Published in final edited form as:

J Child Psychol Psychiatry. 2015 September ; 56(9): 936–948. doi:10.1111/jcpp.12442.

Thresholds and accuracy in screening tools for early detection of psychopathology

R. Christopher Sheldrick¹, James C. Benneyan², Ivy Giserman Kiss³, Margaret J. Briggs-Gowan⁴, William Copeland⁵, and Alice S. Carter³

¹Developmental-Behavioral Pediatrics, Tufts University School of Medicine, Boston, MA, USA

²Healthcare Systems Engineering Institute, Colleges of Engineering and Health Sciences, Northeastern University, Boston, MA, USA

³Department of Psychology, University of Massachusetts Boston, MA, USA

⁴Department of Psychiatry, University of Connecticut Health Center, CT, USA

⁵Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, NC, USA

Abstract

Background—The accuracy of any screening instrument designed to detect psychopathology among children is ideally assessed through rigorous comparison to ‘gold standard’ tests and interviews. Such comparisons typically yield estimates of what we refer to as ‘standard indices of diagnostic accuracy,’ including sensitivity, specificity, positive predictive value (PPV) and negative predictive value. However, whereas these statistics were originally designed to detect binary signals (e.g., diagnosis present or absent), screening questionnaires commonly used in psychology, psychiatry, and pediatrics typically result in ordinal scores. Thus, a threshold or ‘cut score’ must be applied to these ordinal scores before accuracy can be evaluated using such standard indices. To better understand the tradeoffs inherent in choosing a particular threshold, we discuss the concept of ‘threshold probability.’ In contrast to PPV, which reflects the probability that a child whose score *falls at or above* the screening threshold has the condition of interest, threshold probability refers specifically to the likelihood that a child whose score *is equal to* a particular screening threshold has the condition of interest.

Method—The diagnostic accuracy and threshold probability of two well-validated behavioral assessment instruments, the Child Behavior Checklist Total Problem Scale and the Strengths and Difficulties Questionnaire total scale were examined in relation to a structured psychiatric interview in three de-identified datasets.

Results—Although both screening measures were effective in identifying groups of children at elevated risk for psychopathology in all samples (odds ratios ranged from 5.2 to 9.7), children who

Correspondence: Developmental-Behavioral Pediatrics, Tufts University School of Medicine, Boston, MA, USA; rsheldrick@tuftsmedicalcenter.org.

Conflict of interest statement: No conflicts declared.

The authors have declared that they have no competing or potential conflicts of interest.

scored at or near the clinical thresholds that optimized sensitivity and specificity were unlikely to meet criteria for psychopathology on gold standard interviews.

Conclusions—Our results are consistent with the view that screening instruments should be interpreted probabilistically, with attention to where along the continuum of positive scores an individual falls.

Keywords

Assessment; screening; psychopathology; developmental psychopathology; methodology

Introduction

As many studies document, few children who experience psychopathology are detected as part of routine care, and fewer still receive appropriate treatment (Briggs-Gowan, Horwitz, Schwab-Stone, Leventhal, & Leaf, 2000; Sheldrick, Merchant, & Perrin, 2011; U.S. Department of Health and Human Services, 2000; Wang et al., 2005). Use of validated screening instruments is commonly recommended to improve identification of psychopathology in community settings. It is well-recognized that to be considered valid, screening instruments must display acceptable levels of accuracy in detecting psychopathology. Moreover, to be used effectively, it is critically important that the strengths and limitations of screening instruments be understood to inform implementation.

Psychologists, psychiatrists, pediatricians and others who work in healthcare or health sciences are generally familiar with concepts such as sensitivity and specificity. By extension, they are therefore familiar with the fundamental concepts used to evaluate diagnostic accuracy. However, a full understanding of diagnostic accuracy—which includes the ability to effectively communicate results—requires a high level of facility with these concepts. We therefore begin with an in-depth discussion of the basic premises of diagnostic accuracy. Our hope is that this introduction will serve as an in-depth review, while also providing a useful example for how these concepts can be communicated.

Review of diagnostic accuracy

There is strong consensus regarding the qualities of studies that offer the best evidence of screening accuracy. A rich literature (e.g., Jaeschke, Guyatt & Sackett, 1994a) and two detailed expert consensus statements, the *STARD* (Bossuyt et al, 2003) and the *QUADAS-2* (Whiting et al, 2011), detail criteria for conducting and reporting diagnostic accuracy studies. This ‘gold standard’ methodology prescribes comparison of screening results to ‘gold standard’ tests and structured interviews that are considered to be our best indices of psychopathology, typically yielding what we will hereafter refer to as the *standard* indices of diagnostic accuracy:

Sensitivity: the proportion of children with psychopathology who are correctly classified as having a disorder by the screening test

Specificity: the proportion of children without psychopathology who are correctly classified as not having a disorder by the screening test

Positive predictive value (PPV): the proportion of children with positive screening results who have psychopathology (and are therefore correctly classified)

Negative predictive value (NPV): the proportion of children with negative screening results who do not have psychopathology (and are therefore correctly classified)

Consistent with QUADAS-2 recommendations, these indices are usually presented together with a ‘four-fold’ or ‘ 2×2 ’ table of screening and criterion test results from which these statistics can be easily calculated. Figure 1 depicts such a table and reviews how all four indices are calculated.

For review, standard indices of diagnostic accuracy are not independent of one another. All represent ratios that rely on the same four cells within the 2×2 table, and all four indices include only correct classifications in their numerator: either true positives or true negatives. As many recall from introductory lessons on chi-square tests, 2×2 tables have only 3 degrees of freedom. Thus, if prevalence is known, knowledge of any two standard indices of diagnostic accuracy is sufficient to calculate all other indices. For example, if sensitivity, specificity, and prevalence are known, PPV can be readily calculated. It is therefore not possible to raise or lower PPV without altering at least one of the other indices. For a given prevalence, PPV cannot be changed without also changing sensitivity and/or specificity.

Standard indices of diagnostic accuracy and 2×2 tables are sufficient for describing the results of binary statistical tests. However, the results of most tests in psychology, psychiatry and pediatrics are not purely binary. Most produce ordinal raw scores that are transformed into binary results using a ‘cut score’ or ‘threshold.’ As an example, consider a screener with 20 true/false questions that yields raw scores from 0 to 20. A threshold might indicate that individuals who receive a score of 12 or higher are classified as screening positive, whereas those with a score of less than 12 are classified as screening negative. The values for standard indices of diagnostic accuracy would vary if the threshold on the same screening instrument was raised (e.g., to 13) or lowered (e.g., to 11). This is the basis of assessing diagnostic accuracy by applying receiver operating characteristic (ROC) curves, which depict sensitivity and (1 minus) specificity for every possible threshold on a screening instrument.

STARD and QUADAS-2 criteria make at least two references to screening tests that have ordinal results. First, they state that to achieve the highest quality evidence, screening thresholds should be stipulated before an accuracy study is conducted. Second, they state that the full distribution of screening scores, (i.e., not just a 2×2 table), should be presented for individuals who test positive and for those who test negative on the gold standard criterion against which a screening test is evaluated. Figure 2a offers a hypothetical example of what these distributions might look like. All individuals are assumed to be members of one of two populations: those with clinically significant psychopathology and those without. Inspection of this figure reveals that, on average, children with disorders receive higher scores on the screening test (as indicated on the horizontal axis) than do children without such disorders; however, the tails of the two distributions overlap considerably. Figure 2b presents the same distributions, except that the population with psychopathological disorders

has been flipped below the x-axis to avoid overlap in the figure, and one of the many possible thresholds has been added. Figure 2c is identical to Figure 2b, except that a higher threshold has been chosen.

Comparison of Figures 2b and 2c with Figure 1 reveals many similarities. Children who screen positive are depicted on the right, while those who screen negative are depicted on the left. Similarly, those with disorders are represented below the horizontal axis, while those without disorders are depicted above. Likewise, the meaning of indices like sensitivity, specificity, PPV and NPV has not changed. Note that figures 1b and 2c both include normal distributions and depict a population of children with a 15% prevalence of disorders. The threshold set in Figure 2b yields sensitivity = 73% and specificity = 73%, which is identical to the example in Figure 1. In contrast, the threshold in Figure 2c yields sensitivity=20% and specificity = 98%. This illustrates how moving the threshold along the screening score distribution can affect sensitivity and specificity.

There is also an important difference between Figures 2b and 2c—namely, the threshold is higher in Figure 2c than in Figure 2b. What is immediately obvious is that with the higher threshold in Figure 2c, fewer individuals will score positive on the screening test (reflected by the fact that there is less area under the curves to the right of the threshold). The effect on correct and incorrect classifications requires somewhat closer inspection. Any clinical threshold or cut score will yield correct classifications (i.e., true positives and true negatives), as well as errors (i.e., false positive and false negatives). As the threshold changes from Figure 2b to 2c, the proportions of both true positives (upper right quadrant) and false positives (lower right quadrant) fall, while the proportions of true negatives (upper left quadrant) and false negatives (lower left quadrant) rises.

Attending to these changes, one can discern the effect of raising a threshold on standard indices of diagnostic accuracy. Among children with disorders (depicted by the curve below the horizontal axis), those who are correctly classified are the ones to the right of the threshold (i.e., the shaded area). Thus, it is simple to visualize the fact that sensitivity, (i.e., the proportion of children with disorders who are correctly classified), falls as the threshold rises (in this case, from 73% to 20%). The scores of children without disorders are depicted by the curve above the horizontal axis, and those who are correctly classified as negatives are the ones to the left of the threshold. Thus, it is simple to visualize the fact that specificity, (i.e., the proportion of children without disorders who are correctly classified), rises as the threshold rises (in this case, from 73% to 98%).

Visualizing PPV, (i.e., the proportion of children with positive screening results who are correctly classified), requires even closer inspection. As the threshold rises from Figure 2b to 2c, the absolute number of both true positives and false positives decrease. Thus, both the numerator and the denominator in the equation for PPV decrease as the threshold rises. To visualize the effect, consider all children who screen positive with scores to the right of the thresholds in Figures 2b and 2c, whether they have disorders (and are below the horizontal axis) or do not have disorders (and are above the horizontal axis). Now consider the overall proportion who are correctly classified among those who screen positive—i.e., the true positives in the lower right quadrant. Clearly, the proportion of true positives relative to total

screened positives is greater in Figure 2c (where PPV=63%) than in 2b (where PPV=32%). Thus assuming normal distributions of scores, whenever the prevalence of the disorder identified by the screener is below 50%, PPV rises as the threshold rises.

To summarize, as thresholds rise, specificity and PPV also rise, but sensitivity falls. This dependency, or negative correlation, between sensitivity and specificity has long been recognized in meta-analyses of screening tests and assessment instruments, thus necessitating the use methods that account for this interdependency for accurate analyses (Trikalinos, Siebert, & Lau, 2009). It also highlights an important point about any screening test that has ordinal or continuous results: if higher sensitivity is desired, it can often be realized by lowering the threshold, albeit at the cost of lower specificity and PPV. Conversely, if higher PPV is required, it can often be realized by raising the threshold, albeit at the cost of lower sensitivity.

The concept of threshold probability

While sensitivity, specificity, PPV, and NPV offer a great deal of information about the diagnostic accuracy of a screening test, all are group-level statistics, describing the performance of a screening test across the full distribution of scores. To assist in making an informed decision about an individual patient who scored positive on a screening test, one would ideally know the probability that *that particular patient* had been correctly classified given the screening score. Clearly, there is significant heterogeneity among positive scores. For example, children with very high scores on a valid screening test are more likely to be affected by psychopathology than those who score at or near the threshold. Unfortunately, the clinical utility of PPV as a metric is limited in that it offers a weighted average probability across all individuals who score positive, regardless of their underlying score.

We therefore highlight an extremely useful (but less commonly known) statistic referred to as ‘threshold probability,’ which is prominently used in the field of decision analysis (Pauker & Kassirer, 1980; Vickers & Elkin, 2006), and in our case refers to the probability of being in the population affected by psychopathology for an individual whose score falls at the threshold. Figure 3a offers a visual depiction of a single threshold probability. Because standard indices of diagnostic accuracy apply to groups of individuals with a range of screening scores, they can be conceptualized by visualizing the area bounded by different curves. Because threshold probability refers to individuals who receive a single score on a screening instrument, it can be visualized by attending to the proportion of individuals who score at the threshold who are in the affected population—or, to be more specific, the length of the line reflecting the frequency of affected individuals compared to the total line indicating the frequency of scores at the threshold (see Figure 3a). Thus, if a screener displayed properties like those depicted in Figure 3a, 32% of all children who score positive would be expected to actually experience psychopathology (i.e., PPV=32%). However, only 15% of children scored at the threshold would be expected to experience psychopathology (i.e., threshold probability = 15%).

It is possible, and often preferable, to have more than one screening threshold for a given instrument. Different thresholds can correspond to different levels of risk for the condition of interest and each would be associated with different threshold probability. Taken a step

further, just as an ROC curve depicts sensitivity and specificity for each score on a screening instrument, it is possible to determine the threshold probability for each score on a screening instrument. For very low scores, we would expect the threshold probability to approach 0; for very high scores, we would expect the threshold probability to approach 100, and along the continuum, we would expect the full range of probabilities.

Comparable to the depiction of sensitivity and specificity for each possible screening score in an ROC curve, Figure 3b depicts the threshold probabilities for every possible score on the screening instrument depicted in the examples above. As reference values, the PPV is included, as is the probability of being in the affected population among those who screen negative (i.e., 1-NPV). Note that because PPV represents a weighted average for all individuals who score positive (i.e., everyone who scores above the threshold), its value is greater than threshold probability. The threshold probability at any one point along the distribution is determined by two factors: the accuracy of the screening instrument (in this case represented by the standardized difference between two normally distributed populations, which is fixed in this example), and the choice of where to place a threshold (which can vary anywhere along the curve as desired).

Threshold probability has strong implications for screening implementation in clinical settings. For example, if guidelines recommend an evidence-based screening instrument that in practice has a threshold probability of 15% (as is the case in the example depicted by Figure 3b), then clinicians must be willing and able to ensure that patients who screen positive and have a 15% chance of suffering from psychopathology receive treatment and/or further evaluation. If this is not the case, whether because of lack of training or lack of resources, then it is unlikely that evidence-based screening instruments will be used as recommended. Moreover, if clinicians choose to use a higher threshold than is recommended, sensitivity will be lower, and it is therefore unlikely that as many children with psychopathology will be detected as was initially expected.

Unfortunately, estimating threshold probability based on empirical data is not straightforward. Whereas calculation of standard indices of diagnostic accuracy requires only a 2×2 table, threshold probability requires an estimate of the distribution of screening scores among both healthy individuals and among those affected by psychopathological disorders. In this paper, we explore methods to analyze threshold probability in three large datasets. All three include structured diagnostic interviews conducted by interviewers trained to research reliability. All three also include well-regarded evidence-based screening instruments: the Strengths & Difficulties Questionnaire (SDQ), the Child Behavior Checklist (CBCL), or both. Thus, these datasets offer a unique opportunity to explore threshold probability for well-accepted behavioral screening instruments.

Methods

We sought to assess the psychometric properties of the total scales of two well-validated behavioral assessment instruments. The Total Problems scale of the CBCL (Achenbach & Rescorla, 2001), and the total scale of the SDQ (Goodman, 1997; Goodman, 2001; Goodman & Scott, 1999). With approval from two university Institutional Review Boards,

we analyzed three de-identified datasets, each of which included a structured psychiatric interview designed to assess psychopathology among children (i.e., a gold standard) as well as the CBCL and/or the SDQ.

Samples

We analyzed data from three samples:

1. The *National Comorbidity Study-Adolescent Supplement* (NCS-A) is a nationally-representative survey of adolescents aged 13–18 years (Mean = 15.2; SD= 1.5) in the continental U.S. that focuses on psychiatric disorders and related constructs. A sample of 6483 adolescents were selected from a dual sampling frame of households and schools and were interviewed face-to-face by trained lay interviewers using the World Health Organization (WHO) Composite International Diagnostic Interview Version 3.0 (CIDI) (Kessler & Üstün, 2004; Merikangas, Avenevoli, Costello, Koret, & Kessler, 1999). One parent or surrogate was mailed a questionnaire including demographic questions and the SDQ. The survey was administered by staff from the Institute for Social Research at the University of Michigan. Study procedures (Kessler et al., 2009a; Kessler et al., 2009b; Merikangas, Avenevoli, Costello, Koret, & Kessler, 1999); and analyses regarding the diagnostic accuracy of the SDQ (He, Burstein, Schmitz, & Merikangas, 2013) have been detailed in numerous publications.
2. The *North Carolina (NC)* sample included 636 adolescents aged 7-17 (Mean=12.8; SD=2.4) who were recruited from the Duke Pediatric Primary Care Clinics in Durham, North Carolina. Parents received five questionnaires including the CBCL and the SDQ. Because the purpose of the study was to compare several ‘gold standard’ psychiatric interviews, parents then completed two separate interviews over two visits. For the purposes of this study, all analyses were conducted using results from the Child and Adolescent Psychiatric Assessment (CAPA) (Angold & Costello, 2000). Further details regarding study procedures are reported elsewhere (Angold et al., 2012).
3. The *Greater New Haven (NH)* sample was part of a longitudinal representative population study of children selected from birth records provided by the State of Connecticut Department of Public Health (N=8404) and born healthy between July 1995 and September 1997 in the New Haven–Meriden Standard Metropolitan Statistical Area of the 1990 Census (n=1,329). Parents of a subsample enriched for child psychopathology (n=442; 77.6% response rate, 69.5% of eligible sample) were interviewed in the child's kindergarten or first-grade year with the Diagnostic Interview Schedule for Children, Version IV (DISC-IV) (Schaffer et al., 1996). Analyses in this study include 402 children aged 5-7 (Mean=6.0; SD=.3) whose parents completed the CBCL (Achenbach & Rescorla, 2001). Sampling weights were applied to adjust for selection and retention biases. Study details are reported elsewhere (Briggs-Gowan, Carter, Skuban, & Horwitz, 2001; Carter et al., 2010).

Thus, data were available to support four analyses: two samples were available for the CBCL and two samples were available for the SDQ. For the CAPA, the CIDI, and the DISC-IV, all diagnostic categories were included except specific phobia.

Analyses

Analysis of threshold probability necessitates estimation of the probability of having versus not having a diagnosis associated with specific scores on an assessment instrument. Because sampling variation can make point estimates associated with individual screening scores unstable, we sought to harness full sample data by estimating separate probability density functions (pdf) for children with diagnoses and for those without. Specifically, we attempted to fit normal, lognormal and Weibull curves to each distribution (including a shift parameter in the latter two cases), estimating all parameters via maximum likelihood, and then assessing goodness of fit using both Kolmogorov-Smirnov and chi square methods with a rejection criteria of $\alpha = 0.05$. Pdfs were converted to cumulative density functions (cdf) to assess sensitivity and PPV. Threshold probability (i.e., the probability of having a diagnosis conditioned on a particular screening score) was estimated using the equation $P_t = \text{pdf}_{\text{diagnosis}} / (\text{pdf}_{\text{diagnosis}} + \text{pdf}_{\text{no diagnosis}})$. For purposes of comparison, sensitivity and PPV were also analyzed directly from frequency tables using the standard non-parametric equations. Finally, after applying recommended thresholds to both the CBCL and the SDQ, we calculated risk ratios to reflect the relative risk of having a diagnosis conditional on screening results. To facilitate comparison with other studies, associated odds ratios (OR), which are more commonly reported, were also calculated.

Results

Observed frequencies and probability density functions are presented for the CBCL in upper portions of Figure 4 and for the SDQ in upper portions of Figure 5. Whereas normal distributions offered optimal fit for the CBCL scores among children with diagnoses in both the North Carolina ($\mu=57.6$, $\sigma=10.0$) and New Haven samples ($\mu=54.5$, $\sigma=10.3$), all other distributions displayed significant skew. As a result, Weibull distributions offered optimal fit for the CBCL scores of children without diagnoses in both the North Carolina ($\alpha=3.6$, $\beta=36.9$, $\text{shift}=-12.6$) and New Haven ($\alpha=3.9$, $\beta=29.8$, $\text{shift}=-16.8$) samples. Log-normal distributions offered optimal fit for all distributions of SDQ scores, including the scores of children with diagnoses in the North Carolina sample ($\mu=3.0$, $\sigma=0.3$, $\text{shift}=-8.5$) and the NCS-A ($\mu=3.2$, $\sigma=0.3$, $\text{shift}=-13.5$), and for SDQ scores of children without diagnoses in the North Carolina sample ($\mu=2.0$, $\sigma=0.5$, $\text{shift}=-3.2$) and the NCS-A ($\mu=2.0$, $\sigma=0.5$, $\text{shift}=-2.3$).

Both instruments displayed strong ability to stratify children according to risk (see Table 1 for risk ratios, threshold probabilities and standard indices of diagnostic accuracy). In the NCS-A, children scoring above the threshold on the SDQ were 2.1 times as likely to receive a diagnosis than were children who scored negative ($\text{OR} = 6.1$), whereas in the North Carolina sample, children were 2.4 times as likely to receive a diagnosis if they screened positive on the SDQ ($\text{OR} = 6.5$). Children who screened positive on the CBCL were 2.7

times as likely to receive a diagnosis in the North Carolina sample (OR = 6.5), and 5.1 times as likely to receive a diagnosis in the New Haven sample (OR = 9.7).

Despite substantial differences in prevalence, estimates of PPV and threshold probability were relatively consistent across samples; however, sensitivity was uniformly low. For the SDQ, PPV was estimated at 73.7% and 80.5%, threshold probability was estimated at 77.7% and 77.9%, and sensitivity was estimated at 19.2% and 19.1% in the North Carolina sample and NCS-A, respectively. For the CBCL, PPV was estimated at 69.8% and 52.5% and threshold probability was estimated at 56.9% and 40.8% in the North Carolina sample and New Haven sample, respectively. However, sensitivity was poor, estimated at 40.2% and 36.2%. This apparent consistency across samples may mask significant sources of uncertainty. For example, confidence intervals for estimates of PPV grow rapidly as thresholds rise and number of children with positive results falls (see Figure 4 and 5). Moreover, estimates of PPV actually fall for higher thresholds of the SDQ in both samples (see Figure 4), and estimates of threshold probability therefore follow a similar pattern. Uncertainty in estimates in the tails of distributions may be due to low sample size and the presence of outliers, similar to observations made in previous studies (Dujardin et al., 1994), or possibly for substantive reasons, such as if high scores detect the problems that are not identified by the gold standard (e.g., developmental disorders or parent psychopathology).

In an effort to account for sample specific differences in prevalence and screening score distributions, we also examined psychometric characteristics at thresholds determined post hoc using Youden's index. Such thresholds maximize the sum of sensitivity and specificity. For the SDQ, PPV was estimated at 57.5% and 66.6%, threshold probability was estimated at 42.5% and 54.2%, and sensitivity was estimated at 70.3% and 61.1% in the North Carolina sample and NCS-A, respectively. For the CBCL, PPV was estimated at 57.0% and 30.7%, threshold probability was estimated at 36.4% and 9.4%, and sensitivity was estimated at 72.6% and 72.4% in the North Carolina sample and New Haven sample, respectively. In comparison to recommended thresholds, these post hoc thresholds achieved lower levels of PPV and threshold probability but displayed desired levels of sensitivity.

Discussion

The results of this study confirm the utility of behavioral screening instruments for risk stratification. In each case, children who screened positive were approximately 2-5 times more likely than children who screened negative to qualify for a psychiatric diagnosis. However, the precise interpretation of screening result was highly dependent on the threshold chosen, and results were not uniform across samples. Moreover, thresholds that offered acceptable sensitivity also offered diminished PPV and threshold probabilities. These findings are consistent with previous publications that have noted that the probability of misclassification rises as scores approach the decision threshold (Hummel, 1999; Robins, 1985; Swets, Dawes & Monahan, 2000).

We again emphasize that for evidence-based screening instruments to be implemented with success and fidelity, resources must be sufficient for children who score positive to receive treatment or further assessment, regardless of whether they score at or well above the

threshold. It is important to remember that for a screening instrument with a given ROC curve, high sensitivity comes with a tradeoff—that is, comparatively low PPV and even lower threshold probability. We were struck by how low a threshold probability was required to achieve adequate sensitivity for two well-validated screening instruments. If clinicians have a similar perception, it is reasonable to hypothesize that they might not trust scores that fall close to the threshold. If so, clinicians may in effect shift the values of *standard indices of diagnostic accuracy* by implementing higher thresholds than recommended. Such a strategy would offer clinicians a higher PPV and threshold probability, but also lower sensitivity. This hypothesis is consistent with evidence that when pediatricians identify developmental and behavioral disorders in standard care, specificity far outweighs sensitivity (Sheldrick, Merchant, & Perrin, 2011). It is also consistent with several implementation trials of screening instruments where clinicians have referred or treated only a proportion of children who score positive (Guevara et al, 2013).

Moreover, while screening instruments often yield binary results—either ‘positive’ or ‘negative’—our findings underscore the importance of understanding screening results probabilistically. A particular threshold or score on a screening instrument defines groups of children who range in their risk of experiencing psychopathology. While we can say with confidence that children who score positive on validated screening instruments are at much higher risk for having psychopathology than are children who score negative, specifying the probabilities is difficult given variation across samples. Thus, in making decisions when using validated screening instruments, clinicians may confront the fact that not only are the predictive validities lower than they might prefer, but that there is also considerable uncertainty regarding the validity of estimates of diagnostic accuracy of the screening tools in their own setting. Many have persuasively argued that over-reliance on clinical intuition often leads to error, and that identification of psychopathology can be improved through use of valid screening instruments over clinical intuition alone (Dawes, Faus, & Meehl, 1989). However overconfidence in the course of standard clinical decision making is also a primary source of medical error (Berner & Graber, 2008), and failure to appreciate the likelihood of error in screening results and misinterpretation of their probabilistic nature creates the risk of over-valuing the information screening instruments provide. There is a rich and diverse literature on the interpretation of objective tests and their incorporation with clinical findings (e.g., Hummel, 1999), and opinions regarding the relative value of objective ‘actuarial’ methods versus clinical intuition vary widely across scientific fields (Swets, Dawes & Monahan, 2000). In medical and mental health fields, many have argued for an evidence-based approach that uses multiple thresholds (each associated with different diagnostic likelihood ratios) to account for variable predictive values among positive screening results (Jaeschke, Guyatt & Sackett, 1994; Straus, Glasziou, Richardson & Haynes, 2011; Youngstrom, 2013; Youngstrom, 2014; Youngstrom, Choukas-Bradley, Calhoun & Jensen-Doss, 2014). In addition, we argue that training that includes information about the tradeoffs between threshold probabilities and sensitivity might counter the rational use of higher than recommended screening thresholds.

It is important to note many limitations to this study and to the evaluation of diagnostic accuracy of behavioral screening instruments more generally. First and foremost, there are

both theoretical and practical limits to the evaluation of the accuracy of screening instruments. Although they are commonly referred to as ‘gold standards,’ the structured diagnostic interviews by which we assess psychopathology in children lack perfect reliability. This fact is commonly reflected by inter-rater reliabilities with moderate kappa scores, and by modest retest reliabilities (Angold et al., 2012). For this reason, they are described in the statistical literature as examples of ‘fuzzy’ gold standards, and there exists a large literature on this topic (Kraemer, 1992; Pepe, 2003; Phelps & Huston, 1995; Zhou, Obuchowski & McClish, 2002). While the imperfect nature of structured diagnostic interviews is widely recognized, its influence on evaluations of screening accuracy is not. It is well established that less-than-perfect reliability in an independent variable attenuates the magnitude of the correlation that will be observed with a dependent variables. Thus, if a diagnostic interview displays a reliability of r_{xx} , the maximum correlation that should be expected with a screening instrument that measures the same construct is $\text{SQRT}(r_{xx})$. However, imperfect reliability among reference tests can also inflate associations with screening instruments. In particular, this can occur if the error variances of two tests are correlated (Trikalinos, Siebert, & Lau, 2009), for example if both a diagnostic test and a screening instrument tend to make the same errors for the same reasons. Many screeners and interview rely on the same, single, informant. In this scenario, any response bias on the part of the parent will influence the results of both tests, thus inflating their association for reasons that have nothing to do with the child's experience of psychopathology. While accounting for these biases analytically is difficult at best, a recent study offers practical guidance. In a study in which Angold and colleagues assessed the accuracy with which each of three different ‘gold standard’ interviews for psychopathology in children detected the other, the most accurate example was consistent with a sensitivity and specificity of approximately 85% (Angold et al., 2012). If this is the highest accuracy with which one structured interview can detect the results of another, then it may be reasonable to consider these values as a practical limit on the observed accuracy of screening instruments.

Other limitations are also apparent. As is clear from our results, the precise shapes of distributions of screening scores cannot be assessed with certainty. This finding is consistent with prior investigations that note differences in test score distributions across populations (Brenner & Gefeller, 1997; Dans et al., 1998; Moons & Karel et al., 1997; Sox et al., 1990). Thus, analyses of threshold probabilities are prone to error, and sensitivity analyses are recommended to specify a range of plausible values. Moreover, threshold probability and PPV are both directly influenced by underlying prevalence. Thus, differences in prevalence suggested by the use of different ‘gold standard’ diagnostic tools, or the use of different impairment criteria within the same ‘gold standard’ diagnostic tool, will influence results. Moreover, attention to differences in prevalence between local populations and standardization samples is critical. For example, prevalence might be expected to be lower in general pediatrics than in a specialty referral clinic, and this will influence both PPV and threshold probability.

Overall, analysis of standard indices of diagnostic accuracy in general and of threshold probability in particular has significant limitations. Despite these limitations, the concept of threshold probability offers a framework for rationally considering and evaluating optimal screening thresholds. This is in contrast to standard approaches, which often rely on

arbitrary standards that do not consider the costs and benefits of available assessment or intervention options. For example, in recent years, 70% has often been referred to as a minimally acceptable value for sensitivity. However, there is evidence to suggest that standards for the accuracy of screening instruments have fallen over time, perhaps in implicit recognition of the limits on sensitivity and specificity in studies of accuracy. For example, in 1974, a review of pediatric screening instruments commented that the Denver Developmental Screening Test displayed ‘low sensitivity (81%) and specificity (77%), with a resultant large problem of overreferral’ (Bailey et al, 1974). Today, a screening test with this level of accuracy would be widely accepted as falling well within commonly accepted limits. However, neither standard makes an informed argument to support any given threshold.

To make informed choices about screening thresholds, it is useful to understand their relationship to cost-benefit analysis. Setting a threshold represents a decision, and this decision reflects the perceptions and values of the person or group of people setting the threshold. Any given threshold implies that for all children who score above it, the expected benefits of being classified as positive outweigh the expected costs. Conversely, it implies that for all children who score below the threshold, expected costs of being classified as positive outweigh expected benefits. Expected costs and benefits result directly from the consequences of classification—for example, the amount and type of treatment and/or further assessment that a ‘positive’ label triggers. Costs and benefits may include multiple factors, including expected changes in symptoms associated with available interventions, economic costs associated with interventions and maintenance or exacerbation of symptoms, risk of side effects, stigma, or family distress, and impact on child and family quality of life (Godoy & Carter, 2013). How individuals weigh these multiple factors can vary considerably and is dependent on their perspectives and values. Thus, decisions about where and how to set clinical thresholds reflect both perceptions of consequences and value judgments regarding the costs and benefits of identification, diagnosis, and treatment.

Assuming that quantitative estimates of costs and benefits are both valid and available, then threshold probability has a straightforward interpretation in cost-benefit analysis. Screening instruments classify children according to risk for psychopathology. As described above, for children who score positive, the risk or probability of having a disorder is considered to be high enough that the expected benefits of treatment or further evaluation outweigh the expected costs. For children who score negative, the probability of having a disorder is low enough that the expected costs of treatment or further evaluation exceed the expected benefits. Therefore, for children who score precisely at the threshold, the probability of having a disorder is such that expected costs and benefits are equivalent (Pauker & Kassirer, 1980). Thus, a threshold and its associated threshold probability can reveal important information about the perceptions of costs and benefits among those who use or recommend it. Moreover, this framework provided an explicit rationale for setting multiple thresholds (Pauker & Kassirer, 1980). For example, one threshold could be set to reflect the costs and benefits of further assessment, whereas a more stringent threshold might be set to reflect the costs and benefits of typical interventions. Others have methods to optimize thresholds that are functionally similar (Swets, Dawes & Monahan, 2000) or have suggested reasonable approaches to setting thresholds when complete information on costs and benefits are

unavailable, as is typical in most cases (Kraemer, 1992; Straus et al., 2011). To conclude, we recommend further research on the psychometric properties of instruments and the utility of using different thresholds with a particular focus on external validity across samples. In addition, we recommend close attention to the best available evidence on diagnostic accuracy and cost effectiveness when setting and conceptualizing screening thresholds. As important, we recommend even closer attention to what those data do—and do not—mean. Results of even the best screening instruments should be interpreted as probabilities, not as definitive findings, and even our understanding of those probabilities is subject to uncertainty. Thus, we argue that if empirically-validated screening instruments are to fulfill their potential for improving the detection of childhood psychopathology, understanding the limits of evidence for diagnostic accuracy is at least as important as understanding its strengths.

Acknowledgments

This original article was invited by the journal as part of a special issue, it has undergone full, external peer review. This study was made possible in part by funding from the National Institute of Mental Health (R01MH104400).

References

- Achenbach, TM.; Rescorla, LA. Manual for the ASEBA school-age forms & profiles. University of Vermont, Research Center for Children, Youth, and Families; Burlington: 2001.
- Angold A, Costello EJ. The Child and Adolescent Psychiatric Assessment (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry*. 2000; 39(1):39–48. [PubMed: 10638066]
- Angold A, Erkanli A, Copeland W, Goodman R, Fisher PW, Costello EJ. Psychiatric diagnostic interviews for children and adolescents: A comparative study. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2012; 51(5):506–517. [PubMed: 22525957]
- Bailey EN, Kiehl PS, Akram DS, Loughlin HH, Metcalf TJ, Jain R, Perrin JM. Screening in pediatric practice. *Pediatric Clinics of North America*. 1974; 21(1):123–165. [PubMed: 4590155]
- Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*. 2008; 121(5):S2–S23. [PubMed: 18440350]
- Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in medicine*. 1997; 16(9):981–991. [PubMed: 9160493]
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HCW, Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*. 2003; 138(1):W1–12. [PubMed: 12513067]
- Briggs-Gowan MJ, Carter AS, Skuban EM, Horwitz SM. Prevalence of social-emotional and behavioral problems in a community sample of 1- and 2-year-olds. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2001; 40(7):811–819. [PubMed: 11437020]
- Briggs-Gowan MJ, Horwitz SM, Schwab-Stone ME, Leventhal JM, Leaf PJ. Mental health in pediatric settings: Distribution of disorders and factors related to service use. *Journal Of The American Academy Of Child & Adolescent Psychiatry*. 2000; 39(7):841–849. [PubMed: 10892225]
- Carter AS, Wagmiller RJ, Gray SAO, McCarthy KJ, Horwitz SM, Briggs-Gowan MJ. Prevalence of DSM-IV disorder in a representative healthy birth cohort at school entry: Sociodemographic risks and social adaptation. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2010; 49(7):686–98. [PubMed: 20610138]
- Dans AL, Dans LF, Guyatt GH, Richardson S, Evidence-Based Medicine Working Group. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. *JAMA*. 1998; 279(7):545–549. [PubMed: 9480367]

- Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science*. 1989; 243(4899):1668–1674. [PubMed: 2648573]
- Dujardin B, Van den Ende J, Van Gompel A, Unger JP, Van der Stuyft P. Likelihood ratios: a real improvement for clinical decision making?. *European journal of epidemiology*. 1994; 10(1):29–36. [PubMed: 7957786]
- Godoy L, Carter AS. Identifying and addressing mental health risks and problems in primary care pediatric settings: A model to promote developmental and cultural competence. *American Journal of Orthopsychiatry*. 2013; 83(1):73–88. [PubMed: 23330625]
- Goodman R, Scott S. Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology*. 1999; 27:17–24. [PubMed: 10197403]
- Goodman R. The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*. 1997; 38:581–586. [PubMed: 9255702]
- Goodman R. Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry*. 2001; 40:1337–1345. [PubMed: 11699809]
- Guevara JP, Gerdes M, Localio R, Huang YV, Pinto-Martin J, Minkovitz CS, Pati S. Effectiveness of developmental screening in an urban setting. *Pediatrics*. 2013; 131(1):30. [PubMed: 23248223]
- He J, Burstein M, Schmitz A, Merikangas KR. The Strengths and Difficulties Questionnaire (SDQ): The factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology*. 2013; 41(4):583–595. [PubMed: 23183936]
- Hummel, TJ. The usefulness of tests in clinical decisions.. In: Lichtenberg, JW.; Goodyear, RK., editors. *Scientist-practitioner perspectives on test interpretation*. Allyn & Bacon; 1999. p. 59-112.
- Jaeschke R, Guyatt GH, Sackett DL, Guyatt G, Bass E, Brill-Edwards P, Wilson M. 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*. 1994; 271(9):703–707.
- Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, Zaslavsky AM. National Comorbidity Survey Replication Adolescent Supplement (NCS-A): II. Overview and design. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2009a; 48(4):380–385. [PubMed: 19242381]
- Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*. 2004; 13(2):93–121. [PubMed: 15297906]
- Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, Zaslavsky AM. Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). *International Journal of Methods in Psychiatric Research*. 2009b; 18(2):69–83. [PubMed: 19507169]
- Kraemer, HC. *Evaluating medical tests: Objective and quantitative guidelines*. Sage publications; Newbury Park, CA: 1992.
- Merikangas KR, Avenevoli S, Costello EJ, Koretz D, Kessler RC. National Comorbidity Survey Replication Adolescent Supplement (NCS-A): I. Background and measures. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2009; 48(4):367–379. [PubMed: 19242382]
- Moons KG, van Es GA, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology*. 1997:12–17. [PubMed: 9116087]
- Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *The New England Journal of Medicine*. 1980; 302(20):1109–1117. [PubMed: 7366635]
- Pepe, MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press; 2003.
- Phelps CE, Hutson A. Estimating diagnostic test accuracy using a 'fuzzy gold standard'. *Medical Decision Making*. 1995; 15(1):44–57. [PubMed: 7898298]
- Robins LN. Epidemiology: reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry*. 1985; 42(9):918–924. [PubMed: 3899050]

- Schaffer D, Fisher P, Dulcan MK, Davies M, Piacentini J, Schwab-Stone ME. The NIMH Diagnostic Interview Schedule for Children, Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA study. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1996; 35:865–877. [PubMed: 8768346]
- Sheldrick RC, Merchant S, Perrin EC. Identification of developmental-behavioral problems in primary care: A systematic review. *Pediatrics*. 2011; 128(2):356–363. [PubMed: 21727101]
- Sox HC, Hickman DH, Marton KI, Moses L, Skeff KM, Sox CH, Neal EA. Using the patient's history to estimate the probability of coronary artery disease: a comparison of primary care and referral practices. *The American Journal of Medicine*. 1990; 89(1):7–14. [PubMed: 2242131]
- Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. *Psychological science in the public interest*. 2000;1–26. [PubMed: 26151979]
- Straus, SE.; Glasziou, P.; Richardson, WS.; Haynes, RB. Evidence-based medicine: How to practice and teach EBM. 4th ed.. Churchill Livingstone; New York, NY: 2011.
- Trikalinos TA, Uwe S, Joseph L. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. *Medical Decision Making*. 2009
- U.S. Department of Health and Human Services. Mental Health: A Report of the Surgeon General. US Government Printing Office; Washington, DC: 2000.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006; 26(6):565–574. [PubMed: 17099194]
- Wang PS, Lane M, Olfson M, Pincus HA, Wells KB, Kessler RC. Twelve-Month Use of Mental Health Services in the United States: Results From the National Comorbidity Survey Replication. *Archives Of General Psychiatry*. 2005; 62(6):629–640. [PubMed: 15939840]
- Whiting PF, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*. 2011; 155(8):529–536. [PubMed: 22007046]
- Youngstrom EA. Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology*. 2013; 42(1):139–159. [PubMed: 23153181]
- Youngstrom EA. A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*. 2014; 39:204–221. [PubMed: 23965298]
- Youngstrom EA, Choukas-Bradley S, Calhoun CD, Jensen-Doss A. Clinical guide to the Evidence-Based Assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*. 2014; 22(1):20–35.
- Zhou, XH.; McClish, DK.; Obuchowski, NA. Statistical methods in diagnostic medicine. Vol. 569. John Wiley & Sons; 2009.

Key points

- Questionnaires designed to screen for psychopathology typically yield ordinal scores. One or more clinical thresholds or ‘cut scores’ is typically applied to yield ‘positive’ and ‘negative’ screening results.
- ‘Threshold probability’ refers to the likelihood that a child whose score is equal to a particular clinical threshold experiences psychopathology.
- Despite being effective in identifying groups of children at elevated risk for psychopathology, two well-validated screening instruments displayed low threshold probabilities—i.e., children who scored at or near the recommended clinical thresholds were unlikely to meet criteria for psychopathology on gold standard interviews.
- Screening instruments should be interpreted probabilistically, with attention to where along the continuum of positive scores an individual falls. Threshold probabilities offer a useful metric to evaluate and optimize clinical thresholds.

2 x 2 Table for Hypothetical Screener

		Screening Results			
		-	+		
Diagnosis	-	TN	FP	TN+FP	→ Specificity = $\frac{TN}{(TN+FP)}$
	+	FN	TP	FN+TP	→ Sensitivity = $\frac{TP}{(FN+TP)}$
		$\frac{TN+FN}{(TN+FN)} \downarrow NPV = \frac{TN}{(TN+FN)}$		$\frac{FP+TP}{(FP+TP)} \downarrow PPV = \frac{TP}{(FP+TP)}$	

For example:

		Screener			
		-	+		
Diagnosis	-	62	23	85	Sensitivity=73% Specificity=73% PPV=32% NPV=94%
	+	4	11	15	
		66	34		

Note: ■■ = correctly classified

Figure 1.

2 × 2 Table for Hypothetical Screener

Note: ■■ = correctly classified

Figure 2a.
Hypothetical distributions of screening scores

Figure 2b.
Application of a screening threshold

Figure 2c.
Affect of a higher screening threshold

Figure 3a.
Visualizing threshold probability

Figure 3b.
 Threshold probability & PPV

Note: CBCL = Child Behavior Checklist; pdf = probability density function; PPV = Positive predictive value. Vertical lines represent recommended screening thresholds. In the **upper figures**, bars indicate observed frequencies, and lines indicate pdfs that have been fitted to these frequencies. In the **lower figures**, threshold probability has been calculated for each possible screening score based on the pdfs. Sensitivity and PPV have been calculated based on frequency tables of observed results. The area shaded gray represents the 95% confidence interval of the PPV.

Figure 4.
Child Behavior Checklist (CBCL) results

Note: SDQ = Strengths & Difficulties Questionnaire; pdf = probability density function; PPV = Positive predictive value. Vertical lines represent recommended screening thresholds. In the upper figures, bars indicate observed frequencies, and lines indicate pdfs that have been fitted to these frequencies. In the lower figures, threshold probability has been calculated for each possible screening score based on the pdfs. Sensitivity and PPV have been calculated based on frequency tables of observed results. The area shaded gray represents the 95% confidence interval of the PPV.

Figure 5.
Strengths & Difficulties Questionnaire (SDQ) results

Table 1

Standard indices of diagnostic accuracy for CBCL and SDQ across samples

True Positives (n)	True Negatives (n)	False Negatives (n)	Relative Risk Ratio	95% CI	Odds Ratio	95% CI	Sensitivity (%)	95% CI (%)	Specificity (%)	95% CI (%)	PPV (%)	95% CI (%)	Estimate of Threshold Probability (%)
RECOMMENDED SCREENING THRESHOLDS													
38	367	131	2.7	2.2 - 3.2	6.5	4.2 - 10.0	40.2	33.6 - 47.0	90.6	87.3 - 93.3	69.8	61.0 - 77.7	56.9
19	325	37	5.1	3.4 - 7.9	9.7	4.8 - 19.7	36.2	24.0 - 49.9	94.5	91.5 - 96.6	52.5	36.1 - 68.5	40.8
15	389	177	2.4	1.9 - 2.9	6.2	3.3 - 11.4	19.2	14.2-25.0	96.3	94.0-97.9	73.7	60.3-84.5	77.7
130	3554	2266	2.1	2.0 - 2.2	6.5	5.3 - 7.9	19.1	17.7-20.6	96.5	95.8-97.0	80.5	77.3 - 83.4	77.9
SAMPLE-SPECIFIC SCREENING THRESHOLDS													
120	285	60	3.3	2.5 - 4.2	6.3	4.4 - 9.1	72.6	66.2 - 78.4	70.4	65.7 - 74.8	57.0	51.0 - 62.9	36.4
95	249	16	5.1	3.0 - 8.7	6.9	3.7 - 12.8	72.4	59.1 - 83.3	72.4	67.3 - 77.0	30.7	23.5-38.8	9.4
114	290	65	3.1	2.5-4.0	6.0	4.2-8.7	70.3	63.8-76.3	71.8	67.1-76.1	57.5	51.3-63.5	42.5
858	2826	1090	2.4	2.3-2.5	5.2	4.6-5.8	61.1	59.3-62.9	76.7	75.3-78.1	66.6	64.8-68.4	54.2

NC = North Carolina sample; NCS-A = National Comorbidity Study – Adolescent Supplement; NH = New Haven sample; SDQ = Strengths & Difficulties