Empirical Article

Clinical Psychological Science
1–23

DOI: 10.1177/2167702617741845
www.psychologicalscience.org/CPS

Evidence-Based Assessment From Simple Clinical Judgments to Statistical Learning: Evaluating a Range of Options Using Pediatric Bipolar Disorder as a Diagnostic Challenge

Eric A. Youngstrom[1], Tate F. Halverson[1], Jennifer K. Youngstrom[1], Oliver Lindhiem[2], and Robert L. Findling[3]
[1]University of North Carolina at Chapel Hill, [2]University of Pittsburgh, and [3]Johns Hopkins University

Abstract
Reliability of clinical diagnoses is often low. There are many algorithms that could improve diagnostic accuracy, and statistical learning is becoming popular. Using pediatric bipolar disorder as a clinically challenging example, we evaluated a series of increasingly complex models ranging from simple screening to a supervised LASSO (least absolute shrinkage and selection operation) regression in a large ($N = 550$) academic clinic sample. We then externally validated models in a community clinic ($N = 511$) with the same candidate predictors and semistructured interview diagnoses, providing high methodological consistency; the clinics also had substantially different demography and referral patterns. Models performed well according to internal validation metrics. Complex models degraded rapidly when externally validated. Naive Bayesian and logistic models concentrating on predictors identified in prior meta-analyses tied or bettered LASSO models when externally validated. Implementing these methods would improve clinical diagnostic performance. Statistical learning research should continue to invest in high-quality indicators and diagnoses to supervise model training.

Keywords
bipolar disorder, diagnostic accuracy, evidence-based assessment, sensitivity and specificity, open data

Decisions about diagnosis or case formulation are the foundation of the rest of clinical activity (Meehl, 1954; Straus, Glasziou, Richardson, & Haynes, 2011; Youngstrom, 2013). If we do not agree in our definition of the problem, then we are not going to make similar choices about how to treat it (Guyatt & Rennie, 2002), and any similarities in intervention are coincidental. Because of their fundamental importance, diagnostic and classification systems have been a centerpiece of clinical thinking, predating even Aristotle and Galen (Carson, 1996).

Despite the fundamental importance of classification, clinical practice typically relies on imperfect methods. The most commonly used assessment method is the unstructured clinical interview. It relies heavily on intuition, impressionistic weighting of information, and training and expertise to guide the synthesis and probing of data (Garb, 1998). At its best, it swiftly organizes a wealth of multivariate data, recognizes constellations of factors and moderators, and rapidly converges on an appropriate formulation. More often, it is prone to cognitive heuristics that produce biased results (Croskerry, 2003; Jenkins & Youngstrom, 2016), underestimating comorbidity and overdiagnosing some problems while systematically missing others (Jensen & Weisz, 2002; Jensen-Doss, Youngstrom, Youngstrom, Feeny, & Findling, 2014; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). Typical clinical practice shies away from using semistructured interviews,

Corresponding Author:
Eric A. Youngstrom, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270
E-mail: eay@unc.edu

based on the belief that they compromise practitioner autonomy and risk damaging rapport (Bruchmuller, Margraf, Suppiger, & Schneider, 2011), despite quantitative data that patients actually prefer more structured methods (Suppiger et al., 2009). Similarly, practitioners tend to use few rating scales or checklists, and interpretation is often limited to eyeballing total scores and intuitive impressions even when there is a strong research basis for more sophisticated scoring and algorithms (Croskerry, 2003; Garb, 1998).

At this point, it is unambiguous that even relatively simple algorithms equal or better the performance of experienced clinicians (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). But which method should the field prioritize in order to ratchet accurate assessment another notch forward? There are at least four major considerations when comparing methods for interpreting the same data: predictive accuracy—quantified as discrimination and calibration, generalizability, and level of model complexity (Shariat, Karakiewicz, Suardi, & Kattan, 2008). Other factors, such as cost and participant burden as well as cultural factors, should also inform the construction of an assessment battery; here our focus is trying to figure out the best way to analyze and interpret the same data.

Some of these considerations are in tension with each other. In particular, we can always improve predictive accuracy by making our models more complicated, but the tradeoff is that a model that is "overfitted" to one sample will replicate poorly in other samples or when a clinician applies the model to a new case. The core idea has been known for decades, and it is tied to the concepts of Type I statistical error (Silverstein, 1993) as well as more recent discussions of reducing bias versus minimizing variability across replications (James, Witten, Hastie, & Tibshirani, 2013). The question of how to weight scores or items is connected to this as well: The accuracy of predictions using weights based on one sample will shrink when applied to a new set of cases, leading some to advocate simple methods such as unit weighting (Wainer, 1976). Assigning all significant variables the same weight will predict less well than regression weights fit to the same data, but they will shrink less upon cross-validation.

The present article compares a half-dozen different approaches for integrating information. There are two parts to the article. First, we present an overview of progressively more complex solutions to the diagnostic classification problem, followed by a series of applications to two datasets that offer a microcosm of the process of generalization moving from research to practice. We include some simple, classic models that would still improve practice if more widely used clinically, along with some recent, more sophisticated models drawn from the statistical learning model literature and now often incorporated into genetic and imaging research (e.g., Bertocci et al., 2014). The statistical learning models will use what we will call "internal cross-validation," where the sample will be divided 10 times, with one portion used to train the model and the other folds used to test the model. We also used a second independent sample to evaluate reproducibility and generalizability, or external cross-validation (Open Science Collaboration, 2015). External cross-validation is considered best practice when possible, and this also will give readers a sense of the difference between the two methods—especially because external cross-validation is rarely done in practice.

We briefly introduce the clinical demonstration problem, and then we proceed to work through conceptual and practical issues in the application of progressively more complex models, before proceeding to the statistical model building and evaluation.

Clinical Demonstration Problem: Prediction of Pediatric Bipolar Disorder

The detection of pediatric bipolar disorder makes an excellent demonstration scenario for comparing diagnostic models. The condition can occur in childhood or adolescence, as shown by dozens of epidemiological and longitudinal clinical studies from sites around the world (Van Meter, Moreira, & Youngstrom, 2011). It is associated with impairment interpersonally and academically, along with increased risk of substance misuse, incarceration, cardiovascular health problems, and suicide (Youngstrom, Birmaher, & Findling, 2008). Yet the bulk of the research evidence has accumulated in the last 15 years, after the current installed base of practitioners completed their training and became licensed (Goldstein et al., 2017). Lacking any strong research basis for their assessment training, it is not surprising how rife with disagreement clinical opinions are about bipolar disorder. Even so, the data are ugly: Surveys find 100% ranges of opinion about whether or not a given vignette has bipolar (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011), and interrater agreement studies find kappas of 0.1 about bipolar diagnoses (Jensen-Doss et al., 2014; Rettew et al., 2009).

The research on assessment of bipolar has advanced rapidly; there are now at least three well-validated assessments that are public domain and would produce substantial gains in diagnostic accuracy (Youngstrom, Genzlinger, Egerton, & Van Meter, 2015). A key question is how they should be deployed in clinical practice. Make the interpretive approach too simple, and it will not only sacrifice accuracy but also could have serious

unintended consequences, such as overdiagnosis or misdiagnosis. Universal screening for mood disorders combined with simplistic interpretation could do more harm than good (U.S. Preventive Services Task Force, 2009). Make it too complex, and it will not be feasible to use in most clinical settings, and it also may generalize less well. Overfitting complex models provides tight fit to the idiosyncrasies of a sample, not a general method that will be effective across a range of settings (James et al., 2013).

Increasingly Complex Approaches to Clinical Diagnostic Classification

Bet the base rate

Meehl (1954) suggested that the first piece of information to consider is the base rate of the disorder. For clinicians, the base rate can anchor decision-making in a way that avoids the errors of either never considering a diagnostic possibility or overdiagnosing it; but clinicians encounter challenges with applying this strategy at the level of the individual case. Clinicians might not know the base rate at their practice, or the local estimate may be inaccurate, depending on local practices (cf. Jensen-Doss, Osterberg, Hickey, & Crossley, 2013; Rettew et al., 2009). Published rates offer benchmarks that clinicians could substitute for local rates or use to evaluate the accuracy of local practices (Youngstrom et al., 2017). Also, a base rate is a "one size fits all" first approximation, not adjusting probability for any individual risk or protective factors. We will illustrate the practical implications using the local base rates from the two different samples.

Take the best

Gigerenzer and Goldstein (1996) suggested that fast and frugal cognitive heuristics can help improve decision making by using simple strategies and concentrating on powerful variables. For example, name recognition provides a heuristic for predicting which city has a larger population: Berlin or Bremen? Cochin or Calcutta? When confronted with a complex clinical case, "take the best" could suggest focusing on the single variable with the largest validity coefficient and making the decision based on it. This could be a symptom (e.g., mood swings or elated mood), a risk factor (e.g., family history), or a positive score on a test. Consider a case with a family history of bipolar disorder, a caregiver score of 17 on a valid checklist (e.g., the Parent General Behavior Inventory-10 Item Mania scale, PGBI10M; Youngstrom, Frazier, Findling, & Calabrese, 2008), and a self-reported score of 9 on the same scale, plus prior history of anxiety and depression. Which piece of information to consider "best" will change depending on the criteria. In terms of making a decision about a potential youth bipolar diagnosis, family history is probably the variable with the largest research base. If a clinician picked variables based on familiarity (name recognition), then family history might be the "best" variable. In contrast, using the research and effect sizes to guide the choice would lead the clinician to focus on the PGBI10M instead: The effect size for it substantially outperforms youth self-report (Youngstrom et al., 2015), and it is larger than effect sizes for the other variables, too.

Most clinicians will use a single threshold or rule of thumb to decide whether the score is "high" (a positive test result) or "low." Researchers can estimate the tradeoff between diagnostic sensitivity and specificity in a sample. There are different algebraic definitions of the "optimal" threshold, depending on whether the goal is to maximize accuracy for detecting true cases (sensitivity), minimizing false alarms (i.e., maximize specificity), or maximizing the overall chances of being correct (which also needs to consider the base rate of the condition; Kraemer, 1992; Swets, Dawes, & Monahan, 2000). Clinicians usually learn a single published rule of thumb and then interpret the scores as either being positive or negative.

The probability nomogram (a feasible way of applying Bayes theorem)

Bayes theorem is a centuries-old method for revising probabilities conditional on new information. In clinical decision-making, it provides a method for combining the base rate or prior probability with the information offered by an assessment finding. Evidence-based medicine (EBM) advocates it as a way of interpreting diagnostic tests, coupling information about prior probability with the information conveyed by the diagnostic sensitivity and specificity of the result (Jaeschke, Guyatt, & Sackett, 1994; Straus et al., 2011).

Bayes theorem combines a prior probability, such as the disorder's base rate, with new information to revise the probability. It has become widely applied to a variety of prediction and classification problems, including forecasting weather, elections, defaulting on loans, sporting events, or forensic recidivism (Baumer, Kaplan, & Horton, 2017; Silver, 2015; Swets et al., 2000), as well as clinical diagnosis. The practical problem that Bayes theorem addresses is how to re-estimate the posterior probability for a local setting or an individual patient. Different variations of the formula accommodate different inputs. The posterior probability estimate attached to a positive test result is the positive predictive value

(PPV), and the posterior probability for a low-risk or negative test result is the negative predictive value (NPV). These are also estimates of the accuracy of a positive or negative assessment finding.

To improve the feasibility for clinicians, EBM recommends the use of a "probability nomogram," analogous to a slide rule that uses geometry to accomplish the algebraic transformations (Straus et al., 2011; do an online search for "probability nomogram" to find numerous examples and illustrations). The nomogram turns applying Bayes theorem into an exercise in connecting the dots. It sacrifices some precision compared to using an online probability calculator (of which there are now several free versions; do an online search for "EBM probability calculator"), but it also is visual and interactive and may appeal to those who are not "numbers people" (Gigerenzer & Muir Gray, 2011). For example, family history of a mood disorder increases an individual's risk for developing a mood disorder, and higher scores on a symptom checklist often correspond to a greater risk for a diagnosis.

To use a nomogram, start with an initial probability estimate such as the base rate of the condition, using published benchmarks most resembling the setting where the clinician is working, or if available, the actual local base rate. Next, the diagnostic likelihood ratio (DiLR) corresponding to a second source of information (e.g., positive family history) is plotted on the nomogram. A line connecting the base rate, or pretest probability, through the likelihood ratio, extends across a third line on the nomogram to estimate the new posttest probability (Van Meter et al., 2016). To use a probability nomogram, the effect size attached to a particular test result needs to be rescaled as a DiLR, which is the fraction of cases that have the diagnosis scoring in that range divided by the fraction of cases without the diagnosis scoring in that range. In the simple case where there is only one threshold, a positive test result has a DiLR equal to sensitivity/(false alarm rate). A negative test result would have a DiLR of (false negative rate)/specificity (Pepe, 2003).

Naive Bayesian Algorithms, or Using the Nomogram With Multiple Assessment Findings

Two more refinements fully articulate the model EBM advocates: (a) segmenting tests into more than two ranges, milking more informative DiLRs from the segments, and (b) sequentially adding tests to update probabilities. Segmenting makes intuitive sense: Rather than dividing scores on the PGBI10M into two chunks of *low risk* versus *high risk*, perhaps five chunks—*very low, low, neutral, high*, and *very high*—might be more informative. Each segment would have its own DiLR, which would be the fraction of cases with bipolar divided by the fraction without bipolar scoring in the same segment. For example, the clinician would give the PGBI10M and then pick the DiLR matching the score range and plug that into the nomogram. The original PGBI10M validation study divided the scores into six segments with DiLRs that ranged from 0.01 for the segment defined by a raw score of 0 up to a DiLR of 7.25 for extreme high scores of 18+ (Youngstrom, Frazier et al., 2008).

If more than one assessment result is available, then the clinician could use the updated probability after interpreting the PGBI10M as the new "starting" probability and then combine it with the DiLR for the next assessment result. In the case of bipolar disorder, family history is a well-established risk factor (Tsuchiya, Byrne, & Mortensen, 2003) that has been suggested for clinical interpretation (Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012; Youngstrom & Duax, 2005). A rule of thumb would be that bipolar disorder in a first-degree relative would increase the odds of bipolar by 5×. It does not matter whether the clinician applies the family history or the PGBI10M DiLR first. The results would be the same in either scenario, and if several DiLRs from different tests are available simultaneously, they can be multiplied to combine them into one summary DiLR.

The sequential application of DiLRs makes the big assumption that the predictors are uncorrelated (i.e., that there is no association between family bipolar history and scores on the PGBI10M). This is unlikely to be true in practice, which is why the algorithm is also known as the "Naive Bayes" algorithm in statistical learning circles (Baumer et al., 2017). If the correlation among predictors is less than .3, the increases in predictive consistency and accuracy often offset the bias introduced, but when correlations become large (> .5), then picking the best and ignoring the others will avoid more severe bias (Youngstrom, 2013). The Naive Bayesian approach and sequential application of the nomogram also allow for the clinician to add new information in different orders for different patients. Several vignettes illustrating sequential application of nomograms in clinical settings are available (e.g., Ong et al., 2016; Van Meter et al., 2016; see also Wikiversity, n.d.).

Logistic regression

Researchers often combine multiple predictors in a regression model. For classification problems, logistic regression is a widely used regression approach. Logistic regression models the probability of having a diagnosis, using a transformation to let probability (which is bounded between 0 and 1) be the dependent variable in a linear model (Hosmer & Lemeshow, 2013). The regression coefficients indicate the change in the probability corresponding to a one-unit change in each

predictor, with the probability transformed into the log of the odds of the diagnosis. A logistic regression with a single predictor produces the same probability estimates and effect sizes as a receiver operating characteristic (ROC) analysis. When building a model with a single predictor, one should again "take the best," picking the variable with the largest effect size to use as the predictor.

Logistic regression differs in three important respects from the simpler models we have reviewed so far: (a) It makes a separate prediction for each one-point change in the predictor—for example, in the case of the PGBI10M, it would make 31 predictions for a score ranging from 0 to 30, instead of just two probability estimates for a traditional test positive/negative interpretation; (b) it can adjust for covariates, such as some of the demographic and clinical variables that might differ between the academic and community clinics; and (c) it adjusts for the covariance among predictors, rather than making the Naive assumption that family history and the PGBI10M score are independent. The weights from the regression model could be saved to build a prediction equation for out-of-sample cases, providing the components for a "probability calculator."

In principle, there is no reason that the predicted probability could not get dropped into the evidence-based assessment decision-making framework, where the clinician and patient compare the updated probability to the wait-test and test-treat thresholds and share the decision about what clinical action is next. Feasibility changes when using a logistic regression model: No human being can apply the weights in their head. The prediction equation requires a computer or app—raising the base of the natural logarithm to an exponent defined by the regression is not a human skill. The probability nomogram represents the furthest down a clinical interpretive path that we can get without transferring the flag to a vehicle that requires a computer as part of the decision support system. A second consideration is that a regression equation requires complete data on all of the variables in the model. Consider a clinician working with a youth in foster care: Family history simply may not be available. With the nomogram framework, the clinician could still apply the PGBI10M and other assessment findings. A logistic regression equation built with family history as one of the predictors is not usable when the variable is missing (and the other regression weights are all organically adjusted based on the assumption that family history was also in the model). Thus, logistic regression is a familiar technique that represents a tipping point in clinical application that creates new demands for software to implement the scoring and decision support. Again, the increases in accuracy could be worth the trade.

Logistic Regression With Multiple Predictors—Traditional Modeling

A strength of logistic regression is that it can incorporate multiple predictors, assigning them weights that are optimized in the sense of maximizing accuracy in the training sample, as well as accommodating the covariance among predictors (Hosmer & Lemeshow, 2013). The weights also have heuristic value for theory testing or model building, and traditional approaches to regression can use block entry to explicitly describe different conceptual roles for variables (e.g., hypothesized risk factor, covariate, or exploratory identification of new predictors; Hosmer & Lemeshow, 2013). Regression models can include multiple predictors, interactions, nonlinear terms—all increasing fit and reducing "bias," but paying a tax in terms of greater variance of weights and shifting accuracy across replications. In the statistical learning literature, this is called the bias-variance trade-off: Greater predictive accuracy reduces the bias of predictions, but the more complex models may overfit the training data and show great variance in performance when cross-validated (James et al., 2013). Forward stepwise model selection is an old algorithm for maximizing model fit to the data, peeking at all the available candidates and choosing one based on the observed effect size in the data. It overfits the data. If several variables have similar performance, one gets picked as "best," even if only by a whisker. The whiskers are often chance patterns that will not replicate in external data. Stepwise is prone to Type I error (false positive selection of a variable). There also will be a large number of statistical models that provide similar fit while using quite different combinations of variables and weights—called the "Rashomon Problem" in homage to Kurosawa's film where four protagonists offer highly different interpretations of the same facts (Breiman, 2001).

In contrast, clinicians could pick the best based on published studies, clinical training, or other heuristics rather than an exploratory data-mining approach, and researchers can prioritize variables based on past findings or theory and push them into the model using block entry. In a traditional modeling approach, the researcher would identify specific predictors to include, such as a valid assessment and family history, and they might also include some covariates for demographic or clinical features. The optimized weights and adjustment for correlation among the predictors mean that logistic regression will tie or outperform the nomogram in the same sample in terms of both discrimination and calibration. The accuracy of the regression weights will shrink on cross-validation in external data, and the shrinkage would be larger as models include more

predictors, and especially if stepwise methods were used. A good model with robust predictors that generalize well to external samples should continue to outperform Naive Bayesian approaches in terms of accuracy, but in terms of feasibility of application, multivariable models are clearly far past the tipping point where computers would be required.

Controlling for the overlap between predictors also is a two-edged sword in clinical application: It produces more accurate (less biased) estimates, but it only can apply to cases with complete data on the predictor set. All of the weights are organically linked and assume the presence of the other predictors in the model. Using a probability calculator built on a multivariable logistic regression effectively dictates the use of a core battery to gather all the components of the model, and it will break whenever one of the variables is unavailable.

Least Absolute Shrinkage and Selection Operation

The least absolute shrinkage and selection operation (LASSO; James et al., 2013) is an increasingly popular form of statistical learning method that tries to hit the sweet spot between fitting well (reducing bias in predictions) and also being likely to generalize (showing low variance across replications). It is a form of regression, and it can handle multiple predictors (to the extreme of having more variables than cases!), interactions, and nonlinear transformations. With LASSO, several of the normal considerations about logistic regression models get relaxed. In a statistical learning framework, there is no requirement that there be a strong theoretical justification for predictors. It also is easy to include collinear predictors, such as full-length and carved scales of the PGBI at the same time, as well as interaction terms. Because the computer is doing the interpretive hard work instead of a human, the complexity of the model is not a concern in the same way as in the simpler models above. The computer will estimate and compare a staggering set of permutations. Unlike some other statistical learning models (such as support vector machines and random forests), LASSO is not a "black box"—it indicates which variables are contributing to the model (James et al., 2013), which is helpful within the context of discovery in research and also in informing case formulation.

How does it avoid overfitting, as would be the bane of stepwise regression models? Internal cross-validation is its answer: LASSO takes the existing data, shuffles them, and divides them into a training and a testing sample, building the model in the training subset and then evaluating its performance in the testing sample. It then repeats, resampling k times and looking at the stability of the model estimates across the k-fold replications ($k = 10$ resamples is common convention). The LASSO algorithm searches for the best fitting model in terms of balancing the bias reduction (i.e., accuracy of prediction) versus variance (i.e., fluctuation of weights across the cross-validations) to select an accurate but stable model.

One final adjustment is that LASSO applies a shrinkage function, penalizing all of the regression weights and forcing some to zero. Why take optimized weights and shrink them? Killing off the weaker weights simplifies model interpretation, and it protects against Type I error and overfitting when repeated over the cross-validations. The penalty function is a form of *regularization*—simplifying the description of the weights across models and cross-validations (James et al., 2013). The performance of the tuning parameter gets averaged over the 10-fold internal validations to decide what is the best fitting model. In practice, LASSO is often used as a model-building procedure, and then the "winning" model gets repeated in the full sample to provide the final weights for clinical application.

To illustrate how LASSO contrasts with other more familiar approaches, we will build a model that examines more variables than would be the norm in research, adding highly multicollinear predictors, and also interactions terms. LASSO, like ridge regression, is designed to work well in high-collinearity scenarios, and it can also consider a set of variables that is larger than the number of cases in the dataset (James et al., 2013).

Some consider it a flaw in studies designed to compare methods when one algorithm has access to additional variables that the clinician or other algorithms did not. It certainly would feel unfair in a traditional competition if one contestant had access to a trove of information not available to the others. However, proponents of statistical learning models point to data mining as a natural progression, extracting information from variables that might have escaped identification based on prior theory or application. The computer's indifference to complexity makes it possible to build models that add middleweight variables to explain small incremental amounts of variance, cumulatively offering big gains in predictive performance (Breiman, 2001; Walsh, Ribeiro, & Franklin, 2017). It will be crucial to see how much incremental value might accrue from such "brute force" applications of computing power, as well as getting a sense of how much these methods might shrink when applied to new out-of-sample cases.

Aims and Hypotheses

Our aim is to compare a series of increasingly complex models for the purpose of identifying which youths seeking outpatient mental health services might have

a bipolar spectrum disorder. We use a consensus diagnosis following *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; *DSM–IV–TR*; American Psychiatric Association, 2000) criteria as the criterion variable. We test a series of models, building up to the LASSO model that uses machine-learning methods to incorporate far more predictors than traditional methods. The LASSO method will use 10-fold internal cross-validation, designed to avoid overfitting and yield generalizable estimates.

However, we also have a second dataset for external cross-validation, with diagnoses established using similar methods but with a different referral pattern and demography (Youngstrom et al., 2005). We use the second sample to evaluate how each model "published" based on the academic sample would generalize when applied in a different community clinical setting. Unlike a traditional "Study 1, Study 2" structure, we zigzag between the samples when presenting the results. This organization heightens the focus on how external validation affects each model, illustrating how realistic changes in sampling patterns might challenge generalization and clinical application. Finally, we flipped the order and used the community data to build the model, externally validating it on the academic data, as well as looking at how a model trained using clinical diagnoses from the medical record performed. These offer strong tests of whether starting with different samples would converge on similar models.

We hypothesized that methods that took local base rates into account would perform better in terms of model calibration. We also expected that the more complex models would fit better in terms of discriminative accuracy in the samples where they were built. The question of which models fared best during external cross-validation was an exploratory aim, because so little prior work has directly compared these methods in diagnostic assessment. We expected that the more complex models would all confirm statistical validity of the PGBI10M and family history but would differ in the choice of additional predictors.

Method

Participants

Youths 5 to 18 years of age and youth caregivers were recruited from outpatient mental health centers. Families were recruited from outpatient clinic sites at Case Western Reserve University (i.e., academic clinic) and Applewood Centers (i.e., community clinics) in Cleveland, Ohio. Institutional Review Boards at both institutions approved all study procedures. Both youths and caregivers had to be fluent in English. Youths with pervasive developmental disorder or cognitive disability were excluded. Families received compensation for study participation.

The academic clinic sample ($N = 550$) included families presenting to a clinic located within a university psychiatry department (Findling et al., 2005). Families were referred to the clinic from within the psychiatry department or from outside referrals, primarily for concerns regarding youth mood symptoms. However, there also were treatment studies for a variety of other diagnoses, and there were episodes of recruitment for other diagnoses. The community clinic sample ($N = 511$) was a random subset of families presenting for services for the youth's youth mental health and/or behavior who completed both regular intakes and the enhanced research study interview (Youngstrom et al., 2005).

Measures

Bipolar spectrum disorder diagnosis—target variable for classification. Highly trained research assistants completed the Schedule for Affective Disorders and Schizophrenia for School-Age Children–Epidemiologic Version (K-SADS-E; Orvaschel, 1994) or the Present and Lifetime–Version (K-SADS-PL; Kaufman et al., 1997) with the youth and primary caregiver. Interrater reliability kappas were ≥ 0.85. Final diagnoses used a consensus review process with a licensed psychologist or psychiatrist (see Findling et al., 2005; Youngstrom et al., 2005, for more details). The checklist results were masked from the diagnostic process. Present analyses dummy coded the presence or absence of any bipolar spectrum disorder. This served as the dependent, target variable and was a "yes" for any youth meeting *DSM-IV* criteria for Bipolar I, Bipolar II, cyclothymic disorder, or bipolar Not Otherwise Specified (NOS), regardless of other comorbid conditions. The most common presentation for bipolar NOS was insufficient duration of hypomanic or manic episode (Youngstrom, 2009). We also extracted the diagnoses from the medical record for the community cases, providing a similar dummy code that captured diagnosis as usual. This variable would be what statistical learning might use to supervise the training of models built in archival data and electronic medical records.

Clinical characteristics of samples

Comorbid or cognate diagnoses. Diagnoses that involve similar clinical presentations are more likely to generate false-positive scores on predictor variables. Thus, differences in the diagnostic mix of cases could significantly change the validity of predictor variables. Therefore, we also created dummy codes for the presence or absence of any anxiety disorder, attention-deficit/hyperactivity disorder, oppositional defiant disorder, conduct disorder, or posttraumatic stress disorder (PTSD), as well as a count of how many diagnoses each youth carried (Kim & Miklowitz, 2002).

Current severity of mood symptoms. Current mood symptom severity also might change the performance of predictors. If youths in one sample were more symptomatic, that would increase the apparent sensitivity of measures, as more of the target cases would score above threshold on a predictor (Pepe, 2003). We used the Young Mania Rating Scale (YMRS; Young, Biggs, Ziegler, & Meyer, 1978) to compare the manic symptom severity. The YMRS is an interview rating of the severity of 11 symptoms of mania. In youths, it is based on interviews with both the caregiver and child. Similarly, the Child Depression Rating Scale–Revised (CDRS-R; Poznanski, Miller, Salguero, & Kelsh, 1984) measured depressive symptom severity. Raters completed the YMRS and CDRS-R during the same interviews as the KSADS, making the ratings highly collinear (e.g., AUC > .94 if used in a ROC analysis to identify bipolar disorder).

Predictor variables

PGBI. Caregivers completed the PGBI about the youth. The full PGBI has 73 items, with scores ranging from 0 to 3 (Youngstrom, Findling, Danielson, & Calabrese, 2001). It shows exceptional internal consistency estimates and high discriminative validity (Youngstrom et al., 2015). The depression scale has 46 items and alphas > .96 in both samples. The hypomanic/biphasic scale has 28 items (one item is included on both scales) and an alpha > .92 in both samples.

Because of a high reading level and length, several carved shorter forms are available. The 10-item mania scale (PGBI10M) focused on the items best discriminating bipolar from nonbipolar diagnoses using parent report (Youngstrom, Frazier et al., 2008), and it has continued to perform in the top tier of available checklists in terms of discriminative validity (Youngstrom et al., 2015). Another carved scale pulled seven items focusing on sleep disturbance, again with good internal consistency and identification of cases with mood disorder (Meyers & Youngstrom, 2008). A third pair of scales is the 7 Up–7 Down, consisting of seven hypomanic/biphasic and seven depressive items selected for optimal psychometrics in a self-report format (Youngstrom, Murray, Johnson, & Findling, 2013). Interestingly, only one of the 7 Up items overlaps with the 10 most discriminating caregiver report items, reflecting the differences in informant perspective about mood symptoms (Freeman, Youngstrom, Freeman, Youngstrom, & Findling, 2011). The scoring instructions prorate items if one is skipped, making fractions of a point possible. We used the PGBI10M in all models except the LASSO, which considered all of the above.

Family history of bipolar disorder. Caregivers also reported about family history of bipolar disorders (see Jenkins et al., 2012, for full description of family history assessment). Family history was translated into a yes/no variable for presence of any family history of bipolar disorders. The probability nomogram/Naive Bayes models used the same weights and DiLRs as used in Jenkins et al. (2012), duplicating how a clinician would use published estimates in their own clinic.

Demographic variables. Some models included child and parent demographic variables such as child age, sex, race, and number of comorbid *DSM–IV* diagnoses, all based on caregiver and child interviews. Family income was calculated based on parent report of household income and then grouped into ordinal categories.

Procedure

We used the academic sample as the training sample, both because it was larger and because that is typically how research would flow: Initial work would be done in academic settings and later applied in community settings. We fit each successively more complex model in the academic sample. Table 1 lists the candidate predictor values and shows how the models increase in complexity. The final LASSO models used 10-fold cross-validation for the statistical learning models to illustrate how internal cross-validation *estimates* out-of-sample performance. Then, we took the coefficients based on the academic sample and applied them to the community sample, examining *external* cross-validation. External validation is how clinicians need to apply research findings in their own practice—taking published weights and using them with the assumption that performance will not change substantially. Next, comparing these estimates and a third set optimized for the community sample provides a sense of how much differences in demographics or clinical referral patterns may alter performance beyond what a machine-learning approach might anticipate during internal cross-validation.

As a final test of external validation, we flipped the script and used the community sample to build a LASSO model and then examined how that model fared when externally cross-validated in the academic clinic. Last, we ran a LASSO model using the diagnoses from the medical record in the community clinic to supervise the model training, instead of the research KSADS diagnoses. This last scenario is the version where large existing electronic medical records are used as the input for machine learning. If the three approaches to LASSO converge on similar models, that would provide reassuring evidence that statistical methods developed for internal cross-validation offer reasonable assurance about external validation.

Listwise deletion removed participants with missing values on any predictors used in models, ensuring that the same participants were used in successively complex models to allow for model comparison both within and

Table 1. Candidate Variables Included in Each Prediction Model

Variable	Take the best screener	Probability nomogram	Multilevel and multipredictor nomogram	Logistic regression (1 *df*)	Augmented logistic regression (5 *df*)	LASSO (136 candidate variables)
PGBI10M	X	X	X	X	X	X
Family bipolar history			X		X	X
Sex (female)					X	X
Youth age (years)					X	X
Race (White yes/no)					X	X
PGBI–depression						X
PGBI–hypo/biphasic						X
PGBI–sleep						X
PGBI 7 Up						X
PGBI 7 Down						X
Diagnosis count						X
Other diagnoses[a]						X
Two-way interactions						X

LASSO = least absolute shrinkage and selection operation; PGBI = Parent General Behavior Inventory; PGBI10M = PGBI 10-item mania scale.
[a]Dummy codes for attention-deficit/hyperactivity disorder, oppositional defiant disorder, conduct disorder, anxiety, and posttraumatic stress disorder.

between clinic groups. Correlations between a dummy code for participants removed from the sample and all predictors yielded very small R^2s < .001–.05, supporting tenability of the missing-at-random assumption.

Results

Preliminary analyses: Mood symptom benchmarking and sample comparison

Cases with bipolar disorder showed similar manic symptom severity across both samples, $t(304) = 0.12$, $p = 0.902$. Community cases with bipolar had higher depression levels, $t = 2.83$, $p < .005$, and more comorbidity, $t = 5.04$, $p < .0005$, indicating a higher rate of mixed-mood states and complex presentations. Significantly higher rates of anxiety ($p < .0005$), conduct disorder ($p < .05$), and PTSD ($p < .05$) contributed to the higher comorbidity in the bipolar cases at the community clinic. Of the cases with bipolar disorder, 50% had bipolar I in the academic clinic, versus 21% in the community clinic, and 45% of the academic cases had cyclothymic disorder or bipolar NOS, versus 68% of the community cases with bipolar. As Table 2 summarizes, the overall demography and clinical characteristics of the two samples differed significantly and often substantially.

Current practice and the best case scenario

The chart diagnoses from the community clinic provide a good snapshot of the starting point at many clinics. The clinical intake interview was unstructured. The agency used general checklists of behavior problems, mandated either by Medicaid or by the agency leadership. No checklists with manic symptom content were used. The cases included in this article also consented to interviews by the research team that also generated academic center diagnoses. When comparing billing diagnoses to the results of the research KSADS consensus diagnoses, agreement was $K = .28$, $p < .05$, when using a broad definition of bipolar spectrum (including mood disorder NOS). The clinical diagnoses had 31% sensitivity and 94% specificity compared to the research diagnoses (Jensen-Doss et al., 2014). The kappa is similar to that found by Regier et al. (2012) in the field trials for the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–5*; American Psychiatric Association, 2013) and better than that in a meta-analysis of agreement between clinical and structured diagnostic interviews (Rettew et al., 2009).

At the other extreme, what would be the best case scenario for the accuracy of a predictive model? Kraemer (1992) showed that if our diagnoses are not perfectly accurate, then they will constrain the accuracy of our predictive models. Imagine an academic test with errors on 5% of the key—a perfectly prepared student would score around 95%, instead of 100%. Using $K = .85$ for the KSADS consensus diagnoses imposes a ceiling of around AUC ~.925 for predictive models, instead of the theoretical ceiling of 1.000 (Kraemer, 1992, pp. 82–91).

Bet the base rate

Academic clinic. The academic clinic uses semistructured diagnostic interviews (SDIs) done by highly trained and supervised raters. The resulting diagnoses are highly

Table 2. Demographics and Clinical Characteristics by Clinic Setting

	Academic clinic (N = 550)	Community clinic (N = 511)	Effect size[a]
Youth demographics			
Male, % (n)	60% (217)	60% (205)	.01[n.s.]
Age, M (SD)	11.40 (3.23)	10.53 (3.41)	.26***
White, % (n)	79% (433)	6% (31)	.74***
Family income[b]	2.45 (1.21)	1.28 (0.64)	1.20***
Clinical characteristics			
Family history of bipolar	35% (194)	32% (165)	.03[n.s.]
YMRS	11.65 (11.86)	6.05 (8.41)	.54***
CDRS-R	35.49 (16.08)	29.95 (13.20)	.38***
PGBI10M	10.13 (7.88)	7.47 (6.35)	.37***
PGBI–hypo/biphasic	24.66 (16.84)	19.70 (14.22)	.32***
PGBI–depression	36.19 (25.67)	24.48 (21.49)	.49***
7 Up	5.16 (4.61)	4.11 (3.83)	.25***
7 Down	6.24 (5.28)	3.21 (4.04)	.64***
PGBI–sleep scale	5.87 (4.74)	4.06 (4.18)	.41***
Number Axis I diagnoses	2.15 (1.34)	2.69 (1.38)	–.39***
Bipolar spectrum diagnosis	44% (241)	13% (65)	.34***
Any attention-deficit/hyperactivity	54% (295)	66% (338)	–.13***
Any oppositional defiant disorder	30% (167)	38% (196)	–.08**
Any conduct disorder	8% (44)	12% (61)	–.07*
Any anxiety disorder	8% (45)	27% (138)	–.25***
Any posttraumatic stress disorder	2% (11)	11% (54)	–.18***

PGBI = Parent General Behavior Inventory; PGBI10M = PGBI 10-item mania scale; YMRS = Young Mania Rating Scale; CDRS-R = Child Depression Rating Scale–Revised.
[a]Phi-squared for categorical variables (sex, race, diagnostic group), and Cohen's d for continuous variables (age, number of diagnoses, rating scales). A positive coefficient means the effect was larger in the academic sample, and a negative coefficient means that the effect was larger in the community—the academic parameter would underestimate the corresponding value in the community. [b]Income level of 2 = $20,000–$40,000; Level 3 = $40,000–$60,000. [c]Equal variances not assumed, Levene's test $p < .05$.
n.s., not significant. $p > .05$. *$p < .05$. **$p < .005$. ***$p < .0005$.

reliable and valid (reaching the longitudinal expert evaluation of all available data, or LEAD, standard; Spitzer, 1983). Those psychometric features make it attractive to use the academic base rate as a benchmark for other clinics. However, referral patterns and inclusion/exclusion criteria may result in idiosyncratic rates that differ markedly from community clinics. Table 2 shows that will be a stern challenge moving from the academic to the community clinic. Bipolar spectrum disorders (n = 241 of 550) composed 44% of the sample at the academic clinic. Because the clinic specialized in mood disorders, the referral pattern was highly enriched, and cases without bipolar disorder often were not entered into the research database for some projects.

Community clinic. Performing a chart review of the cases seen at the community mental health center found that no cases were clinically diagnosed with bipolar I, and 9% were diagnosed with bipolar spectrum disorders broadly defined (almost entirely NOS). Using the research interview process, bipolar spectrum disorders (n = 65 of 511) accounted for 12.7% of cases in the community clinics. As noted above, the agreement about which cases had bipolar was only modest, but the similarity in base rate estimates will have good consequences for the calibration of several of the following models.

Based on what is well known about the reliability and validity of both SDIs and of unstructured diagnosis as usual, it might make sense to switch to the academic base rate as a better estimate of bipolar prevalence (leaving aside the referral pattern problem for a moment). But how would a clinician use that data? Randomly applying the academic base rate in the community clinic would result in an AUC = .500 and K = 0, by definition (see Table 3). This is not satisfactory.

Without adding any more assessment data, clinicians still could glean some guidance from a study directly comparing SDIs and clinical diagnoses. The results

Table 3. Accuracy Statistics for Discrimination (AUC) and Calibration (Spiegelhalter's *z*) for Internal Validation and Cross-Validation in an Academic Sample and External Cross-Validation in the Community Sample

	Academic sample ($N = 550$)		External cross-validation: Academic weights in community sample ($N = 511$)	
Model	AUC	Spiegelhalter's *z*	AUC	Spiegelhalter's *z*
Bet the base rate	.500 (.025)	0.01[n.s.]	.500 (.038)	–14.16****
Take the best (dichotomize PGBI10M)	.781 (.020)	–0.01[n.s.]	.729 (.029)	5.27****
Nomogram	.781 (.020)		.729 (.029)	0.01[n.s.]
Multilevel and two-variable nomogram	.882 (.014)	0.19[n.s.]	.775 (.025)	2.09*
Logistic regression (1 *df*)	.857 (.016)	0.13[n.s.]	.799 (.024)	0.04[n.s., b]
Logistic regression (5 *df*)	.890 (.014)	–0.06[n.s.]	.775 (.026)	4.47****
LASSO (136 candidates)	.902 (.013)	–3.72***	.801 (.024)	0.32[n.s., b]
Reversed LASSO (community weights)	.864 (.015)	20.55****	.830 (.023)	–0.62[n.s.]
Diagnosis upper limit	.925[a]	—	.925[a]	—

[a]The KSADS diagnosis kappa of .85 imposes an upper bound on the AUC (Kraemer, 1992).
[b]The calibration plot (Fig. 2) and Hosmer–Lemeshow test both indicated marked calibration problems, X^2 (10 *df*) > 200.00, $p < .00005$. In other models, the Spiegelhalter *z* and other calibration diagnostics agreed.
n.s., not significant. $p > .05$. *$p < .05$. **$p < .005$. ***$p < .0005$. ****$p < .00005$.

presented in Jensen-Doss et al. (2014), for example, suggest that clinicians underestimate rates of comorbidity, tend to be highly specific but less sensitive to most diagnoses, and were particularly conservative about diagnosing bipolar. However, the results from an academic sample, without a direct linkage to clinical diagnoses, are much less useful than they could be, because the sampling differences make differences in base rate ambiguous—Are they due to inaccuracy in local diagnoses, or varying referral patterns, or a combination of factors?

Take the best

A next step would be to add a checklist to screen or assess for potential bipolar disorder. A recent meta-analysis comparing all published pediatric mania scales found three performing in the top tier (Youngstrom et al., 2015). Clinicians will not want to add redundant tests, and practical considerations push towards using short and free scales. The PGBI10M satisfies all of these criteria.

Academic clinic. A traditional way of interpreting a test is to pick a data-driven threshold and treat all cases scoring at or above the threshold as having a "positive" test result. ROC analysis identified a cutoff score of 6.25 as producing the highest combination of sensitivity (92%) and specificity (64%) in the academic sample; 67% of cases testing positive (scoring at 6.25 or above) had a bipolar diagnosis (the PPV). Of cases scoring below the threshold, 91% did not have bipolar disorder (the NPV). Figure 1 shows a back-to-back histogram with the PGBI10M score distributions for both bipolar and nonbipolar cases, and Table 3 reports the AUC as an effect size for the accuracy of the PGBI10M. Of note, the act of splitting the scores into "test positive" and "test negative" defines an ROC curve with an AUC of .781; this is lower than for the full PGBI10M AUC of .857 because a threshold model is using less information about the score.

Community clinic. A diligent clinician could read the meta-analysis, pick the PGBI10M, get a free copy of the measure and scoring instructions (Wikipedia, n.d.), and use the threshold of 6.25 based on the "published" academic sample result. In theory, the diagnostic sensitivity and specificity of the test should be stable, and they are algebraically unconnected to the base rate of the disorder in the sample. In practice, test sensitivity can change as a function of factors that affect illness severity or the range of presentations (Zhou, Obuchowski, & McClish, 2002). Looking back at the histograms (Fig. 1), anything that shifts the distribution of scores for the bipolar group higher will move a larger percentage of the cases above any given threshold, increasing the sensitivity, and factors affecting the spread of scores also could change performance. Conversely, factors shifting the score distribution for nonbipolar cases will change the specificity. Meta-analyses (Youngstrom et al., 2015) and direct comparisons (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006) have shown that using healthy controls increases the apparent specificity of tests, and building a comparison group with cases seeking services and with higher levels of impairment or higher rates of cognate diagnoses reduces the specificity estimate. Because the academic sample used clinical cases as the comparison, and both are outpatient settings, we did not expect big changes in the specificity; but if the community clinic has

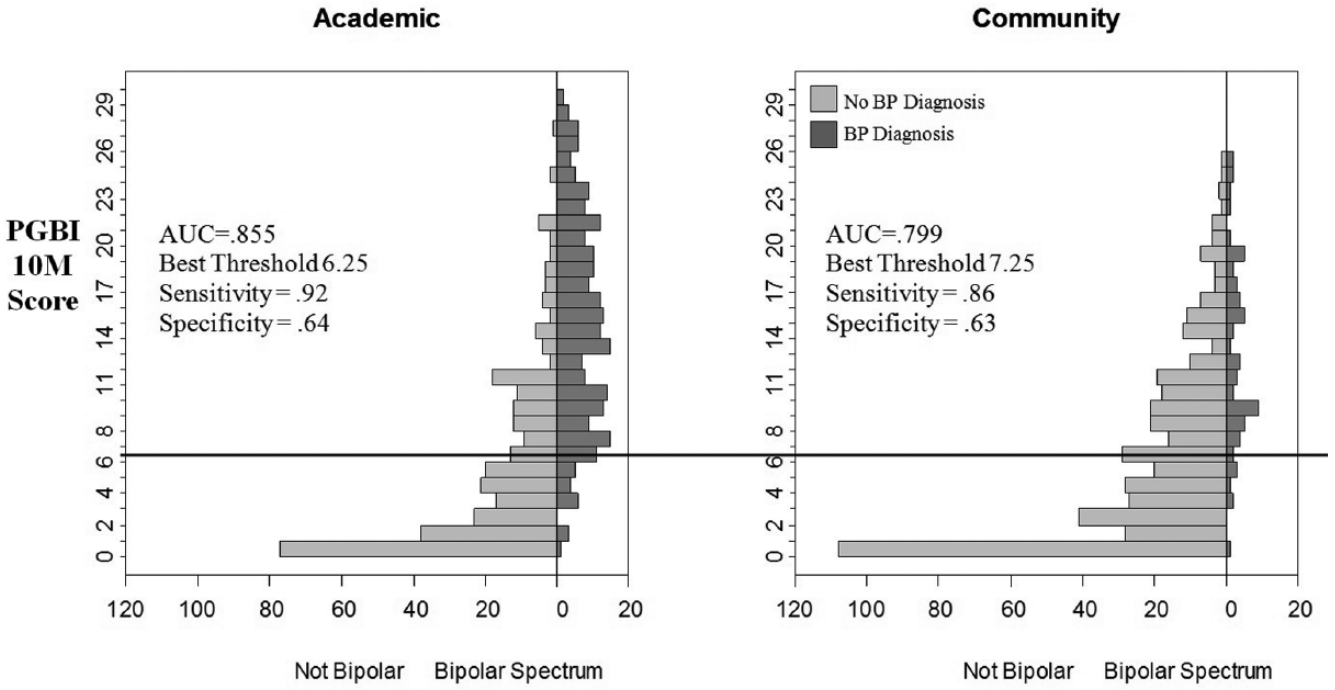

Fig. 1. Back-to-back histograms of the distribution of scores on the Parent General Behavior Inventory 10-item mania scale (PGBI10M) for youths with and without a diagnosis of bipolar disorder (BP) in both academic (N = 550) and community (N = 511) clinic settings. AUC = area under the curve.

fewer cases with mania, or less severe manic presentations on average, then we would expect the sensitivity to drop.

Clinicians will not usually have access to SDIs or to other ways of checking their accuracy. Meehl (1973) identified the lack of corrective feedback as one of the main reasons years of experience tend to improve clinical acumen little, if at all. In practice, clinicians will not be able to re-estimate the accuracy of a test in their setting; they need to trust the generalizability of the researcher's estimate. Using the threshold of 6.25 from the academic sample identifies 43% of the cases as "testing positive" for bipolar disorder and, conversely, 57% of the cases as testing negative. How should a clinician interpret these results?

Comparison to external benchmarks or local estimates about the base rate of bipolar disorder both indicate that the test positive rate is too high, containing a large number of false-positives. The clinician has a sensitivity estimate from the academic sample (92%) but that is not the same thing as how accurate a positive test result would be (i.e., sensitivity and PPV are not the same thing). The "published" academic PPV (67%) suggests that two out of three of the cases scoring above threshold actually have bipolar disorder. However, the academic clinic had a much higher base rate of bipolar. Both PPV and NPV are algebraically linked to the base rate of the sample, meaning that estimates cannot generalize to settings with different base rates. EBM suggests a mnemonic to help clinicians interpret positive and negative results: SnNOut and SpPIn (Straus et al., 2011). On a Sensitive test, a Negative result rules the diagnosis Out (SnNOut). Conversely, on a Specific test, a Positive result rules the diagnosis In (SpPIn). Here, the published sensitivity is good, whereas the specificity is mediocre. Counterintuitively, a negative result (scoring 6 or lower) would be more decisive than a positive result: According to this mnemonic, it would meet the SnNOut criteria.

If the clinician had diagnostic X-ray vision, or a researcher could re-evaluate a representative set of cases to re-estimate the accuracy, the PGBI10M generalized fairly well, with good sensitivity (.87) and adequate specificity (.52). However, the accuracy of positive test results would be poor, PPV = .25. On the other hand, the NPV would be excellent: .97. This is what the SnNOut rule of thumb approximates. SnNOut does not quantify the accuracy, but it reminds the clinician that the negative result is the decisive one (especially combined with a low base rate), whereas positive results need further evaluation.

The probability nomogram or Naive Bayesian Model

Academic clinic. The academic clinic can directly estimate its base rate, because the research protocol gathered SDIs for all cases. Using a nomogram to estimate the predictive values would duplicate the PPV, because the combination of the prior probability (44%) and the DiLR attached to a high PGBI10M score [.92 sensitivity/(1 – .64 specificity) = 2.56] would produce the same estimate (91% probability of having bipolar). Using the nomogram with a low score would produce an estimate of a 9% chance of having bipolar disorder (44% base rate combined with a DiLR of 0.125), which is the converse of an NPV of .91 (i.e., a 91% chance that a person with a low score *does not* have bipolar, versus a 9% chance that the PGBI10M score was wrong, and he or she *does* have bipolar).

Community clinic. A clinician working in the community might be able to estimate a local base rate, either by querying the local medical records or by selecting a reasonable benchmark from published rates in similar settings. Using the 9% rate from local billing diagnoses combined with the published DiLRs yields a PPV of 20% for a high score and a bipolar probability of 1.2% for a low score (corresponding to an NPV of 98.8%).

All of the numbers used above are suboptimal for the local sample. The billing diagnoses were based on unstructured interviews by clinicians working under time pressure and pragmatically constrained to select diagnoses that would be reimbursable by Medicaid. The broad bipolar definition using the chart diagnoses happens to yield an estimate similar to the number based on research interviews of the families participating in the grant, but the agreement about which 9% to 13% of cases had bipolar disorder was only moderately better than chance. The DiLRs from the academic sample also might not be accurate, nor based on the optimal threshold, for the community clinic, given differences in severity of bipolar presentation, case mix, or demographic features.

Multilevel DiLRs and multiple assessments

Academic clinic. Using the six-level DiLRs published for the PGBI10M, combined with a 5× increase in the odds of bipolar if the patient has a first-degree relative with bipolar disorder defines 12 different predictions. Starting with a base rate of 44%, the lowest risk combination (PGBI10M score of 0 and no family history) gets updated to a probability of 0.8%, and the highest risk combination moves to 96.6%. The reader can confirm these estimates by using a nomogram, connecting the .44 to the DiLR of 5 for family history and then iterating with the 7.25 (or multiplying 5 × 7.25 = 36.25 for the combined high-risk DiLR).

Community clinic. Using the same published DiLRs but combining with the lower base rate at the community clinic results in a spread of revised probabilities ranging from 0.1% for the lowest risk segment to 84.1% for the highest risk combination using the 12.7% base rate (or 78.1% using the billing estimate of 9% prevalence).

Logistic regression—single and multiple predictors, traditional model building

Academic clinic. To show the potential advantages of logistic regression, we use the PGBI10M as a predictor by itself (to illustrate the more fine-grained prediction) and then build a second model combining it with family history, while also adjusting for age, race, and sex. The PGBI10M predicted bipolar diagnosis, $B = .21$, $p < .0005$, and Nagelkerke $R^2 = .46$. Adding the family history and youth sex, race, and age to the model further improved model fit, $p < .0005$, with the R^2 rising to .55. Family history and PGBI10M made significant incremental contributions to the model, both $p < .0005$; the demographic variables were not significant. Of note, the PGBI10M and family history correlated $r = .26$, $p < .0005$, consistent with concerns about the Naive Bayesian assumption of independence. Predicted probabilities ranged from 8.0% to 98.2% using the PGBI10M alone and 2.7% to 99.5% for the five-predictor model; the average predicted probabilities were 43.8% (the base rate in the academic sample).

Community clinic. If an enterprising researcher started a small business and made a smartphone application to apply the logistic regression results (or if an industrious clinician made a scoring spreadsheet), then it would be possible to take a patient's score on the PGBI10M and estimate his or her probability of a bipolar diagnosis using the published weights. Applying the academic sample weights to the community cases yields probability estimates spanning from 8.0% to 95.8%, with an average predicted probability of 33.6%. Using the out-of-sample weights is exactly what we do any time we use scoring software or norms. However, the typical research study does not have a rigorously defined representative sample, and so it is a larger leap to apply the weights in a new setting.

If the clinicians could re-evaluate the same predictors in the community mental health sample compared to an SDI criterion diagnosis, they would get different weights for every variable. For the simple model with the PGIB10M as the lone predictor, the R^2 is .21, and for the five-predictor model, the R^2 is only .22—less than half of the variance explained compared to the academic sample. Attempting to covary for the clinical or demographic variables that differed between the samples did not improve the fit in the community sample. The researcher would be mistaken to assume that

including the covariates in the regression model protected against problems in generalizability.

Using the local, optimized weights (which would normally not be available to a clinician) produces predicted probabilities ranging from 3.0% to 65.5% for the PGBI10M as a single predictor, and 2.4% to 67.5% for the five-predictor model, both with average probabilities of 12.7% (anchored to the sample's actual base rate). The correlation between the software's prediction and the locally optimized regression was $r = .85$, $p < .0005$, but there is a significant difference in the average probability, $p < .005$. This is a problem in model *calibration*. More on this below.

LASSO—a statistical learning model

Academic clinic. Table 1 summarizes the variables considered in the LASSO regression. We limited ourselves to two-way interactions, resulting in a matrix of 136 candidate predictors in a training sample of 550 cases. In principle, we could have made the analytic space even more complicated, going with higher order interactions, nonlinear transformations, and an even more extensive list of covariates. This is enough to give the flavor. Using 10-fold internal cross-validation and the tuning lambda within 1 standard error of the minimum (1se) rule to select the final model (James et al., 2013) resulted in a prediction equation that retained six predictors (see Table S1 in the Supplemental Material available online; versus a 28-predictor model using the model with the minimum lambda). Using these in the full academic sample produced probability estimates ranging from 6.9% to 98.0% for individual cases, with an M of 43.8% (the observed base rate). The AUC for these predictions was .902 (95% confidence interval [CI] = [.88, .93]), reaching the upper bound of .925 imposed by the accuracy of the criterion diagnosis and base rate (Kraemer, 1992). The more augmented model selected by the minimum lambda criterion had an AUC of .926—right at the upper bound, and also suggestive of overfitting.

Community clinic. Using the six-predictor model built by LASSO with the regression weights from the academic sample produces probability estimates ranging from 6.9% to 97.3% for individual cases in the community clinic, with $M = 36.6\%$. The AUC for the predictions compared to the SDI diagnoses (normally not available to the clinician) drops to .801 (95% CI = [.75, .85]). This is significantly lower than in the academic sample, $p < 2E^{-16}$ based on a Hanley and McNeil test, and the confidence interval is double what the internal validation in the academic data suggested. More alarmingly, the accuracy is also lower than the simpler models using the nomogram and two predictors (PGBI10M and family history) or even the PGBI10M alone in a regression.

Looking at the list of six predictors from the academic sample, only three are main effects (and the situation gets more extreme with the 28-predictor model: 26 are interactions). LASSO does not follow the convention of retaining all the main effects that build the interaction terms. In a purely predictive application, this may be fine (Breiman, 2001), but it is highly problematic from a model-building perspective (Cox, 2001), and the weights are hard to interpret from a theory-building perspective. This undermines one of the advantages of LASSO compared to "black box" machine-learning methods.

Flipping the script: What if LASSO had trained in the community setting? As an additional way of evaluating external cross-validation, we reversed the sequence we have been using so far and used LASSO to evaluate the same 136 candidate variables in the community data, again with 10-fold internal cross-validation and identical model selection procedures. This produced a model with an AUC of .830 (95% CI = [.79, .87]), indicating that the community data are harder to fit well than the academic sample (with a corresponding AUC of .902). Table 3 shows that, once again, the LASSO-generated model produces the best discrimination accuracy of any model in the data used to build it. But the community LASSO model keeps only one predictor (an interaction not included in the model built in the academic sample) instead of six. Using the minimum lambda algorithm does not improve consistency: That indicates 9 variables versus 28 in the academic version, and only 3 are the same in both models. If the researcher had started in the community, built a prediction model based on those data, and then applied those weights in the academic clinic, the AUC would be .864 (95% CI = [.83, .89])—again, similar or significantly lower than using simpler models (see Table 3; the 5 *df* logistic regression performed significantly better, $p = .03$). The LASSO model built in the community sample omits family history of bipolar disorder as a predictor, as well as any main effect for a PGBI scale, raising questions about the value of the statistical learning method from a variable discovery perspective, too. Simpler models that included these main effects performed better in terms of accuracy, whereas the LASSO method obscured their role by replacing them with a set of interaction terms.

Calibration

Another important accuracy metric is *calibration*, defined as the degree to which predicted probabilities match the actual "true" probabilities (e.g., Jiang, Osl, Kim, & Ohno-Machado, 2012; Spiegelhalter, 1986). For dichotomous outcomes (e.g., bipolar diagnosis), this "true" probability can be operationalized as the proportion of observed outcomes for a group of cases. For all

predictions of .30, a bipolar diagnosis should be confirmed 30% of the time. Calibration can be evaluated visually via calibration plots and using inferential statistics such as Spiegelhalter's (1986) z statistic. Calibration is a particularly important metric of accuracy for risk calculators (e.g., Hafeman et al., 2017) and other decision-support tools used in clinical settings.

The models were generally well calibrated for the "original" (i.e., model-building) datasets (i.e., academic weights in the academic sample, community weights in the community sample), sometimes referred to as "apparent calibration." Calibration plots for these models looked reasonable, with all data points falling on or close to the diagonal line (see Fig. 2) and mostly nonsignificant Spiegelhalter's z statistics (see Table 3).

External cross-validation was a different story. When the academic weights were used to predict bipolar disorder in the community sample, most models were poorly calibrated. The calibration plots show that the models generally overpredicted bipolar for predicted values above .20 (see Fig. 2). In an EBM framework, this would increase the number of cases in the assessment zone (yellow zone), leading to unnecessary follow-up evaluation but not necessarily to overly aggressive treatment (Youngstrom et al., 2017); the more complicated regression and the LASSO using academic weights could lead to overtreatment in the community sample as well. All models had either a significant Hosmer–Lemeshow χ^2 or Spiegelhalter's z statistic ($p < .05$, see Table 3), except for the nomogram model that applied local base rate.

Coda: What if we used clinical diagnoses to supervise LASSO training?

A final model used the diagnoses from the medical record at the community clinic to supervise a LASSO model. The same 136 candidate variables got considered in the $N = 511$ community cases, using the same 10-fold internal cross-validation and model tuning parameters. The 1se algorithm excluded all predictors, keeping only the intercept (collapsing to "bet the base rate"), and the minimum lambda algorithm model kept three predictors, none of which overlapped with either of the previous LASSO models. All were interactions; none were main effects. Using the clinical diagnoses to supervise the machine learning failed to recover any of the PGBI variables or family bipolar history as predictors, despite their extensive support in meta-analyses (Hodgins, Faucher, Zarac, & Ellenbogen, 2002; Youngstrom et al., 2015). However, the PGBI10M by itself would have been a valid predictor of the chart diagnoses, AUC = .675, $p < .0005$. Family bipolar history did not discriminate among chart diagnoses, AUC = .569, $p = .147$.

Discussion

The goal of this article was to compare a series of progressively more complicated approaches to evidence-based assessment and clinical decision-making. The first part of the article discussed both theoretical and pragmatic considerations, and the second part contrasted the empirical performance of the models. We used a vexing clinical challenge for the empirical case study: the accurate identification of pediatric bipolar disorder. We began with simple and familiar models, such as giving one of the best available screening tests, and then worked through newer and less widely implemented methods, such as using multilevel likelihood ratios and a probability nomogram to apply a Naive Bayesian algorithm, as recommended in EBM (Straus et al., 2011) and evidence-based assessment (Youngstrom et al., 2017). We contrasted these with logistic regression models using the screening test as a single predictor and then with a five-predictor model that combined two well-established predictors (the psychometric scale and family history of bipolar disorder) with three demographic covariates that vary to different extents across clinical settings. Finally, we used a statistical learning procedure, LASSO, to consider 136 candidate variables—including everything in the prior logistic regression, plus additional information about comorbidity, plus alternate ways of slicing the scores from the symptom questionnaire, plus second-order interactions among all of the above. We followed recommended practices, using 10-fold cross-validation to select the optimal model in terms of the bias-variance trade-off. We evaluated model accuracy in terms of AUC from ROC analyses and also calibration of the probabilities (Spiegelhalter's z). As hypothesized, the more complex models showed better discriminative accuracy.

Perhaps most importantly, though, we had a second independent sample with the same inputs and the same criterion diagnoses available but with markedly different demography and referral patterns. This provided an opportunity to examine *external* cross-validation, versus the more common practice of estimating generalizability using *internal* cross-validation methods. We used the data from the academic medical center as the first run and then evaluated the performance of the same model and "published" weights if applied in an urban community mental health center. This flow mimics the way that scientific research is supposed to propagate into clinical practice: A research team gathers a sample, evaluates models, and publishes recommendations (including cut scores, diagnostic accuracy estimates, or regression weights) that others are then supposed to apply in their clinics to new patients who were not in the original published sample. Unlike practicing clinicians, we were also able to re-evaluate the same

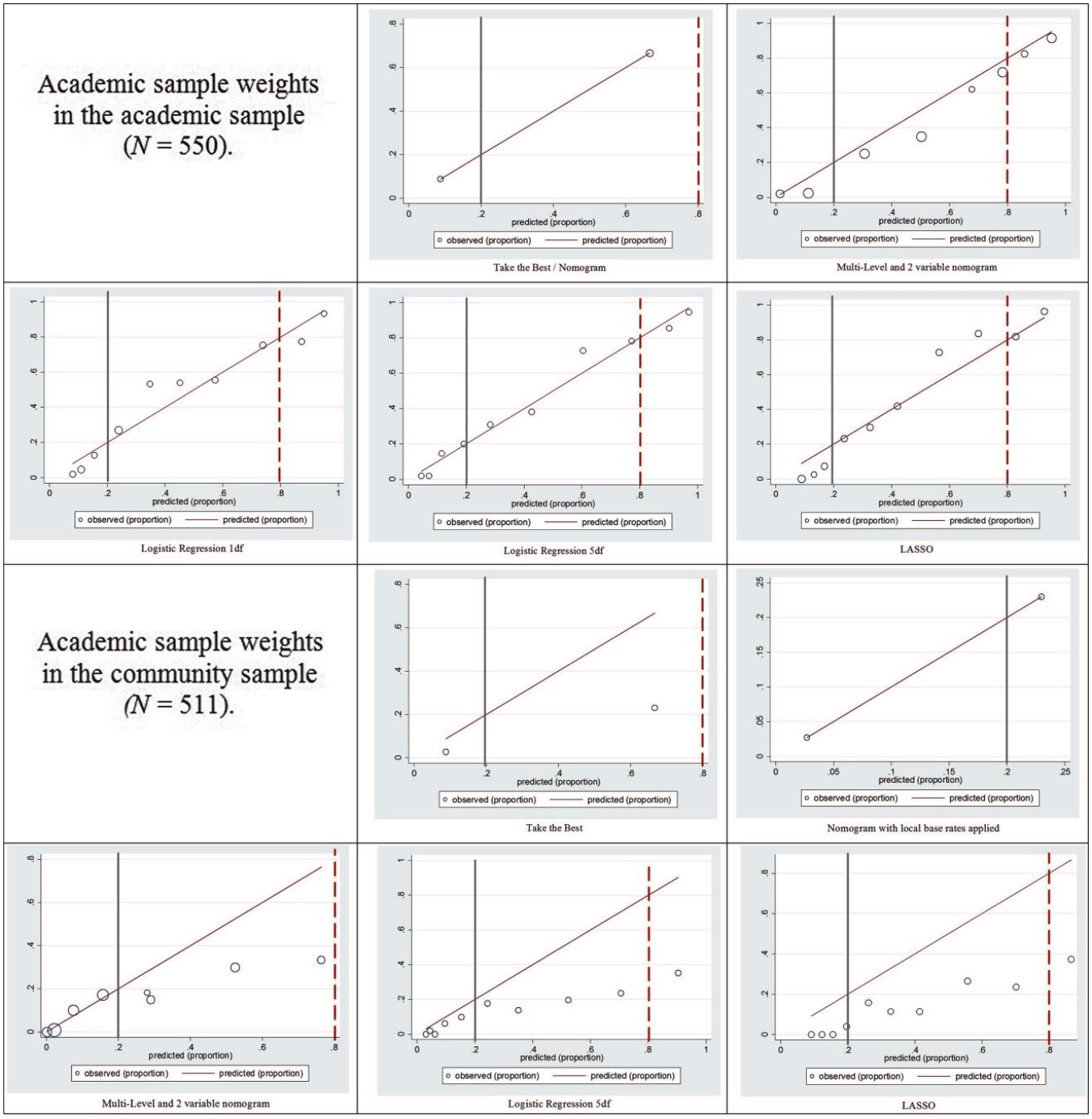

Fig. 2. Calibration plots. Perfect calibration is represented by the diagonal line. The vertical lines represent the wait-test decision threshold (solid green) and test-treat threshold (dashed red) in an evidence-based-assessment approach to clinical decision-making. The values of 20% and 80% were chosen as an approximation for heuristic purposes, not as a set threshold. Predicted probabilities below 20% would result in a clinician ruling out the diagnosis, "waiting" until other findings raised the probability again. Probabilities between 20% and 80% would be in the "assessment zone," suggesting more intensive evaluation, and probabilities above 80% would be in the "treatment zone." Most models overestimate the probability of bipolar disorder in external validation.

variables using the same statistical methods. Clinicians are not in a position to run LASSO regression on all their cases—at least not yet—even if the same data inputs were available. We framed this as the clinician having "research X-ray vision" as a way of reminding readers that this information is not usually available to clinicians, nor would it be available to end-users of statistical-learning algorithms unless the designers took extra steps to gather the criterion diagnosis using a research interview and then to re-evaluate the algorithm.

As the effect sizes comparing the samples show, the academic and community samples differed significantly on most variables considered, with the effect sizes spanning from small to very large. The performance of the prediction models was always worse in the community sample. Some degree of shrinkage is expected whenever there is out-of-sample cross-validation. We refit several of the models in the community data to provide a best case scenario for how the same model would have fared if the weights had been optimized to the community setting, and these were consistently less accurate than their counterparts built in the academic setting. The community data are fundamentally more challenging, probably due in part to the type of bipolar presentation seen there (with higher rates of mixed mood and higher comorbidity with anxiety, conduct, and PTSD). What is striking, though, is that the LASSO model showed by far the largest shrinkage, dropping from an AUC of .93 in the academic setting to .80 in the community—this despite following best practices for model cross-validation within the academic sample. In contrast, simpler models showed much less shrinkage.

Perhaps most concerning is the performance of LASSO from a variable discovery perspective. Given 136 candidate predictors, LASSO built a model with three main effects and three interactions in the academic sample and one interaction in the community data. None of the chosen variables were the same. LASSO failed to identify two of the most well-established predictors, family history of bipolar disorder or the PGBI, when built in the community data, despite them contributing to predictions in simpler models. This is the Rashomon problem: The model selected interaction terms and predictors with slight advantages in the current sample, with no recourse to the literature, theory, or established model-building heuristics (e.g., include main effects when examining interactions to improve interpretability) to increase the consistency across external validations. Most disappointing of all was the performance of a model built using the clinical diagnoses from the medical records to supervise the LASSO model training. The parsimonious version of the model excluded all predictors (turning into a "bet the base rate" model), and the more elaborate version still produced poor accuracy, did not identify any of the same predictors as the other two LASSO models, and failed to recover either of the well-established predictors included in the candidate list—family bipolar history (Smoller & Finn, 2003) and the PGBI (Youngstrom et al., 2015).

We examined a second aspect of accuracy, the calibration of the predictions based on each model. In general, calibration was good during the internal validation scenarios and problematic across all external validation scenarios unless the algorithm incorporated accurate information about the local base rate. Poor calibration has important consequences in clinical decision making. When the models were miscalibrated, the probability estimates were too high in the community setting (see Fig. 2). EBM talks about two decision thresholds for clinicians—whether to consider a diagnosis low enough probability that it is ruled out ("green zone") versus intermediate probability, requiring further assessment (yellow zone), and the assess versus treat threshold, past which the probability is high enough that the diagnosis is functionally confirmed and becomes a focus of intervention (red zone; see Youngstrom, 2013, for further elaboration). In this framework, the nomogram approach combined with local base rates would functionally rule out a bipolar disorder, putting the probability in the green zone, and high scores would move the probability into the yellow zone, warranting further assessment. Poorly calibrated models put too many cases into the assessment zone, leading to increased follow-up evaluation that would increase fiscal costs (Kraemer, 1992). The more complicated models, including logistic regression with five predictors and any version of LASSO, were the ones with calibration problems that misclassified cases into the treatment zone. Based on present findings, using the LASSO models would lead to overdiagnosis and overtreatment of bipolar disorder; the nomogram approach recommended by EBM would not.

Model complexity improves accuracy, to a point

Diagnosing bipolar disorder is challenging, and reviews of interrater agreement are sobering (e.g., Rettew et al., 2009). Using any of these methods could improve diagnostic accuracy for pediatric bipolar disorder. In terms of balancing feasibility, accuracy, and calibration, the Naive Bayes approach and the probability nomogram perform well—the accuracy remained good upon external cross-validation, and the calibration was good if there was reasonable information about the base rate. Given the high stakes associated with the diagnosis and the low rate of accuracy of current practices, it would be valuable to implement some of these methods. Because they were well calibrated if given reasonable base rate estimates, using these methods will not lead to overdiagnosis of bipolar and will improve decision-making about when bipolar can be considered "ruled out" versus when to follow up with more systematic (and ideally structured) interviewing (Youngstrom et al., 2017).

The traditional logistic regression and LASSO models clearly would require software to score them and provide the probability to the clinician. Both tended to be poorly calibrated (although the tuning algorithm selecting simpler models tended to show milder calibration issues),

and using weights from the academic sample in the community would lead to overdiagnosis despite good discrimination. Further, the apparent improvements in discrimination disappeared for LASSO in the external cross-validation scenarios compared to simpler models.

Model complexity creates barriers to implementation

Model complexity and lack of familiarity create barriers to implementation. How much can we change assessment practices without losing the clinician? Jaeschke et al. (1994) published the nomogram approach in *JAMA* in 1994. Subsequent books about EBM have consistently included it as a core method for medical diagnosis, but it is still unfamiliar to most psychologists. It being a visual, not an algebraic, approach could make it more appealing to people who go into health care professions, as they tend to be "people people," rather than "numbers people" (Gigerenzer, 2002). Logistic regression and anything more complicated will require a computer program to score and provide the feedback (Lindhiem, Yu, Grasso, Kolko, & Youngstrom, 2015). IBM's Watson is doing the computations and using probability dashboards as a way of providing decision support for the clinician. Even as these computer-assisted algorithms become more available, the nomogram will probably remain a good teaching tool, as well as being a robust interim option when the software is not available.

How robust will the more complicated algorithms be?

Initially, not very. When we flipped the script and used the community clinic to build the LASSO model, only three of the two dozen variables LASSO identified as important in the academic sample were picked for the model. It is worth noting that while the demography and referral patterns for the two settings were quite different, both were located in the same city, and both used the same criterion diagnostic methods (e.g., KSADS interviews conducted by the same research team). In those respects, the external cross-validation may further degrade if the model were applied in a different part of the county (e.g., consider Texas, with very different demography and language issues). As the contrast in performance between the academic and community samples vividly illustrates, sampling matters. It affects the generalization of models and algorithms across clinical settings. Settings may differ in ways both known and unknown. The first pass of statistical learning models is being applied to convenience samples and convenient sets of variables (Campbell & Stanley, 1963). These will not generalize well. Sample folding cross-validation, because it is taking the academic sample and using it to train and to test the model, cannot accurately forecast the size of the challenge when applied in a different setting, nor what the key moderators of performance will be.

The choice of criterion diagnosis is going to be especially challenging for machine-learning applications. Where the models have been applied with success, they are working with a much more objective criterion, such as defaulting on debt or predicting a purchase (Breiman, 2001). Even in these more concrete applications, the models change over time and require recalibration (Hoadley, 2001; Konig, Malley, Weimar, Diener, & Ziegler, 2007). Mental health diagnoses are a different order of complexity (Cronbach & Meehl, 1955; Meehl, 1954, 1973). Using clinical diagnoses to supervise statistical learning will encounter snares with local variation in training and practice. The Dartmouth Atlas of Health Care project consistently finds that geography is destiny in terms of diagnostic and treatment practices for other areas of medicine (Mulley & Wennberg, 2011); that will be even more the case when predicting mental health diagnoses. Consider using chart data from New York City or Paris (where psychoanalysis remains common) versus Boston or Seattle, where cognitive behavioral therapy might be the mode. Shifting to focus on more basic components of the medical record, such as textual analysis of the written notes or mining the other medical test results, will also initially stumble over the variations in practice—an analyst and a behaviorist will record different information. Furthermore, as Meehl (1957) noted, when something is not indicated in the chart, it is ambiguous whether it was not present in the case or just not recorded in the chart; statistical learning as a computerized chart review will be hampered by missing data that will not be missing at random. Geographic transportability of models is often a bigger challenge than other aspects of generalizability (Konig et al., 2007).

Pediatric bipolar disorder was an interesting test case because the bulk of the research on the topic was published after the current cohort of clinicians had completed their training (Goldstein et al., 2017). Other diagnoses will create major challenges for different reasons: Substance misuse and physical or sexual abuse will often be underestimated in clinical records because of issues of stigma and concerns about reporting. Conduct disorder will often be systematically underdiagnosed because third-party payers may not provide reimbursement for psychological services when it is used as a billing diagnosis. Both stigma and fiscal issues will add systematic bias to the billing diagnoses, heightening the need for caution if considering using these to supervise statistical learning methods.

This is not to say that statistical learning approaches should be abandoned. They clearly will be a helpful

tool for improving prediction of clinically important targets (Bone et al., 2016; Chekroud et al., 2016; Walsh et al., 2017), and they can integrate information sources that previously were difficult or impossible to use, such as using vocal acoustics and verbal content analysis to predict suicide attempt (Pestian et al., 2016). But relying on chart diagnoses to build the models will lead to some fundamental misspecification and challenges in generalization, as shown in the abysmal performance of the last LASSO model. Similar issues will apply when Google mines search history or test publishers mine their clouds of user-generated data. These large datasets lack a good mental health diagnostic criterion variable, and 23AndMe's efforts to enroll patients with mood disorder diagnoses is implicitly using diagnosis as usual, not a structured approach (although it will supplement the data with online surveys and cognitive tasks). Gathering high-quality SDIs and using well-established checklists and theoretically guided predictors will accelerate model development by providing more valid supervision of the algorithm and higher quality input. Alert model builders may find ways to add expert review (Brynjolfsson & McAfee, 2014) or otherwise refine the information gathered from charts (e.g., Walsh et al., 2017). Conversely, we should be cautious about generalizing complex models without strong evidence of external cross-validation in settings similar to where we are using them. When test publishers mine their cloud-based data and develop scoring routines that they then market to clinicians, it will be imperative that they gather an external sample with high-quality criterion diagnoses and publish the cross-validation results. It will be tempting to skip that step: It is costly (note that most technical manuals rely on chart diagnoses for the clinical validity studies, not SDIs), and it may not make the product look good. It remains an essential step for evaluating the validity of the assessment method (Buros, 1965; Konig et al., 2007).

Limitations

Although we covered an ambitious range of models, there are many other approaches to classification, and there are rapid advances in the area. Interested readers can find a good introduction and overview (Baumer et al., 2017) or more technical treatments if they want to actually apply some of the methods (James et al., 2013). We did not consider any of the "black box" approaches, such as nearest neighbor, deep neural networks, or support vector machines, which do not indicate which variables they are using to make the prediction. Though powerful in terms of prediction, these are even less helpful in terms of theory building. Similarly, we did not evaluate unsupervised methods, which use principal components analysis or clustering methods to develop ad hoc targets for prediction. These will obviously be sensitive to the choice and quality of indicators available, and the importance of comparing them to high-quality criterion diagnoses seems paramount. Because of space constraints and concerns about making the article too complicated, we did not model missing data. We want to emphasize that in real-life applications, missing data will be a major consideration, and they are not likely to be missing at random. We also only focused on one diagnosis. Clinicians need to consider multiple hypotheses; the hypotheses are not only competing explanations for symptom clusters or presenting problems, but patients frequently have comorbidity. Any reasonable complete approach to clinical diagnosis and decision support will need to handle algorithms for multiple diagnostic targets simultaneously and present clinicians with interpretable dashboards that summarize the information and guide next actions (Few, 2006; Powsner & Tufte, 1994).

Conclusions

Consistent with calls for adopting an evidence-based assessment approach (Norcross, Hogan, & Koocher, 2008; Straus et al., 2011; Youngstrom et al., 2017), any of the models investigated improve discrimination accuracy over typical clinical practice with regard to bipolar disorder. This is not impugning clinicians, but rather it indicates that there has been progress in research and that some clinically feasible algorithms showed external validity and could lead to big gains. Based on present results, a Naive Bayesian approach or software using relatively simple models with adjustments for calibration—using a small set of well-established predictors—currently appear to be the place to concentrate our training and upgrade our practices. The first iterations of statistical learning need to be treated with caution because (a) the differences in referral pattern and demography are more substantial than bootstrapping or folding can augur during internal cross-validation and (b) there are systematic biases in the conveniently available data that will distort the criterion variables used to supervise the machine learning. This will be important to keep in mind as companies with large convenience samples build and release scoring tools built with statistical learning models.

For the foreseeable future, a blended approach seems most productive. Researchers should continue to invest in high-quality diagnostic and indicator variables and pay heed to sampling issues. Clinicians should aim to use the best approach available for the question at hand, relying on Cochrane reviews and meta-analyses to help find the best available tool.

Relying on meta-analyses implicitly focuses on methods that have been trained and tuned across different settings. Naive Bayes and the probability nomogram (Straus et al., 2011) or simple calculators offer a big advance over current practice and may be the best option available until scoring applications have been rigorously externally cross-validated. The next wave of studies should include Naive Bayes using high-quality predictors as the incumbent model until a contender claims victory in an external dataset. Watson and Google will not replace the mental health diagnostician quickly, giving us time to figure out the best hybrid in our evolving profession (Susskind & Susskind, 2015).

Author Contributions

E. A. Youngstrom developed the study concept. E. A. Youngstrom, J. K. Youngstrom, and R. L. Findling supervised data collection. E. A. Youngstrom, T. F. Halverson, and O. Lindhiem planned and performed the analyses. E. A. Youngstrom, T. F. Halverson, and O. Lindhiem drafted sections of the manuscript. All authors provided critical revisions and approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported in part by NIH Grant R01 MH066647 (PI: E. Youngstrom) and a grant from the Stanley Medical Research Institute (PI: R. L. Findling).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2167702617741845.

Open Practices

All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/n9sd6/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2167702617741845. This article has received the badge for Open Data. More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382. doi:10.1177/0011000005285875.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Baumer, B. S., Kaplan, D. T., & Horton, N. R. (2017). *Modern data science with R*. Boca Raton, FL: Taylor & Francis.

Bertocci, M. A., Bebko, G., Olino, T., Fournier, J., Hinze, A. K., Bonar, L., . . . Phillips, M. L. (2014). Behavioral and emotional dysregulation trajectories marked by prefrontal-amygdala function in symptomatic youth. *Psychological Medicine*, *44*, 2603–2615. doi:10.1017/S0033291714000087

Bone, D., Bishop, S., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology & Psychiatry*. doi:10.1111/jcpp.12559

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–231.

Bruchmuller, K., Margraf, J., Suppiger, A., & Schneider, S. (2011). Popular or unpopular? Therapists' use of structured interviews and their estimation of patient acceptance. *Behavior Therapy*, *42*, 634–643. doi:10.1016/j.beth.2011.02.003

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York, NY: Norton.

Buros, O. K. (1965). Foreword. In O. K. Buros (Ed.), *The mental measurements yearbooks* (6th ed., p. xxii). Lincoln: University of Nebraska.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Carson, R. C. (1996). Aristotle, Galileo, and the DSM taxonomy: The case of schizophrenia. *Journal of Consulting and Clinical Psychology*, *64*, 1133–1139.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., . . . Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*. doi:10.1016/s2215-0366(15)00471-x

Cox, D. R. (2001). Comment: Statistical modeling: The two cultures. *Statistical Science*, *16*, 216–218.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*, 775–780. doi:10.1097/00001888-200308000-00003

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. Cambridge, MA: O'Reilly Press.

Findling, R. L., Youngstrom, E. A., McNamara, N. K., Stansbrey, R. J., Demeter, C. A., Bedoya, D., . . . Calabrese, J. R. (2005). Early symptoms of mania and the role of parental risk. *Bipolar Disorders*, *7*, 623–634.

Freeman, A. J., Youngstrom, E. A., Freeman, M. J., Youngstrom, J. K., & Findling, R. L. (2011). Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? *Journal of Child and Adolescent Psychopharmacology*, *21*, 425–432. doi:10.1089/cap.2011.0033

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York, NY: Simon and Schuster.

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

Gigerenzer, G., & Muir Gray, J. A. (Eds.). (2011). *Better doctors, better patients, better decisions*. Cambridge, MA: MIT Press.

Goldstein, B. I., Birmaher, B., Carlson, G. A., DelBello, M. P., Findling, R. L., Fristad, M., . . . Youngstrom, E. A. (2017). The International Society for Bipolar Disorders Task Force report on pediatric bipolar disorder: Knowledge to date and directions for future research. *Bipolar Disorders*. doi:10.1111/bdi.12556

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.

Guyatt, G. H., & Rennie, D. (Eds.). (2002). *Users' guides to the medical literature*. Chicago, IL: AMA Press.

Hafeman, D. M., Merranko, J., Goldstein, T. R., Axelson, D., Goldstein, B. I., Monk, K., . . . Birmaher, B. (2017). Assessment of a person-level risk calculator to predict new-onset bipolar spectrum disorder in youth at familial risk. *JAMA Psychiatry*. doi:10.1001/jamapsychiatry.2017.1763

Hoadley, B. (2001). Comment: Statistical modeling: The two cultures. *Statistical Science*, *16*, 220–224.

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.

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (3rd ed.). New York, NY: Wiley.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature: III. How to use an article about a diagnostic test: B: What are the results and will they help me in caring for my patients? *JAMA*, *271*, 703–707.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jenkins, M. M., & Youngstrom, E. A. (2016). A randomized controlled trial of cognitive debiasing improves assessment and treatment selection for pediatric bipolar disorder. *Journal of Consulting & Clinical Psychology*, *84*, 323–333. doi:10.1037/ccp0000070

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. doi: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. doi:10.1037/a0025775

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*, *70*, 158–168. doi:10.1037//0022-006x.70.1.158

Jensen-Doss, A., Osterberg, L. D., Hickey, J. S., & Crossley, T. (2013). Agreement between chart diagnoses and standardized instrument ratings of youth psychopathology. *Administration and Policy in Mental Health*, *40*, 428–437. doi:10.1007/s10488-012-0436-6

Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2014). Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *Journal of Consulting & Clinical Psychology*, *82*, 1151–1162. doi:10.1037/a0036657

Jiang, X., Osl, M., Kim, J., & Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, *19*, 263–274. doi:10.1136/amiajnl-2011-000291

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 and Adolescent Psychiatry*, *36*, 980–988. doi:10.1097/00004583-199707000-00021

Kim, E. Y., & Miklowitz, D. J. (2002). Childhood mania, attention deficit hyperactivity disorder and conduct disorder: A critical review of diagnostic dilemmas. *Bipolar Disorders*, *4*, 215–225.

Konig, I. R., Malley, J. D., Weimar, C., Diener, H. C., & Ziegler, A. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine*, *26*, 5499–5511. doi:10.1002/sim.3069

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

Lindhiem, O., Yu, L., Grasso, D. J., Kolko, D. J., & Youngstrom, E. A. (2015). Adapting the posterior probability of diagnosis index to enhance evidence-based screening: An application to ADHD in primary care. *Assessment*, *22*, 198–207. doi:10.1177/1073191114540748

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

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268–273.

Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). Minneapolis: University of Minnesota Press.

Meyers, O. I., & Youngstrom, E. A. (2008). A Parent General Behavior Inventory subscale to measure sleep disturbance in pediatric bipolar disorder. *Journal of Clinical Psychiatry*, *69*, 840–843. doi:ej07m03594 [pii]

Mulley, A. G., & Wennberg, J. E. (2011). Reducing unwarranted variation in clinical practice by supporting clinicians and patients in decision-making. In G. Gigerenzer & J. A. Muir Gray (Eds.), *Better doctors, better patients, better decisions* (pp. 45–52). Cambridge, MA: MIT Press.

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

Ong, M. L., Youngstrom, E. A., Chua, J. J., Halverson, T. F., Horwitz, S. M., Storfer-Isser, A., . . . Group, L. (2016). Comparing the CASI-4R and the PGBI-10 M for differentiating bipolar spectrum disorders from other outpatient diagnoses in youth. *Journal of Abnormal Child Psychology*. doi:10.1007/s10802-016-0182-4

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi:10.1126/science.aac4716

Orvaschel, H. (1994). *Schedule for affective disorders and schizophrenia for school-age children-epidemiologic version* (5th ed.). Ft. Lauderdale, FL: Nova Southeastern University.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Pestian, J. P., Sorter, M., Connolly, B., Bretonnel Cohen, K., McCullumsmith, C., Gee, J. T., . . . Group, S. T. M. R. (2016). A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial. *Suicide & Life Threatening Behavior*. doi:10.1111/sltb.12312

Powsner, S. M., & Tufte, E. R. (1994). Graphical summary of patient status. *The Lancet*, *344*, 368–389. doi:10.1016/S0140-6736(94)91406-0

Poznanski, E. O., Miller, E., Salguero, C., & Kelsh, R. C. (1984). Preliminary studies of the reliability and validity of the Children's Depression Rating Scale. *Journal of the American Academy of Child Psychiatry*, *23*, 191–197.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2012). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, *170*, 59–70. doi:10.1176/appi.ajp.2012.12070999

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. doi:10.1002/mpr.289

Shariat, S. F., Karakiewicz, P. I., Suardi, N., & Kattan, M. W. (2008). Comparison of nomograms with other methods for predicting outcomes in prostate cancer: A critical analysis of the literature. *Clinical Cancer Research*, *14*, 4400–4407. doi:10.1158/1078-0432.CCR-07-4713

Silver, N. (2015). *The signal and the noise: Why so many predictions fail-but some don't*. New York, NY: Penguin.

Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment*, *5*, 72–74.

Smoller, J. W., & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics*, *123C*, 48–58. doi:10.1002/ajmg.c.20013

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, *5*, 421–433.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, *24*, 399–411.

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.

Suppiger, A., In-Albon, T., Hendriksen, S., Hermann, E., Margraf, J., & Schneider, S. (2009). Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy*, *40*, 272–279. doi:S0005-7894(08)00088-9 [pii]

Susskind, R., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. New York, NY: Oxford University Press.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26. doi:10.1111/1529-1006.001

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.

U.S. Preventive Services Task Force. (2009). Screening for depression in adults: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, *151*, 784–792. doi:10.7326/0003-4819-151-11-200912010-00006

Van Meter, A., Moreira, A. L., & Youngstrom, E. A. (2011). Meta-analysis of epidemiological studies of pediatric bipolar disorder. *Journal of Clinical Psychiatry*, *72*, 1250–1256. doi:10.4088/JCP.10m06290

Van Meter, A. R., You, D. S., Halverson, T., Youngstrom, E. A., Birmaher, B., Fristad, M. A. . . . The Lams Group. (2016). Diagnostic efficiency of caregiver report on the SCARED for identifying youth anxiety disorders in outpatient settings. *Journal of Clinical Child and Adolescent Psychology*. doi:10.1080/15374416.2016.1188698

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*, 457–469. doi:10.1177/2167702617691560

Wikipedia. (n.d.). *General Behavior Inventory*. Retrieved from https://en.wikipedia.org/wiki/General_Behavior_Inventory

Wikiversity. (n.d.). *Evidence based assessment/Vignettes*. Retrieved from https://en.wikiversity.org/wiki/Evidence_based_assessment/Vignettes

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity, and sensitivity. *British Journal of Psychiatry*, *133*, 429–435.

Youngstrom, E. A. (2009). Definitional issues in bipolar disorder across the life cycle. *Clinical Psychology: Science & Practice*, *16*, 140–160. doi:10.1111/j.1468-2850.2009.01154.x

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

Youngstrom, E. A., Birmaher, B., & Findling, R. L. (2008). Pediatric bipolar disorder: Validity, phenomenology, and recommendations for diagnosis *Bipolar Disorders*, *10*, 194–214.

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 and Adolescent Psychiatry*, *44*, 712–717. doi:10.1097/01.chi.0000162581.87710.bd

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R. (2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment*, *13*, 267–276.

Youngstrom, E. A., Frazier, T. W., Findling, R. L., & Calabrese, J. R. (2008). Developing a ten item short form of the Parent General Behavior Inventory to assess for juvenile mania and hypomania. *Journal of Clinical Psychiatry*, *69*, 831–839. doi:10.4088/JCP.v69n0517

Youngstrom, E. A., Genzlinger, J. E., Egerton, G. A., & Van Meter, A. R. (2015). Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology*, *3*, 112–137. doi:10.1037/arc0000024

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., . . . Calabrese, J. R. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders*, *7*, 507–517. doi:10.1111/j.1399-5618.2005.00269.x

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry*, *60*, 1013–1019. doi:10.1016/j.biopsych.2006.06.023

Youngstrom, E. A., Murray, G., Johnson, S. L., & Findling, R. L. (2013). The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment*, *25*, 1377–1383. doi:10.1037/a0033975

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M., Ong, M.- L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science & Practice*. Advance online publication. doi:10.1111/cpsp.12207

Zhou, X.- H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley.