Published in final edited form as:
J Am Acad Child Adolesc Psychiatry. 2009 April ; 48(4): 386–399. doi:10.1097/CHI.0b013e31819a1cbc.

The National Comorbidity Survey Adolescent Supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments

Ronald C. Kessler, Ph.D., **Shelli Avenevoli, Ph.D.**, **Jennifer Greif Green, Ph.D.**, **Michael J. Gruber, M.S.**, **Margaret Guyer, Ph.D.**, **Yulei He, Ph.D.**, **Robert Jin, M.S.**, **Joan Kaufman, Ph.D.**, **Nancy A. Sampson, B.A.**, **Alan M. Zaslavsky, Ph.D.**[*], and **Kathleen R. Merikangas, Ph.D.**
Department of Health Care Policy, Harvard Medical School (Kessler, Green, Gruber, He, Sampson, Zaslavsky); the Division of Developmental Translational Research, National Institute of Mental Health (Avenevoli); the Massachusetts Mental Health Center (Guyer); the Department of Psychiatry, Yale Medical School (Kaufman); and the Section on Developmental Genetic Epidemiology, Intramural Research Branch, National Institute of Mental Health (Merikangas)

Abstract

OBJECTIVE—To report results of the clinical reappraisal study of lifetime DSM-IV diagnoses based on the fully-structured lay-administered WHO Composite International Diagnostic Interview Version 3.0 (CIDI) in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A).

METHOD—Blinded clinical reappraisal interviews with a probability sub-sample of 347 NCS-A respondents were administered using the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS) as the gold standard. DSM-IV/CIDI cases were over-sampled and the clinical reappraisal sample was weighted to adjust for this over-sampling.

RESULTS—Good aggregate consistency was found between CIDI and K-SADS prevalence estimates, although CIDI estimates were meaningfully higher than K-SADS estimates for specific phobia (51.2%) and oppositional-defiant disorder (38.7%). Estimated prevalence of any disorder, in comparison, was only slightly higher in the CIDI than K-SADS (8.3%). Strong individual-level CIDI vs. K-SADS concordance was found for most diagnoses. Area under the ROC curve (AUC, a measure of classification accuracy not influenced by prevalence) was .88 for any anxiety disorder, .89 for any mood disorder, .84 for any disruptive behavior disorder, .94 for any substance disorder, and .87 for any disorder. Although AUC was unacceptably low for alcohol dependence and bipolar I and II disorders, these problems were resolved by aggregation with alcohol abuse and bipolar I disorder, respectively. Logistic regression analysis documented that consideration of CIDI symptom-level data significantly improved prediction of some K-SADS diagnoses.

[*] Statistical expert and we would also like to acknowledge Steven Heeringa, Ph.D. as an additional statistical expert for this manuscript.

Address comments to RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; Tel. 617-432-3587, Fax 617-432-3588, Kessler@hcp.med.harvard.edu.

Disclosure: Dr. Kessler has been a consultant for GlaxoSmithKline Inc., Kaiser Permanente, Pfizer Inc., Sanofi-Aventis, Shire Pharmaceuticals, and Wyeth-Ayerst; has served on advisory boards for Eli Lilly & Company and Wyeth-Ayerst; and has had research support for his epidemiological studies from Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson Pharmaceuticals, Ortho-McNeil Pharmaceuticals Inc., Pfizer Inc., and Sanofi-Aventis. The remaining authors report no conflicts of interest.

CONCLUSIONS—These results document that the diagnoses made in the NCS-A based on the CIDI have generally good concordance with blinded clinical diagnoses.

Keywords

National Comorbidity Survey Adolescent Supplement (NCS-A); Composite International Diagnostic Interview (CIDI); mental disorders; epidemiology; validity

INTRODUCTION

While clinician-administered research diagnostic interviews are generally considered the gold standard for making diagnoses of mental disorders in research settings, gold standard interviews generally are not financially or logistically feasible in large-scale epidemiological research.[1] Diagnostic assessments of mental disorders in most large-scale community surveys are consequently based on fully-structured interviews administered by trained lay interviewers. Considerable controversy exists about whether these lay interviews generate diagnoses consistent with those based on gold standard clinician-administered semi-structured research diagnostic interviews.[2–4] The ability of epidemiological studies to produce clinically meaningful diagnoses is critical to their credibility and utility for clinicians, researchers, and policymakers.[5] A number of fully-structured lay-administered diagnostic interviews have been developed for adolescent mental illness, such as the Child Assessment Schedule (CAS),[5] the Child and Adolescent Psychiatric Assessment (CAPA),[6] the Diagnostic Interview for Children and Adolescents-Revised (DICA-R),[7] and the Diagnostic Interview Schedule for Children (DISC).[8] Psychometric studies of the validity of diagnoses based on these lay-administered interviews using parent and child reports compared to diagnoses based on blinded clinician-administered interviews have generally yielded only moderate levels of agreement (κ in the range .3–.6).[5,9,10] Agreement levels were higher, particularly for attention-deficit hyperactivity disorder, when ancillary information from parents and-or teachers is included in the diagnostic assessments.

The current paper presents data from a clinical reappraisal study carried out in conjunction with the US National Comorbidity Survey Adolescent Supplement (NCS-A). Companion papers discussion the background, rationale, and instruments used in the NCS-A[11] and present a description of the NCS-A study design.[12] A separate report presents a detailed statistical analysis of NCS-A design effects and an evaluation of the effects of weight trimming on the design bias-efficiency trade-off.[13]

We examined the concordance of diagnoses based on parent and child responses to the fully-structured instrument used in the NCS-A, the WHO Composite International Diagnostic Interview Version 3.0 (CIDI),[11,14] with diagnoses based on blinded clinical reappraisal interviews using a modified version of the semi-structured clinician-administered Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS).[15] Previous clinical reappraisal studies of earlier versions of the CIDI in adult community samples found that the CIDI generated DSM diagnoses that were generally consistent with those obtained in clinical reappraisal interviews based on the Structured Clinical Interview for DSM Disorders (SCID).[16–19] The results of earlier CIDI clinical reappraisal studies that generated diagnoses based on ICD-10 criteria using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN)[20] as the clinical gold standard have been more variable depending on the source of the sample.[21–23] CIDI 3.0 was developed to update earlier versions of CIDI by operationalizing DSM-IV criteria and improving validity with the benefit of insights gained from these earlier CIDI clinical reappraisal studies.[23] Subsequent clinical reappraisal studies of CIDI 3.0 with adults have documented generally good concordance of DSM-IV/CIDI diagnoses with diagnoses based on blinded clinical reappraisal interviews.[3, 23]

As discussed in more detail in a companion paper,11 pilot studies of CIDI 3.0 concordance with independent clinical assessments were carried out in clinical and community samples of early adolescents (ages 13–17) in preparation for the NCS-A. These studies used cognitive interviewing methods24, 25 to study participants' understanding of survey questions and procedures. Results of these cognitive interviews led to numerous CIDI modifications in question wording to make the language and characterization of symptoms more appropriate for adolescents. In addition, the instrument was expanded to include informant reports from a parent or parent-surrogate (henceforth referred to as a parent) for diagnoses where previous research has shown informant reports to be most valuable.26, 27

In order to assess concordance of diagnoses based on this modified version of the CIDI with blinded clinical diagnoses in the NCS-A sample, a nationally representative probability sub-sample of NCS-A respondents was administered a K-SADS clinical reappraisal interview. This clinical reappraisal sub-sample over-sampled NCS-A respondents with DSM-IV/CIDI diagnoses. Data analysis had three phases. First, we weighted the clinical reappraisal sub-sample to adjust for the over-sampling of CIDI cases and evaluated aggregate consistency of prevalence estimates based on the CIDI and K-SADS interviews. Second, we evaluated consistency of individual-level diagnostic classifications between the two instruments. Third, we examined whether concordance could be improved by modifying CIDI symptom-level criteria and developing prediction equations for K-SADS diagnoses based on both CIDI diagnoses and CIDI item-level data.

METHODS

The NCS-A sample

The NCS-A is a nationally representative face-to-face survey of 10,148 adolescents (ages 13–17) in the continental US carried out between February 2001 and January 2004 in a dual-frame sample that included a household sub-sample and a school sub-sample. The household sub-sample consisted of adolescent residents of the households that participated in the National Comorbidity Survey Replication (NCS-R), a nationally representative household survey of adults.[28] This sub-sample was interviewed between April 2001 and April 2003 and yielded 904 interviews, with a conditional (on adult participation in the NCS-R) response rate 85.9%. The school sub-sample was drawn from students in a probability sample of schools in the same counties as the NCS-R sample. The response rate of adolescents in the school sample was 74.7%, yielding 9,244 interviews. Each respondent (one adolescent and one parent or parent surrogate in each household) was given a $50 incentive for participation. Written informed consent from a parent or parent surrogate and written informed assent from the adolescent were obtained before collecting any data. The Human Subjects Committees of both Harvard Medical School (HMS) and the University of Michigan approved these recruitment, consent, and field procedures. Once collected, the data were weighted for within-household probability of selection (only in the household sample) and for residual discrepancies between the sample and population on a wide range of Census socio-demographic and geographic variables. A more detailed description of the NCS-A sample design and field procedures is presented elsewhere.13

The clinical reappraisal sample

The clinical reappraisal sample included a probability sample of 347 adolescent NCS-A respondents in addition to one parent of each adolescent drawn within strata defined by DSM-V/CIDI diagnoses. The sample was confined to households with telephones because the K-SADS clinical reappraisal interviews were administered by phone. Telephone administration is now widely accepted in clinical reappraisal studies based on evidence of comparable validity to in-person administration in both adults[29, 30] and adolescents.[31, 32] A great advantage of telephone administration is that a centralized and closely supervised clinical interview staff can carry out the interviews throughout the country without the geographic restrictions required for face-to-face clinical assessment. A disadvantage is that the small part of the population without telephones cannot be included in clinical calibration studies when interviews are done by telephone. In addition, telephone interviews cannot as readily use non-verbal communications to facilitate probing and scoring.

Respondents who met DSM-IV/CIDI criteria for one or more of the more uncommon disorders assessed in the NCS-A (e.g., agoraphobia, bipolar disorder, panic disorder, substance dependence with abuse) were sampled at a higher rate than respondents in a second sampling stratum who met criteria only for more common disorders. The lowest sampling fraction was for a third stratum of respondents who did not meet criteria for any lifetime DSM-IV/CIDI disorder. Respondents were selected into the clinical reappraisal sample with probabilities proportional to sample weight so as to reduce the effects of weighting. Each adolescent and parent respondent was given a $50 incentive for participation in the clinical reappraisal survey (over and above the $50 incentive for participation in the main survey). K-SADS interviews were conducted over the phone an average of 77 days after the CIDI interviews. Clinicians first conducted the interview with adolescents and then a parent. The focus was on lifetime prevalence, which is one reason the time interval between interviews was made much longer than the 1–3 days typical for validation studies that focus on the assessment of current prevalence. A long time interval was used in order to avoid respondent reluctance to participate in an intensive re-interview within a shorter time period from the original interview. A danger in doing this, though, is that the time interval is long enough that there might have been true change in lifetime diagnostic status, introducing a conservative bias into estimates of concordance between CIDI and K-SADS assessments.

Disorder assessment

CIDI 3.0 is described in detail elsewhere[14] and the modifications of CIDI 3.0 for the NCS-A are described in a companion paper.[11] In brief, CIDI 3.0 is a fully-structured research diagnostic interview that is designed to be administered by a trained lay interviewer. CIDI questions for the most part have a yes-no response format. The K-SADS, in comparison, is a semi-structured research diagnostic interview that is designed to be administered by trained clinical interviewers. K-SADS questions to respondents are designed to elicit rich verbal responses that form the basis of interviewer ratings about the presence versus absence of symptoms. The standard K-SADS was modified by deleting disorders not assessed in the NCS-A, focusing on a lifetime time frame, and somewhat streamlining the initial screening section of the interview to include information about respondent endorsement of diagnostic stem questions in the earlier CIDI interview.

The disorders assessed in this version of the K-SADS included six DSM-IV anxiety disorders (panic disorder with or without agoraphobia, agoraphobia without a history of panic disorder, generalized anxiety disorder, specific phobia, social phobia, post-traumatic stress disorder), three mood disorders (bipolar spectrum disorder, major depressive disorder, dysthymic disorder), three disruptive behavior disorders (attention/deficit-hyperactivity disorder, conduct disorder, oppositional-defiant disorder), and four substance use disorders (alcohol abuse with or without dependence, illicit drug abuse with or without dependence, alcohol dependence with a history of abuse, illicit drug dependence with a history of abuse). We also considered summary measures of any anxiety disorder, any mood disorder, any disruptive behavior disorder, any substance use disorder, and any disorder. All disorders in both the CIDI and K-SADS were diagnosed using DSM-IV organic exclusions and diagnostic hierarchy rules.

Study design

The clinical reappraisal study was designed to determine whether diagnostic classifications based on the CIDI are different from those made by carefully trained clinical interviewers using the K-SADS. The diagnoses included in the assessment were limited to those in the K-SADS, which means that some of the disorders assessed in the CIDI (e.g., separation anxiety disorder, intermittent explosive disorder) were not included in the clinical reappraisal study. It would have been desirable to assess these other disorders with separate clinical interviews, but this exceeded the resources of the project. As the entry questions (i.e., the diagnostic stem questions) in the CIDI and K-SADS are very similar, the distinction between the two types of interview hinges largely on differences in probes after endorsement of a diagnostic stem question. A major impediment to assessing this difference is that respondents are inconsistent in their endorsement of diagnostic stem questions over time. Indeed, previous methodological research has shown clearly that respondents in community surveys tend to report less and less as they are interviewed more and more, both within a single interview[33] and across multiple interviews,34 due to respondent fatigue. A major part of this pattern is a tendency for respondents to endorse a smaller number of diagnostic stem questions in follow-up interviews than in initial interviews even when the questions are identical,16 leading to the biased perception that initial structured interviews over-estimate prevalence compared to second clinical interviews.

In order to address this problem, we modified the conventional blinded clinical re-interview design in two important ways. First, the clinical interviewers were informed whether the respondent endorsed CIDI diagnostic stem questions, but not if the respondent met full criteria for diagnoses, in the main interview, and respondents were reminded of this endorsement of stem questions in their clinical reappraisal interviews so as to make sure they entered the diagnostic sections they entered in the initial interview. This approach reduced the problem of respondents denying diagnostic stems in the re-interviews of sections that they entered in the main interview. It is important to note, in this context, that the vast majority of community survey respondents who endorse CIDI stem questions do not go on to meet full DSM-IV/CIDI criteria for the associated disorder, which means that this partial un-blinding of interviewers did not introduce a bias towards clinical interviewers assuming that initial endorsement of a stem question meant that the CIDI diagnosis was positive. This bias might have existed if the over-sampling approach typically used in clinical reappraisal studies had been used; that is, if we only over-sampled respondents who were classified by the CIDI as meeting criteria for one or more DSM-IV disorders. The problem here is that the ratio of stem-positive to diagnosis-positive cases is increased in the conventional approach to over-sampling, which means that positive stem endorsement in such a sample usually means that the respondent met criteria for the disorder in the CIDI. We protected against this possibility in our design with a second modification of the conventional design: over-sampling not only CIDI cases but also respondents who endorsed the CIDI stem questions but failed to meet full DSM-IV/CIDI diagnostic criteria. This second kind of over-sampling was done at half the rate used to over-sample respondents who met full DSM-IV criteria for the disorder in the CIDI. This design modification made it impossible for clinical interviewers to infer CIDI diagnoses from stems. In cases where the respondent had not endorsed the CIDI stem question in the original interview, the K-SADS probing for a diagnostic stem endorsement was carried out in the conventional fashion in order to discover false negative responses in the CIDI. The clinical interviewers also had complete flexibility in entering diagnostic sections that had been previously skipped if any new information emerged subsequently in the interview.

Clinical interviewer training and supervision

Nine clinical interviewers were used to administer the K-SADS interviews. Six were PhD-level clinical psychologists (with 20, 11, 6, 4, 4, and 4 years of clinical experience), one an MA-level clinical psychologist (with 5 years of clinical experience), and two MSW psychiatric social workers (with 2 and 5 years of clinical experience). The clinical interviewers were trained by an experienced K-SADS trainer. The interviewers were allowed to participate in production interviewing only after they were tested to confirm that they were reliable in their administration and scoring of the K-SADS. All production interviews were closely supervised by an experienced clinical interviewer supervisor. Quality control monitoring included bi-weekly clinical interviewer meetings to prevent drift, supervisor review of all hard copy completed interviews, supervisor consultation with the trainer to resolve uncertainties regarding rating rules, re-contact of respondents whenever the supervisor felt that more information was needed to make a rating, periodic consultation with diagnostic experts who served as consultants for complex cases, and review of a random sub-sample of interview audiotapes. K-SADS diagnoses were made using the same best estimate diagnostic procedures used in clinical studies.[35]

In addition, independent reviews of more than half the clinical interviews were conducted by a team of experienced child and adult clinical interviewers (primarily doctoral-level psychologists and licensed clinical social workers with more than 5 years of experience working with adolescents) led by an experienced clinical interview supervisor (KRM) to establish standards for administration of diagnostic supplements and to maintain comparability with the adult CIDI interviews. This review raised questions about the thresholds used by the K-SADS interviewers for the diagnoses of conduct disorder (CD), phobias, and bipolar spectrum disorder (BPSD) being too high. Based on these concerns, 90 clinical reappraisal survey respondents for whom uncertainties existed regarding one or more of these diagnoses were administered a second clinical reappraisal interview to reassess this subset of disorders. The clinical interviewers who carried out these second clinical reappraisal interviews received additional special training in the assessment of CD, phobias, and BPSD and close supervision. Final K-SADS diagnoses of the disorders for these 90 respondents were based on a review and synthesis of information across the pair of clinical reappraisal interviews. In five cases where concerns about one or more of these three original diagnoses could not be resolved due to our inability to obtain a second reappraisal interview, clinical diagnoses were based on multiple imputation from reappraised cases who had similar CIDI and original K-SADS assessments.[36]

Analysis methods

After weighting the clinical reappraisal sample data to be representative of the full NCS-A sample on the stratification variables used in sub-sampling, we investigated whether CIDI prevalence estimates are biased in comparison to K-SADS prevalence estimates using McNemar χ^2 tests. McNemar tests are explicitly designed for paired comparisons of dichotomies. As with all significance tests reported in this paper, McNemar tests were carried out using the Taylor series design-based estimation method to adjust for the effects of weighting and clustering and over-sampling of CIDI cases.[37]

Individual-level concordance between diagnoses based on the CIDI and K-SADS was evaluated using two different descriptive measures, the area under the receiver operator characteristic curve (AUC)38 and Cohen’s κ, 39 although the main focus is on AUC. The κ statistic is presented as well because it is the most widely used measure of concordance in validity studies of psychiatric disorders, but it has been criticized because it is dependent on prevalence and consequently is often low in situations where there appears to be high agreement between low-prevalence measures.40¯42 An important implication is that κ varies across populations that differ in prevalence even when the populations do not differ in sensitivity (SN; the percent of true cases correctly classified) or specificity (SP; the percent of true non-cases correctly classified). As sensitivity and specificity are considered to be fundamental parameters of agreement, the comparison of κacross different populations cannot be used to evaluate cross-population variation in performance of a test. Critics of κ prefer to assess concordance with measures that are a function of SN and SP. The odds-ratio (OR) meets this requirement, as OR is equal to [SN × SP]/[(1-SN) × (1 - SP)].43 However, the upper end of the OR is unbounded, making it difficult to use the OR to evaluate the extent to which CIDI diagnoses are consistent with clinical diagnoses. Yules Q resolves this problem,44 as Q is a bounded transformation of OR [Q = (OR − 1)/(OR + 1)] that ranges between −1 and +1. Q can be interpreted as the difference in the probabilities of a randomly selected clinical case and a randomly selected clinical non-case that differ in their classification on the K-SADS being correctly versus incorrectly classified by the CIDI. The difficulty with Q, though, is that “tied pairs” (i.e., clinical cases and non-cases that have the same CIDI classification) are excluded, which means that Q does not tell us about actual prediction accuracy. The AUC resolves this problem by identifying half the tied pairs as correctly classified and the other half as incorrectly classified, based on the assumption that a 50-50 split of correct and incorrect classifications would be expected in the event of a tie. Although the AUC was developed to study the association between a continuous predictor and a dichotomous outcome, it can be used in the special case where the predictor is a dichotomy, in which case AUC equals (SN + SP)/2. As a result of this useful identity, we focus on AUC in our evaluation of diagnostic concordance between the CIDI and the K-SADS. We also report SN and SP, the key components of AUC in the dichotomous case, as well as κ and total classification accuracy (TCA). TCA is reported because it, like κ, is usually reported in studies of this type. We do not emphasize TCA in our interpretations, though, because it is biased toward high levels of agreement in situations such as ours in which the vast majority of the population does not have the disorder.

The final phase of analysis involved modifying the way in which the CIDI data were combined to generate estimates of diagnoses to improve concordance with K-SADS diagnoses. This was done using regression methods to predict K-SADS diagnoses from CIDI symptom-level data. By estimating CIDI-K-SADS concordance using several modifications of the CIDI data used, we were able to determine which CIDI data best estimated the K-SADS diagnosis. For example, several diagnoses require that symptoms are not better accounted for by another medical or mental disorder. While this criterion was included in the CIDI interviews, we found that our lay interviewers rarely reported this exclusion and that using the responses to the CIDI question about this exclusion in making diagnoses worsened rather than improved our prediction of clinician-assigned K-SADS diagnoses. As a result, we eliminated this diagnostic criterion from our algorithms for overall CIDI diagnosis for some disorders.

We also estimated a series of stepwise logistic regression equations that began in the first step with CIDI yes-no diagnoses for a given disorder predicting K-SADS diagnoses for the same disorder and then added CIDI symptom-level data in subsequent steps. This was done in an effort to determine if CIDI symptom-level data can be used to increase concordance of CIDI diagnoses with K-SADS diagnoses. As discussed in more detail elsewhere,[23] similar analyses in samples of adults showed clearly that concordance can be improved by taking into consider CIDI symptom-level data. Our goal in replicating this kind of analysis here was to see if the same was true among adolescents. The AUC statistic was again used to characterize concordance. A desirable feature of AUC in this regard is that it can be used when the predictor is a continuous measure of predicted probability of a dichotomous outcome, as it was when we converted the results of the logistic regression analysis into individual-level predicted probabilities. We used these predicted probabilities to impute a yes-no diagnosis for each respondent who was not in the clinical reappraisal sample by drawing a separate random number for each respondent from a binomial distribution that has a prevalence equal to the respondent's predicted probability. These imputed diagnoses were then compared to the K-SADS diagnoses to see if they had higher concordance than did the original CIDI diagnoses.

Significance tests for differences in prevalence estimates based on the CIDI and K-SADS interviews were consistently made using design-based .05-level, two-sided tests. We also comment on a number of non-significant trends in the data.

RESULTS

Aggregate concordance

It should be noted that the prevalence estimates presented here are *not* identical to the NCS-A prevalence estimates. Instead, they represent best estimates in the clinical reappraisal sample based on an attempt to weight the latter sample as best we could to represent the larger sample. There are inevitable limitations to this weighting procedure, though, based on the small size of the clinical reappraisal sample. Within the context of these limitations, the McNemar tests of CIDI vs. K-SADS prevalence differences are significant for three of the six anxiety disorders (agoraphobia, generalized anxiety disorder, specific phobia), one of the mood disorders (major depressive disorder or dysthymic disorder), one of the disruptive behavior disorders (oppositional-defiant disorder), and two substance disorders (alcohol and illicit drug dependence). (Table 1) Differences are significant for three of four classes of disorders (any anxiety, any mood, and any impulse). Prevalence estimates do not differ significantly for any of the other 9 disorders or for overall substance disorders.

The weighted CIDI prevalence estimate in this clinical reappraisal sample is higher than the K-SADS estimate for 5 of the 7 significant disorder differences. The largest differences involve anxiety disorders despite the ranking of anxiety prevalence estimates being very similar in the two instruments (r_s = .88), with specific phobia by far the most prevalent anxiety disorder in both the CIDI and K-SADS followed by social phobia and then PTSD. The CIDI prevalence estimate is substantially higher than the K-SADS estimate for specific phobia (19.2% vs. 12.7%, a 51.2% higher proportional prevalence in the CIDI than the K-SADS), although comparable for social phobia (9.8% vs. 9.2%) and PTSD (4.4% vs. 4.2%). The other three anxiety disorders – panic disorder, generalized anxiety disorder (GAD), and agoraphobia -- have much lower prevalence estimates in both instruments (1.5–3.3%), with the rank-order of these estimates differing across instruments due to a higher estimated prevalence of agoraphobia (2.6% vs. 1.5%) and lower estimated prevalence of GAD (2.6% vs. 3.3%) in the CIDI than K-SADS. The significantly higher prevalence estimate of any anxiety disorder in the CIDI than the K-SADS (31.4% vs. 25.0%) is due largely to specific phobia.

The CIDI prevalence estimate is significantly lower than the K-SADS estimate, in comparison, for overall mood disorders (21.9% vs. 23.7%), although this difference is modest in substantive terms. Prevalence differences vary across individual mood disorders, though, with the BPSD prevalence estimate marginally higher (6.6% vs. 6.2%) and the depression-dysthymia prevalence estimate somewhat lower (18.0% vs. 19.8%) in the CIDI than K-SADS. The major depression prevalence estimate does not differ significantly between the two instruments (17.7% vs. 17.5%), but the inclusion of dysthymic disorder with major depression results in a higher K-SADS estimate due to the CIDI under-diagnosing dysthymic disorder.

The CIDI prevalence estimate of any disruptive behavior disorder is significantly higher than the K-SADS estimate, as the CIDI estimate is substantially higher than the K-SADS estimate for oppositional-defiant disorder (14.7% vs. 10.6%, a 38.7% higher proportional prevalence in the CIDI than the K-SADS), although there are no meaningful differences in the two instruments for ADHD (7.9% vs. 7.8%) or conduct disorder (8.8% vs. 7.8%). The between-instrument difference is insignificant, in comparison, for any substance disorder (11.1% vs. 11.1%) despite prevalence estimates being significantly higher in the CIDI than the K-SADS for both alcohol and illicit drug dependence (1.2–1.9% vs. 0.5–0.9%). This inconsistency can be traced to the fact that substance dependence was assessed only among respondents with lifetime abuse in the CIDI and to the fact that between-instrument differences in the estimated prevalence of abuse are not significant (6.7–8.4% vs. 6.4–8.9%). The lifetime prevalence estimate of any disorder, finally, is higher in the CIDI (56.9%) than the K-SADS (52.5%), although this difference is modest in substantive terms.

Individual-level concordance

Using descriptors employed for roughly comparable values of κ,[45] individual-level concordance between lifetime CIDI and K-SADS diagnoses can be described as almost perfect (AUC greater than or equal to .9) or substantial (AUC in the range .8–.9) for all but three disorders. The exceptions are PTSD and ADHD, where concordance is moderate (AUC in the range .7–.8), and alcohol dependence, where concordance is slight (AUC in the range .5–.6). (Table 2) Individual-level concordance is also either substantial or almost perfect for having at least one diagnosis in each class or any diagnosis overall. Classifications would be identical if we focused on κ rather than AUC, with the exception of illicit drug dependence with abuse, where concordance would be considered moderate based on the κ value of .50 but substantial based on the AUC of .90.

Turning to the two main components of AUC (SN and SP), the vast majority of respondents classified by the K-SADS as having any lifetime disorder (92.0%) are also detected as cases by the CIDI (high SN). In addition, the great majority of K-SADS non-cases (81.9%) are also classified by the CIDI as not meeting lifetime criteria for any disorder (high SP). Inspection of our results indicates that CIDI sensitivity within classes is highest for substance (89.0%) and anxiety (88.5%) disorders, lower for mood disorders (81.4%), and lowest, but still very good, for disruptive behavior disorders (77.9%). CIDI classifications of K-SADS non-cases as not having a disorder (i.e., SP) are highest for substance (98.7%) and mood (96.6%) disorders, lower for disruptive behavior disorders (90.9%), and lowest for anxiety disorders (87.6%). The comparatively low SP for anxiety disorders is consistent with the earlier observation that the estimated prevalence of anxiety disorders is higher in the CIDI than the K-SADS.

At the disorder level, CIDI detection of K-SADS cases is greater than 90% for four disorders, greater than 80% for three others, and greater than 50% for all but one other. The single instance in which the CIDI fails to detect the majority of K-SADS cases involves alcohol dependence (AD), where a low K-SADS prevalence in conjunction with a high weight for a single outlier K-SADS case results in a very low estimate of SN (12.4%). When combined with the fact that an extremely low proportion of CIDI cases of AD are confirmed by the K-SADS (5.4%), the low SN of the CIDI AD diagnosis leads us to conclude that analysis of CIDI alcohol disorders in the NCS-A should focus on abuse with or without dependence, where both SN and SP are excellent (96.5% and 99.4%, respectively), rather than on dependence. It should be noted that even though the CIDI skip rule of not assessing dependence among respondents who deny any lifetime symptoms of abuse introduces the possibility of false negative assessments of dependence, the high SN for the diagnosis of abuse or dependence shows that this kind of error is very rare in the NCS-A.

The proportion of K-SADS non-cases confirmed as such in the CIDI is consistently higher than 90% for each of the disorders considered, with the vast majority higher than 98%. Interpretation of this finding, though, should take into account the fact that SP of 90% would be expected by chance with an unbiased but completely random diagnosis of a disorder that has 10% prevalence, as the expected proportion of non-cases from any random sample would be 90% in such a case. It is consequently worth noting that the quantity 1–SP, the proportion of K-SADS non-cases classified by the CIDI as cases, is consistently less than half as high as the CIDI prevalence estimate, indicating good separation between true positives and false positives, with the exception of two disorders: AD and drug dependence (DD). We already noted above that the CIDI diagnosis of AD is unreliable. This new observation raises questions about DD as well. As with AD, the proportion of CIDI cases confirmed by a K-SADS diagnosis of DD is very low (37.1%). Indeed, the reappraisal interviews suggest that even the K-SADS interviewers found it difficult to assess substance dependence in this population, presumably due to the duration of alcohol and drug use being relatively brief compared to adult samples. Based on these observations, we conclude that analysis of CIDI illicit drug disorders in the NCS-A should focus on abuse with or without dependence, where SN is good (85.1%) and SP is excellent (99.0%), rather than on dependence.

The special case of bipolar spectrum disorder

We have a special interest in adolescent sub-threshold bipolar disorder as a correlate of ADHD and as a predictor of the subsequent onset of threshold bipolar disorder in anticipated adult follow-up surveys of the NCS-A sample. Respondents were classified as having sub-threshold BPD if they met any of the following CIDI criteria: (i) a history of recurrent sub-threshold hypomania (at least two Criterion B symptoms, such as grandiosity or decreased need for sleep, along with all other criteria for hypomania) plus a history of major depressive episode (MDE); (ii) a history of recurrent hypomania in the absence of recurrent MDE; or (iii) a history of recurrent sub-threshold hypomania in the absence of MDE. The reduction in number of required symptoms for a determination of sub-threshold hypomania was confined to two Criterion B symptoms (compared to the DSM-IV requirement of three or four if the mood is only irritable) in order to retain the core features of hypomania in the sub-threshold definition. Recurrent hypomania and sub-threshold hypomania absent MDE were included in the definition because hypomania in the absence of MDE is part of the DSM-IV definition of BPD NOS. Bipolar spectrum disorder (BPSD) was defined as a lifetime history of BP-I, BP-II or sub-threshold BPD.

As noted above, the estimated prevalence of BPSD using the CIDI is very similar to the estimate using the K-SADS (6.6% vs. 6.2%). However, the estimated prevalence of BP-I is significantly lower using the CIDI than the K-SADS (0.5% vs. 1.0%), while the estimated prevalence of BP-II is significantly higher using the CIDI than the K-SADS (1.8% vs. 1.3%). (Table 3) The estimated prevalence of sub-threshold BPD, in comparison, does not differ significantly between the two instruments (4.3% vs. 3.9%). Because the lower estimated prevalence of BP-I and higher estimated prevalence of BP-II in the CIDI cancel out, the estimated prevalence of having either BP-I or BP-II is identical using the CIDI or the K-SADS (2.3% vs. 2.3%). Individual-level concordance between the two instruments is also considerably higher for the diagnosis of BP-I or BP-II (AUC = .85) than for either BP-I alone (AUC = .72) or BP-II alone (AUC = .74). This means that between-instrument differences in threshold BPD are due largely to disagreements about whether a case qualifies for a diagnosis of BP-I or BP-II. Agreement for a diagnosis of sub-threshold BPD, in comparison, is quite good (AUC = .91). Based on these observations, we conclude that CIDI analysis of BPSD should focus on the two diagnoses of BP-I/II and sub-threshold BPD and that the distinction between BP-I and BP-II should be considered unreliable.

Lifetime concordance using CIDI symptom-level data

For GAD, ADHD, social phobia, and oppositional-defiant disorder, we found that small modifications to the CIDI diagnostic algorithm using symptom counts, impairment criteria, and available continuous data can meaningfully improve concordance with the K-SADS both at the aggregate level (i.e., increase concordance between CIDI and K-SADS prevalence estimates) and at the individual level (i.e., increase AUC). Diagnoses of these disorders in the full sample are consequently based on these modified diagnostic algorithms. In addition, for each disorder, a stepwise logistic regression analysis was carried out to determine if CIDI symptom questions could significantly improve prediction of K-SADS diagnosis after including the CIDI diagnosis as a dummy variable in the prediction equation. Consistent evidence was found for significant associations of this sort. For example, in the case of predicting K-SADS major depressive disorder, CIDI information about total number of symptoms endorsed and severe symptoms (e.g., suicide attempt) significantly predicted K-SADS diagnoses over and about the dichotomous CIDI diagnosis.

Based on such prediction equations, each respondent in the clinical reappraisal sample was assigned a predicted probability of each K-SADS diagnosis. The AUC for the predicted probability in relation to the K-SADS diagnosis is meaningfully higher than for the dichotomous CIDI diagnosis classification for eight disorders: three anxiety disorders (social phobia, GAD, PTSD), two disruptive behavior disorders (ADHD, ODD), and three substance disorders (alcohol dependence, drug abuse, drug dependence). (Table 4) Four of these increases are fairly modest in magnitude (.06–.09%), increasing AUC for disorders where concordance is already either almost perfect (drug abuse and dependence) or substantial (social phobia). In one of these four cases (PTSD), though, AUC increased from being in the moderate range (.79) to the substantial range (.88). Three of the remaining four increases (GAD, ADHD, ODD) are somewhat greater (.11–.16%), increasing AUC from moderate or substantial (.78–.84) to almost perfect (.91–.98). The last increase, finally, is dramatic, with the slight AUC for alcohol dependence based on the dichotomous classification (.56) increasing to an almost perfect value (.98) in the continuous classification.

CONCLUSION

Before turning to substantive interpretation, a number of methodological limitations must be acknowledged. First, K-SADS interviews were carried out by telephone and CIDI interviews were carried out face-to-face. Even though telephone interviews constitute a valid mode of clinical assessment in both adults[29, 30] and adolescents,[31, 32] we do not know what would have happened if the same mode of administration had been used in both interviews. Second, the design of the clinical reappraisal study, in which clinical interviewers were provided information about respondent reports to diagnostic stem questions in the initial interview, might have biased results. Third, findings may be biased by the tendency for respondents to report more symptoms in first interviews than subsequent interviews.[9] Such bias might have been minimized by counter-balancing order of CIDI and K-SADS interviews, but that was not feasible in our design. We tried to minimize this bias by having a substantial period of time between the two interviews, but this bias might nonetheless remain to some extent. In addition, the comparatively long time lag between CIDI and K-SADS interviews might have resulted in some first onsets occurring in the interval, introducing a conservative bias into estimates of concordance between CIDI and K-SADS diagnoses.

It should be noted that we focused only on prevalence rather than severity.[46] The high prevalence of mental disorders in the community makes it more relevant for policy purposes to study disorders with higher-than-average clinical severity.[47] It has also been argued that the clinical relevance of epidemiological studies would be improved by considering dimensional measures of clinical severity.[48, 49] Criticism along these lines might contend that good diagnostic concordance such as documented here is less relevant than information about diagnostic concordance in distinguishing severe cases from mild cases and about consistency of dimensional clinical severity ratings. It is noteworthy in this regard that the CIDI includes fully structured versions of standard clinical severity scales to assess the severity of individual disorders, such as the Quick Inventory of Depressive Symptoms Self-Report Version[50] to assess the severity of major depression and the Panic Disorder Severity Scale Self-Report Form[51] to assess the severity of panic disorder. In addition, the WHO Disability Assessment Schedule[52] is included in the CIDI to assess the severity of overall psychopathology. Although these dimensional measures were not considered in the current report, they are available in the NCS-A to consider disorder severity rather than only disorder prevalence.

Despite the focus of the current study only on diagnoses rather than also on severity, information about diagnostic concordance is useful in determining whether DSM-IV diagnostic thresholds and criteria are defined consistently in the CIDI versus the K-SADS. Our results show that the CIDI diagnostic thresholds are generally consistent with K-SADS thresholds, with the two exceptions of specific phobia and oppositional-defiant disorder. In the latter two cases, the CIDI thresholds are well below the K-SADS thresholds, resulting in proportionally much higher prevalence estimates in the CIDI than K-SADS (51.2% for specific phobia; 38.7% for oppositional defiant disorder). The problem with the CIDI assessments of these two diagnoses involves the fact that both evaluated core symptoms with a yes-no checklist that failed to distinguish symptoms in terms of persistence or severity. We suspect that the use of dimensional rather than dichotomous ratings in future versions of the CIDI would help resolve these problems. The other cases where CIDI diagnoses were significantly (in a statistical, rather than substantive, sense of that term) higher than K-SADS diagnoses all involved either a substantively small proportional difference for a disorder with very high prevalence (major depression/dysthymia) or substantively small absolute differences for disorders with low prevalence (GAD, agoraphobia, alcohol and drug dependence).

We found that biased prevalence estimates in the CIDI could be corrected by using predicted probabilities of K-SADS diagnoses instead of CIDI diagnoses as outcome measures. As discussed in more detail elsewhere21 and illustrated in a series of recent disorder-specific analyses of adult disorders,53–55 it is practical to use predicted probabilities of clinical diagnoses in substantive analyses of CIDI surveys by imputing these predicted probabilities to all survey respondents based on the prediction equations generated in the clinical reappraisal sub-sample. These predicted probabilities can then either be treated as outcomes in substantive analyses or can be used as input to more complex analyses that use the method of multiple imputation (MI)[56] to make estimates of the prevalence and correlates of clinical diagnoses. Comparison with parallel estimates of the prevalence and correlates of CIDI diagnoses can be used in such cases to carry out much more fine-grained analyses of consistency with clinical diagnoses than conventional analyses of diagnostic concordance. We consequently plan to make use of predicted probabilities in substantive analyses of the NCS-A data to correct problems with diagnoses where CIDI and K-SADS prevalence estimates differed substantially (most notably, specific phobia and oppositional-defiant disorder).

Individual-level concordance between diagnoses based on the CIDI and the K-SADS were generally good. In the one case where individual-level concordance was slight, involving alcohol dependence, much higher concordance was found for the broader diagnosis of alcohol abuse. Similarly, although the assessment of illicit drug dependence suffers from low PPV, this problem was addressed by considering the broader diagnosis of illicit drug abuse. We consequently plan to focus on the diagnoses of alcohol and illicit drug abuse rather than on dependence in our substantive analyses. A less extreme version of the same situation occurred in distinguishing Bipolar I from Bipolar II, where we found that concordance of diagnoses based on the CIDI and K-SADS improved when we combined both diagnoses. We therefore plan to combine BP-I and BP-II in our substantive analyses. In all these cases (i.e., substance dependence vs. abuse and BP-I vs. BP-II), the severe form of the disorder is comparatively rare among children and adolescents and the CIDI severity questions are too coarse to make powerful distinctions between the severe and less severe forms. Future versions of the CIDI should modify these sections to increase the ability to make these distinctions.

We also documented that the few cases in which individual-level diagnostic concordance is less than substantial can be corrected by developing dimensional probability-of-disorder measures based on CIDI symptom data. Three cases of this sort exist: PTSD, ADHD, and alcohol dependence. This means that although we are not able to reproduce K-SADS diagnoses of these disorders with high accuracy at the individual level, we can generate predicted probabilities of these diagnoses that have excellent concordance with K-SADS distributions, allowing us to estimate prevalence and correlates of these disorders with good accuracy using statistical methods appropriate to the analysis of predicted probabilities.[55] The one exception is our inability to make accurate distinctions between bipolar I and bipolar II disorders. Given the rarity of threshold bipolar disorder among adolescents, we failed in our attempts to develop a logistic regression equation that had high AUC in distinguishing between these two disorders. As a result, all NCS-A analyses of threshold bipolar disorder will combine bipolar I and bipolar II cases into a single category.

As noted in the introduction, previous validation studies of lay-administered diagnostic interviews with clinician-administered gold standard interviews administered to adolescents generally found relatively low concordance,[5] particularly for disruptive behaviour disorders,[33] although concordance increased when informant reports were obtained from parents and/or teachers.[5, 9] As noted in the introduction, those studies generally found concordance in the range κ = .3–.6. The aggregate κ estimates documented in our study are generally above this range. This might be due to the fact that numerous features of CIDI 3.0 improve on earlier fully-structured research diagnostic interviews in interviewer training, quality control, question wording, and interview flow. The inclusion of a separate section to review lifetime diagnostic stem questions for all disorders might have played an especially important part in this respect, as previous research has shown that this approach leads to a substantial increase in the endorsement of lifetime stem questions.[16] The modification of CIDI 3.0 questions based on cognitive interviewing might also have been involved.[14] Our clinical reappraisal study design, most notably the separation of the clinical reappraisal interview by two months to minimize respondent fatigue, and the un-blinding of clinical interviewers to CIDI diagnostic stem questions to encourage reluctant respondents who reported episodes in the CIDI to discuss those episodes rather than conceal them, might also have contributed to the good concordance, although, as noted above, these design features can be seen as limiting external validity.

Although the word *validation* is often used to characterize the kinds of results reported here, this term is not entirely accurate due to the fact that the K-SADS diagnoses cannot be taken as perfect representations of true DSM disorders. This is true both because K-SADS test-retest reliability is imperfect[15] and because some respondents in community surveys consciously hide information about their mental disorders from clinical interviewers.[57] This imperfect validity, which characterizes not only the K-SADS but all “gold standard interviews,”15, 58 presumably attenuates associations with diagnoses based on fully-structured diagnostic interviews. Consistent with this thinking, the application of external criteria of validity, such as measures of impairment and service use, generally yield evidence of stronger associations than those found with independent diagnostic interviews.59 Based on these considerations, the estimates of concordance reported here should be considered lower bound estimates of CIDI validity. A good empirical illustration of this thinking can be found in the work of Booth et al.,[60] who compared lifetime diagnoses of major depression based on an earlier version of CIDI administered to an adult sample with diagnoses based on SCID clinical reappraisal interviews, where κ was .53. However, when CIDI diagnoses were compared with more accurate LEAD standard diagnoses (longitudinal, expert, and all data)[61] that used not only the SCID, but also all the clinical information available, to arrive at an improved estimate of clinical diagnoses, κ increased to .67.

In conclusion, the results reported here demonstrate that lifetime DSM-IV diagnoses based on the CIDI as implemented in the NCS-A have good individual-level concordance with diagnoses based on blinded clinical reappraisal interviews using the K-SADS, that prevalence estimates based on the two instruments are fairly similar in substantive terms for most disorders, and that symptom-level modifications can be used to correct prevalence estimates in most cases where between-instrument differences in prevalence estimates are substantively meaningful. As noted in the first paper in this series[11] there is considerable need for national data on the prevalence and correlates of psychiatric disorders in adolescents.[62] The practical utility of such data relies on accurate classification of disorders, a complex task given inconsistencies in diagnostic decision-making by clinicians.[63, 64] Substantial efforts were made to ensure that the CIDI provided clinically meaningful diagnoses of adolescents, including the use of cognitive interviewing strategies described elsewhere[11] that built on earlier iterative CIDI revisions and refinements.[14, 16, 17, 65] The results of the current study show that the CIDI has good concordance with clinician diagnoses, providing a solid foundation for later substantive analyses of the NCS-A data.

Acknowledgments

The NCS-A is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We thank the staff of the WMH Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. The WMH Data Coordination Centres have received support from NIMH (R01-MH070884, R13-MH066849, R01-MH069864, R01-MH077883), NIDA (R01-DA016558), the Fogarty International Center of the National Institutes of Health (FIRCA R03-TW006481), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, and the Pan American Health Organization. The WMH Data Coordination Centres have also received unrestricted educational grants from Astra Zeneca, BristolMyersSquibb, Eli Lilly and Company, GlaxoSmithKline, Ortho-McNeil, Pfizer, Sanofi-Aventis, and Wyeth. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/.

FINANCIAL SUPPORT

The National Comorbidity Survey Replication Adolescent Supplement (NCS-A) is supported by the National Institute of Mental Health (NIMH; U01-MH60220) with supplemental support from the National Institute on Drug Abuse (NIDA), the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044780), and the John W. Alden Trust. The work of Dr. Merikangas and her staff on the NCS-A is additionally supported by the NIMH Intramural Research Program, while the work of Dr. Zaslavsky and his staff on the validity of the NCS-A measures is supported by NIMH grant R01-MH66627. The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or U.S. Government. A complete list of NCS-A publications can be found at http://www.hcp.med.harvard.edu/ncs. Send correspondence to ncs@hcp.med.harvard.edu.

REFERENCES

1. Jewell J, Handwerk M, Almquist J, Lucas C. Comparing the validity of clinician-generated diagnosis of conduct disorder to the diagnostic interview schedule for children. J Clin Child Adolesc Psychol 2004;33:536–546. [PubMed: 15271611]

2. Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Psychol Med 2001;31:1001–1013. [PubMed: 11513368]

3. Haro JM, Arbabzadeh-Bouchez S, Brugha TS, et al. Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. Int J Methods Psychiatr Res 2006;15:167–180. [PubMed: 17266013]

4. Narrow WE, Rae DS, Robins LN, Regier DA. Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. Arch Gen Psychiatry 2002;59:115–123. [PubMed: 11825131]

5. Schwab-Stone ME, Shaffer D, Dulcan MK, et al. Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3). J Am Acad Child Adolesc Psychiatry 1996;35:878–888. [PubMed: 8768347]

6. Angold A, Prendergast M, Cox A, Harrington R, Simonoff E, Rutter M. The Child and Adolescent Psychiatric Assessment (CAPA). Psychol Med 1995;25:739–753. [PubMed: 7480451]

7. 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). J Am Acad Child Adolesc Psychiatry 1987;26:649–653. [PubMed: 3667494]

8. Costello EJ, Edelbrock CS, Costello AJ. Validity of the NIMH Diagnostic Interview Schedule for Children: a comparison between psychiatric and pediatric referrals. J Abnorm Child Psychol 1985;13:579–595. [PubMed: 4078188]

9. Boyle MH, Offord DR, Racine Y, et al. Evaluation of the Diagnostic Interview for Children and Adolescents for use in general population samples. J Abnorm Child Psychol 1993;21:663–681. [PubMed: 8126319]

10. Ezpeleta L, de la Osa N, Domenech JM, Navarro JB, Losilla JM, Judez J. Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents--DICA-R--in an outpatient sample. J Child Psychol Psychiatry 1997;38:431–440. [PubMed: 9232488]

11. Merikangas KR, Avenevoli S, Costello EJ, Koretz D, Kessler RC. Background and measures in the National Comorbidity Survey Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry. in press.

12. Kessler RC, Avenevoli S, Costello EJ, et al. The Design of the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry. in press.

13. Kessler RC, Avenevoli S, Costello EJ, et al. Design and field procedures in the US National Comorbidity Survey Replication Adolescent (NCS-A) Supplement. Int J Methods Psychiatr Res. in press.

14. Kessler, RCc; Üstün, TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 2004;13:93–121. [PubMed: 15297906]

15. Kaufman J, Birmaher B, Brent D, et al. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 1997;36:980–988. [PubMed: 9204677]

16. Kessler RC, Wittchen H-U, Abelson JM, et al. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. Int J Methods Psychiatr Res 1998;7:33–55.

17. Wittchen HU. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res 1994;28:57–84. [PubMed: 8064641]

18. Wittchen HU, Kessler RC, Zhao S, Abelson J. Reliability and clinical validity of UM-CIDI DSM-III-R generalized anxiety disorder. J Psychiatr Res 1995;29:95–110. [PubMed: 7666382]

19. Wittchen HU, Zhao S, Abelson JM, Abelson JL, Kessler RC. Reliability and procedural validity of UM-CIDI DSM-III-R phobic disorders. Psychol Med 1996;26:1169–1177. [PubMed: 8931163]

20. Wing JK, Babor T, Brugha T, et al. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch Gen Psychiatry 1990;47:589–593. [PubMed: 2190539]

21. Andrews G, Peters L, Guzman AM, Bird K. A comparison of two structured diagnostic interviews: CIDI and SCAN. Aust N Z J Psychiatry 1995;29:124–132. [PubMed: 7625961]

22. Jordanova V, Wickramesinghe C, Gerada C, Prince M. Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. Psychol Med 2004;34:1013–1024. [PubMed: 15554572]

23. Kessler RC, Abelson J, Demler O, et al. Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). Int J Methods Psychiatr Res 2004;13:122–139. [PubMed: 15297907]

24. DeMaio, T.; Landreth, A. Do different cognitive interview techniques produce different results?. In: Presser, S.; Rothgeb, J.; Couper, M., et al., editors. Methods for Testing and Evaluating Survey Questionnaires. New York, NY: Wiley; 2004.

25. Willis, G. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage Publications; 2004.

26. Grills AE, Ollendick TH. Issues in parent-child agreement: the case of structured diagnostic interviews. Clin Child Fam Psychol Rev 2002;5:57–83. [PubMed: 11993545]

27. Johnston C, Murray C. Incremental validity in the psychological assessment of children and adolescents. Psychol Assess 2003;15:496–507. [PubMed: 14692845]

28. Kessler RC, Merikangas KR. The National Comorbidity Survey Replication (NCS-R): background and aims. Int J Methods Psychiatr Res 2004;13:60–68. [PubMed: 15297904]

29. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A population-based twin study of major depression in women. The impact of varying definitions of illness. Arch Gen Psychiatry 1992;49:257–266. [PubMed: 1558459]

30. Sobin E, Weissman MM, Goldstein RB, Adams P. Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. Psychiatr Genet 1993;3:227–233.

31. Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Measuring depression in the community: a comparison of telephone and personal interviews. Public Opin Q 1982;46:110–121. [PubMed: 10256145]

32. Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. Am J Psychiatry 1997;154:1593–1598. [PubMed: 9356570]

33. Jensen PS, Watanabe HK, Richters JE. Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. J Abnorm Child Psychol 1999;27:439–445. [PubMed: 10821625]

34. Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC. Long-term reliability of diagnosing lifetime major depression in a community sample. Arch Gen Psychiatry 1986;43:435–440. [PubMed: 3964022]

35. Leckman JF, Sholomskas D, Thompson WD, Belanger A, Weissman MM. Best estimate of lifetime psychiatric diagnosis: a methodological study. Arch Gen Psychiatry 1982;39:879–883. [PubMed: 7103676]

36. Rubin DB. The Bayesian bootstrap. The Annals of Statistics 1981;9:130–134.

37. Wolter, KM. Introduction to Variance Estimation. New York: Springer-Verlag; 1985.

38. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. [PubMed: 7063747]

39. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measur 1960;20:37–46.

40. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–429. [PubMed: 8501467]

41. Cook, RJ. Kappa and its dependence on marginal rates. In: Armitage, P.; Colton, T., editors. The Encyclopedia of Biostatistics. New York: Wiley; 1998. p. 2166-2168.

42. Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. J Am Acad Child Adolesc Psychiatry 2003;42:1524–1529. [PubMed: 14627890]

43. Agresti, A. An Introduction to Categorical Data Analysis. New York: John Wiley and Sons; 1996.

44. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985;42:725–728. [PubMed: 4015315]

45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]

46. Brugha TS. The end of the beginning: a requiem for the categorization of mental disorder? Psychol Med 2002;32:1149–1154. [PubMed: 12420884]

47. Regier DA. Community diagnosis counts. Arch Gen Psychiatry 2000;57:223–224. [PubMed: 10711907]

48. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996;26:477–486. [PubMed: 8733206]

49. Shear MK, Brown TA, Barlow DH, et al. Multicenter collaborative panic disorder severity scale. Am J Psychiatry 1997;154:1571–1575. [PubMed: 9356566]

50. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54:573–583. [PubMed: 12946886]

51. Houck PR, Spiegel DA, Shear MK, Rucci P. Reliability of the self-report version of the panic disorder severity scale. Depress Anxiety 2002;15:183–185. [PubMed: 12112724]

52. Rehm J, Üstün TB, Saxena S, et al. On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. Int J Methods Psychiatr Res 1999;8:110–123.

53. Kessler RC, Adler L, Barkley R, et al. The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. Am J Psychiatry 2006;163:716–723. [PubMed: 16585449]

54. Kessler RC, Birnbaum H, Demler O, et al. The prevalence and correlates of nonaffective psychosis in the National Comorbidity Survey Replication (NCS-R). Biol Psychiatry 2005;58:668–676. [PubMed: 16023620]

55. Lenzenweger MF, Lane MC, Loranger AW, Kessler RC. DSM-IV personality disorders in the National Comorbidity Survey Replication. Biol Psychiatry 2007;62:553–564. [PubMed: 17217923]

56. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons; 1987.

57. Kranzler HR, Tennen H, Babor TF, Kadden RM, Rounsaville BJ. Validity of the longitudinal, expert, all data procedure for psychiatric diagnosis in patients with psychoactive substance use disorders. Drug Alcohol Depend 1997;45:93–104. [PubMed: 9179511]

58. Ambrosini PJ. Historical development and present status of the schedule for affective disorders and schizophrenia for school-age children (K-SADS). J Am Acad Child Adolesc Psychiatry 2000;39:49–58. [PubMed: 10638067]

59. Angold A, Costello EJ. The Child and Adolescent Psychiatric Assessment (CAPA). J Am Acad Child Adolesc Psychiatry 2000;39:39–48. [PubMed: 10638066]

60. Booth BM, Kirchner JE, Hamilton G, Harrell R, Smith GR. Diagnosing depression in the medically ill: validity of a lay-administered structured diagnostic interview. J Psychiatr Res 1998:353–360. [PubMed: 9844951]

61. Spitzer RL. Psychiatric diagnosis: are clinicians still necessary? Compr Psychiatry 1983;24:399–411. [PubMed: 6354575]

62. Report of the Surgeon General's Conference on Children's Mental Health: A National Action Agenda. Washington, DC: Department of Health and Human Services; 2000. U.S. Public Health Service.

63. Galanter CA, Patel VL. Medical decision making: a selective review for child psychiatrists and psychologists. J Child Psychol Psychiatry 2005;46:675–689. [PubMed: 15972065]

64. Lewczyk CM, Garland AF, Hurlburt MS, Gearity J, Hough RL. Comparing DISC-IV and clinician diagnoses among youths receiving public mental health services. J Am Acad Child Adolesc Psychiatry 2003;42:349–356. [PubMed: 12595789]

65. Composite International Diagnostic Interview. Geneva, Switzerland: World Health Organization; 1990. World Health Organization.

66. Hasin DS, Grant BF. The co-occurrence of DSM-IV alcohol abuse in DSM-IV alcohol dependence: results of the National Epidemiologic Survey on Alcohol and Related Conditions on heterogeneity that differ by population subgroup. Arch Gen Psychiatry 2004;61:891–896. [PubMed: 15351767]

Table 1

Consistency of lifetime prevalence estimates of DSM-IV disorders based on the CIDI and the K-SADS in the NCS-A clinical reappraisal sample (n=347)

	CIDI		K-SADS		McNemar
I. Anxiety disorders	**%**	**(se)**	**%**	**(se)**	**$\chi^2$1**
Panic disorder	2.4	(0.5)	2.1	(0.7)	1.0
Agoraphobia without panic disorder	2.6	(0.6)	1.5	(0.7)	8.0*
Specific phobia	19.2	(3.1)	12.7	(2.4)	41.6*
Social phobia	9.8	(1.4)	9.2	(1.7)	1.5
Generalized anxiety disorder	2.6	(0.8)	3.3	(1.0)	3.3*
Post-traumatic stress disorder	4.4	(1.0)	4.2	(1.2)	0.9
Any anxiety disorder	31.4	(3.3)	25.0	(3.0)	40.5*
II. Mood disorders					
Major depressive episode (MDE)	17.7	(2.1)	17.5	(2.4)	1.4
Major depressive episode or dysthymic disorder	18.0	(2.2)	19.8	(2.5)	8.3*
Bipolar spectrum disorder [1]	6.6	(1.7)	6.2	(1.7)	3.7
Any mood disorder	21.9	(2.6)	23.7	(2.8)	10.2*
III. Disruptive behavior disorders					
Attention-deficit/hyperactivity disorder	7.9	(1.6)	7.8	(1.6)	0.1
Conduct disorder	8.8	(3.6)	7.8	(3.5)	2.5
Oppositional-defiant disorder	14.7	(4.1)	10.6	(4.0)	112.0*
Any disruptive behavior disorder	20.8	(2.7)	17.0	(2.5)	55.5*
IV. Substance disorders[2]					
Alcohol abuse with or without dependence	6.7	(1.6)	6.4	(1.7)	0.8
Alcohol dependence with abuse	1.2	(0.5)	0.5	(0.4)	24.9*
Illicit drug abuse with or without dependence	8.4	(1.7)	8.9	(1.8)	1.9
Illicit drug dependence with abuse	1.9	(0.7)	0.9	(0.5)	62.8*
Any substance disorder	11.1	(2.1)	11.1	(2.2)	2.1
V. Any					
Any lifetime diagnosis	56.9	(4.5)	52.5	(4.0)	27.1*

	CIDI		K-SADS		McNemar
I. Anxiety disorders	**%**	**(se)**	**%**	**(se)**	**$\chi^2$1**
Two or more lifetime diagnoses	22.5	(2.7)	21.5	(2.7)	2.4
Three of more lifetime diagnoses	9.5	(1.4)	9.7	(1.5)	0.9

[*] Significant at the 0.05 level, two-sided test.

[1] Bipolar spectrum disorder includes BP-I, BP-II, and sub-threshold BPSD. See the section of the text on Disorder Assessment for our operation definition of sub-threshold BPSD.

[2] Substance abuse was diagnosed in both the CIDI and K-SADS with or without dependence. The CIDI assessment of substance dependence was made only among respondents who met lifetime criteria for abuse based on the finding in an early study that the prevalence of dependence without abuse is very uncommon.[12] This result has recently been called into question.[66] The K-SADS assessment of substance dependence was made with or without a history of abuse. The fact that the estimated prevalence of any substance disorder in the CIDI is identical to the estimate in the K-SADS confirms the assumption that dependence seldom occurred in the absence of a history of abuse in this sample.

Table 2

Individual-level concordance of lifetime diagnoses of DSM-IV disorders based on the CIDI with clinical diagnoses based on the K-SADS, where the K-SADS is treated as the gold standard, in the NCS-A clinical reappraisal sample (n=347)

	SN[1]		SP[1]		PPV[1]		NPV[1]		TCA[1]		κ[1]		
	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	AUC[1]
I. Anxiety disorders													
Panic disorder	74.4	(13.8)	99.2	(0.4)	65.8	(17.0)	99.4	(0.4)	98.6	(0.6)	.69	(0.1)	.87
Agoraphobia without panic disorder	81.9	(15.1)	98.7	(0.5)	48.9	(19.1)	99.7	(0.3)	98.4	(0.6)	.60	(0.1)	.90
Specific phobia	96.9	(2.8)	92.1	(2.2)	63.9	(7.3)	99.5	(0.5)	92.7	(1.9)	.73	(0.0)	.94
Social phobia	65.5	(9.2)	95.8	(1.2)	61.3	(9.3)	96.5	(1.3)	93.0	(1.6)	.59	(0.0)	.81
Generalized anxiety disorder	60.1	(13.7)	99.3	(0.4)	75.2	(13.8)	98.7	(0.6)	98.1	(0.7)	.66	(0.0)	.80
Post-traumatic stress disorder	59.9	(14.2)	98.0	(0.8)	57.0	(12.7)	98.2	(0.9)	96.4	(1.2)	.56	(0.1)	.79
Any anxiety disorder	88.5	(4.1)	87.6	(2.7)	70.5	(5.1)	95.8	(1.7)	87.9	(2.3)	.63	(0.0)	.88
II. Mood disorders													
Major depressive episode (MDE)	78.4	(7.3)	95.3	(1.6)	77.8	(7.0)	95.4	(1.8)	92.3	(2.1)	.74	(0.0)	.87
Major depressive episode or dysthymic disorder	76.5	(6.4)	96.4	(1.4)	84.2	(5.7)	94.3	(1.9)	92.5	(2.0)	.75	(0.0)	.86
Bipolar spectrum disorder [2]	100.0	(0.0)	99.6	(0.2)	94.3	(3.6)	--	(0.0)	99.6	(0.2)	.97	(0.0)	1.00
Any mood disorder	81.4	(5.4)	96.6	(1.4)	88.2	(4.6)	94.4	(2.0)	93.0	(1.9)	.80	(0.0)	.89
III. Disruptive behavior disorders													
Attention-deficit/hyperactivity disorder	58.5	(11.3)	96.5	(1.0)	58.5	(9.7)	96.5	(1.2)	93.5	(1.4)	.55	(0.0)	.78
Conduct disorder	96.8	(2.4)	98.7	(0.9)	86.3	(10.2)	99.7	(0.2)	98.5	(0.9)	.90	(0.0)	.98
Oppositional-defiant disorder	77.3	(12.1)	92.7	(2.1)	55.8	(14.7)	97.2	(1.3)	91.1	(2.2)	.60	(0.0)	.85
Any disruptive behavior disorder	77.9	(7.1)	90.9	(1.8)	63.7	(6.4)	95.2	(1.6)	88.7	(1.9)	.63	(0.0)	.84
IV. Substance disorders[2]													
Alcohol abuse with or without dependence	96.5	(2.7)	99.4	(0.8)	91.9	(10.1)	99.8	(0.2)	99.2	(0.7)	.94	(0.0)	.98
Alcohol dependence with abuse	12.4	(21.1)	98.9	(0.5)	5.4	(8.6)	99.6	(0.4)	98.5	(0.7)	.07	(0.1)	.56
Illicit drug abuse with or without dependence	85.1	(7.9)	99.0	(0.7)	89.6	(6.8)	98.6	(0.9)	97.8	(1.0)	.86	(0.0)	.92
Illicit drug dependence with abuse	80.9	(17.6)	98.8	(0.6)	37.1	(19.7)	99.8	(0.2)	98.7	(0.6)	.50	(0.0)	.90
Any substance disorder	89.0	(6.8)	98.7	(1.0)	89.8	(7.4)	98.6	(0.9)	97.6	(1.2)	.88	(0.0)	.94
V. Any													
Any lifetime diagnosis	92.0	(3.1)	81.9	(4.2)	84.9	(3.3)	90.2	(3.5)	87.2	(2.5)	.74	(0.0)	.87

	SN[1]		SP[1]		PPV[1]		NPV[1]		TCA[1]		κ[1]		
	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	AUC[1]
Two or more lifetime diagnoses	77.6	(6.4)	92.6	(1.6)	74.2	(4.2)	93.8	(2.2)	89.4	(2.3)	.69	(0.0)	.85
Three of more lifetime diagnoses	70.0	(7.0)	96.9	(0.9)	70.9	(7.3)	96.8	(1.0)	94.3	(1.2)	.67	(0.0)	.83

[1] SN: sensitivity; SP: specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; TCA: total classification accuracy; κ: Cohen's κ[39]; AUC: area under the receiver operating characteristic curve.

[2] Bipolar I, bipolar II, or bipolar spectrum disorder. See the text for a definition of the latter.

Table 3

Consistency of lifetime prevalence estimates and individual-level concordance of lifetime DSM-IV bipolar I disorder, bipolar II disorder, and sub-threshold bipolar disorder based on the CIDI with clinical diagnoses based on the K-SADS, where the K-SADS is treated as the gold standard, in the NCS-A clinical reappraisal sample (n = 347)

	CIDI		K-SADS		McNemar	SN[1]		SP[1]		PPV[1]		NPV[1]		TCA[1]		κ[1]		
	%	(se)	%	(se)	$\chi^2$1	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	Est	(se)	AUC[1]
Bipolar I disorder	0.5	(0.3)	1.0	(0.4)	--	44.2	(21.8)	100.0	(0.0)	100.0	(0.0)	99.4	(0.3)	99.4	(0.3)	.61	(0.2)	.72
Bipolar II disorder	1.8	(0.7)	1.3	(0.7)	1.6	48.6	(25.7)	98.8	(0.5)	34.1	(21.0)	99.3	(0.4)	98.1	(0.6)	.39	(0.1)	.74
Bipolar I or II disorder	2.3	(0.8)	2.3	(0.8)	0.0	70.0	(14.8)	99.3	(0.4)	70.6	(15.5)	99.3	(0.4)	98.6	(0.6)	.70	(0.1)	.85
Sub-threshold bipolar disorder	4.3	(1.6)	3.9	(1.6)	0.9	83.7	(11.3)	98.9	(0.5)	76.0	(12.2)	99.3	(0.4)	98.3	(0.6)	.79	(0.0)	.91

[1] SN: sensitivity; SP: specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; TCA: total classification ccuracy; κ: Cohen's κ[39]; AUC: area under the receiver operating characteristic curve.

Table 4

Area under the ROC curve (AUC) for dichotomous DSM-IV/CIDI diagnostic classifications (DICH) versus CIDI-based predicted probabilities (CONT) predicting lifetime K-SADS diagnoses in the NCS-A clinical reappraisal sample (n=347)[1]

	DICH[2]	CONT[2]
I. Anxiety disorders		
Social Phobia	.81	.87
Generalized anxiety disorder	.80	.91
Post-traumatic stress disorder	.79	.88
II. Disruptive behavior disorders		
Attention-deficit/hyperactivity disorder	.78	.94
Oppositional-defiant disorder	.84	.98
IV. Substance disorders		
Alcohol dependence with abuse	.56	.98
Illicit drug abuse with or without dependence	.92	.98
Illicit drug dependence with abuse	.90	.99

[1] Results are reported only for the eight disorders for which the AUC of the predicted probabilities of K-SADS diagnoses based on the symptom-level prediction equation is meaningfully higher than the AUC of the dichotomous CIDI diagnostic classification.

[2] DICH = AUC values for dichotomous CIDI diagnostic classification. CONT = AUC values for continuous CIDI-based predicted probabilities of K-SADS diagnoses derived from the logistic regression equations; The DICH values are identical to the AUC values reported in Table 2 and are repeated here only to facilitate comparison with the CONT values.