



Published in final edited form as:

J Consult Clin Psychol. 2014 December ; 82(6): 1151–1162. doi:10.1037/a0036657.

Predictors and Moderators of Agreement between Clinical and Research Diagnoses for Children and Adolescents

Amanda Jensen-Doss,

University of Miami

Eric A. Youngstrom,

University of North Carolina at Chapel Hill

Jennifer Kogos Youngstrom,

University of North Carolina at Chapel Hill

Norah C. Feeny, and

Case Western Reserve University

Robert L. Findling

Johns Hopkins University

Abstract

Objective—Diagnoses play an important role in treatment planning and monitoring, but extensive research has shown low agreement between clinician-generated diagnoses and those from structured diagnostic interviews. However, most prior studies of agreement have not used research diagnoses based on gold standard methods, and research needs to identify characteristics of diagnostically challenging clients. This study examined agreement between youth diagnoses generated through the research-based LEAD (Longitudinal, Expert, and All Data) Standard to clinician diagnoses.

Method—Participants were 391 families seeking outpatient community mental health services for youths ages 6-18 (39.1% female, 88.2% African American). Youths and parents completed research interviews and clinic diagnoses were extracted from clinic records. LEAD diagnoses synthesized results of the Schedule for Affective Disorders and Schizophrenia for School-Age Children- Present and Lifetime (KSADS-PL) and the youth's developmental, family, and psychiatric history.

Results—Agreement between the LEAD and chart diagnoses was low, not exceeding “poor” agreement for most diagnostic categories ($\kappa = .10-.46$, median = .37). Disagreement was largely driven by missed diagnoses, although clinicians also did assign extra diagnoses for some clients. Fewer diagnostic errors occurred when the youth's clinical picture was more clear (e.g., high or low symptom severity, lower comorbidity), when the youth was older, when the family was higher functioning, and when the parent had more depression. However, youth and family characteristics explained very little of the variability in diagnostic errors.

Conclusions—Results support the need to investigate strategies to improve clinician diagnostic accuracy.

Keywords

Diagnosis; clinical decision-making; diagnostic agreement; structured diagnostic interviews

It is clear that we are in an era of evidence-based practice (EBP), with state mental health agencies (e.g., Sigel et al., 2013), professional organizations (e.g., APA Presidential Task Force on Evidence-Based Practice, 2006), and research funders (e.g., National Institute of Mental Health, 2010) all striving to improve mental health services through the dissemination of EBPs to clinical settings. Now perhaps more than ever, it is essential that researchers and clinicians are speaking the same language when it comes to constructs needed to ensure appropriate use of EBPs. One dialect that plays a key role in both research and practice is diagnosis. A diagnostic label is useful in clinical communication, as it provides a shorthand summary of a cluster of symptoms. Diagnoses also provide categories to organize reviews and they offer a heuristic for matching the needs of a patient with techniques with known validity. Even as the field moves toward matching treatments to problem clusters (e.g., Chorpita, Bernstein, & Daleiden, 2011) or transdiagnostic treatments cutting across multiple diagnoses (e.g., Bilek & Ehrenreich-May, 2012), concordance between researcher- and clinician-generated diagnoses, either at the level of specific diagnoses or at the level of diagnostic clusters (e.g., depressive disorders, internalizing disorders), helps ensure that EBPs are being applied to the correct clients.

Unfortunately, despite their importance, questions have been raised about the accuracy of clinician-generated diagnoses (Garb, 1998), which are often generated through unstructured interviews (e.g., Anderson & Paulosky, 2004; Cashel, 2002). Without a pre-determined set of questions or scoring algorithms, clinicians are susceptible to several biases that can influence the diagnostic process. These can include prematurely deciding on a diagnosis before collecting all relevant data, seeking information to confirm that diagnosis while ignoring inconsistent information, and terminating the interview before exploring all alternatives (Croskerry, 2003). Additional biases include a tendency to perceive psychopathology over normative behavior and biases based on stereotypes about gender, ethnicity, and/or age (Garb, 1998). Not surprisingly, diagnoses generated through clinicians' "as usual" procedures show poor agreement with those generated by standardized diagnostic interviews (SDIs; i.e., research interviews with standard rules for gathering and interpreting data). Eleven studies of diagnostic agreement in youth samples in a recent meta-analysis (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009) had a mean kappa of .39, which is considered "poor" agreement (Landis & Koch, 1977).

However, extant youth studies have several limitations that preclude drawing firm conclusions regarding the validity of clinician diagnoses. First, the "clinician" diagnoses examined in many of these studies do not represent "usual care" practices. For example, studies have provided clinicians with checklists of diagnostic possibilities (Pellegrino, Singh, & Carmanico, 1999), or asked "expert" clinicians to generate diagnoses specifically for the study, rather than for clinical care (e.g., Ghanizadeh, Mohammadi, & Yazdanshenas,

2006). Second, many studies only utilized either parent- (Jensen & Weisz, 2002) or youth-report SDIs (Jewell, Handwerk, Almquist, & Lucas, 2004; Weinstein, Stone, Noam, Grimes, & Schwab-Stone, 1989; Welner, Reich, Herjanic, Jung, & Amado, 1987), or analyzed data from different reporters separately (Kramer, Robbins, Phillips, Miller, & Burns, 2003; Vitiello, Malone, Buschle, & Delaney, 1990). In contrast, clinicians typically based their diagnoses on information from multiple reporters. Third, many studies were done in inpatient settings (e.g., Vitiello et al., 1990; Weinstein et al., 1989; Welner et al., 1987); their findings may not generalize to diagnostic practices conducted in outpatient settings, where clinicians often have less contact with clients.

In addition, while existing studies have studied whether clinicians and SDIs concur, this is not necessarily the same as examining the validity of the clinician diagnoses. Several authors have discussed the limitations of relying solely on SDIs, particularly highly-structured diagnostic interviews that do not incorporate clinical expertise (e.g., Brugha, Bebbington, & Jenkins, 1999). Respondents may not understand the intention of questions, creating false positives (e.g., endorsing “recurrent thoughts” that do not constitute clinical obsessions). Fully structured interviews limit the opportunity for probing or clarifying (Kaufman et al., 1997). Misinterpretation of content and false positives may lead to pseudo-diagnoses that are not associated with clinical impairment (Bird et al., 1990; Brugha et al., 1999). Consequently, it is considered best practice to utilize a combination of SDIs, expert clinical opinion, and auxiliary information, such as information from medical records, to generate best estimate diagnoses (Garb, 1998; Pilkonis, Heape, Ruddy, & Serrao, 1991). This approach is best operationalized by Spitzer's (1983) LEAD (Longitudinal, Expert, and All Data) Standard, which is widely used in psychopathology research, including the present study. Very few studies have compared clinician diagnoses to best estimate diagnoses; to our knowledge only one study has done so in a youth sample. Vitiello and colleagues (1990) examined agreement between inpatient chart diagnoses and diagnoses utilized through a “review” process wherein two psychiatrists reviewed clients' charts and the results of SDIs to formulate a comprehensive diagnosis. These “review diagnoses” demonstrated poor agreement with chart diagnoses for most categories, although demonstrated “fair to good” agreement (Fleiss, 1981) for “major affective disorder,” adjustment disorder, and enuresis. Clearly, additional research is needed utilizing LEAD diagnoses to examine the validity of clinician diagnoses, particularly in youth outpatient settings.

Finally, knowing about low concordance between clinician diagnoses and other validity indicators only serves to identify a problem. Solving that problem requires examination of why agreement might be low. In addition to the biases discussed above, another important source of information would be data regarding “hard to diagnose” youths. For example, a key challenge in youth diagnosis is the integration of often discrepant reports from children and their parents (De Los Reyes & Kazdin, 2005); any youth or family characteristics that might lead to increased discrepancy might also lower diagnostic accuracy. Also, any factors that might impact the quality of the reports from any individual reporter, such as the child's developmental level, could also make diagnosis more difficult (Youngstrom et al., 2011). In addition, to the degree that the youth's clinical picture is less clear, accurate diagnosis becomes more difficult. These factors likely impact the quality of both researcher- and clinician-generated diagnoses. However, without the use of standardized questions to ensure

coverage of topics that the interviewee does not choose to disclose and algorithms and consensus procedures to aid in the interpretation of results, clinician-generated diagnoses might be less robust to these challenges.

Few studies have examined predictors of diagnostic agreement between researcher and clinician diagnoses, and findings have been inconsistent. One might expect that agreement would be better for older youths, as they might be more psychologically minded and better at communicating with adults about their symptoms. However, support for this notion is mixed, as older age predicts higher diagnostic agreement (Kramer et al., 2003), lower agreement (Lewczyk, Garland, Hurlburt, Gearity, & Hough, 2003), or appears unrelated to agreement (Jensen & Weisz, 2002). A rival hypothesis would be that maturation leads to adolescents having a more independent view of themselves, more informed by internal states that are less visible to outside observers. Teens are also more likely to engage in covert behaviors, such as substance use. To the extent that the teen is successful at avoiding detection, other informants will remain unaware of the behavior (e.g., Loeber & Schmalting, 1985). Gender might influence agreement in different ways: girls might be more likely to talk to adults about their symptoms, but boys are more likely to experience psychopathology that is more easily observable by adults (Achenbach & Rescorla, 2001). Consistent with the former, agreement has been found to be lower for boys in one study (Lewczyk et al., 2003), but gender was not related to agreement in another (Jensen & Weisz, 2002). Given literature suggesting that ethnic minority families might be less engaged in treatment (Garland et al., 2005) or experience more stigma related to disclosing symptoms (Hinshaw & Cicchetti, 2000), one might expect agreement to be lower for ethnic minority youths. Ethnicity was not related to agreement in any prior youth studies (Jensen & Weisz, 2002; Kramer et al., 2003; Lewczyk et al., 2003), although some adult studies have found lower agreement for minorities (e.g., Ramirez Basco, et al., 2000).

Studies on predictors of diagnostic agreement have also examined clinical characteristics that might make the clinical picture less clear. For example, more comorbidity has been linked to lower agreement (Lewczyk et al., 2003), suggesting that more complex psychopathology may lead to more missed diagnoses, and perhaps less agreement about which diagnoses are present. On the other hand, more severe youth psychopathology has also been found to be associated with higher agreement (Pellegrino et al., 1999), as has higher impairment (Kramer et al., 2003), suggesting that perhaps more “obvious” symptoms might lead to higher agreement. In support of this notion, Lewczyk and colleagues (2003) found that agreement was better for youths with “extreme psychopathology,” defined as either very high *or* very low symptom severity.

Finally, parental and family factors might be related to diagnostic agreement, although efforts to examine these variables have also yielded mixed findings. Some researchers have examined whether agreement might be lower in the presence of parental psychopathology or other stressors that might interfere with a parent's ability to accurately report on a child's behavior. Some studies found that higher parental psychopathology was predictive of lower agreement (Lewczyk et al., 2003), but others have found no relationship (Jensen & Weisz, 2002; Kraemer et al., 2003). Negative affective states in the adult may feed some exaggeration in their perceptions of negative child behaviors, but make them more accurate

about positive qualities (Youngstrom, Ackerman, & Izard, 1999). Examinations of other indicators of stress, including low income (Jensen & Weisz, 2002), insurance status (Kramer et al., 2003), and child welfare system involvement (Lewczyk et al., 2003) have yielded null results. More work needs to address this issue, including examination of additional variables that might be unique to child samples. For example, agreement might be better for families with multiple children because parents would have a better sense of normative child behavior; on the other hand, these parents might have less time to pay attention to an individual child's symptoms, decreasing agreement.

This study addressed these gaps in the literature by utilizing a sample of youths seeking outpatient services in a large community mental health center to examine agreement between chart diagnoses and diagnoses generated through a LEAD standard consensus conference (Spitzer, 1983). Given that adult studies have shown that predictors can differ depending on definitions of agreement (e.g., Klinkman, Coyne, Gallo, & Schwenk, 1998), we used two operational definitions. First, *missed diagnoses* represented the number of diagnoses assigned by the LEAD team that were present in the charts. Second, *extra diagnoses* represented the number of chart diagnoses not assigned by the LEAD team. We predicted that the LEAD diagnoses, which incorporated an SDI, would generate significantly more diagnoses, by avoiding premature discontinuation of interviews once an initial diagnosis was confirmed (Croskerry, 2003). This greater sensitivity to comorbidity would therefore generate more missed diagnoses than extra diagnoses. To better understand patterns of agreement, we tested several predictors of diagnostic errors. We hypothesized that more errors would be associated with higher functioning (due to the clinical picture being less obvious than in more impaired cases) and more comorbidity (due to the clinical picture being more complex). Consistent with the findings of Lewczyk and colleagues (2003), we predicted a curvilinear relationship between symptom severity and errors, such that very low and very high severity would be associated with fewer errors. We also explored several predictors either not examined, or with inconclusive findings in prior studies, including: 1) youth age, 2) gender, 3) ethnicity, 4) parental depression, 5) presence of multiple children in the home, 6) family stress, and 7) caregiver education level.

Method

Participants

Participants were 391 families seeking services at a youth-serving community mental health center. The clinic was the largest provider of outpatient services to children and families in a large Midwestern state, located in one of the poorest urban regions in the USA based on the most recent Census. Table 1 details participant characteristics and descriptive statistics for all study variables. Youth participants ranged in age from 6 to 18 ($M = 11.0$) and were primarily male, attending grades K through 12 (median = 4th grade). Most participants were African Americans (89%), followed by Non-Hispanic Caucasians (6%), and Hispanics (2%), while 4% self-identified as "other ethnicity" and one participant refused to provide ethnicity data. Primary caregiver participants were predominantly mothers (82%; $n = 322$), followed by other relatives (13%; $n = 52$), fathers (4%; $n = 15$), and non-relatives (0.5%; $n = 2$).

Procedures

Participants were drawn from a larger assessment study of more than 800 families presenting for treatment between September, 2003, and March, 2008, recruited to the study at the time of their clinic intake evaluation. Medical records were not consistently available from most private practices and other clinics, so these analyses concentrate on the youths seen at the urban community mental health center, where there was access to the intake record as part of the research protocol. As a result, the subsample used here is significantly younger (~1 year on average), with lower levels of caregiver education and income, higher rates of externalizing problems, and lower levels of anxiety and mood disorder than the excluded cases seeking services elsewhere. Youths were included in the larger project if they were between 4 years 11 months and 17 years 11 months of age and if both the youth and the primary caregiver were available for the assessment. Youths were excluded if they (or their caregivers) could not communicate at a conversational level in English or had suspected moderate, severe, or profound mental retardation. These exclusion criteria were included in the study design to ensure consistency with pilot data, but in practice involved <1% of cases otherwise eligible for the project. For the present analyses, youths were included if they were age 6 or above (the minimum age for the Child Behavior Checklist—see below). Clinicians diagnosed multiple clients in the sample, creating dependencies in the data that needed to be modeled. This sample was therefore restricted to the 391 participants who could be matched to clinicians via record review. *T* tests indicated that those participants did not differ significantly on any of the study variables from the 76 excluded participants without therapist information.

The research interview took place an average of 8.0 days ($SD = 7.1$) after the clinic intake. Caregivers provided written consent, and all youths gave written assent, to participate in the research assessment and record review. During the research assessment, both caregivers and youths participated in the KSADS interview and completed a series of questionnaires (see Measures, below), receiving an incentive of \$25. After the family completed the assessment and the research team reviewed the clinic records, research diagnoses were generated through a LEAD (Spitzer, 1983) conference conducted by an expert consensus team consisting of at least one licensed psychologist with expertise in youth psychopathology and the members of the research interview team. The person conducting the SDI always participated in the LEAD consensus team, presenting the SDI findings, and also noting any clinical impressions or other history that fell outside of the SDI items. A second interview team member presented the family history based on separate direct interview of the caregiver, as well as the youth's prior treatment and forensic history--if any--based on a review of their medical record. All study procedures were approved by the Case Western Reserve/University Hospitals of Cleveland and the Applewood Centers Institutional Review Boards.

Measures

Schedule for Affective Disorders and Schizophrenia for School-Age Children—Present and Lifetime (KSADS; Kaufman et al., 1997) plus the mood modules from the Washington University KSADS (WASH-U; Geller et al., 2001)—The KSADS-PL is a semi-structured interview assessing symptoms of over 30 DSM-IV

diagnoses, including systematic inquiry about current and lifetime diagnoses. Extensive data exist regarding the reliability of the KSADS (Ambrosini, 2000). The same interviewer administered the KSADS to the youth and caregiver and generated a single set of diagnoses combining information from both reporters. When informants provided discrepant information, they were re-interviewed, and remaining discrepancies resolved using clinical judgment. KSADS Interviewers were highly trained graduate students or predoctoral interns who attained item-level agreement exceeding κ 's of .85 on 10 cases prior to conducting interviews independently. KSADS interviews also generated a Global Assessment of Functioning (GAF) score, ranging from 1 to 100, with higher scores indicating better functioning (American Psychiatric Association, 2001).

LEAD Diagnoses—LEAD diagnoses synthesized: a) KSADS results, b) developmental history, c) family history of mental illness, and d) the youth's prior psychiatric history. We grouped diagnoses into 7 clusters for analysis: Depression, Bipolar, Anxiety, Posttraumatic Stress, Attention-Deficit Hyperactivity Disorders (ADHD), Disruptive Behavior, and Elimination Disorders. Other potential clusters of diagnoses (e.g., eating disorders, substance use disorders) were assigned to fewer than 5% of participants, so were not utilized in the analyses. We analyzed agreement at the cluster level because disagreement within a cluster (e.g., a research diagnosis of major depressive disorder versus a chart diagnosis of dysthymia) might not change treatment decisions substantially. In keeping with this approach, the clusters had broad definitions that included similar symptom presentations that might lead clinicians to make similar treatment decisions (e.g., adjustment disorders were grouped with other diagnoses with similar symptoms). Table 2 details the disorders falling into each category.

Chart Diagnoses—The youths' medical records provided DSM-IV intake diagnoses. The clinic's assessment procedures consisted of a 90 minute interview collecting a developmental history, exploring the presenting problem, and using an unstructured clinical interview to assign diagnoses for billing and treatment. Clinic diagnoses were generated by 15 clinicians who were either licensed masters level clinicians employed by the agency as intake specialists or predoctoral psychology interns completing an assessment intake rotation at the agency. Although assignment of a diagnosis was required for billing purposes, there were few restrictions on the specific diagnoses that could be billed and the clinic had funds available to cover unreimbursed diagnoses. Clinicians did not have access to the research assessment results.

Youth Symptom Severity—The caregiver- (Child Behavior Checklist, CBCL) and youth-report (Youth Self-Report, YSR) versions of the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001) forms measured symptom severity. The ASEBA scales are designed to facilitate multi-informant assessment of youth psychopathology; they have extensive reliability and validity data (Achenbach & Rescorla, 2001). Informants rate 118 behavior problems on a scale from 0 (*Not True*) to 2 (*Very True or Often True*). Scale scores are converted to *T* scores, normed for age and gender. The YSR is only normed for children ages 11 and up, so only participants in that age range ($n = 209$)

completed it. *T* scores for the CBCL and YSR Total Problem scales (both α s = .95 in this sample) were our measures of severity.

Parental Depression—Caregivers completed the Beck Depression Inventory (BDI; Beck, Steer, & Carbin, 1988), a 21 item self-report measure of depression with numerous reports of reliability and validity (Beck et al., 1988). In the current sample, the BDI Total had α = .89.

Family Stress—Interviewers completed the Global Family Environment Scale, rating the quality of the family environment on a continuum from 1 to 90, with higher scores indicating a more stable, nurturing environment (Rey et al., 1997). Interviewers also recorded whether there were multiple children in the home.

Analysis Plan

Rates of missing data were low (< 5% for all variables), and the missing completely at random assumption was tenable using Little's (1988) MCAR test ($\chi^2 = 78.43$, $df = 105$, $p = .98$), suggesting that listwise deletion of missing data was an acceptable strategy that was more parsimonious than solutions such as multiple imputation (Tabachnick & Fidell, 2007). Cohen's (1960) κ quantified agreement between the chart and LEAD diagnoses about each of the 7 diagnostic categories. Following Fleiss (1981), κ s below .40 reflect “poor” agreement, κ s between .40 and .74 reflect “fair to good” agreement, and κ s .75 and higher reflect “excellent” agreement. McNemar's test compared rates of assignment for each diagnostic category. High values on the *missed diagnoses* variable indicated that clinicians made errors of omission, failing to detect diagnoses assigned through the LEAD procedures. High values on the *extra diagnoses* variable indicated that clinicians made errors of commission, assigning extra diagnoses that were not validated through the LEAD process. Because clients were nested within diagnosing clinicians, we used Generalized Estimating Equations (GEE) to predict these two variables from client demographic and clinical information (see Table 3 for a list of predictors), with clients nested within clinicians (Hanley, 2003). The total number of LEAD diagnoses assigned was used as a control variable in analyses predicting missed LEAD diagnoses and the total number of chart diagnoses in the analyses of extra chart diagnoses. Given the large sample size, an alpha level of $p < .005$ was employed to avoid type I errors.

Results

Agreement Between Research and Chart Diagnoses

Table 2 details agreement between the LEAD and chart diagnoses for the seven diagnostic clusters. For all clusters, agreement was significantly greater than chance (all $ps < .005$; range = .10-.46; median = .37); however, for nearly all clusters, agreement was below Fleiss's (1981) cutoff of .40 for “poor” agreement. The only two clusters for which agreement reached .40 for “fair to good” were the Depression ($\kappa = .43$) and Posttraumatic Stress clusters ($\kappa = .46$). McNemar's tests indicated that all categories were assigned at a higher rate by the LEAD team (all $ps < .005$), showing that *missed diagnoses* (i.e., clinicians failing to assign a diagnosis identified by the LEAD team, or the “LEAD+/Charts-” column

in Table 2) were a significant contributor to the low agreement. However, examination of the “LEAD-/Charts+” column in Table 2 indicates that *extra diagnoses* (i.e., clinicians assigning diagnoses not assigned by the LEAD team) also occurred from 1% to 12% of the time. Clinicians assigned an average of 1.4 diagnostic categories per child ($SD = .75$, range = 0-4) and the LEAD team assigned an average of 2.3 ($SD = 1.1$, range = 0-6); as hypothesized, the LEAD time assigned more diagnoses per child than the clinicians [$t(390) = 14.3$, $p < .005$, $d = .95$]. The average number of diagnostic errors per child was 2.2 missed diagnoses ($SD = 1.1$; range = 0-6) and 1.4 extra diagnoses ($SD = 0.8$; range = 0-4). Treating the LEAD diagnosis as the criterion, clinical diagnoses showed a wide range of sensitivity, detecting the majority of cases with ADHD and disruptive behavior disorders, but missing more than half of depression, ~70% of bipolar spectrum, and more than 90% of anxiety diagnoses. On the other hand, clinical diagnoses showed strong specificity in most categories, indicating few false positive diagnoses compared to the LEAD criterion.

To examine the possibility that agreement was lowered by our creation of “broad” diagnostic categories (e.g., including adjustment disorders), we re-examined agreement for more “narrow” categories by excluding mood disorder, NOS, from the Bipolar category, excluding adjustment disorder diagnoses from the Depression, Anxiety, and Disruptive Behavior categories, and excluding the abuse and neglect codes from the Posttraumatic Stress category. Agreement for these “narrow” categories was nearly identical to agreement for the “broad” categories for Depression ($\kappa = .41$), Anxiety ($\kappa = .08$), and Disruptive Behavior ($\kappa = .24$). Agreement for the Bipolar ($\kappa = .04$) and Posttraumatic Stress ($\kappa = .12$) clusters was lower when they were defined narrowly. For both clusters, narrow definitions resulted in markedly lower rates of chart diagnoses (0.3% vs. 9.0% for Bipolar; 2.8% vs. 10.0% for Posttraumatic Stress). The rates of LEAD diagnoses remained essentially unchanged for Bipolar (12.3% vs. 12.5%), suggesting that the higher agreement for the “broad” definition for this category was driven by clinicians assigning mood disorder NOS, in lieu of assigning bipolar disorder. For Posttraumatic Stress, rates of LEAD diagnoses also were lower under the “narrow” definition (10.2% vs. 19.9%), suggesting that both researchers and clinicians may be using the abuse and neglect codes somewhat interchangeably with stress disorder diagnoses.

Given the increasing use of transdiagnostic treatments that cut across diagnostic categories (e.g., depression and anxiety; Bilek & Ehrenreich-May, 2012), as well as the utility of similar treatment strategies such as behavioral parent training for ADHD and disruptive behavior disorders, we also examined whether agreement would be better if considered at the level of “internalizing” (Depression, Anxiety, Posttraumatic Stress) and “externalizing” (ADHD, Disruptive Behavior) clusters. Even at this very inclusive level, agreement remained poor (Internalizing $\kappa = .34$, $p < .005$; Externalizing $\kappa = .36$, $p < .005$).

Demographic Predictors of Diagnostic Errors

Table 3 details the relations between youth and family characteristics and diagnostic errors. Because only youths 11 and older completed the YSR, we ran two models for each dependent variable: one that included the YSR total score and one that did not. The models that did not include the YSR therefore utilized the whole sample, whereas the models that

included the YSR were restricted to participants ages 11 and up. Older age was associated with both fewer missed and extra diagnoses in both samples ($ps < .005$). Female gender was also associated with fewer errors ($ps < .005$), but only in the analysis with the older participants. African American ethnicity was associated with fewer missed diagnoses and fewer extra diagnoses in the older sample and fewer missed diagnoses in the whole sample. It was not a significant predictor of extra diagnoses in the whole sample.

To understand whether differences in findings between the whole sample and the older sample reflected age-related differences in the strength of the predictors, a post-hoc analysis examined whether there was a significant interaction between age and any of the other predictors listed in Table 3. To probe significant interactions, simple slopes for children (1 standard deviation below the mean sample age = 7.7 years) and adolescents (1 standard deviation above the mean sample age = 14.1 years) were calculated following Aiken & West (1991). The interaction between age and gender was significant ($p < .005$; see Table 4). Among children, being female was associated with more diagnostic errors (both $ps < .005$); the reverse was true among adolescents (both $ps < .005$). The interaction between age and ethnicity was also significant ($p < .005$). For both types of errors, African American ethnicity was significantly associated with fewer diagnostic errors among adolescents ($p < .005$), but not among children. As discussed below, age also significantly moderated some clinical and family predictors.

Clinical Predictors of Diagnostic Errors

As detailed in Table 3, results supported our hypotheses regarding the relationships between youth clinical characteristics and errors. As predicted, there were negative, curvilinear relationships between both types of errors and both measures of symptom severity (the CBCL and the YSR), indicating that errors were less prevalent at very low and very high symptom levels, and more prevalent at moderate symptom levels. Also as hypothesized, missed diagnoses were more frequent among higher functioning youths, as evidenced by positive relationships between the GAF rating and both types of errors in the whole sample, although this relationship was not significant in the older sample or in analyses of extra diagnoses.

However, as detailed in Table 4, age moderated the predictive nature of both the CBCL and the GAF. For the CBCL, the interaction between the linear CBCL Total Score and age was significant for both types of errors; the interactions for the quadratic CBCL score were not. This indicates that, while the relationship between errors and severity is curvilinear in both groups, the steepness of that relationship is stronger among adolescents than among children. The interactions between age and GAF were significant for both types of errors; higher functioning was associated with more diagnostic errors among children, but not among adolescents.

We also hypothesized that more comorbidity would be related to more errors, because the greater complexity in comorbid cases might make them more difficult to diagnose. The number of LEAD diagnoses predicted more missed diagnoses and the number of chart diagnoses predicted more extra diagnoses. These variables also interacted with age, such that higher numbers of diagnoses were more strongly predictive of errors among children

than among adolescents (Table 4). However, these findings were difficult to interpret, given that more diagnoses provided more opportunities for errors. To further explore the relationship between comorbidity and errors, follow-up analyses used the proportion of missed (i.e., the number of missed diagnoses divided by the number of LEAD diagnoses assigned) and extra (i.e., the number of extra diagnoses divided by the number of chart diagnoses assigned) diagnoses. GEE then predicted these proportions from the total number of diagnoses assigned by either source. If the relationship between errors and comorbidity was purely a function of increased opportunity for errors, we would expect that these proportions would remain constant as the number of diagnoses increased. However, consistent with the idea that diagnostic complexity is associated with increased errors, the total number of diagnoses was positively associated with both the proportions of missed diagnoses ($B = .056, p < .005$) and extra diagnoses ($B = .060, p < .005$).

Family Predictors of Diagnostic Errors

The last set of analyses tested parental depression (BDI), multiple children in the home, family functioning (GFES), and parental education level as predictors of errors (see Table 3). BDI was not a significant predictor of errors; however, there was a significant interaction between BDI and age in the prediction of extra diagnoses, such that higher BDI scores were associated with fewer extra diagnoses among adolescents ($p < .005$), but not children (Table 4).

The presence of multiple children in the home was significantly associated with fewer errors. However, there was a significant interaction between age and the presence of multiple children. Among adolescents, there were fewer diagnostic errors for families with multiple children; among children, the presence of multiple children in the home was not associated with errors. Higher family functioning was also associated with fewer errors. There were no significant interactions between the GFES scores and age.

Parental education level was significantly associated with both types of errors, with more errors observed for parents with some post-high school education, compared to parents who did not complete high school. The analysis in the older sample also showed more missed diagnoses for parents with high school degrees. In addition, there were significant interactions between post-high school education and age for both types of errors, such that the positive relationship between parental education and errors was stronger among children than among adolescents.

Discussion

The goal of this study was to examine agreement between clinical diagnoses and best estimate diagnoses derived from integrating semi-structured diagnostic interviews with prior treatment history and family mental health history through a consensus diagnostic process. The study used a large sample to investigate agreement across seven diagnostic clusters, as well as to test potential demographic and clinical predictors of agreement. Consistent with prior work (Rettew et al., 2009), agreement between LEAD and clinical diagnoses was statistically significant, but poor, with kappas ranging from .10 for anxiety disorders to .46 for PTSD. The median κ of .38 was roughly the same as the mean of .39 for the meta-

analysis of the agreement between SDIs and clinical diagnoses for youths (Rettew et al., 2009). As hypothesized, the LEAD method identified significantly more diagnoses than clinicians, and the rate of cases identified with each of the seven diagnostic clusters was higher for the LEAD than the chart diagnoses. Missed diagnoses were therefore the primary driver of disagreement. The results extend prior work because the research diagnoses were based on a semi-structured diagnostic interview—the KSADS-PL—and a LEAD consensus review (Spitzer, 1983). Both of these methods are likely to overturn false positive diagnoses that could occur when using a fully structured diagnostic interview that does not incorporate clinical judgment about whether reported symptoms are associated with impairment or constitute a clinically meaningful pattern.

These findings reinforce concerns that clinicians using unstructured interviews may be prone to “search satisficing,” or discontinuing consideration of alternate explanations or comorbidity once a plausible diagnosis is confirmed (Galanter & Patel, 2005). Clinical diagnoses also may underestimate diagnoses due to a lack of structure and failing to elicit important details despite excellent intentions. The finding that clinicians also assigned extra diagnoses not assigned by the LEAD team, particularly for the more commonly-assigned diagnoses of depression and disruptive behavior disorder, also supports the notion that clinicians may be relying on “availability heuristics,” or over-estimating the likelihood of salient diagnoses (Galanter & Patel, 2005). There are other factors that also might influence clinical diagnoses, such as concerns about stigma attached to particular diagnoses, or whether payers will reimburse for services billed under particular diagnoses. However, the fact that clinicians used all of the diagnostic categories suggests that these were not the main drivers of the diagnoses; the clinic also had ways of funding treatment for any diagnoses. Although there are some disorders that clinicians reported they wanted more certainty prior to diagnosing (such as conduct disorder) it is not clear how concerns about stigma would pertain more to anxiety disorders (which had the lowest kappa) compared to depression or PTSD (which had the highest kappas).

Another aim was to test predictors of disagreement between research and clinical diagnoses. As hypothesized, better agreement was associated with more obvious and clearly delineated symptoms, as in the case of both high and low symptom severity, lower functioning, and less comorbidity. Better agreement was also associated with factors that might improve the quality of reporting, such as older youth age, having more children in the home (which could give the caregiver a better sense of normative behavior), and better family functioning. Conversely, error rates were higher for more educated caregivers and those with less depression.

However, it is important to note that all of these characteristics have effects that are relatively small in magnitude. The combination of large sample size and statistical methods accounting for nesting within clinicians afforded sufficient statistical power to detect small effects. While LEAD methods generated an average of 0.9 more diagnoses than the unstructured clinical interviews, all of the demographic, clinical, and family characteristics accounted for small fractions of difference in diagnoses. For example, an increase of 8.6 points (1 standard deviation in our sample) on the CBCL predicted a decrease of .02 in the number of missed diagnoses; even 20 point differences in CBCL or YSR *T* scores, or 25

point differences in caregiver BDI scores, would account for much less than half a point difference in predicted diagnostic agreement on average. Although these effects are statistically significant, it is hard to make a case that they are clinically meaningful. In contrast, the difference between using LEAD consensus procedures versus diagnosis as usual are large and meaningful at the level of the individual case, where multiple diagnoses often go undetected and extra diagnoses are frequently assigned. These diagnostic errors have the potential to radically change the focus of treatment and have been found to be associated with worse treatment engagement and client outcomes (Jensen-Doss & Weisz, 2008; Pogge et al., 2001). The fact that agreement was poor even at the level of internalizing versus externalizing disorders suggests that lack of agreement on target problems will likely be an issue even as EBPs are developed that cut across diagnostic groups.

Strengths of the study include the large sample size, the use of an SDI implemented by highly trained raters coupled with the use of a LEAD consensus review process to further refine the research diagnoses, statistical methods that modeled the nesting of clients among clinicians, exploration of interaction and quadratic effects, and the coding of clinical diagnoses that came directly from the chart and actually guided treatment for the cases—as opposed to rating vignettes or mock cases that might not be as generalizable to clinical practice. Limitations include that chart diagnoses were only available from one clinic, albeit the largest provider of outpatient mental health services to youths in the state at the time of data collection. Because the clinic was an urban community mental health center, families were mostly low income; the referral pattern had high rates of externalizing behavior problems and disruptive behavior disorders, as well as relatively lower rates of anxiety disorders. It would be helpful to replicate and extend findings in other settings with different demographic and clinical characteristics. However, given the modest effect of demographic and clinical characteristics on diagnostic agreement in the present data, it seems unlikely that sample differences would lead to much higher levels of diagnostic agreement. These findings are also need of replication with clinicians using DSM-5, although low agreement has been robust across multiple revisions of the DSM (Rettew et al. 2009). The low inter-rater agreement found for many diagnoses in the DSM-5 field trials (Regier et al., 2012) suggests agreement may not improve.

An additional study limitation was that we were not able to examine clinician-level predictors of agreement, such as years of experience. Given that the LEAD diagnoses were generated by a team that included highly trained doctoral level experts and the chart diagnoses were generated by master's level clinicians or trainees, it is possible that agreement was driven not only by data collection and synthesis procedures, but also by the quality of the clinical judgment involved. To our knowledge, only one prior study has examined whether clinician experience, level of training, and professional discipline predicted diagnostic agreement in a youth sample (Jensen & Weisz, 2002); this study did not find any significant effects, but additional research on this topic is clearly needed.

Another limitation suggestive of future research is the single snapshot nature of this study, which examined intake diagnoses only. It is possible that, as clinicians become more familiar with a case, their diagnoses become more accurate. We are aware of only one study that has examined this question with a child sample, finding that agreement between

researchers and clinicians' discharge diagnoses was no better than agreement with their intake diagnoses (Aronen, Noam, & Weinstein, 1993). However, it is possible that discharge diagnoses are not the best source of data to examine this question. Although studies have found that clinicians often do not update their chart diagnoses (Powsner & Tufte, 1994), they do often change their treatment targets over the course of treatment (Young, Daleiden, Chorpita, Schiffman, & Mueller, 2007). These findings suggest that future studies should examine: 1) the extent to which both intake and later diagnoses reflect clinicians' treatment targets and 2) whether these treatment targets become more accurate as clinicians become more familiar with clients.

Overall, findings reinforce the view that clinician diagnoses lack accuracy, even at the level of clusters such as depression or even internalizing disorders. A similar message has been reiterated for decades (Meehl, 1954; Spengler et al., 2009), but the use of a LEAD diagnosis in this pediatric study advances the field by helping rule out the possibility that previous findings of diagnostic differences were driven by false positives in the SDIs. Additionally, our extensive analysis of child and family factors suggests that diagnostic errors extend across different types of clients. Although authors have rightfully pointed out that additional work is needed to establish the “treatment utility” of specific diagnoses (Nelson-Gray, 2003), problem clusters such as depression or anxiety play a central role in the dissemination of EBPs to practice settings. These findings therefore raise questions about whether those practices are being applied to appropriate clients and suggest a need for efforts to improve clinician diagnostic practices. Despite data indicating that patients actually prefer structured approaches (Suppiger et al., 2009), practitioners rarely use SDIs (Jensen-Doss & Hawley, 2011), likely due to time and funding challenges associated with training in and administering these time intensive measures. A hybrid approach, combining checklists and other brief assessments to indicate targets for intensive interviewing, might offer the benefits of structured approaches at less cost in terms of time, while preserving some flexibility for the clinician (Ebesutani, Bernstein, Chorpita, & Weisz, 2012; Youngstrom, 2013). Given research suggesting that clinicians may revise their treatment targets over time (Young et al., 2007), these models also need to incorporate methods for re-assessment over time (e.g., Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, in press). Future research should also explore whether other components of the LEAD approach, such as consensus procedures and review of auxiliary information, might be feasible or useful in practice settings. Clinical trainings that target decision-making biases might also be useful. Given the significant time and money being invested in increasing the use of EBPs, it is essential that feasible and effective strategies are developed to help clinicians identify the appropriate clients with whom to apply those practices.

Acknowledgments

Amanda Jensen-Doss, Department of Psychology, University of Miami; Eric A. Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill; Jennifer Kogos Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill, Norah C. Feeny, Department of Psychological Sciences, Case Western University; Robert L. Findling, Department of Psychiatry and Behavioral Sciences, Johns Hopkins University. This research was supported in part by NIH R01 MH066647 (PI: E. Youngstrom). Dr. Findling receives or has received research support, acted as a consultant and/or served on a speaker's bureau for Alexza Pharmaceuticals, American Academy of Child & Adolescent Psychiatry, American Physician Institute, American Psychiatric Press, AstraZeneca, Bracket, Bristol-Myers Squibb, Clinsys, Cognition Group, Coronado Biosciences,

Dana Foundation, Forest, GlaxoSmithKline, Guilford Press, Johns Hopkins University Press, Johnson & Johnson, KemPharm, Lilly, Lundbeck, Merck, NIH, Novartis, Noven, Otsuka, Oxford University Press, Pfizer, Physicians Postgraduate Press, Rhodes Pharmaceuticals, Roche, Sage, Seaside Pharmaceuticals, Shire, Stanley Medical Research Institute, Sunovion, Supernus Pharmaceuticals, Transcept Pharmaceuticals, Validus, and WebMD.

References

- Achenbach, TM.; Rescorla, LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: Research Center for Children, Youth, and Families; 2001.
- Aiken, L.; West, S. Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage; 1991.
- Ambrosini PJ. Historical development and present status of the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS). *Journal of the American Academy of Child & Adolescent Psychiatry.* 2000; 39:49–58. [PubMed: 10638067]
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th--Text Revision. Washington, DC: Author; 2001.
- Anderson DA, Paulosky CA. A survey of the use of assessment instruments by eating disorder professionals in clinical practice. *Eating and Weight Disorders.* 2004; 9:238–241. [PubMed: 15656022]
- APA Presidential Task Force on Evidence-Based Practice. Evidence-based practice in psychology. *American Psychologist.* 2006; 61:271–285. [PubMed: 16719673]
- Aronen ET, Noam GG, Weinstein SR. Structured diagnostic interviews and clinicians' discharge diagnoses in hospitalized adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry.* 1993; 32:674–681.10.1097/00004583-199305000-00027 [PubMed: 8496132]
- Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review.* 1988; 8:77–100.
- Bilek EL, Ehrenreich-May J. An open trial investigation of a transdiagnostic group treatment for children with anxiety and depressive symptoms. *Behavior Therapy.* 2012; 43:887–897. [PubMed: 23046789]
- Bird HR, Yager TJ, Staghezza B, Gould MS, Canino G, Rubio-Stipec M. Impairment in the epidemiological measurement of childhood psychopathology in the community. *Journal of the American Academy of Child & Adolescent Psychiatry.* 1990; 29:796–803. [PubMed: 2228936]
- Brugha TS, Bebbington PE, Jenkins R. A difference that matters: Comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychological Medicine.* 1999; 29:1013–1020. [PubMed: 10576294]
- Cashel ML. Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice.* 2002; 33:446–453.10.1037/0735-7028.33.5.446
- Chorpita BF, Bernstein A, Daleiden EL. Empirically guided coordination of multiple evidence-based treatments: An illustration of relevance mapping in children's mental health services. *Journal of Consulting and Clinical Psychology.* 2011; 79:470–480. [PubMed: 21787050]
- Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960; 20:37–46.10.1177/001316446002000104
- Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine.* 2003; 78:775–780.10.1097/00001888-200308000-00003 [PubMed: 12915363]
- De Los Reyes A, Kazdin AE. Informant Discrepancies in the Assessment of Childhood Psychopathology: A Critical Review, Theoretical Framework, and Recommendations for Further Study. *Psychological Bulletin.* 2005; 131:483–509. doi:10.1037/. [PubMed: 16060799]
- Ebesutani C, Bernstein A, Chorpita BF, Weisz JR. A transportable assessment protocol for prescribing youth psychosocial treatments in real-world settings: Reducing assessment burden via self-report scales. *Psychological Assessment.* 2012; 24:141–155. [PubMed: 21859220]
- Fleiss, JR. Statistical methods for rates and proportions. 2nd. New York: Wiley; 1981.

- Galanter CA, Patel VL. Medical decision making: A selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry*. 2005; 46:675–689. [PubMed: 15972065]
- Garb, HN. *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association; 1998.
- Garland AF, Lau AS, Yeh M, McCabe KM, Hough RL, Landsverk JA. Racial and ethnic differences in utilization of mental health services among high-risk youths. *American Journal of Psychiatry*. 2005; 162:1336–1343. [PubMed: 15994717]
- Geller B, Zimerman B, Williams M, Bolhofner K, Craney JL, DelBello MP, Soutullo C. Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2001; 40:450–455. [PubMed: 11314571]
- Ghanizadeh A, Mohammadi MR, Yazdanshenas A. Psychometric properties of the Farsi translation of the Kiddie Schedule for Affective Disorders and Schizophrenia-Present and Lifetime Version. *BMC Psychiatry*. 2006; 610.1186/1471-244x-6-10
- Hanley JA. Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*. 2003; 157:364–375. [PubMed: 12578807]
- Hinshaw SP, Cicchetti D. Stigma and mental disorder. *Development & Psychopathology*. 2000; 12:555–598.10.1017/S0954579400004028 [PubMed: 11202034]
- Jensen-Doss A, Hawley KM. Understanding clinicians' diagnostic practices. *Administration and Policy in Mental Health and Mental Health Services Research*. 2011; 38:476–485.10.1007/s10488-011-0334-3 [PubMed: 21279679]
- Jensen-Doss A, Weisz JR. Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology*. 2008; 76:711–722.10.1037/0022-006x.76.5.711 [PubMed: 18837589]
- Jensen AL, Weisz JR. Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology*. 2002; 70:158–168. [PubMed: 11860042]
- Jewell J, Handwerk M, Almqvist J, Lucas C. Comparing the validity of clinician-generated diagnosis of conduct disorder to the Diagnostic Interview Schedule for Children. *Journal of Clinical Child and Adolescent Psychology*. 2004; 33:536–546. [PubMed: 15271611]
- Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Ryan N. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1997; 36:980–988.10.1097/00004583 [PubMed: 9204677]
- Klinkman MS, Coyne JC, Gallo S, Schwenk TL. False positives, false negatives, and the validity of the diagnosis of major depression in primary care. *Archives of Family Medicine*. 1998; 7:451–461. [PubMed: 9755738]
- Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ. A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*. 2003; 160:1566–1577. [PubMed: 12944328]
- Kramer TL, Robbins JM, Phillips SD, Miller TL, Burns BJ. Detection and outcomes of substance use disorders in adolescents seeking mental health treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2003; 42:1318–1326. [PubMed: 14566169]
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
- Lewczyk CM, Garland AF, Hurlburt MS, Gearity J, Hough RL. Comparing DISC-IV and clinician diagnoses among youth receiving public mental health services. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2003; 42:349–360. [PubMed: 12595789]
- Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. 1988; 83:1198–1202.
- Loeber R, Schmalting KB. Empirical evidence for overt and covert patterns of antisocial conduct problems. *Journal of Abnormal Child Psychology*. 1985; 13:337–353. [PubMed: 4008760]

- Meehl, PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis, MN: University of Minnesota Press; 1954.
- National Institute of Mental Health. The National Institute of Mental Health Strategic Plan. 2010. Retrieved February 17, 2010, 2010, from <http://www.nimh.nih.gov/about/strategic-planning-reports/index.shtml>
- Nelson-Gray RO. Treatment Utility of Psychological Assessment. *Psychological Assessment*. 2003; 15:521–531.10.1037/1040-3590.15.4.521 [PubMed: 14692847]
- Pellegrino JF, Singh NN, Carmanico SJ. Concordance among three diagnostic procedures for identifying depression in children and adolescents with EBD. *Journal of Emotional and Behavioral Disorders*. 1999; 7:118–127.
- Pilkonis PA, Heape CL, Ruddy J, Serrao P. Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment*. 1991; 3:46–54.
- Pogge DL, Wayland-Smith D, Zaccario M, Borgaro S, Stokes J, Harvey PD. Diagnosis of manic episodes in adolescent inpatients: Structured diagnostic procedures compared to clinical chart diagnoses. *Psychiatry Research*. 2001; 101:47–54. [PubMed: 11223119]
- Powsner SM, Tufte ER. Graphical summary of patient status. *The Lancet*. 1994; 344:368–389.10.1016/S0140-6736(94)91406-0
- Ramirez Basco M, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, Barnett V. Methods to improve diagnostic accuracy in a community mental health setting. *American Journal of Psychiatry*. 2000; 157:1599–1605. [PubMed: 11007713]
- Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ. DSM-5 Field Trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*. 2012
- Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*. 2009; 18:169–184. [PubMed: 19701924]
- Rey JM, Singh M, Hung Sf, Dossetor DR, Newman L, Plapp JM, Bird KD. A global scale of measure the quality of the family environment. *Archives of General Psychiatry*. 1997; 54:817–822.10.1001/archpsyc.1997.01830210061006 [PubMed: 9294372]
- Sigel BA, Kramer TL, Conners-Burrow NA, Church JK, Worley KB, Mitrani NA. Statewide dissemination of trauma-focused cognitive-behavioral therapy (TF-CBT). *Children and Youth Services Review*. 2013; 35:1023–1029.10.1016/j.childyouth.2013.03.012
- Spengler PM, White MJ, Ægisdóttir S, Maugherman AS, Anderson LA, Cook RS, et al. Rush JD. The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist*. 2009; 37:350–399.
- Spitzer RL. Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*. 1983; 24:399–411.10.1016/0010-440x(83)90032-9 [PubMed: 6354575]
- Suppiger A, In-Albon T, Hendriksen S, Hermann E, Margraf J, Schneider S. Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy*. 2009; 40:272–279. doi:S0005-7894(08)00088-9 [pii]. [PubMed: 19647528]
- Tabachnick, BG.; Fidell, LS. *Using Multivariate Statistics*. 5th. Boston, MA: Pearson Education, Inc; 2007.
- Vitiello B, Malone R, Buschle PR, Delaney MA. Reliability of DSM-III diagnoses of hospitalized children. *Hospital & Community Psychiatry*. 1990; 41:63–67. [PubMed: 2295479]
- Weinstein S, Stone K, Noam G, Grimes K, Schwab-Stone M. Comparison of DISC with clinicians' DSM-III diagnoses in psychiatric inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1989; 28:53–60. [PubMed: 2914836]
- Welner Z, Reich W, Herjanic B, Jung KG, Amado H. Reliability, validity, and parent-child agreement studies of the diagnostic interview for children and adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry*. 1987; 26:649–653. [PubMed: 3667494]
- Young J, Daleiden EL, Chorpita BF, Schiffman J, Mueller CW. Assessing stability between treatment planning documents in a system of care. *Administration and Policy in Mental Health and Mental Health Services Research*. 2007; 34:530–539. [PubMed: 17846879]

- Youngstrom EA. Future directions in psychological assessment. *Journal of Clinical Child & Adolescent Psychology*. 2013; 42:139–159.10.1080/15374416.2012.7363 [PubMed: 23153181]
- Youngstrom EA, Ackerman BP, Izard CE. Dysphoria-related bias in maternal ratings of children. *Journal of Consulting and Clinical Psychology*. 1999; 67:905–916. [PubMed: 10596512]
- 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*. in press.
- Youngstrom EA, Youngstrom JK, Freeman AJ, De Los Reyes A, Feeny NC, Findling RL. Informants are not all equal: predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology*. 2011; 21:407–415.10.1089/cap.2011.0032 [PubMed: 22040186]

Table 1
Participant demographic, clinical, family, and diagnostic error characteristics

Variable	N	M (SD) or n (%)
<i>Youth Demographic Characteristics</i>		
Age	391	10.9 (3.3)
Gender	391	
Female		153 (39.1%)
Male		238 (60.9%)
Ethnicity	390	
Black/African American		344 (88.2%)
Hispanic		7 (1.8%)
White, Not Hispanic		23 (5.9%)
Other		16 (4.1%)
<i>Youth Clinical Characteristics</i>		
Number of LEAD diagnoses	391	2.3 (1.1)
Number of Chart diagnoses	391	1.4 (0.7)
CBCL Total Score	380	69.2 (8.6)
YSR Total Score	209	57.9 (12.5)
GAF Rating from the KSADs	388	52.5 (7.9)
<i>Family Characteristics</i>		
Parent BDI Score ⁵	391	9.1 (8.1)
Multiple Children in Family	393	
No, participant is only child		24 (6%)
Yes, participant had siblings		369 (94%)
Global Family Environment Scale ⁵	379	66.9 (11.6)
Parent educational level	380	
Did Not Finish High School		113 (29.7%)
High School Graduate/GED		125 (32.0%)
Some Post-High School		142 (36.3%)
<i>Diagnostic errors</i>		
Missed Diagnoses	391	2.2 (1.1)
Extra Diagnoses	391	1.4 (.8)

Table 2
Agreement between longitudinal expert integration of all available data (LEAD) research and chart diagnoses

Cluster	Diagnoses included in Cluster/ ¹	Agreement: LEAD +/ Charts + % (n)	Missed: LEAD +/ Charts - % (n)	Extra: LEAD -/ Charts + % (n)	LEAD -/ Charts - % (n)	Which source assigned more?	Kappa	Sensitivity	Specificity
Depression	Major Depressive D/O; Dysthymic D/O; Adjustment D/O with Depressed Mood; Depressive D/O- NOS	15% (58)	16% (62)	6% (25)	63% (246)	LEAD*	.43*	48%	91%
Bipolar- Broad NOS)	Bipolar I D/O; Bipolar II D/O; Cyclothymic D/O; Bipolar D/O- NOS; Mood D/O- NOS	4% (15)	9% (34)	5% (20)	82% (322)	LEAD	.28*	31%	94%
Anxiety	Specific Phobia; Social Phobia; Obsessive-Compulsive D/O; Generalized Anxiety D/O; Separation Anxiety D/O; Selective Mutism; Adjustment D/O with Anxiety; Anxiety D/O NOS	2% (6)	18% (70)	1% (3)	80% (312)	LEAD*	.10*	8%	99%
Attention Deficit/ Hyperactivity	Attention-Deficit/Hyperactivity D/O (ADHD)- Inattentive Type; ADHD- Hyperactive-Impulsive Type; ADHD- Combined Type; ADHD- Unspecified Type; ADHD- NOS	37% (144)	28% (110)	4% (16)	31% (121)	LEAD*	.39*	57%	88%
Disruptive Behavior	Oppositional Defiant D/O; Conduct D/O; Adjustment D/O With Disturbance of Conduct; Disruptive Behavior D/O NOS	44% (173)	23% (90)	12% (48)	21% (80)	LEAD*	.26*	66%	63%
Elimination Disorders	Enuresis	4% (14)	9% (35)	1% (5)	86% (337)	LEAD*	.37*	29%	99%
Posttraumatic Stress	Posttraumatic Stress D/O; Acute Stress D/O; Physical Abuse of Child; Sexual Abuse of Child; Child Neglect	8% (31)	12% (47)	2% (8)	78% (305)	LEAD*	.46*	40%	97%

Note. LEAD + = Diagnosis assigned by the LEAD team, LEAD - = Diagnosis not assigned by the LEAD team; Chart + = Diagnosis present in the charts; Chart - = Diagnosis not present in the charts; Sensitivity and Specificity are accuracy rates for chart diagnoses using LEAD as the criterion; D/O = Disorder. NOS = Not Otherwise Specified;

¹ Table includes diagnoses that were actually assigned by either source. For some categories, additional diagnoses would have been included had they been assigned (e.g., Panic Disorder would have been coded as an Anxiety Disorder).

* $p < .005$

Table 3
Multiple Regression Analyses Examining Demographic, Clinical, and Family Characteristics as Predictors of Errors

Predictor variable	Number of LEAD Diagnoses Missed		Number of Extra Chart Diagnoses	
	Whole Sample ¹ <i>B</i>	Age 11-18 ²	Whole Sample ¹ <i>B</i>	Age 11-18 ²
Youth Demographic Characteristics				
Age ³	-.02*	-.04*	-.02*	-.05*
Gender (female = 1)	-.01	-.02*	-.01	-.02*
Ethnicity (African American = 1)	-.02*	-.08*	.01	-.07*
Youth Clinical Characteristics				
Number of LEAD diagnoses	.98*	.98*	----	----
Number of Chart diagnoses	----	----	1.01*	.99*
CBCL Total Score ³ - linear	-.02*	-.03*	-.03*	-.04*
CBCL Total Score ³ - quadratic	-.01*	-.01*	-.01*	-.01*
YSR Total Score ³ - linear	----	-.03*	----	-.03*
YSR Total Score ³ - quadratic	----	-.01*	----	-.01*
GAF Rating from the KSADS ³	.01*	-.00	.01	-.00
Family Characteristics				
Parent BDI Score ³	.00	-.01	.000	-.01
Multiple Children in Family (yes = 1)	-.04*	-.09*	-.05*	-.10*
Global Family Environment Scale ³	-.02*	-.04*	-.02*	-.03*
Parent education level ⁴				
High School Graduate	.01	.02*	.01	.02
Some Post-High School Education	.05*	.04*	.04*	.03*

* $p < .005$

¹ $n = 344$;

² $n = 186$;

³ Continuous predictors were standardized;

⁴ Parent education level was dummy coded with Did not Finish High School as the reference group

Table 4
Exploratory Analysis of the Interaction between Age and Other Predictors

Predictor variable	Number of LEAD Diagnoses Missed				Number of Extra Chart Diagnoses							
	Interaction term ¹		Simple slope for children ²		Simple slope for adolescents ³		Interaction term ¹		Simple slope for children ²		Simple slope for adolescents ³	
	B		B		B		B		B		B	
Age × Gender (Female = 1)	-.05*		.05*		-.06*		-.05*		.07*		-.05*	
Age × Ethnicity(African American = 1)	-.04*		.02		-.05*		-.02*		.01		-.03*	
Age × Number of LEAD diagnoses	-.02*		.99*		.96*		-----		-----		-----	
Age × Number of Chart diagnoses	-----		-----		-----		-.02*		1.02*		.99*	
Age × CBCL Total Score- linear	-.01*		-.01*		-.03*		-.01*		-.02*		-.04*	
Age × CBCL Total Score- quadratic	.00		.00		.00		.00		.04*		-.00	
Age × GAF Rating	-.02*		.03*		-.004		-.02*		.00		-.01*	
Age × Parent BDI Score	.00		.00		.00		-.004*		.00		-.01*	
Age × Multiple Children in Family (yes = 1)	-.04*		.00		-.08*		-.04*		-.01		-.09*	
Age × Global Family Environment Scale	.00		.00		.00		.00		.07*		.03	
Age × High School Graduate	.01		.01		.00		.00					
Age × Some Post-High School Education	-.02*		.07*		.04*		-.01*					

* $p < .005$

¹ Interaction term from a multiple predictor model including main effects of age and all other predictors;

² Simple slope for 1 standard deviation below the mean sample age (7.7 years);

³ Simple slope for 1 standard deviation above the mean sample age (14.1 years)