	**30%**	**20%**

Note. Statistics adapted from (Kessler, Berglund, Demler, Jin, & Walters, 2005; Merikangas et al., 2010; Substance Abuse and Mental Health Services Administration, 2012).

the rate of ADHD at our clinic is substantially higher than 8% or 9%, as would be true at most clinics. The SAMHSA Medicaid data illustrate how clinical issues tend to be more prevalent in clinical settings, although the comparison is not straightforward because the SAMHSA numbers use unstructured clinical diagnoses, along with confounding poverty and treatment seeking.

Using the epidemiological rates provides a conservative starting point for the EBA approach. Because the rates are likely to be lower than what presents at the clinic, they are essentially taking a skeptical stance and requiring that the assessment process build a case in favor of the diagnosis. If good local data about rates of diagnoses are available, then one would start with these instead. It is important to bear in mind that emotional and behavioral concerns may be intrinsically more difficult to assess than some medical conditions, and variations in clinical conceptualization and training make interrater agreement about diagnoses quite low in clinical practice. EBM authorities sometimes object to using local chart or billing diagnoses in place of rates based on structured diagnostic interviews, because the agreement between them can be low with regard to psychiatric diagnoses (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). However, there are pragmatic advantages to tying the assessment process to the diagnoses already in use at the clinic; and connecting the EBA methodology to existing practices will iteratively refine the local base rates. If a particular condition was under- or overdiagnosed locally, but valid assessment strategies are fed into the decision-making, then they will push the rates to converge over time on the accurate estimates. Incorporation of structured or semi-structured approaches into clinical assessment (Step G) will accelerate the process of building reliable local estimates.

Steps C to H: Assessment of the Client Before Treatment

The next several steps involve gathering and interpreting data regarding symptoms and risk factors in order to determine whether a client falls in the Red, Yellow, or Green zones. Clinicians typically do this intuitively and impressionistically, changing formulations while listening to the client describe the presenting problem and looking at the scores on checklists they completed before starting the session. The EBM approach makes the interpretation more formal and systematic, but not much slower. The clinician still decides what additional information is

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

needed, and what action to take next; the EBM algorithms make the interpretation more accurate, less biased, more consistent, and less prone to distorting effects of cognitive heuristics (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011; Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012) that otherwise assail clinical decision-making (Galanter & Patel, 2005), just as they beset any complex mental activity (Gigerenzer & Goldstein, 1996). We will use a case example to illustrate these steps. Our client, "Lea," was a White 18-year-old female who presented with concerns about difficulties maintaining attention at school and high stress levels.

C. Evaluate Relevant Risk and Moderating Factors

Table 2 suggests that, based on epidemiological rates, the most common clinical issues in our client's age range are anxiety disorders (affecting 30% of 18- to 29-year-olds), impulse control disorders (27%), mood disorders (21%) and substance misuse (17% of the general population). The most frequently occurring diagnoses are major depressive disorder, social and specific phobia, conduct disorder, ADHD, and ODD. Therefore, anxiety, impulse control, mood, and substance issues all start in the "Assessment Zone" – the probability is too high to ignore, but not enough to indicate treatment (see Figure 1). Helpful tests and findings will be able to rule these issues out for the majority of unaffected cases, and ideally they would raise the index of suspicion to prompt more inquiry for affected cases. Other diagnoses, such as PTSD, might not need routine screening for a client in this age range, but would move into the Assessment Zone if there were other risk factors or cues.

Family history of mood disorders is the most well–established risk factor for mood disorders (Tsuchiya, Byrne, & Mortensen, 2003), and it also increases risk of anxiety, ADHD, and substance misuse (Hodgins, Faucher, Zarac, & Ellenbogen, 2002). Family history of substance misuse also increases the risk of substance misuse in the client. When there are known risk factors that at least double the odds of the client's having the condition, then they are worth asking about during the intake interview or when checking responses on a checklist. Statistically significant relationships linked with smaller changes in odds are unlikely to have clinically meaningful impact on decision-making about individuals (Straus et al., 2011). The Family Index of Risk for Mood (FIRM) is an example of a brief checklist designed to be a feasible tool for gathering family history at the beginning of an evaluation (Algorta et al., 2013).

Lea reported that one of her biological parents has a history of bipolar disorder. Bipolar disorder in a first-degree relative increases the risk by a factor of 5. Combining the change in likelihood due to family history with the base rate of bipolar disorder of 6% (Table 2) leads to a revised probability of 32% of Lea having a bipolar disorder; the nomogram (Figure 2) shows the posterior probability.

D. Synthesize Intake Instruments Into Revised Probability Estimates

Our clinic uses the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001) to assess internalizing problems (relevant to both the anxiety and mood clusters), externalizing problems (relevant to the impulse control), and more specific scales providing information about potential attention problems, social problems and other clinical syndromes. The Achenbach instruments do not include a separate substance use scale, but they do have items embedded within the "Rule Breaking Behavior" scale that ask about "drinks alcohol" (#2), "uses tobacco" (#99), and "uses drugs" (#105). The clinician can check the score on these items separately in order to evaluate substance use.

The client in question completed the ASEBA Youth Self Report (YSR). Her *T* scores were 73 for Internalizing Problems, 61 for Externalizing Problems, and 78 for Attention Problems. In order to illustrate the discrepancy between estimates based on clinical judgment versus estimates derived mathematically from the EBM approach, take a moment to write down your best guess of the probabilities of each of the client's potential diagnoses. Specifically, before you examine the EBM approach for this case outlined in Table 3, write down your estimate, from 0% to 100%, of the probability that the client has ADHD based on the presenting problem and this profile of scores. Next, jot down estimates for major depression and anxiety, and also note any other disorders that you think are likely. We will compare these initial impressions with the results from the EBA approach to give a sense of whether these methods might alter your decision-making in practice (see Table 3).

Not all of these scores will be equally informative about different clinical hypotheses. A major contribution of research studies is that they quantify the diagnostic validity of different assessment methods or informants, as well as determine the incremental validity of combining multiple methods. Self-report on an instrument like the YSR has good diagnostic validity for internalizing problems and weak validity for attention problems (Pelham, Fabiano, & Massetti, 2005). For our client, we used the parent-report ASEBA Child Behavior Checklist (CBC) as a way of quickly gathering complementary information about attention problems (see Table 3).

How does one find the best tests to use for a particular question? A review of the literature on the YSR, using Google Scholar or PubMed with the search terms "Achenbach" AND ("sensitivity and specificity") AND ("anxiety" OR "depression") helps limit the results to those likely to be relevant to the immediate question

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

Table 3
Scores and Interpretive Information for Applying EBA Approach to Lea (18-year-old White female presenting to an outpatient clinic)

		Broad Measure (Step D)			*Cross-Informant (Step E)*				*Treatment Phase*		
Common Diagnostic Hypotheses (Step A)	*Starting Probability (Step B)*	*Scale & Score*	*DLR (Source)*	*Revised Probability*	*Next Test Score*	*DLR (Source)*	*Revised Probability* [b]	*Confirmation (Step G)*	*Process (Step I)*	*Outcome (Step J)*	*Maintenance (Step K)*
Depression	21%	YSR *T* Internal: 73	2.43 (local data)	39%	CBC Internal Raw: 14	0.90 (E. A. Youngstrom, 2013b)	37%	MINI (Sheehan et al., 1998): Major Depressive Episode	Youth Top Problems (Weisz et al., 2011)	Beck Depression Inventory (Beck & Steer, 1987)	Worsening of mood or energy symptoms
Hypomania/ Mania	32% [a]	YSR *T* External: 61	1.15 (Youngstrom et al., 2004)	37%	CBC *T* External 56	0.53 (Youngstrom et al., 2004)	16%	MINI: Hypomanic Episode➔ Bipolar II	Smartphone mood app		" "
ADHD	8%	YSR *T* Attention Probs: 78	1.36 (local data)	11%	CBC *T* Attention Probs: 70	2.19 (local data)	21% [c]	MINI: ADHD Predominantly Inattentive Type	CAARS	CAARS	Monitor schoolwork completion rate
Anxiety	29%	YSR *T* Internal: 73	2.35 (Van Meter et al., under review)	49%	CBC *T* Internal 63	0.98 (Van Meter et al., under review)	48%	–	–	Not a primary focus	–
Substance Issues	15%	YSR #2: 0 YSR #99: 2 YSR #105: 1.5	3.4 (local data)	37%	CBC #2: 0 CBC #99: 1 CBC #105: 1 (marijuana)	5.6 (local data)	77%	MINI: Substance Abuse – past cannabis and Xanax™ abuse	Check in at therapy sessions	Not agreed as a treatment goal	Contact therapist if usage back at prior level

Note. Steps H (finish treatment planning and goal setting) and L (seek and use client preferences) are discussed in text though not mentioned in Table 3.

[a] Our starting probability was based on the prevalence of bipolar spectrum disorder in the NCS in Lea's age range (6%, see Table 2), then adjusted for the history of bipolar disorder in a first degree relative (DLR = 5.0), resulting in a revised probability of 32% (see marked up nomogram in Figure 2; Step C).

[b] Readers can compare their impressions based on the presenting problem and test scores with the EBA estimates in this column. The estimates often are different, but the EBA approach is much more consistent across sets of clinicians as well as often being less biased (Jenkins et al., 2011).

[c] We could replace the CBC and YSR with the CAARS scores, as the CAARS provides more coverage of ADHD symptoms, and more information about severity (Step F). Van Voorhees et al. (2011) report that the combination of CAARS $T > 65$ from both self and observer had a DLR 2.6 for the inattentive subtype. Combining the initial base rate estimate of 8% for ADHD with a DLR of 2.6 yields a revised probability of 18%, essentially confirming the estimate of 21% obtained via the CBC and YSR.

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

(Straus et al., 2011). If the search still produces a large number of hits, then crossing it with "review" may help find relevant summaries. The goal is to find articles that evaluate the utility of the measure while providing details for applying results to individual cases. The information will most likely be in the form of diagnostic sensitivity and specificity, but diagnostic likelihood ratios (DLRs) make it even easier to integrate the results. If the clinic uses a standard intake battery, then it is efficient to do searches for articles applying the core measures to the most common diagnoses and issues. This builds a reference sheet of DLRs for applying the main tools to the most frequent topics. Because finding the DLRs can be the most labor-intensive part of the process, concentrating on the regular measures and presenting problems produces the greatest return on investment and makes it much easier to implement the EBA approach in real time.

Using this search strategy, we identified review articles that evaluated the Achenbach instruments for a variety of diagnoses (Warnick, Bracken, & Kasl, 2008), and which also compared several different self-report measures for ADHD ratings (Taylor, Deb, & Unwin, 2011). Receiver operating characteristic (ROC) analyses are an exceptionally useful way of evaluating assessments for the purpose of helping with diagnosis (McFall & Treat, 1999), and the area under the curve (AUC) from these analyses provides a summary of the discriminative power of the test. Most software estimates the AUC by plotting the sensitivity of the test (i.e., out of 100 cases that have the diagnosis, how many does the assessment classify correctly?) versus the false alarm rate (i.e., out of 100 cases without the diagnosis, how many does the assessment misclassify?). AUC values of .50 reflect chance discrimination, and 1.00 would be perfect.

Once we have identified good instruments for our assessment purposes and have our client complete them, how do we translate their scores into revised probability estimates? If the test has norms, then it may provide a percentile rank for a score, but this is not the same thing as the probability that a person has a particular diagnosis. Some studies report the diagnostic sensitivity and specificity attached to a set cut score. If the case scores above that threshold, then they "tested positive" for the diagnosis; scores below the threshold would be negative test results. Again, these are not the same thing as establishing the diagnosis. Not everyone who scores high on an attention problems scale has ADHD; attention problems are associated with stress, hypomania, substance use, and a variety of other conditions besides ADHD. These will sometimes produce false alarms on a measure designed to catch ADHD. Conversely, not all of the negative results will be accurate: Cases that truly have ADHD might score below threshold, perhaps because it is mild, or the person has taken a stimulant, or drunk a lot of coffee. . . .

The EBM approach repackages the diagnostic sensitivity and specificity, making it possible to combine the information from a test with the prior probability of the diagnosis to generate a new, revised probability. As with the family history, one can either use a probability nomogram, an online calculator, or an app to synthesize the information. When combining multiple tests or findings, the posterior probability from the previous step becomes the new starting probability entered into the nomogram or calculator. The DLR corresponding to the appropriate score gets entered next, generating an updated probability. The DLR on the middle line comes from comparing two percentages: the percentage of cases with the disorder scoring in this range (i.e., diagnostic sensitivity for a positive test result) divided by the percentage of cases without the disorder that also score in the same range (i.e., 1 minus specificity for a positive test). If high scores on a test are common among those with ADHD, but rare among those that have other conditions, then the test result indicates that the probability of an ADHD diagnosis should increase substantially. If the score is rare among cases with ADHD, but common among other cases, then the assessment helps rule out ADHD as a concern. If roughly equal percentages of cases with ADHD and without ADHD score in the same range, then the assessment result is ambiguous and does not change the probability estimate much. Some rules of thumb about the diagnostic likelihood ratios are that values greater than 10 are powerful enough to raise a 50% prior probability to a revised estimate greater than 90%; and values of 2 to 7 are helpful. The inverse ratios are equally powerful at reducing the revised probability: a DLR of 0.10 changes a 50% probability to less than 10%, for example. DLRs close to 1.0 mean that the result does not add new information.

Table 3 lays out the short list of common diagnostic issues, what we selected as a starting base rate, and the DLR that we used for the YSR for our case. The "revised probability" column provides the posterior probabilities, estimated via calculator. Readers can use the nomogram or a calculator to combine the base rate and DLR and then check their answer against the tabled value. The posterior values reveal how the probability of the different diagnoses changes based on the YSR. Note that different YSR scores are most relevant for different issues, with the empirical literature guiding the selection of each. Based on Lea's YSR responses, the leading hypotheses so far are depression (possibly on the bipolar spectrum, due to her family history), anxiety, and substance issues. Posterior probabilities based on the YSR for hypomania/mania and ADHD do not significantly differ from prior probabilities.

E. Interpret Cross-Informant Patterns

Particularly when working with youths or with couples, clinicians often gather checklists from multiple

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

informants. Scores on cross-informant scales can be integrated in the same way as any other assessment information – pick the most relevant score for each diagnostic issue based on the literature, find the DLR attached to the observed score, and then synthesize this information with the current probability based on prior information. The optimal choice of scale on the CBC need not be the same as on the YSR, due to differences in typical validity of each person's perspective for varying topics. If multiple different pieces of information are simultaneously available, such as having the YSR, CBC, and risk factor information all together at intake, then the DLRs can be multiplied to make a single combined DLR to use. The algebra works out to be the same whether these are combined as a product of DLRs versus iterating sequentially through the nomogram.

Technically, Bayes' Theorem assumes that the inputs are independent, i.e., that CBC and YSR scores are not correlated. In practice, correlations of the magnitude typically seen across informants ($r \sim .20$ to .35) can be treated as independent without distorting the accuracy of the predictions substantially. The correlation between scores becomes problematic when asking the same informant to complete multiple rating scales, as these often correlate $r \sim .5$ or higher, even when putatively measuring different constructs. For example, the Externalizing and Internalizing scores on the CBC correlate with each other $r = .54$ in the standardization data (Achenbach & Rescorla, 2001). To avoid bias, take the single best score from each informant for each purpose, and only use that in the nomogram or calculator.

Table 3 shows which CBC scores are most relevant for each clinical hypothesis, along with a citation for the source of the DLR values. Adding the CBC scores made substance issues the leading clinical hypothesis for this case, with a posterior probability well above our Wait-Test Threshold. Anxiety, depression, bipolar, and ADHD are also still contenders. Although the revised probabilities for these disorders are below 50%, they are still above our Wait-Test Threshold; thus, we want to gather more data about them during the next steps in the assessment process.

F. Add Narrow and Incremental Assessments to Clarify Diagnoses

We decided to ask Lea and her parent to both fill out the Long Version of the Conners (2008) Adult ADHD Rating Scale (CAARS) to gather more detail about attention problems. Lea reported extremely high levels of concern on the Inattention/Memory problems scale ($T = 80$), and her parent confirmed these on the Observer form ($T = 75$). A recent paper evaluated the sensitivity and specificity of the CAARS in an outpatient clinical sample and found that a rule of having both the self-report T-score >65 and observer T >65 had a sensitivity of 58% and a specificity of 78% (Van Voorhees, Hardy, & Kollins, 2011). This translates into a DLR of 2.55. We could choose to substitute the CAARS parent and youth scores for the CBC and YSR for the purposes of estimating the ADHD probability; in Lea's case the CAARS scores generated a similar probability estimate of 18% versus the 21% based on the Achenbach (see Table 3).

G. Add Necessary Intensive Methods to Finalize Diagnoses and Formulation

The next step in the EBM assessment process is to deploy tools that are specific enough to confirm a diagnosis, raising the probability above the Test-Treat Threshold. For formal diagnoses of disorders, a structured or semistructured diagnostic interview is often the method of choice. The structured component ensures systematic inquiry about all relevant symptoms and impairment, as well as consideration of exclusion criteria. The "semistructured" provides more latitude about phrasing and probing, making it possible to adapt the language to better engage the client, and to verify that the information provided constitutes clinically meaningful symptoms rather than developmentally or culturally appropriate phenomena. Clinicians tend to prefer semistructured methods over fully structured ones because they allow greater autonomy; but conversely, these methods also risk lowering the interrater reliability to the extent that clinicians interpret and query content in idiosyncratic ways. Structured interviews also can be faster.

Why not skip the earlier steps and start directly with a structured or semistructured interview? First, the preceding steps do not take more than a few minutes of the clinician's time once the infrastructure is in place, and they provide valuable context for interpreting symptoms. The best semistructured interviews include open-ended sections asking about the developmental history and history of the presenting problem to help generate a macro-level view of the problems. Without the big picture, it is difficult to know whether a nonspecific symptom, such as irritability, is more indicative of anxiety, mood disorder, a response to trauma, or a proclivity towards instrumental aggression and antisocial behavior. Second, having the big picture painted by screening tools and context will guide choices about semistructured approaches. There are more than 360 diagnoses in the DSM-5 (American Psychiatric Association, 2013) and ICD (World Health Organization, 1992), and no interview includes all of them. It would be prohibitively cumbersome to try. Thus, there will be gaps in coverage. Furthermore, certain tools

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

are recognized to be "best in class" for particular clusters of diagnoses. For example, experts generally pick the Schedule for Affective Disorders and Schizophrenia for Children and Adolescents (Kaufman et al., 1997) for mood disorders in youths, and the ADIS for anxiety disorders. Significant concerns about a personality disorder based on the preceding assessment steps would warrant an additional module or entirely different semistructured interview. Third, even semistructured interviews are neither completely reliable nor perfectly valid (Kraemer, 1992). If the results of the interview are consistent with the findings from checklists and risk factors, they deserve greater confidence than they would get in the absence of support or in the face of countervailing evidence. Fourth, most semistructured interviews do not capture information about the severity of problems, and so they require supplementation with other rating scales to establish a current baseline against which to define treatment goals and measure progress. Fifth, integration of the semistructured interview and other assessment data makes a hybrid model possible, wherein the clinician selects specific modules to confirm or disconfirm indicated competing hypotheses (Ebesutani, Bernstein, Chorpita, & Weisz, 2012). Using broad measures and an awareness of the base rates at a particular setting avoids the traps of confirmation bias and search satisficing, two cognitive heuristics that bedevil unstructured clinical case formulation. Selecting modules allows some abbreviation of the structured component, reducing burden for the clinician and client. Basing selection on prior assessment results also establishes medical necessity, increasing the likelihood that a third-party payer would reimburse for the service. Adding these methods as part of the assessment process also improves the reliability of diagnoses, speeding up the process of refining local estimates of base rates. These in turn can provide updated baseline probability estimates for future cases, iteratively improving the whole approach.

Based on the combination of mood, attention problems, and substance issues as leading candidates for Lea, we selected to do the MINI (Sheehan et al., 1998). The MINI was designed to be faster than competing interviews. We opted for the basic version, which covers 19 disorders, augmenting with the Attention Problems module from the MINI-Plus based on the elevated concerns from her CBC. Because we were using the MINI clinically and not for research purposes, we were willing to rephrase questions and probe if Lea was unclear in her responses, essentially shifting to a semistructured format if needed.

During the MINI interview, Lea endorsed symptoms and impairment consistent with both a major depressive episode and a hypomanic episode. This combination meets criteria for bipolar II disorder. The MINI also confirmed the presence of sufficient symptoms outside the context of a mood episode to also meet criteria for ADHD, predominantly inattentive type. Exploration of substance use issues during the MINI interview identified past abuse of Xanax and cannabis, but Lea denied current concerns about her usage (i.e., reported infrequent and nonimpairing use). The MINI also includes a detailed assessment of suicidal ideation and behavior, consistent with best practices for assessing and documenting suicidal ideation and behavior (Cukrowicz, Wingate, Driscoll, & Joiner, 2004), especially when working with depression or bipolar II, which are particularly strong risk factors for suicidality (Berk & Dodd, 2005; Cukrowicz et al., 2004). Lea described some past instances of self-injurious behaviors, including bouts of cutting that required stitches on at least one occasion, but she denied intent to kill herself or harm others.

H. Finish Assessment for Treatment Planning and Goal Setting

Based on the assessment results so far, the treatment plan should include goals for addressing mood and attention problems as well as monitoring substance use and self-injury. Before plunging in to treatment, it is important to consider the context of the client's symptoms. For example, other possible medical conditions, medication usage, and physical conditions (including what used to be called "Axis III" conditions in DSM-IV) could contribute to the client's stress and functioning, as well as treatment outcomes. Similarly, clinicians want to have a good sense of environmental factors that might change the formulation or moderate treatment selection. It might be helpful to have a short checklist of key things to address routinely with clients, which could be tailored for the common issues at the clinic. Using a checklist will provide big benefits in terms of consistency and coverage (Gawande, 2010), and could be even more important with the discontinuation of the multi-axial system in DSM-5, as there may be fewer cues to assess physical conditions.

A second objective at this phase of assessment is to quantify the severity of problems and establish some goals for progress and outcome. The CAARS indicates the current severity of Lea's attention problems, including a separate score for the severity of her primary presenting complaint of inattention. The CBC and YSR also provide some sense of severity, but the 6-month rating period in the instructions may make them less sensitive to treatment effects, and they were not designed to be repeated often. We asked Lea to complete a Beck Depression Inventory (Beck & Steer, 1987) to gauge the current severity of her distress and depressive symptoms. She scored a 29, which is often considered in the "severely depressed" range.

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

A third objective is to assess overall functioning and quality of life, so that treatment also can aim to improve functioning and not solely focus on symptom reduction. Axis V in DSM-IV provided a simple metric for summarizing this on a 1 to 100 scale, with higher scores reflecting better functioning. We rated Lea's functioning a 55, reflecting: (a) moderate to severe symptom levels, (b) some problems at school and with her family, (c) a moderately strong social support network, and (d) above average academic performance, despite her stress.

L. Solicit and Integrate Client Preferences

Engagement is a crucial ingredient for a successful evaluation or course of treatment. The best recommendations do no good if the client does not want to hear them, and the best therapy does no good if the client does not show. As we develop the case formulation and recommendations, we want to gather information about values, beliefs, preferences, and prior experiences that may change receptivity to the formulation and treatment plan. Does the person perceive the behavior as a problem? Have they been in therapy before? Or taken medication? What did they like about it? What did they dislike? Are there religious or cultural factors that would reinforce some approaches or conflict with others? Issues of cultural sensitivity (Sue, 1998), as well as attention to readiness to change (Prochaska, 2000), come to the fore. Although this step is listed last in the table, it is important to note that it permeates the assessment cycle.

As mentioned above, the EBM model incorporates the client beliefs and values by negotiating where to put the Wait-Test and Test-Treat thresholds (Straus et al., 2011). Done well, this becomes a conversation between the clinician and client: Here is the current probability of this particular problem based on the available information. Is the probability high enough to agree that this should be a focus of treatment? If not, then more assessment still is needed about that issue. Is the probability low enough to consider the issue ruled out? If so, then we can attend to other issues in assessment and treatment. This approach empowers the client by giving them a say in shaping both the assessment package and the treatment planning.

It also is possible to make more formal yet individualized probability estimates about the risks and benefits of different approaches to treatment. EBM uses the "Number Needed to Treat" (NNT) as an effect size for re-expressing treatment outcomes as a probability estimate, and the "Number Needed to Harm" (NNH) as a similar format for conveying risks. It is possible to combine these two into a "Likelihood of Help versus Harm" (LHH), which in turn can factor in client preferences about different risks and benefits (Straus et al., 2011). Other approaches allow direct estimation of the probability of a good outcome for different treatments, helping weigh potential moderating factors (Beidas et al., 2013; Lindhiem, Kolko, & Cheng, 2012).

Though Lea reported feeling initially overwhelmed by the assessment findings and diagnoses assigned, she agreed that the assessment findings indicated a significant mood problem. She was not happy with the word "bipolar," because of the stigma surrounding the diagnosis. Lea was much more amenable to framing it as a type of depression that involved more mood swings versus feeling down all the time, and she was open to the idea that she had a form of depression that often responds differently to some types of treatment. She agreed that the assessment provided an adequate basis for proceeding with treatment for the mood issues. On the other hand, she did not see her substance use as problematic, and she did not want it to be a focus of treatment. She was agreeable to the idea of monitoring her substance use, and she agreed to discuss with her treatment provider any future increases in use or interference with functioning.

Assessment During and After Treatment

We were doing a comprehensive psychological evaluation with Lea, and not providing treatment, so this case example ends here. However, there are ways to lay a good foundation for assessment during active treatment and maintenance in the recommendations section of assessment reports. At our clinic, we provide a "care package" of EBA methods that the client can take with them, or that we can send directly to their other care providers.

I. Measure Progress and Process

After the initial treatment plan is clear, then the role of assessment shifts from diagnostic clarification to measuring progress. Borrowing the metaphor of therapy as a journey, the prior stages of assessment are akin to getting oriented, picking a mutually agreeable destination, and then charting a route that should reach the goal reasonably directly. Once under way, assessment becomes the dashboard for monitoring changes and alerting to critical events.

Progress measures need to be brief enough and easy to use so that the client will tolerate repetition. Using dieting as a concrete example, a person might use intensive assessment, with skin calipers or a water immersion tank, to determine baseline body composition and establish some desired goals, but they would not repeat these on a daily or weekly basis to measure progress. Stepping on a bathroom scale is less informative and accurate, but highly feasible. Repeating it regularly also still provides helpful information about general trends, especially as more data points become available and smooth out the daily fluctuations.

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

What would be psychotherapy analogs to the bathroom scale? If doing cognitive behavioral therapy or other skills-based interventions, then completion rates for homework assignments are a behavioral indicator of engagement as well as content learning. Working with anxiety disorders, subjective units of distress (SUDs) ratings are a common in-session measure of distress. With ADHD, daily report cards would be another example (Pelham et al., 2005). For Lea, a life charting smartphone app could be a convenient way of tracking mood and energy daily. The Youth Top Problems is another brief, practical method for defining the youth's primary concerns, and then monitoring them over the course of treatment (Weisz et al., 2011). The Youth Top Problems readily accommodates inattention as a concern, again keeping a focus on one of Lea's primary presenting complaints.

At a minimum, process assessment should involve asking about what the client and therapist have agreed are the primary problems each session using a consistent scale, whether it be 1 to 10, 1 to 100, or whatever, and then writing it down. Each case should also have a short list of other key things to ask about regularly, such as changes in suicidal ideation, increases in drug or alcohol consumption, or emergence of side effects due to medications. The exact list will vary by client, but having a written list increases the odds that clinicians will remember to ask consistently. The therapeutic process also generates a lot of "meta-data" of processes influencing therapy outcomes: late arrivals, cancellations, and "no shows" all are data that can inform about motivation, conscientiousness (Barnett et al., 2011), or therapy alliance. The power of ongoing assessment emerges from measuring a few key things consistently, often, and writing them down. Doing therapy without ongoing measures would be as silly as attempting to diet without stepping on a scale: Progress could still happen, but it is much less likely, and it would take longer to recognize. Similarly, a dieter who tracks "process" information like exercise and food intake might make more progress or be better able to identify why progress is not being made.

J. Chart Progress and Outcome

The next role of assessment is to chart progress and measure outcomes. Having a method of reviewing trends in the progress and process measures helps to illustrate trajectories of change. Functional behavior analysis relies heavily on charts as a way of visualizing change, and Excel and GoogleDocs make it easy to build line graphs of SUDs ratings or daily report cards.

Other assessment strategies also may be helpful for defining end goals of therapy. In research, "loss of diagnosis" is one common operationalization, which could be measured by repeating a semistructured interview. Repeated interviewing would rarely be done in practice; it feels cumbersome and unnecessary. Yet going with a completely informal and impressionistic approach may lose reliability and precision (Christon, McLeod, & Jensen-Doss, this issue). Jacobson and colleagues developed a psychometrically informed framework for evaluating clinically significant change that was intended to be sophisticated yet more practical than a repeated interview. There are two parts to their definition: (a) reliable change, and (b) moving past a benchmark defined by comparisons with norms for clinical and nonclinical reference groups (Jacobson, Roberts, Berns, & McGlinchey, 1999). Reliable change is driven by the precision of the measure; Jacobson suggested dividing the observed change by the standard error of the difference to create a Reliable Change Index (RCI).

Jacobson et al. defined three possible thresholds for clinically significant change, which the mnemonic "ABC" can help us to remember (Youngstrom & Frazier, 2013): moving *Away* from the clinical distribution (defined as scoring two standard deviations better than the average for a sample with the target condition), *Back* into the normal range (defined as scoring within two standard deviations of the nonclinical average), or *Crossing* closer to the nonclinical distribution (moving past the weighted mean, pooling the means and standard deviations of the clinical and nonclinical groups). Similar to the DLRs, researchers can publish these, and clinicians can gather them for widely used instruments. These become mileposts for treatment.

Many of the rating scales and checklists would be best suited for occasional use. Making an analogy to teaching, these are more like "midterm" and "final exams" that assess accumulated learning rather than the briefer "quizzes" described in the previous step to assess learning along the way. The ideal "exam" instrument is short enough that it could be repeated several times without becoming burdensome, but precise enough to be sensitive to early treatment response. Clinicians want to be able to make course corrections if treatment is not moving in the right direction (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, Hansen, & Finch, 2001). Unlike personal progress measures, these "exams" also can use normative data to make nomothetic comparisons. Measures such as the BDI may hit the sweet spot, combining reasonable brevity with sensitivity to treatment effects. Based on data in the technical manual, reductions of 8 points would be 90% likely to reflect reliable change, and 10 points would be 95% likely. The *Back*, *Crossing closer*, and *Away* thresholds would be scores of 22, 14, and 4 (Beck & Steer, 1987). Lea scored a 29 on the BDI during her evaluation. The recommendations in the assessment report could suggest repeating the BDI four to six sessions into treatment as a "midterm" or progress check, with a reduction of 8 points providing strong

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

evidence of treatment response. Passing the threshold of 22 points could be an early benchmark, and reaching the more ambitious threshold of 14 points could be a longer term treatment goal.

Failure to make anticipated progress should trigger at least two things: a frank discussion about engagement and congruence of goals, and also a review of formulation and serious consideration of whether there is some additional hypothesis that needs to be considered. Framing treatment goals using the Jacobson approach also sends the message that perfection is not the goal: People living without depression still experience stress in their daily lives, and people without ADHD still have trouble concentrating or are sometimes forgetful.

K. Monitor Maintenance and Relapse

When treatment is successful, then the focus can shift to planning for how to maintain the gains and how to recognize cues of relapse. As anyone who has tried dieting knows well, consolidating the success and maintaining it over the long term are also challenging, and they require a plan. Some of the progress measures may serve well in this role, too. If the person is using a smartphone application to chart their mood and energy, it may well become a habit after several weeks or months of treatment. As a convenient habit, it would offer an easy way of continuing to monitor the situation and recognize when things were worsening again.

Another approach would be to make a short list of critical events or behaviors. For Lea, the list might include keeping track of risky behaviors associated with bipolar disorder such as binge drinking, having unprotected sex, and staying up all night. In isolation, none of these is desirable, but none clearly signifies a relapse into a mood episode, either. However, two of those events would be concerning, and three could be a good sign that seeking treatment would be helpful. It is important to collaboratively define a set of personalized warning signs—ahead of time and while the person is functioning relatively well—that they agree to heed later (Newman, Leahy, Beck, Reilly-Harrington, & Gyulai, 2002). This may be particularly important when working with bipolar disorder, ADHD, or substance issues, where insight into one's behavior is more likely to be compromised precisely when things are starting to get worse and when early intervention would be most beneficial. Having a clear "game plan" written down for managing transitions and addressing signs of relapse increases the chance of successful implementation.

Discussion

The evidence-based assessment model we describe integrates many different sources of information, including the types of problems seen at different clinical settings, and information about risk factors and test results, to provide updated probability estimates of different clinical hypotheses. The same assessment approaches also can provide information about treatment targets and define intermediate and outcome goals. Changes in technology also make it possible to use innovative tools, such as smartphone applications, to enhance the measurement of progress in a more fine-grained way.

What is surprising is that this approach need not add a lot of time or expense to clinical work. Much of the power of the methods comes from laying a foundation, selecting good tools for each hypothesis, and optimizing the sequence of assessments. EBA involves working smarter, not harder: After the initial investment of gathering the measures and making the "cheat sheets" with the key information, the approach adds less than 5 minutes to the evaluation time for most cases, and less than $5 in expense (if choosing to use a typical smartphone application; Youngstrom et al., 2012). The probabilistic framework underpinning the approach feels different than what most of us were taught in graduate school, yet also familiar inasmuch as clinical practice approximates detective work, building a case in favor of, or against, different hypotheses about formulation and treatment.

The reality of clinical practice is constantly challenging. Real cases, like Lea, do not fit neatly into research boxes. Lea is entering "emerging adulthood," still in school, becoming increasingly independent. Which assessments are most age-appropriate? Which norms make sense to use? What are treatment goals that would engage Lea and motivate her to continue in therapy? All of these decisions require reflection and skill from the practitioner (Schon, 1983). However, they are not insurmountable hurdles, and they often can be navigated in a principled way (Straus et al., 2011).

Adopting the EBA approach improves practice by revealing gaps in coverage, where the assessment toolkit needs upgrading to address common referral issues, as well as highlighting redundancy or points of obsolescence in the typical battery. It also illustrates ways that assessment and treatment can be interwoven all the way through treatment termination and maintenance, promoting long-term success. Unfortunately there is not yet an "off-the-shelf" assessment battery that will work in all settings, but there are now principles that can guide customization to tailor a battery to suit each of our practices. EBA at its core is relentlessly focused on the individual client – identifying their needs, crafting an intervention with the highest probability of success, and then reaching measurable goals. If we were the client, would we want anything less?

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont.

Algorta, G. P., Youngstrom, E. A., Phelps, J., Jenkins, M. M., Youngstrom, J. K., & Findling, R. L. (2013). An inexpensive family index of risk for mood issues improves identification

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

of pediatric bipolar disorder. *Psychological Assessment*, *25*, 12–22. http://dx.doi.org/10.1037/a0029225

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Barnett, J. H., Huang, J., Perlis, R. H., Young, M. M., Rosenbaum, J. F., Nierenberg, A. A., . . . Smoller, J. W. (2011). Personality and bipolar disorder: Dissecting state and trait associations between mood and personality. *Psychological Medicine*, *41*, 1593–1604. http://dx.doi.org/10.1017/s0033291710002333

Beck, A. T., & Steer, R. A. (1987). *Beck depression inventory manual*. San Antonio, TX: Psychological Corporation.

Beidas, R. S., Lindhiem, O., Brodman, D. M., Swan, A., Carper, M., Cummings, C., . . . Sherrill, J. (2013). A probabilistic and individualized approach for predicting treatment gains: An extension and application to anxiety disordered youth. *Behavior Therapy*. http://dx.doi.org/10.1016/j.beth.2013.05.001

Berk, M., & Dodd, S. (2005). Bipolar II disorder: a review. *Bipolar Disorders*, *7*, 11–21.

Burr, J. T. (1990). The Tools of Quality, Part VI: Pareto Charts. *Quality Progress*, *23*, 59–61.

Christon, L.M., McLeod, B.D., & Jensen-Doss, A. (this issue). Evidence-based assessment meets evidence-based treatment: A science-informed approach to case conceptualization. *Cognitive and Behavioral Practice*.

Conners, C. K. (2008). *Conners, 3rd Edition*. North Tonawanda, NY: Multi-Health Systems.

Cukrowicz, K. C., Wingate, L. R., Driscoll, K. A., & JoinerJr., T. E. (2004). A standard of care for the assessment of suicide risk and associated treatment: The Florida State University psychology clinic as an example. *Journal of Contemporary Psychotherapy*, *34*, 87–100.

Derogatis, L. (1977). *SCL-90: Administration, scoring, and procedures manual for the R(evised) version*. Baltimore, MD: Johns Hopkins University School of Medicine.

Ebesutani, C., Bernstein, A., Chorpita, B. F., & Weisz, J. R. (2012). A transportable assessment protocol for prescribing youth psychosocial treatments in real-world settings: Reducing assessment burden via self-report scales. *Psychological Assessment*, *24*, 141–155. http://dx.doi.org/10.1037/a0025176

Frazier, T. W., & Youngstrom, E. A. (2006). Evidence-based assessment of attention-deficit/hyperactivity disorder: Using multiple sources of information. *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*, 614–620. http://dx.doi.org/10.1097/01.chi.0000196597.09103.25

Galanter, C. A., & Patel, V. L. (2005). Medical decision making: A selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry*, *46*, 675–689. http://dx.doi.org/10.1111/j.1469-7610.2005.01452.x

Gawande, A. (2010). *The Checklist Manifesto*. New York, NY: Penguin.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Glasziou, P. (2006). The EBM journal selection process: How to find the 1 in 400 valid and highly relevant new research articles. *Evidence-Based Medicine*, *11*, 101. http://dx.doi.org/10.1136/ebm.11.4.101

Hodgins, S., Faucher, B., Zarac, A., & Ellenbogen, M. (2002). Children of parents with bipolar disorder. A population at high risk for major affective disorders. *Child & Adolescent Psychiatric Clinics of North America*, *11*, 533–553.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist*, *51*, 1059–1064. http://dx.doi.org/10.1037/0003-066X.51.10.1059

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307.

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice*, *42*, 121–129. http://dx.doi.org/10.1037/a0022506

Jenkins, M. M., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2012). Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychological Assessment*, *24*, 269–281. http://dx.doi.org/10.1037/a0025775

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., . . . Ryan, N. (1997). 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*, *36*, 980–988. http://dx.doi.org/10.1097/00004583-199707000-00021

Kessler, R. C., Berglund, P., Demler, O., Jin, R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*, 593–602.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. New York, NY: Academic Press.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting & Clinical Psychology*, *69*, 159–172. http://dx.doi.org/10.1037/0022-006X.69.2.159

Lindhiem, O., Kolko, D. J., & Cheng, Y. (2012). Predicting psychotherapy benefit: A probabilistic and individualized approach. *Behavior Therapy*, *43*, 381–392. http://dx.doi.org/10.1016/j.beth.2011.08.004

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology*, *50*, 215–241. http://dx.doi.org/10.1146/annurev.psych.50.1.215

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). 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*, 49, 980-989. http://dx.doi.org/S0890-8567(10)00476-4 [pii]

Newman, C. F., Leahy, R. L., Beck, A. T., Reilly-Harrington, N. A., & Gyulai, L. (2002). *Bipolar disorder: A cognitive therapy approach*. Washington, DC: American Psychological Association.

Norcross, J. C., Hogan, T. P., & Koocher, G. P. (2008). *Clinician's guide to evidence based practices: Mental health and the addictions*. London: Oxford.

Pelham, W. E., Jr. Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*, 449–476. http://dx.doi.org/10.1207/s15374424jccp3403_5

Prochaska, J. O. (2000). Change at differing stages. In C. R. Snyder & R.E. Ingram (Eds.), *Handbook of psychological change* (pp. 109–127). New York, NY: Wiley.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*, *18*, 169–184. http://dx.doi.org/10.1002/mpr.289

Sackett, D. L., Straus, S. E., Richardson, W. S., & Rosenberg, J. (1998). *Evidence-based medicine: How to practice and teach EBM*. New York, NY: Churchill Livingstone.

Schon, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005

for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59*, 22–33. http://dx.doi.org/10.4088/JCP.09m05305whi

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of Clinical Psychology, 63*, 611–631.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York, NY: Churchill Livingstone.

Substance Abuse and Mental Health Services Administration (2012). *Mental Health, United States, 2010*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Sue, S. (1998). In search of cultural competence in psychotherapy and counseling. *American Psychologist, 53*, 440–448.

Taylor, A., Deb, S., & Unwin, G. (2011). Scales for the identification of adults with attention deficit hyperactivity disorder (ADHD): A systematic review. *Research in Developmental Disabilities, 32*, 924–938. http://dx.doi.org/10.1016/j.ridd.2010.12.036

Tsuchiya, K. J., Byrne, M., & Mortensen, P. B. (2003). Risk factors in relation to an emergence of bipolar disorder: A systematic review. *Bipolar Disorders, 5*, 231–242.

Van Meter, A., Youngstrom, E. A., Youngstrom, J. K., Ollendick, T., Demeter, C., & Findling, R. L. (under review). Clinical decision-making about child and adolescent anxiety disorders using the Achenbach System of Empirically Based Assessment. *Journal of Clinical Child* & Adolescent Psychology.

Van Voorhees, E. E., Hardy, K. K., & Kollins, S. H. (2011). Reliability and validity of self- and other-ratings of symptoms of ADHD in adults. *Journal of Attention Disorders, 15*, 224–234. http://dx.doi.org/10.1177/1087054709356163

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening Efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A Systematic Review. *Child and Adolescent Mental Health, 13*, 140–147. http://dx.doi.org/10.1111/j.1475-3588.2007.00461.x

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., . . . Hoagwood, K. E. (2011). Youth Top Problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology, 79*, 369–380. http://dx.doi.org/10.1037/a0023307

World Health Organization (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. London: World Health Organization.

Youngstrom, E. A. (2013a). Future directions in psychological assessment: Combining Evidence-Based Medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology, 42*, 139–159. http://dx.doi.org/10.1080/15374416.2012.736358

Youngstrom, E. A. (2013b). A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*. http://dx.doi.org/10.1093/jpepsy/jst062

Youngstrom, E. A., & Duax, J. (2005). Evidence based assessment of pediatric bipolar disorder, part 1: Base rate and family history. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 712–717. http://dx.doi.org/10.1097/01.chi.0000162581.87710.bd

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., & Price, M. (2004). 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, 43*, 847–858. http://dx.doi.org/10.1097/01.chi.0000125091.35109.1e

Youngstrom, E. A., & Frazier, T. W. (2013). Evidence-based strategies for the assessment of children and adolescents: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (2nd ed., pp. 36-79). New York: Wiley.

Youngstrom, E. A., Jenkins, M. M., Jensen-Doss, A., & Youngstrom, J. K. (2012). Evidence-based assessment strategies for pediatric bipolar disorder. *Israel Journal of Psychiatry & Related Sciences, 49*, 15–27.

Youngstrom, E. A., & Kogos Youngstrom, J. (2005). Evidence based assessment of pediatric bipolar disorder, part 2: Incorporating information from behavior checklists. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 823–828. http://dx.doi.org/10.1097/01.chi.0000164589.10200.a4

Address correspondence to Eric A. Youngstrom, Ph.D., Department of Psychology, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270; e-mail: eay@unc.edu.

Received: October 4, 2013
Accepted: December 3, 2013
Available online xxxx

Please cite this article as: Youngstrom et al., Clinical Guide to the Evidence-Based Assessment Approach to Diagnosis and Treatment, *Cognitive and Behavioral Practice* (2014), http://dx.doi.org/10.1016/j.cbpra.2013.12.005