International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 16(2): 52–65 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.208

Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members

RONALD C. KESSLER,[1] LENARD A. ADLER,[2] MICHAEL J. GRUBER,[1] CHAITANYA A. SARAWATE,[3] THOMAS SPENCER,[4] DAVID L. VAN BRUNT[5]

1 Department of Health Care Policy, Harvard Medical School
2 Departments of Psychiatry and Neurology, New York University School of Medicine
3 HealthCore, Inc.
4 Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School
5 Eli Lilly & Company, US Health Outcomes

Abstract

The validity of the six-question World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener was assessed in a sample of subscribers to a large health plan in the US. A convenience subsample of 668 subscribers was administered the ASRS Screener twice to assess test-retest reliability and then a third time in conjunction with a clinical interviewer for DSM-IV adult ADHD. The data were weighted to adjust for discrepancies between the sample and the population on socio-demographics and past medical claims. Internal consistency reliability of the continuous ASRS Screener was in the range 0.63–0.72 and test-retest reliability (Pearson correlations) in the range 0.58–0.77. A four-category version The ASRS Screener had strong concordance with clinician diagnoses, with an area under the receiver operating characteristic curve (AUC) of 0.90. The brevity and ability to discriminate DSM-IV cases from non-cases make the six-question ASRS Screener attractive for use both in community epidemiological surveys and in clinical outreach and case-finding initiatives. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: attention-deficit/hyperactivity disorder (ADHD), adult ADHD, ASRS screener

Introduction

Although it has long been known that attention-deficit/hyperactivity disorder (ADHD) is one of the most common psychiatric disorders among children and adolescents (Shekim et al., 1985; Bird et al., 1988), the prevalence of adult ADHD has been the subject of controversy. We are aware of only three attempts to estimate the prevalence of adult ADHD in general population surveys, one administering semi-structured clinical interviews to a telephone sample in the US (Faraone and Biederman, 2005), another administering a self-report questionnaire without clinical reappraisal to a sample selected from an automated general practitioner registry in the Netherlands (Kooij et al., 2005), and the third administering an in-person fully structured screen followed by a telephone clinical reappraisal interview to a national sample in the US (Kessler et al., 2006). The third study was the most compelling of the three, as it was based on a large nationally representative sample and featured both a structured screen and a clinical reappraisal. The estimated prevalence of adult ADHD (standard error in parentheses) in this study was 4.4% (0.6).

A point prevalence of 4.4% makes adult ADHD one of the most commonly occurring DSM-IV disorders in the general population of the US. Additional findings from the same survey suggest that adult ADHD is not only common but also important by virtue of being highly comorbid with other disorders, associated with substantial role impairment, and undertreated (Kessler et al., 2005a; Kessler et al., 2005b). The undertreatment is especially unfortunate in light of the availability of treatments with proven efficacy (Safren et al., 2005; Wilens et al., 2005). Based on this confluence of factors, adult ADHD would appear to be an attractive target for public health outreach, screening, and treatment (Kessler and Stang, 2006). However, in order to do this, a short and valid screening measure needs to be developed to pinpoint people who are likely to meet criteria for adult ADHD.

Although a number of candidate measures exist to screen for adult ADHD (Belendiuk et al., 2007), the most promising would appear to be the screening scale used in the US national survey described above, the WHO Adult ADHD Self-Report Scale (ASRS) Screener (Kessler et al., 2005b), as this is a very short scale (six questions) that can be self-administered and that has been shown to have adequate sensitivity (68.7%), excellent specificity (99.5%), excellent total classification accuracy (97.9%), and a good κ (0.76) in the general population of the US (Kessler et al., 2005b). However, the only US validity study of the ASRS was the one carried out in conjunction with the nationally representative US survey described above. An important limitation of that study was that it used the same sample that picked the six ASRS Screener questions from a larger battery to estimate validity, possibly leading to an overestimation of the concordance of the scale with blinded clinical diagnoses. The purpose of the present report is to present the results of a cross-validation of the ASRS in a separate sample.

Methods

Development of the ASRS screener

The ASRS was developed in conjunction with the revision of the WHO Composite International Diagnostic Interview (CIDI) (Kessler and Ustun, 2004) for the WHO World Mental Health (WMH) Survey Initiative (Demyttenaere et al., 2004). The original item pool for the ASRS included questions about the frequency of all 18 DSM-IV Criterion A symptoms of adult ADHD. Clinical calibration compared each of these items to blind clinical ratings of the symptoms of DSM-IV adult ADHD in a sample of 154 respondents ages 18–44 who previously participated in the NCS-R, oversampling NCS-R respondents who reported childhood ADHD and adult persistence. The data were weighted to correct for this over-sampling prior to analysis (Table 1). ASRS symptom-level responses were consistently related to blind clinical symptom ratings, but varied substantially in concordance (κ in the range 0.16–0.81).

Forward stepwise logistic regression of clinical diagnoses on the full set of ASRS questions was carried out in the total sample and in subsamples. Consistent evidence was found based on plots of analysis of AUC plots of the area under the receiver operating characteristic curve (AUC) (Hanley and McNeil, 1982) that the strength of association between all the items in the scale and clinical diagnoses was captured by extracting only six questions from the entire item pool. Additional items did not meaningfully improve the strength of association between the screening scale and the clinical diagnoses (Kessler et al., 2005b). However, a number of different six-question subsets of questions could be found to generate roughly comparable prediction accuracy. Based on this result, all-possible subsets logistic regression analysis was used to generate a complete list of all the six-question subsets of ASRS questions that were roughly equivalent in overall predictive power (Table 2). Subgroup analysis was then used to select one of these scales as the ASRS Screener based on having the most stable psychometric characteristics across major segments of the population defined on the basis of age, gender, education, and urbanicity.

Summation of dichotomous responses across all 18 ASRS questions was the optimal unweighted scoring method to predict clinical syndrome classifications using all information in the ASRS. However, a dichotomous version of the six-question ASRS Screener that distinguished between scores of 0–3 and 4–6 outperformed the best dichotomy in the full 18-question ASRS (0–10 versus 11–18) in terms of sensitivity (68.7% versus 56.3%), specificity (99.5% versus 98.3%), total classification accuracy (97.9% versus 96.2%), and κ (0.76 versus 0.58). See Kessler et al. (2005b) for a more

Table 1. Concordance of optimally dichotomized[1] ASRS questions with blind ADHD-RS clinical symptom ratings in the NCS-R validation study

Scoring[2]	Mean[3]		Sensitivity		Specificity		TCA[4]		McNemar Test[5]	Kappa		AUC
	$\bar{x}$	(se)	%	(se)	%	(se)	%	(se)	χ^2_1	K	(se)	
I. Inattention												
1. 0–2 v. 3–4	1.2	(0.1)	80.6**	(11.7)	99.0	(0.5)	97.8	(0.7)	0.1	0.81	(0.10)	0.90
2. 0–2 v. 3–4	1.6	(0.2)	39.5	(15.0)	86.6	(7.0)	77.2	(7.2)	0.1	0.27	(0.09)	0.63
3. 0–1 v. 2–4	0.9	(0.1)	72.9**	(12.4)	88.2	(5.6)	85.9	(5.1)	3.9*	0.52	(0.09)	0.81
4. 0–1 v. 2–4	0.9	(0.1)	56.3**	(15.2)	90.0	(3.9)	84.3	(4.7)	0.1	0.45	(0.09)	0.73
5. 0–1 v. 2–4	1.0	(0.1)	84.8**	(7.7)	81.7	(6.6)	82.0	(6.0)	18.7*	0.40	(0.09)	0.83
6. 0–2 v. 3–4	1.3	(0.2)	31.8	(12.2)	93.2	(6.2)	83.9	(6.4)	1.9	0.28	(0.11)	0.62
7. 0–2 v. 3–4	1.3	(0.1)	38.9**	(14.1)	98.2	(1.3)	90.3	(3.5)	6.7*	0.47	(0.12)	0.69
8. 0–2 v. 3–4	1.6	(0.1)	53.1**	(13.0)	93.2	(6.0)	88.8	(5.6)	0.1	0.45	(0.11)	0.73
9. 0–1 v. 2–4	0.9	(0.1)	49.3**	(14.1)	87.2	(4.6)	82.0	(5.0)	1.4	0.33	(0.10)	0.68
II. Hyperactivity-impulsivity												
1. 0–2 v. 3–4	1.4	(0.1)	59.7**	(14.7)	95.4	(2.5)	88.7	(4.3)	1.9	0.60	(0.09)	0.78
2. 0–1 v. 2–4	0.3	(0.1)	19.5**	(7.5)	98.6	(0.6)	91.5	(3.0)	6.3*	0.26	(0.14)	0.59
3. 0–2 v. 3–4	1.3	(0.2)	34.6	(10.7)	90.6	(5.9)	85.3	(5.9)	0.6	0.23	(0.11)	0.63
4. 0–2 v. 3–4	1.1	(0.1)	48.1**	(15.1)	92.4	(4.5)	86.6	(4.7)	0.0	0.41	(0.11)	0.70
5. 0–2 v. 3–4	1.4	(0.2)	74.9**	(13.3)	95.5	(2.1)	92.2	(2.7)	0.0	0.71	(0.08)	0.85
6. 0–2 v. 3–4	1.2	(0.2)	32.6	(14.2)	90.7	(6.1)	75.7	(7.5)	6.9*	0.27	(0.09)	0.62
7. 0–1 v. 2–4	1.1	(0.1)	46.1	(15.2)	76.8	(7.1)	72.0	(6.8)	6.6*	0.18	(0.09)	0.61
8. 0–2 v. 3–4	1.3	(0.2)	33.9	(14.7)	92.5	(6.5)	81.7	(6.2)	3.1	0.30	(0.10)	0.63
9. 0–1 v. 2–4	1.2	(0.1)	62.8**	(13.5)	70.8	(8.6)	70.1	(7.9)	26.6*	0.16	(0.07)	0.67

*Significant difference between false positives and false negatives at the 0.05 level, two-sided test.

**Significant difference between the proportions of respondents with versus without clinician-defined symptoms who are positive on the screening question.

***Significant at the 0.05 level.

[1] Optimality was defined as minimizing the difference between the weighted number of false positive and false negative responses.

[2] The nine rows in each of the two parts of the table represent the nine DSM-IV symptoms of Inattention and Hyperactivity-impulsivity. The row headings for each numbered symptom show the optimal dichotomization of the 0–4 response scale in terms of minimizing overall item-level classification error.

[3] Means are for the full 0–4 response scale.

[4] TCA (Total Classification) accuracy is the proportion of all respondents who are classified accurately by the screen.

[5] The McNemar test evaluates the significance of the difference between the frequency of false positives and the frequency of false negatives.

Table 2. Concordance of optimally dichotomized[1] alternative six-question ASRS screeners with blind ADHD-RS clinical syndrome classifications in the NCS-R validation study

Screeners[2]	Questions[3]		Sensitivity		Specificity		TCA[4]		McNemar Test[5]	Kappa		AUC
	IN	H-I	%	(se)	%	(se)	%	(se)	χ^2_1	K	(se)	
1.	4,5,6,9	1,5	68.7	(8.2)	99.5	(0.3)	97.9	(0.6)	0.9	0.76	(0.13)	0.84
2.	5,6,8,9	1,5	63.5	(8.6)	99.2	(0.3)	97.4	(0.6)	0.7	0.70	(0.14)	0.81
3.	4,5,6	2,3,5	68.8	(8.4)	98.7	(0.6)	97.2	(0.8)	0.1	0.70	(0.14)	0.84
4.	4,5,8,9	1,5	67.8	(8.2)	99.1	(0.4)	97.5	(0.6)	0.4	0.72	(0.14)	0.83

* Significant at the 0.05 level.

[1] Optimality was defined as minimizing the difference between the weighted number of false positive and false negative responses.

[2] Scales 1, 2 and 4 were dichotomized at 0–3 versus 4–6. Scale 3 was dichotomized at 0–2 versus 3–6.

[3] The numbers in the first column correspond to the ASRS questions for the DSM-IV inattention (IN) symptoms, while the numbers in the second column correspond to the ASRS questions for the DSM-IV hyperactivity-impulsivity (H-I) symptoms.

[4] TCA (Total Classification) accuracy is the proportion of all respondents who are classified accurately by the screen.

[5] The McNemar test evaluates the significance of the difference between prevalence estimates based on the screener and the clinical interviews. None of these tests was significant at the 0.05 level.

detailed discussion of the initial scale development process.

The primary care cross-validation sample

A three-phase design was used to select the sample for cross-validation of the ASRS Screener. We began with a population of 1.75 million adult (ages 18 and older) subscribers to a managed care plan in the states of California and Georgia in the US. We excluded the small number of subscribers who were in treatment for ADHD. In the first phase, a convenience sample of 20,011 subscribers was interviewed by telephone for purposes of another study (Brod et al., 2006) that included the six-question ASRS Screener. These interviews were carried out between March and June, 2004. The respondents in this time 1 (T1) sample were weighted to adjust for discrepancies between the sample and population on a multivariate profile of socio-demographic characteristics and information about past medical claims. In the second phase, convenience subsamples of 496 T1 screened positives and 172 T1 screened negatives were rescreened by telephone between December 2004 and May 2005 in order to estimate ASRS Screener stability over time. The 668 respondents in this Time 2 (T2) sample were weighted to adjust for the over-sampling of screened positives from the T1 sample as well as for residual discrepancies between the T2 sample and the T1 sample on a multivariate profile of socio-demographic characteristics and information about past medical claims. In the third phase, convenience subsamples of 155 T2 screened positives and 63 T2 screened negatives were administered semi-structured clinical interviews between one and three months after their T2 assessments to make gold standard diagnoses of DSM-IV ADHD. The ASRS Screener was also repeated after the completion of these clinical interviews. The 218 respondents in the Time 3 (T3) sample were weighted to adjust for the over-sampling of T2 screened positives as well as for residual discrepancies between the T3 sample and the T2 sample on a multivariate profile of socio-demographic characteristics and information about past medical claims.

The clinical interview

The clinical interview was the Adult ADHD Clinician Diagnostic Scale (ACDS v1.2) (Adler and Cohen, 2004). The ACDS v1.2 is a semi-structured interview, which uses childhood and adult specific prompts to assess the eighteen DSM IV symptoms of ADHD in both adulthood (past six months) and childhood; it allows the clinician to make a DSM IV diagnosis (with

DOI: 10.1002/mpr

sub-typing) of adult ADHD on the basis of sufficient current symptoms, childhood onset and current impairment from ADHD symptoms. This clinical interview has been used in many clinical studies of adult ADHD (Spencer et al., 1995; Spencer et al., 1998; Spencer et al., 2001). Results from the clinical interview were evaluated to generate a diagnosis of adult ADHD by requiring at least two symptoms, before age six, full criteria in childhood, and also full current criteria in adulthood. Six experienced clinical interviewers (all PhD-level clinical psychologists or MA-level social workers) carried out these interviews. Each interviewer received 40 hours of training from two board certified psychiatrists who specialize in research on adult ADHD (LA, TS) and successfully completed five practice interviews in which their symptom ratings matched those of the trainers before they began production interviewing. All clinical interviews were tape recorded and a random sample was reviewed by one of the two supervising psychiatrists. The supervising psychiatrists also held weekly group interviewer calibration meetings.

A clinical diagnosis of adult ADHD required a respondent to have at least six symptoms of either inattention or hyperactivity-impulsivity during the six months before the interview (DSM-IV Criterion A), at least two Criterion A symptoms of ADHD before age seven (Criterion B), some impairment in at least two areas of living during the past six months (Criterion C), and clinically significant impairment in at least one area of living over the same time period (Criterion D). No attempt was made to operationalize the DSM-IV diagnostic hierarchy rules for ADHD (Criterion E). Nor was any attempt made to diagnose ADHD in partial remission.

Statistical methods

Each ASRS Screener question asks respondents how often a particular symptom of ADHD occurred to them over the past six months on a five-point response scale of never (0), rarely (1), sometimes (2), often (3), and very often (4). We began by examining the mean and standard deviation of responses to these questions along with factor loadings on the first principal factor of a factor analysis at each of the three time points in the three-wave sample. Internal consistency reliability was calculated based on Cronbach's α (Cronbach, 1951).

Test-retest stability was then calculated using Pearson correlations for scales based on two different scoring approaches. The first approach assigned one point to each of the six questions endorsed above a prespecified threshold value established in earlier methodological research (Kessler et al., 2005b) and then summed across the six questions to yield a scale with a theoretical range of 0–6. This is the approach developed in the original construction of the ASRS Screener in an effort to simplify scoring in primary care and workplace screening programmes.

The second approach summed responses to each question using the full 0–4 response scale to yield a summary score with a theoretical range of 0–24. A third approach, similar to the second, was also investigated. This summed the 0–4 responses across the six items by using the parameter estimates in a two-parameter item response theory (IRT) analysis of the 24 nested dichotomies embedded in the six 0–4 responses to generate weights for each response to each question. This was done in order to maximize concordance between the scale and an assumed underlying unidimensional true score measure of ADHD symptoms. As the 0–24 score based on the second approach consistently out-performed the IRT-based scale, though, no results are reported for the latter.

A structural equation model was estimated to distinguish the effects of measurement reliability and true score stability in accounting for the observed intertemporal stability of the ASRS Screener. The LISREL 8.14 software system (Joreskog and Sorbom, 1994) was used to estimate the parameters in this model.

The strength of association between T2 ASRS Screener scores and T3 clinical diagnoses was assessed by calculating the AUC. The AUC is our preferred measure of concordance rather than the more commonly used κ (Cohen, 1960) because AUC, unlike κ, is not influenced by prevalence. AUC can be interpreted as the probability that a randomly selected clinical case would score higher on the ASRS Screener than a randomly selected non-case. AUC was calculated both for continuous and dichotomous versions of the ASRS Screener, as continuous screening scales are generally superior for research purposes while dichotomous scoring rules are needed for clinical purposes.

A range of descriptive statistics was calculated to assess the concordance between dichotomous versions of the T2 ASRS Screener and T3 clinical diagnoses. Included here were estimates of sensitivity (the percentage of respondents with the clinical diagnosis who are classified as having the disorder by the screening scale), specificity (the percent of respondents without the clinical diagnosis who are classified as not having the diagnosis by the screening scale), total classification accuracy (the percent of all respondents who are correctly classified by the ASRS Screener as to whether or not they have the clinical diagnosis), κ (a measure of concordance that adjusts for chance agreement), and AUC.

The conventional clinical screening approach creates a dichotomy out of continuous screening scale scores to differentiate predicted cases from predicted non-cases. However, this dichotomization often discards potentially useful information that would be retained in a polychotomous screening scale, such as the distinction between a nearly definite case and a probable case. As a result, the use of polychotomous screening scales is becoming increasingly popular (Peirce and Cornell, 1993). In order to investigate whether additional information of this sort might be obtained from the ASRS Screener, a polychotomous version of the 0–24 scale was created that collapsed scores with comparable probabilities of being defined as a clinical case into strata. Sensitivity was then calculated for each resulting stratum for purposes of estimating the positive predictive value (PPV) of the stratum (the probability that a given individual in the stratum meets clinical criteria for the syndrome) in any population assumed to have the same sensitivity based on an assumption about the prevalence of the syndrome in that population.

The motivation for calculating strata of this sort is based on the fact that PPV can vary widely depending on the prevalence of the disorder in the population being screened, while estimates of sensitivity and specificity are generally thought to be less variable across populations (Guyatt and Rennie, 2001). This means that an independent estimate of the population prevalence is needed to calculate PPV even if we are willing to assume that sensitivity and specificity are the same in a population under study as in the population from which the calibration sample was drawn.

Once this assumption is made, a four-step process can be used to estimate the stratum-specific PPV (PPV_s) for each screening scale stratum. The first step is to calculate a stratum-specific likelihood ratio (SSLR) for each ASRS Screener stratum. An SSLR is a descriptive statistic defined as the ratio of the sensitivity to the specificity of a screening scale within a given stratum divided by the same ratio for all other strata combined (Guyatt and Rennie, 2001).

The second step is to transform the assumed population prevalence (P_p) into an assumed population odds (O_p) using the formula

$$O_p = P_p/(1-P_p). \tag{1}$$

The third step is to use O_p in conjunction with the SSLR for the stratum (SSLR_s) to define the stratum-specific odds of having the clinical syndrome (O_s) as

$$O_p \times \mathrm{SSLR}_s = O_s. \tag{2}$$

The fourth step, finally, is to transform O_s into an estimate of the stratum-specific PPV (PPV_s) using the formula

$$\mathrm{PPV}_s = O_s/(1+O_s). \tag{3}$$

This four-step process was used to estimate PPV_s for plausible values of P_p within each stratum of the ASRS Screener.

As the sample design features weighting of cases, significance tests that assume a simple random sample will be biased. Design-based methods were consequently used to evaluate statistical significance and to calculate the standard errors and confidence intervals of descriptive statistics. The Taylor series linearization method (Wolter, 1985) implemented in SAS 9.1.3 (SAS Institute, 2003) was used to make these calculations. The Taylor series method was also used to calculate McNemar χ^2 tests. The method of jack-knife repeated replications (Rust and Rao, 1996) was used to calculate the standard error of the κ coefficient.

Results

Distribution and item inter-correlations

Means of the six ASRS Screener questions in the three samples are in the range 0.9–1.8 (Table 3). Principal axis factor analysis found only the first factor in each sample to have an eigenvalue greater than 1.0 (1.4–2.0), with Cronbach's α for the factor-based scales in the range 0.63–0.72. It is noteworthy that we would not expect very high factor loadings or Cronbach's α for two reasons. First, the symptoms of ADHD are known to be two-dimensional (inattention and hyperactivity-impulsivity) even though these two dimensions are

DOI: 10.1002/mpr

Table 3. Distribution and factor loadings of responses to the ASRS Screener questions in the three waves of the cross-validation study[1]

Question	Mean (SD)[2]			Factor Loadings[3]		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
(1) How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?	1.1 (1.0)	1.3 (0.9)	1.1 (0.9)	0.53	0.43	0.36
(2) How often do you have difficulty getting things in order when you have to do a task that requires organization?	1.1 (1.0)	1.4 (0.8)	1.1 (0.9)	0.61	0.57	0.74
(3) How often do you have problems remembering appointments or obligations?	0.9 (0.9)	1.2 (1.0)	1.2 (1.0)	0.45	0.56	0.50
(4) When you have a task that requires a lot of thought, how often do you avoid or delay getting started?	1.3 (1.0)	1.6 (0.9)	1.4 (1.1)	0.52	0.66	0.81
(5) How often do you fidget or squirm with your hands or feet when you have to sit down for a long time?	1.4 (1.2)	1.7 (1.3)	1.8 (1.4)	0.40	0.60	0.53
(6) How often do you feel overly active and compelled to do things, like you are driven by a motor?	1.5 (1.2)	1.7 (1.2)	1.3 (1.2)	0.28	0.45	0.34
Cronbach's Alpha				0.63	0.72	0.70
(n)	(20,011)	(668)	(218)	(20,011)	(668)	(218)

[1] All analyses are based on weighted data that adjusted samples to have the same distribution as the population (1.75 million people) on a multivariate profile of socio-demographic characteristics and information about past medical claims.
[2] Responses were coded on a scale with responses ranging between 0 (never) and 4 (very often).
[3] Based on principal axis factor analysis. Eigenvalues for the first unrotated principal factors were 1.4 and 0.3 (T1), 1.8 and 0.4 (T2), and 2.0 and 0.5 (T3).

strongly related to each other and the majority of the structure among symptoms can be captured in a one-factor model. The fact that a second factor does not emerge in the ASRS Screener is a joint function of the small number of items and the fairly strong inter-correlations among symptoms of inattention and hyperactivity-impulsivity. Second, the ASRS Screener questions were selected by stepwise logistic regression analysis. This method selects the *least redundant* set of symptoms in a set in an effort to maximize prediction of an external criterion (in this case, the DSM-IV diagnosis of adult ADHD), thereby optimizing inconsistency among the items in a way that would be reflected in lower bound estimates of internal consistency.

Test-retest stability

Pearson correlations for stability of scale scores over time are consistently somewhat lower for the 0–6 scoring approach (T1–2 0.63, T2.3 0.67, T1–3 0.47) than for the 0–24 scoring approach (T1–2 0.74, T2–3 0.77, T1–3 0.58). As one would expect based on the fact that the time interval separating T1 from T2 was longer than the time interval separating T2 from T3, the T1–T2 correlations (0.63–0.74) are consistently lower than the T2-T3 correlations (0.67–0.77), while the T1–T3 correlations are lowest of all (0.47–0.58). A simplex structure of change is implied by the correlations in that the product of the T1–T2 and T2–T3 correlations closely approximates the magnitude of the T1–T3 correlation for both the 0–6 scoring system (0.63 × 0.67 = 0.42 compared to the T1–T3 correlation of 0.47) and the 0–24 scoring system (0.74 × 0.77 = 0.57 compared to the T1-T3 correlation of 0.58).

A more formal way to investigate stability is with a structural equation model that separates unreliability of measurement from true score change by positing the existence of an unmeasured true score that changes over time and is imperfectly indicated by the ASRS Screener questions (Figure 1). The standardized parameter estimates based on a version of this model that

DOI: 10.1002/mpr

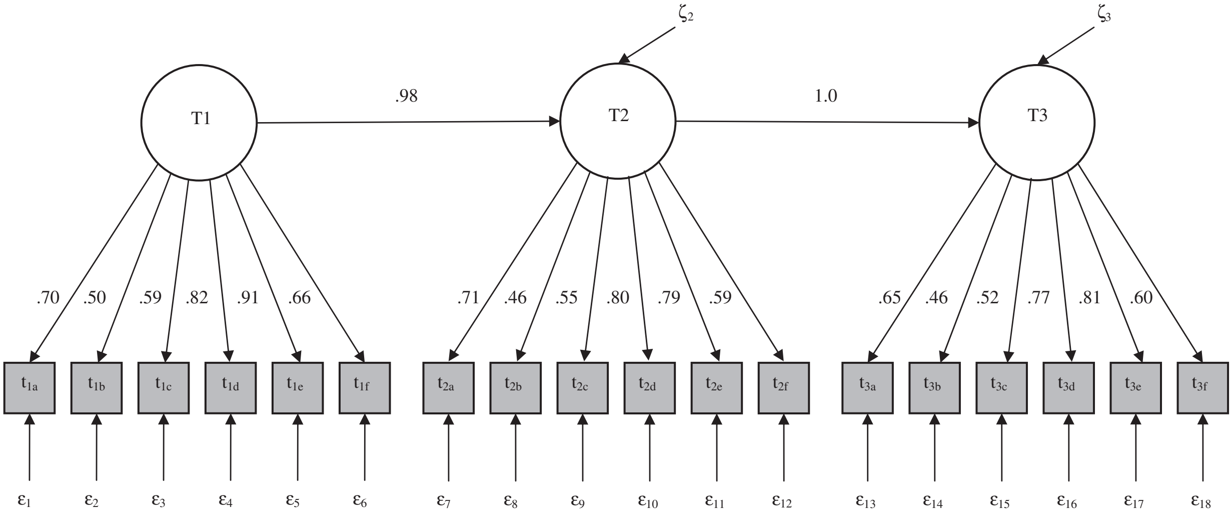

Figure 1. Three-wave structural equation model of stability in the true score underlying the ASRS Screener items (n = 218).
The three variables in circles are unobserved. The 18 variables in squares are observed. Coefficients are standardized. The model was identified by fixing the metric coefficients linking the unobserved variable with the first observed variable to unity at each time. Other metric coefficients linking unobserved to observed variables were constrained to be equal over time. The standardized coefficients vary over time despite the metric coefficients being constrained due to inter-temporal variation in variances of the observed variables. With 153 (18 × 17/2) bivariate correlations and 8 unknowns, the model has 145 degrees of freedom. Model fit (χ^2_{145} = 1575.7) was improved substantially by relaxing the constraint on metric coefficients to be stable over time and by allowing for first-order auto-regressive correlations among measurement errors (χ^2_{127} = 806), but standardized coefficients linking unobserved to observed scores never were higher in these or other specifications than those reported in this initial specification.

constrains the metric slopes of the screener questions on the true scores to be constant over time and that assumes measurement errors to be uncorrelated over time (Figure 1) suggests that virtually all of the inter-temporal instability in the ASRS Screener is due to measurement unreliability rather than to change in the true score. This can be seen by noting that the standardized true-score stability coefficients are estimated to be close to 1.0. Although further analyses (results available on request) found that certain model specifications, such as allowing auto-correlations among measurement error terms, improved the fit of this basic model in some ways, none of these respecifications led to evidence of greater screening scale reliability.

Concordance with clinical diagnoses

It is important to recall the implications noted earlier of the fact that the ASRS Screener items were selected using stepwise logistic regression. As this method selects items to have incremental importance in predicting clinical diagnoses over and above the effects of items that entered the equation in earlier steps, significant correlations between unique information in the individual items and the true score outcome might be induced by the item-selection process. The latter correlations violate the assumption of independence between true scores and measurement errors in the single-factor simplex change model estimated to evaluate the reliability of the screener (Figure 1). Given that these correlations are likely to be positive, reliability will be underestimated in this model.

In light of this potential problem, a better way to evaluate the ASRS Screener is to consider concordance with clinical diagnoses rather than to consider either internal consistency reliability or test-retest stability. We evaluated concordance by focusing on the clinical diagnoses obtained in the T3 sample. The weighted prevalence estimate (with standard error in parentheses) of clinician-diagnosed DSM-IV adult ADHD in this sample was 8.4% (3.9). Concordance of the ASRS Screener with clinician-diagnosed adult ADHD was assessed for the T2 screener rather than for the T3 screener in order to avoid the possibility of confounding between ASRS reports and clinical

DOI: 10.1002/mpr

diagnoses obtained in the same interview due to the priming effect of the T3 clinical probing being carried out before administering the T3 ASRS.

The AUC of the T2 ASRS Screener in predicting T3 clinical diagnoses was 0.82 for the 0–6 scoring approach and 0.87 for the 0–24 scoring approach. It is noteworthy that the prediction accuracy was found to decrease only modestly with the substantial coarsening of the data in the 0–6 scoring approach compared to the 0–24 scoring approach. Both of these results, however, used the full range of the ASRS Screener to predict clinical diagnoses. This is legitimate when the screener is being used for research purposes, where the main goals are to study the prevalence and correlates of disorder, as there is no need to require any single respondent to be classified dichotomously as a case or a non-case to achieve these goals. When the screener is used as a first stage in clinical assessment, in comparison, it is necessary to make dichotomous distinctions at the individual level. The high values of AUC reported in the last paragraph are not the appropriate measures of screener accuracy in situations of this sort. Instead, it is necessary to define dichotomous cut points on the continuous scale to mimic the individual-level triage decisions that would be made using the scale for clinical purposes. To that end, we dichotomized T2 ASRS Screening scale scores using both the 0–6 scoring approach (0–4 versus 4–6) and the 0–24 scoring approach (0–13 versus 14–26). Cut-points were selected to approximate the T3 clinical prevalence estimate of 8.4% as closely as possible.

Test statistics for these optimal dichotomous cut-points were calculated to estimate sensitivity, specificity, and a number of other descriptive statistics that are commonly used to assess screening scale accuracy. (Table 4) As one might expect, the value of AUC was found to be lower in the dichotomous version of the scale than in the full range of the scale both for the 0–6 scoring approach (0.64) and for the 0–24 scoring approach (0.79). All descriptive measures were considerably better for the 0–24 than the 0–6 scoring approach; κ was in the moderate range (0.52) for the dichotomy based on the 0–24 scoring approach and in the poor range (0.21) for the dichotomy based on the 0–6 scoring approach (Landis and Koch, 1977). In terms of cross-validation, the 0–6 scoring approach, which was the recommended approach in the original report on the ASRS Screener (Kessler et al., 2005b), had an AUC considerably higher in that earlier study (0.84) than in the current study (0.64). Based on this result, the 0–24 scoring approach appears to be more robust than the 0–6 scoring approach.

Clinician-diagnosed DSM-IV adult ADHD as a function of ASRS Screener scores

As noted in the section on analysis methods, it is often useful to go beyond simple dichotomous scoring to distinguish possible cases and non-cases with borderline

Table 4. Concordance of optimally dichotomized[1] T2 ASRS Screener responses with T3 clinician-diagnosed DSM-IV adult ADHD (n = 218)[2]

	0–6 scoring approach		0–24 scoring approach*	
	Est	(se)	Est	(se)
Predicted prevalence	14.0	(5.6)	11.0	(4.2)
McNemar Chi Square (p-value)	4.3	(0.04)	1.7	(0.20)
Sensitivity	39.1	(22.6)	64.9	(23.3)
Specificity	88.3	(5.8)	94.0	(3.3)
PPV	23.5	(15.9)	49.9	(20.0)
NPV	94.0	(3.7)	96.7	(2.8)
Total classification accuracy	84.1	(5.9)	91.5	(3.8)
Kappa	0.21	(0.09)	0.52	(0.10)
OR (Lower, Upper)	4.8	(0.5, 42.5)	28.8	(2.8, 292.1)
AUC	0.64		0.79	

[1] The optimal dichotomy was 0–3 v. 4–6 for the 0–6 scoring approach and 0–13 v. 14–24 for the 0–24 scoring approach.
[2] Based on weighted data that adjusted samples to have the same distribution as the population (1.75 million people) on a multivariate profile of socio-demographic characteristics and information about past medical claims.

screening scores from those with more extreme scores. To this end, we compared the PPV of each point on the T2 ASRS Screener using the 0–24 scoring approach and collapsed categories to remove non-monotonic patterns and non-significant differences. This resulted in a four-stratum classification scheme made up of scores in the range 0–9, 10–13, 14–17 and 18–24. The AUC of this four-category ASRS Screener scoring scheme in predicting T3 clinical diagnoses was 0.90.

If we make the conventional assumption that the distribution of screening-scale scores among true cases (sensitivity) and true non-cases (specificity) is more constant across populations than the distribution of clinical diagnoses among respondents with given screening scale scores (positive and negative predictive values), we could project to new samples from the results of this validity study by assuming constant sensitivity and specificity. However, positive and negative predictive values are the more useful statistics to estimate in practical applications. Therefore, we used the methods described in the section on analysis methods to estimate likely values of positive predictive value (PPV) for each of the four ASRS Screener strata based on a range of assumptions about population prevalence. It is important to note that these estimates require an assumption to be made about population prevalence of adult ADHD.

We began by calculating the distribution of T3 clinical cases and clinical non-cases across the four T2 ASRS Screener strata (Table 5). None (0.0%) of the clinician-defined cases had screener scores in the range 0–9, while 58.8% of clinician-defined non-cases had screening scale scores in this range. Virtually identical proportions of clinician-defined cases (35.1%) and non-cases (35.1%), in comparison, had screener scores in the range 10–13. Much higher proportions of clinician-defined cases than non-cases, finally, had screener scores in the ranges 14–17 (59.8% of cases versus. 5.8% of non-cases) or 18–24 (5.1% of cases vs. 0.2% of non-cases). Note that the ratio of these proportions among cases compared to non-cases is much higher for screening scale scores in the range 18–24 (roughly 25:1) than in the range 14–17 (roughly 10:1), justifying the distinction between these two ranges.

Based on these calculations, we used the procedures described in the section on analysis methods to estimate the prevalence of clinician-diagnosed adult ADHD separately among people in each of the four ASRS Screener strata in populations that differ in overall prevalence of the disorder (Table 6). These calculations were based on the assumption that sensitivity and specificity are constant across populations and are identical to their values in Table 5. Prevalence was investigated in a range 2–12%, where 2% is less than half the estimated prevalence of adult ADHD in the US general population (Kessler et al., 2006) and 12% is roughly three times this prevalence. Results show that less than 0.1% of people with screening scores in the range 0–9 would be expected to meet clinical criteria for DSM-IV adult ADHD even if they came from a population with prevalence in the high end of this

Table 5. Distributions of T3 clinical cases (sensitivity) and non-cases (specificity) across the strata of the T2 ASRS Screener (n = 218)[1]

	Sensitivity		Specificity	
	%	(se)	%	(se)
ASRS Screener strata				
0–9	0.0	(0.0)	58.8	(3.8)
10–13	35.1	(7.1)	35.1	(3.7)
14–17	59.8	(7.3)	5.8	(1.8)
18–24	5.1	(3.3)	0.2	(0.3)

[1]Based on weighted data that adjusted samples to have the same distribution as the population (1.75 million people) on a multivariate profile of socio-demographic characteristics and information about past medical claims.

Table 6. Estimated prevalence of clinician-diagnosed DSM-IV adult ADHD within ASRS Screener strata (positive predictive value) as a function of assumed population prevalence[1]

	ASRS Screener Strata			
	0–9	10–13	14–17	18–24
Prevalence (%)				
2	0.0	2.0	17.3	33.6
4	0.0	4.0	30.0	50.8
6	0.0	6.0	39.6	61.2
8	0.0	8.0	47.2	68.3
10	0.0	10.0	53.3	73.4
12	0.0	12.0	58.4	77.2

[1]All results are calculated based on the assumption that sensitivity and specificity are the same as in Table 6.

DOI: 10.1002/mpr

range. The estimated prevalence varies between 2.0% and 12.0%, in comparison, among people with screening scores in the range 10–13 depending on the prevalence in the total population. Estimated prevalence varies even more widely (from 17.3% to 58.4%) among people with screening scores in the range 14–17 selected from populations that vary across the assumed range in overall prevalence. Among people with screening scores in the range 18–24, finally, estimated prevalence ranges from 33.6% to 77.2% based on assumed population prevalence.

Discussion

It is important to recognize that the clinical interviews used as the gold standard in the current study classified cases in partial remission as not being cases. This is a conservative approach. As noted in the introduction, the DSM-IV criteria for ADHD were developed with children in mind and offer only limited guidance regarding the diagnosis of adult ADHD. This lack of guidance is of considerable concern because clinical studies make it clear that symptoms of ADHD are more heterogeneous and subtle in adults than children (DeQuiros and Kinsbourne, 2001; Wender et al., 2001) and that cases in partial remission are quite common (Faraone and Biederman, 2005). These observations have led some clinical researchers to suggest that the valid assessment of adult ADHD might require either an increase in the variety of symptoms assessed (Barkley, 1995), a reduction in the severity threshold for considering a symptom clinically significant (Ratey et al., 1992), or a reduction in the DSM-IV six-of-nine symptom requirement (McBurnett, 1997), any of which would require a complete reconsideration of the findings reported here.

Another limitation of the current report is that both the ASRS Screener and the clinical interviews were based entirely on self-reports. Childhood ADHD is diagnosed largely on the basis of parent and teacher reports rather than self-reports because parents and teachers are both in good positions to observe child behaviour and because children with ADHD often have little insight into the severity of their symptoms (Jensen et al., 1999). The situation is different for adults, where there is great variability in the extent to which other people observe their behaviour and where access to reliable informants varies with the respondent's marital status, occupational status, and social networks, making it necessary as a practical matter to base assessment largely on self-report (Wender et al., 2001). However, methodological studies comparing adult self-reports versus informant reports of ADHD symptoms find some of the same disagreements as in studies of child self-reports versus informant reports, with informants reporting higher symptom levels than focal respondents (Gittelman and Mannuzza, 1985; Zucker et al., 2002). This suggests that self-report scales might under-estimate the true prevalence of adult ADHD, although the one self vs. informant study of adult ADHD carried out in a non-clinical sample found fairly strong associations between the two reports and no self-informant difference in reported symptom severity (Murphy and Schachar, 2000).

Within the context of these limitations, the results reported here suggest that the ASRS Screener is a useful tool both for epidemiological research and for clinical outreach and case-finding, although the 0–24 scoring approach out-performs the 0–6 scoring approach. It is noteworthy that our earlier substantive studies with the ASRS Screener (Fayyad et al., in press; Kessler et al., 2005a; Kessler et al., 2006) were all based on the 0–6 scoring approach, which had somewhat stronger concordance with clinical diagnoses than the 0–24 scoring approach in the NCS-R. The fact that the AUC of the 0–6 scoring approach is lower in the current sample suggests that the high AUC in the NCS-R might have been due to over-fitting. Based on this interpretation, we now recommend using the four-stratum classification of the 0–24 scoring approach, although the 0–6 approach is still valid.

With regard to epidemiological research, the 0.90 AUC using the four-stratum classification of the 0–24 scoring approach is strong enough to support powerful analysis of prevalence and correlates of adult ADHD in samples drawn from the general population, from workplaces, and from primary care. When these research purposes are primary, scores on the screener can be transformed into predicted probabilities of clinical diagnoses and these continuous probabilities can be used either to generate weights for weighted logistic regression analyses or used to estimate individual-level multiple imputations (MI). Although an exposition of these methods is beyond the scope of this report, such an overview has been presented elsewhere (Kessler and Ustun, 2004). In addition, a recent report on the epidemiology of adult ADHD in the US illustrates the use of the MI approach (Kessler et al., 2006).

DOI: 10.1002/mpr

With regard to clinical screening, the results suggest that the ASRS Screener could be of considerable value by virtue of the fact that nearly two-thirds of true clinician-diagnosed cases screen positive and a high proportion of screened positives are true cases under plausible assumptions about the population prevalence of the disorder. For example, if the prevalence of adult ADHD is 4% in a given population (roughly equal to the general population prevalence in the US), 30% of the people who screen in the low positive range (14–17) and 50.8% of those who screen in the high positive range (18–24) on the ASRS Screener will be clinical cases compared to 0.0% who screen in the low negative range (0–9) and 4.0% who score in the high negative range (10–13). In a primary care sample, where the prevalence of ADHD is likely to be two to three times higher than in the general population, as many as 50% of the low positive and 75% of the high positive screens will be true clinical cases compared to none of the low negatives and no more than 10% of the high negative screens. Thus, the screener provides a method to identify the majority of cases quickly with negligible inclusion of non-cases.

A complexity in using the screener is that transformations to predicted probabilities require the user to make a prior assumption about prevalence in the population under study. This assumption is required due to the fact that PPV varies for constant values of sensitivity and specificity as population prevalence changes. In research applications, it would be best to carry out an independent validation in a probability subsample of respondents. When this is not possible, the assumption can be based on an analysis of the distribution of the ASRS Screener using maximum-likelihood methods to select the population prevalence most likely to generate the observed distribution based on the assumption of constant sensitivity and specificity. Clinical experience with the population under study is a more practical basis for making a prior assumption about population prevalence when the ASRS Screener is used for case-finding and intervention. In the absence of any clinical basis for making this assumption, the clinician might choose an initial cutoff score of 14, which corresponds with the 'optimal cutoff' for the 0–24 scoring approach shown in Table 5, and adjust this value after enough information is gathered to generate an empirical estimate of the treatment population prevalence.

The situation with the roughly one-third of clinical cases who are negative on the ASRS screener is important to consider. Not surprisingly, these screened negative clinical cases had an average symptom severity score in the clinical interviews that was lower than the average for the clinical cases that screened positive. A similar result was found in the initial ASRS Screener validation study (Kessler et al., 2005b). In addition, both the ASRS and the clinical interview presumably missed people with adult ADHD who denied having had childhood symptoms or who underestimated the severity of their current symptoms, although we do not know how large a proportion of true cases fall into these categories.

Besides using the ASRS Screener in epidemiological surveys and in screening interventions, the fact that the screener can be self-administered easily and quickly (less than two minutes) might make it a useful secondary measure to include in clinical studies as a complement to the dimensional clinical assessments of ADHD symptom severity typically used in such studies. The screener could be useful in such studies to define a lower bound severity threshold that distinguishes community cases from non-cases. The use of the ASRS Screener in clinical studies would also provide a useful crosswalk between clinical research and community epidemiological research by allowing a comparison of the severity distribution between community and clinical cases. The absence of such comparative data has restricted our ability to interpret the clinical significance of categorical prevalence estimates of most mental disorders in community epidemiological studies up to now. The inclusion of identical short dimensional assessments of adult ADHD in both clinical and community studies would be a useful step in the direction of addressing this important problem for this heretofore under-studied disorder.

Acknowledgements

The ASRS Screener was developed in collaboration with the WHO World Mental Health (WMH) Survey Initiative in conjunction with the revision of the WHO Composite International Diagnostic Interview (CIDI) (Kessler and Ustun, 2004). Development of the ASRS Screener was supported by Eli Lilly & Co. The ASRS Screener cross-validation study was supported by Eli Lilly & Co. and HealthCore Inc. The authors gratefully acknowledge John Barth and Laura Kleckner for administering the T1 survey and Joe Johnston, Steve Able, Doug Faries, and Ralph Swindle for their assistance in conceptualizing and designing the T1 study and the T3 validation study. The larger series of WMH methodological studies in which the ASRS development work was embedded were supported by the National Institute of Mental Health (R01 MH070884), the John D and Catherine T MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Fogarty International Center (FIRCA R01-TW006481), the Pan American Health Organization, Eli Lilly & Co., Ortho-McNeil Pharmaceutical, Inc., GlaxoSmithKline, and Bristol-Myers Squibb. The WMH CIDI Advisory Group that developed the ASRS included Lenard Adler (New York University Medical School), Russell Barkley (Medical College of South Carolina), Joseph Biederman (Massachusetts General Hospital and Harvard Medical School), Keith Conners (Duke University Medical School), Stephen Faraone (Massachusetts General Hospital and Harvard Medical School), Laurence Greenhill (New York State Psychiatric Institute), Molly Howes (Harvard Medical School), Ronald Kessler (Harvard Medical School), Thomas Spencer (Massachusetts General Hospital and Harvard Medical School) and T Bedirhan Ustun (World Health Organization). All WMH instruments and publications are posted at http://www.hcp.med.harvard.edu/wmh.

References

Adler L, Cohen J. "Diagnosis and evaluation of adults with attention-deficit/hyperactivity disorder." Psychiatr Clin North Am 2004; 27(2): 187–201.

Barkley RA. ADHD behavior checklist for adults. The ADHD Report 1995; 3: 16.

Belendiuk KA, Clarke TL, Chronis AM, Raggi VL. Assessing the concordance of measures used to diagnose adult ADHD. J Atten Disord 2007; 10: 276–87.

Bird HR, Canino G, Rubio-Stipec M, Gould MS, Ribera J, Sesman M, Woodbury M, Huertas-Goldman S, Pagan A, Sanchez-Lacay A, Moscoso M. Estimates of the prevalence of childhood maladjustment in a community survey in Puerto Rico. The use of combined measures. Arch Gen Psychiatr 1988; 45: 1120–26.

Brod M, Johnston J, Able S, Swindle R. Validation of the adult attention-deficit/hyperactivity disorder quality-of-life Scale (AAQoL): a disease-specific quality-of-life measure. Qual Life Res 2006; 15: 117–29.

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

Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297–34.

Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, Angermeyer MC, Bernert S, de Girolamo G, Morosini P, Polidori G, Kikkawa T, Kawakami N, Ono Y, Takeshima T, Uda H, Karam EG, Fayyad JA, Karam AN, Mneimneh ZN, Medina-Mora ME, Borges G, Lara C, de Graaf R, Ormel J, Gureje O, Shen Y, Huang Y, Zhang M, Alonso J, Haro JM, Vilagut G, Bromet EJ, Gluzman S, Webb C, Kessler RC, Merikangas KR, Anthony JC, Von Korff MR, Wang PS, Brugha TS, Aguilar-Gaxiola S, Lee S, Heeringa S, Pennell BE, Zaslavsky AM, Ustun TB, Chatterji S. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health surveys. JAMA 2004; 291: 2581-25-90.

DeQuiros GB, Kinsbourne M (2001). Adult ADHD: analysis of self-ratings in a behavior questionnaire. Ann NY Acad Sci 2001; 931: 140–47.

Faraone SV, Biederman J (2005). What is the prevalence of adult ADHD? Results of a population screen of 966 adults. J Atten Disord 2005; 9: 384–91.

Fayyad J, de Graaf R, Kessler RC, Alonso J, Angermeyer M, Demyttenaere K, de Girolamo G, Haro JM, Karam EG, Lara C, Lépine J-P, Ormel J, Posada-Villa J, Zaslavsky AM. The cross-national prevalence and correlates of adult ADHD: Results from the WHO World Mental Health Survey Initiative. Br J Psychiatr, in press.

Gittelman R, Mannuzza S. Diagnosing ADD-H in adolescents. Psychopharmacol Bull 1985; 21: 237–42.

Guyatt G, Rennie D. User's Guide to the Medical Literature: A Manual for Evidence-based Clinical Practice. Chicago, IL: AMA Press, 2001.

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

Jensen PS, Rubio-Stipec M, Canino G, Bird HR, Dulcan MK, Schwab-Stone ME, Lahey BB (1999). Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? J Am Acad Child Adolesc Psychiatr 1999; 38: 1569–79.

Joreskog KG, Sorbom D (1994). LISREL. Chicago, IL: Scientific Software International.

Kessler RC, Adler L, Ames M, Barkley RA, Birnbaum HG, Greenberg PE, Johnston JA, Spencer T, Ustun TB. The prevalence and effects of adult attention-deficit/hyperactivity disorder on work performance in a nationally representative sample of workers. J Occup Environ Med 2005a; 47: 565–72.

Kessler RC, Adler L, Ames M, Demler O, Faraone S, Hiripi E, Howes MJ, Jin R, Secnik K, Spencer T, Ustun TB, Walters EE. The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. Psychol Med 2005b; 35: 245–56.

Kessler RC, Adler L, Barkley RA, Biederman J, Connors K, Demler O, Greenhill L, Howes MJ, Secnik K, Spencer T, Ustun TB, Walters EE, Zaslavsky AM (2006). The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. Am J Psychiatry 2006; 163: 716–23.

Kessler RC, Stang PE. Future directions in health and work productivity research. In RC Kessler and PE Stang (eds) Health and Work Productivity: Making the Business Case for Quality Health Care. Chicago, IL: University of Chicago Press, 2006, pp 271–88.

Kessler RC, Ustun 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.

Kooij JJ, Buitelaar JK, van den Oord EJ, Furer JW, Rijnders CA, Hodiamont PP. Internal and external validity of attention-deficit hyperactivity disorder in a population-based sample of adults. Psychol Med 2005; 35: 817–27.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–74.

McBurnett K. Attention-deficit/hyperactivity disorder: a review of diagnostic issues. In TA Widiger, AJ Francis, HA Pincus, R Ross, MB First and W Davis (eds) DSM-IV Sourcebook. Washington, DC: American Psychiatric Association, 1997, pp 111–143.

Murphy P, Schachar R. Use of self-ratings in the assessment of symptoms of attention deficit hyperactivity disorder in adults. Am J Psychiatr 2000; 157: 1156–59.

Peirce JC, Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making 1993; 13: 141–51.

Ratey J, Greenberg S, Bemporad. JR, Lindem K. Unrecognized attention-deficit hyperactivity disorder in adults presenting for outpatient psychotherapy. J Child Adolesc Psychopharmacol 1992; 4: 267–75.

Rust KF, Rao JN. Variance estimation for complex surveys using replication techniques. Stat Methods Med Res 1996; 5: 283–310.

Safren SA, Otto MW, Sprich S, Winett CL, Wilens TE, Biederman J. Cognitive-behavioral therapy for ADHD in medication-treated adults with continued symptoms. Behav Res Ther 2005; 43: 831–42.

SAS Institute I. SAS/STAT Software. Versions 8 and 9.1.3. Cary, NC: SAS Institute, Inc, 2003.

Shekim WO, Kashani J, Beck N, Cantwell DP, Martin J, Rosenberg J, Costello A. The prevalence of attention deficit disorders in a rural midwestern community sample of nine-year-old children. J Am Acad Child Psychiatry 1985; 24: 765–70.

Spencer T, Biederman J, Wilens T, Faraone S, Prince J, Gerard K, Doyle R, Parekh A, Kagan J, Bearman SK. Efficacy of a mixed amphetamine salts compound in adults with attention-deficit/hyperactivity disorder. Arch Gen Psychiatry 2001; 58: 775–82.

Spencer T, Biederman J, Wilens T, Prince J, Hatch M, Jones J, Harding M, Faraone SV, Seidman L. Effectiveness and tolerability of tomoxetine in adults with attention deficit hyperactivity disorder. Am J Psychiatry 1998; 155: 693–95.

Spencer T, Wilens T, Biederman J, Faraone SV, Ablon JS, Lapey K. A double-blind, crossover comparison of methylphenidate and placebo in adults with childhood-onset attention-deficit hyperactivity disorder. Arch Gen Psychiatry 1995; 52: 434–43.

Wender PH, Wolf LE, Wasserstein J. Adults with ADHD. An overview. Ann NY Acad Sci 2001; 931: 1–16.

Wilens TE, Haight BR, Horrigan JP, Hudziak JJ, Rosenthal NE, Connor DF, Hampton KD, Richard NE, Modell JG. Bupropion XL in adults with attention-deficit/hyperactivity disorder: a randomized, placebo-controlled study. Biol Psychiatry 2005; 57: 793–01.

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

Zucker M, Morris MK, Ingram SM, Morris RD, Bakeman R. Concordance of self- and informant ratings of adults' current and childhood attention-deficit/hyperactivity disorder symptoms. Psychol Assess 2002; 14: 379–89.

Correspondence: RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
Email kessler@hcp.med.harvard.edu
Telephone 617-432-3587
Fax 617-432-3588

DOI: 10.1002/mpr