Psychological Medicine (2013), **43**, 1625–1637. © Cambridge University Press 2012
doi:10.1017/S0033291712002334

ORIGINAL ARTICLE

Composite International Diagnostic Interview screening scales for DSM-IV anxiety and mood disorders

R. C. Kessler[1]*, J. R. Calabrese[2], P. A. Farley[3], M. J. Gruber[1], M. A. Jewell[3], W. Katon[4], P. E. Keck Jr.[5], A. A. Nierenberg[6], N. A. Sampson[1], M. K. Shear[7], A. C. Shillington[3], M. B. Stein[8], M. E. Thase[9] and H.-U. Wittchen[10]

[1] *Department of Health Care Policy, Harvard Medical School, Boston, MA, USA;* [2] *Department of Psychiatry, University Hospitals Case Medical Center, Case Western Reserve University, Cleveland, OH, USA;* [3] *Clinical Services, EPI-Q, Inc., Oakbrook Terrace, IL, USA;* [4] *School of Medicine, University of Washington, Seattle, WA, USA;* [5] *Lindner Center of HOPE and Department of Psychiatry, University of Cincinnati College of Medicine, Cincinnati, OH, USA;* [6] *Depression Clinical and Research Program and the Bipolar Clinic and Research Program, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA;* [7] *Columbia University School of Social Work, New York, NY, USA;* [8] *Department of Psychiatry, University of California San Diego, San Diego, CA, USA;* [9] *Departments of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia Veterans Affairs Medical Center, and the University of Pittsburgh Medical Center, Philadelphia and Pittsburgh, PA, USA;* [10] *Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany*

Background. Lack of coordination between screening studies for common mental disorders in primary care and community epidemiological samples impedes progress in clinical epidemiology. Short screening scales based on the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI), the diagnostic interview used in community epidemiological surveys throughout the world, were developed to address this problem.

Method. Expert reviews and cognitive interviews generated CIDI screening scale (CIDI-SC) item pools for 30-day DSM-IV-TR major depressive episode (MDE), generalized anxiety disorder (GAD), panic disorder (PD) and bipolar disorder (BPD). These items were administered to 3058 unselected patients in 29 US primary care offices. Blinded SCID clinical reinterviews were administered to 206 of these patients, oversampling screened positives.

Results. Stepwise regression selected optimal screening items to predict clinical diagnoses. Excellent concordance [area under the receiver operating characteristic curve (AUC)] was found between continuous CIDI-SC and DSM-IV/SCID diagnoses of 30-day MDE (0.93), GAD (0.88), PD (0.90) and BPD (0.97), with only 9–38 questions needed to administer all scales. CIDI-SC *versus* SCID prevalence differences are insignificant at the optimal CIDI-SC diagnostic thresholds ($\chi^2_1 = 0.0$–2.9, $p = 0.09$–0.94). Individual-level diagnostic concordance at these thresholds is substantial (AUC 0.81–0.86, sensitivity 68.0–80.2%, specificity 90.1–98.8%). Likelihood ratio positive (LR+) exceeds 10 and LR− is 0.1 or less at informative thresholds for all diagnoses.

Conclusions. CIDI-SC operating characteristics are equivalent (MDE, GAD) or superior (PD, BPD) to those of the best alternative screening scales. CIDI-SC results can be compared directly to general population CIDI survey results or used to target and streamline second-stage CIDIs.

Received 21 June 2012; Accepted 5 September 2012; First published online 18 October 2012

Key words: Bipolar disorder, Composite International Diagnostic Interview (CIDI), generalized anxiety disorder, major depression, panic disorder, screening scales, validity.

Introduction

Although research on the community epidemiology of mental disorders (i.e. general population incidence, prevalence, risk factors, consequences) is an active area of investigation (Susser *et al.* 2006; Kessler & Üstün, 2008*b*; Tsuang *et al.* 2011), research on clinical epidemiology (i.e. prevalence, severity, long-term course in treatment samples) is underdeveloped, especially in primary care settings. Indeed, the most important clinical epidemiological study of mental disorders in primary care remains the World Health Organization (WHO) Collaborative Study on Psychological Problems in General Health Care (Üstün & Sartorius, 1995), a study carried out nearly two decades ago that led to few extensions (e.g. Wittchen *et al.* 2002; Barkow *et al.* 2003; Kisely & Simon, 2005;

* Address for correspondence: R. C. Kessler, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
(Email: Kessler@hcp.med.harvard.edu)

Kisely *et al.* 2006). More sustained long-term clinical epidemiological studies exist in specialty treatment samples (Katz *et al.* 1979; Bruce *et al.* 2005), but those studies are now outdated because of changes in the composition of patient populations since the studies were initiated (Kessler *et al.* 2005).

We know from primary care screening studies that untreated mental disorders are common in primary care (Lowe *et al.* 2008; Gili *et al.* 2011). However, screening studies tell us little about the natural history of these disorders, as screening studies typically focus on current prevalence or treatment response. Yet information is needed on episode recurrence and onset of secondary disorders to understand the public health significance and long-term cost-effectiveness of primary care screening, outreach and treatment quality improvement (Barrett *et al.* 2005; Konnopka *et al.* 2009). This integration of primary care with public health is now an area of considerable policy interest (Committee on Integration to Improve Population Health, 2012).

One way to build a critical mass of such data would be to blend longitudinal clinical epidemiological studies with community epidemiological surveys. For example, several new community epidemiological surveys in the WHO World Mental Health (WMH) Survey Initiative (Kessler & Üstün, 2008*a*) are using a dual-frame sampling approach with parallel samples of (i) patients in primary care (both with and without detected and undetected mental disorders) and (ii) other household residents in the same communities. This design facilitates comparisons of illness prevalence course among treated and untreated cases by collecting successive snapshots of current prevalence of disorders over multiple points in time.

Screening scales will be used in the primary care segment of these surveys as the first stage in a two-stage approach to oversample patients with current mental disorders for second-stage interviews. Screening scale responses are being 'preloaded' in the computerized scripts of the second-stage interviews to control question skip logic (e.g. skipping sections based on negative screening responses; expanding questions based on positive screening responses). The screening scales used for this purposes are based on the WHO Composite International Diagnostic Interview (CIDI; Kessler & Üstün, 2004), the diagnostic interview used in the WMH surveys and most other psychiatric epidemiological surveys throughout the world. Psychometric analyses of these disorder-specific CIDI screening scales (CIDI-SC) have been reported previously for adult attention deficit/hyperactivity disorder (ADHD; Kessler *et al.* 2007), insomnia, (Kessler *et al.* 2010*a*) and overall serious mental illness (SMI; Kessler *et al.* 2010*b*). The current report presents comparable results for the CIDI-SC scales of major depressive episode (MDE), bipolar disorder (BPD), generalized anxiety disorder (GAD) and panic disorder (PD). Although the results presented are for cross-sectional rather than longitudinal analyses, they are relevant for the longitudinal design described above because the latter is made up of a series of cross-sectional snapshots.

Method

Screening scale development

The CIDI

The CIDI is a fully structured research diagnostic interview developed for use by trained lay interviewers to generate diagnoses of lifetime and recent DSM-IV-TR/ICD-10 disorders (Robins *et al.* 1988). Clinical reappraisal studies document generally good concordance of CIDI diagnoses with blinded clinical diagnoses (Wittchen, 1994; Kessler *et al.* 1998). The CIDI uses extensive skip logic to reduce interview length. This skip logic is also used in the CIDI-SC based on the assumption that tablet computers will be used to administer, score and print out summary screening scale results.

Expanding the CIDI item pool

All CIDI symptom questions operationalize DSM/ICD criteria using simple descriptive language (Robins *et al.* 1988). However, validation studies find some CIDI questions less concordant than others with independent clinical assessments (Wittchen *et al.* 1995; Kessler *et al.* 2006; Green *et al.* 2011). We consequently expanded the CIDI item pool in developing the CIDI-SC scales by reviewing a wide range of other diagnostic instruments to generate alternative symptom questions. The expanded question set was reviewed iteratively, with diagnostic experts using their judgment to pinpoint alternative questions they considered potentially useful and to help revise and prioritize indicators. Previous methodological research has shown that such iterative expert review is often the most useful form of pretesting (Converse & Presser, 1986; Presser & Blair, 1994; Groves *et al.* 2009).

Pilot testing

Once preliminary symptom questions were generated, a convenience sample of 15 psychiatric out-patients with each diagnosis was administered the disorder-specific symptom questions. Cognitive debriefing interviews (Willis, 2005) assessed problems in conceptual understanding and question wording. These interviews were conducted by professional cognitive

interviewers using the 'think aloud' method (Presser *et al.* 2004) to elicit initial respondent reactions and collect alternative terminologies for confusing phrases. The results were presented to the diagnostic experts for review and final question revision.

The clinical reappraisal study

The sample

The revised questions were then administered to 3058 patients sampled from 29 primary care offices selected to include practices in both urban and rural areas in all four US Census Regions (Northeast, South, Midwest and West). No other stratification criteria were used in selecting practices. Practices were recruited through the Primary Care Network (www.primarycarenet.org). The original sample design called for a quota sample of 100 completed interviews in each of 30 offices with an unselected consecutive sample of patients. This sample size was selected to allow for the second-stage assessment of at least 30 screened positives for even the least common disorder (BPD) assuming plausible prevalence estimates and second-stage response rates. However, because one selected office dropped out after office recruitment ended, other offices in the same sample stratum were asked to continue data collection for 2 days beyond the time they met their quota, yielding a sample slightly larger than the originally targeted 3000 and based on 29, not 30, offices.

Respondent recruitment began by giving a 'study fact brochure' to patients as they checked in that explained the study as a test of a new screening questionnaire for common anxiety and mood disorders. The brochure explained that the study needed people aged $\geqslant$18 years both with and without the disorders to complete a 15-min laptop computer questionnaire in the waiting room; that participants would be remunerated US$25; that some participants would be asked to participate in a telephone follow-up interview that could take up to 1 h to complete; and that telephone respondents would receive an additional US$50. The brochure emphasized that responses were confidential and decisions about participation would not affect health-care treatment or benefits. Patients who informed the office receptionist that they were interested in the study then provided written informed consent and received a laptop computer to complete the questionnaire in the waiting room. Telephone numbers provided in the questionnaire were used to contact respondents and administer clinical reappraisal interviews within 3 days of the visit. The Human Subjects Committee of the New England Institutional Review Board (www.neirb.com) approved these recruitment, consent and field procedures.

The clinical reappraisal interview

Each CIDI-SC respondent was classified as 'very likely', 'likely', 'possible' or 'no' on each screening scale. A probability subsample of 30 respondents classified 'very likely' and 20 classified 'likely-possible' was selected for each scale with replacement and administered the clinical reappraisal interview. The sampling fraction varied across disorders due to prevalence differences to make the sample well-distributed across practices. Fifty patients who screened 'no' on all screening scales were also interviewed. The total clinical reappraisal sample of 206 is less than $50 \times 5 = 250$ because some respondents were independently selected for multiple disorders.

The clinical interview was an abridged Research Version, Non-Patient Edition of the Structured Clinical Interview for DSM-IV (SCID-I; First *et al.* 2002) focused on the four syndromes under study: 30-day MDE and GAD and lifetime and 30-day PD and mania/hypomania. Experienced SCID interviewers administered the interviews under the supervision of a study collaborator (P.E.K.) blinded to CIDI-SC responses. 30-day PD was defined as lifetime PD with 30-day panic attacks and/or persistent concern about additional attacks, worry about implications/consequences of attacks, or significant change in behavior due to attacks. 30-day BPD was defined as lifetime mania/hypomania with either 30-day MDE or 30-day mania/hypomania. SCID diagnoses were made without diagnostic hierarchy rules but with organic exclusions. Organic exclusions were not made in the screening scales. Each SCID disorder was classified as severe or non-severe to determine whether CIDI-SC could differentially detect more severe cases. BP-I *versus* BP-II defined BPD severity whereas the distinction between severe and non-severe cases of the other disorders was based on SCID interviewer assessments of whether there were (i) many symptoms more than needed for diagnosis, (ii) several symptoms that were particularly severe and/or (iii) marked impairment in social or occupational functioning associated with the disorder.

Analysis methods

The clinical reappraisal sample was weighted to adjust for oversampling of patients screened as 'very likely', 'likely' or 'possible'. Iterative stepwise logistic regression was then used (0.05-level entry criterion) to predict SCID diagnoses from CIDI-SC symptoms to determine the minimum CIDI-SC question set needed to approximate SCID diagnoses. An unweighted

https://doi.org/10.1017/S0033291712002334 Published online by Cambridge University Press

summary CIDI-SC score was created for each diagnosis from this minimum symptom set and receiver operating characteristic curve (ROC) analysis (Margolis *et al.* 2002) was used to estimate the area under the ROC curve (AUC) for each scale. The AUC is the probability of correctly identifying SCID cases from CIDI-SC scores in paired comparisons of randomly selected pairs of SCID cases and non-cases, where CIDI-SC tied scores are assigned a 50% chance of correct classification (Kraemer, 1992). The AUC has a predicted value of 0.5 when the screening scale is completely unrelated to the true score and 1.0 when perfectly related. CIDI-SC scores were not weighted to avoid overfitting in the absence of a large enough sample for cross-validation.

Each CIDI-SC score was then collapsed so that SCID prevalence estimates increased monotonically across screening scale strata but did not differ significantly within strata using the logic of stratum-specific likelihood ratio (LR) analysis (Pepe, 2003). McNemar χ^2 tests then tested the significance of differences between CIDI-SC and SCID prevalence estimates. Significance tests were based on Taylor series design-based standard errors to adjust for data weighting (Wolter, 1985).

Individual-level concordance was evaluated using the AUC and Cohen's κ (Cohen, 1960). Although κ is the traditional measure used in psychiatric research, it is not emphasized here because is varies across populations that differ in prevalence even when sensitivity (SN; the percentage of true cases correctly classified) and specificity (SP; the percentage of true non-cases correctly classified) are constant (Cook, 1998). The AUC, in comparison, is a function of SN and SP, which are considered the fundamental parameters of agreement. The AUC equals (SN + SP)/2 when the screen is dichotomous. AUC scores between 0.5 and 1.0 are often interpreted in parallel with κ as slight (AUC = 0.5–0.6, κ = 0.0–0.2), fair (AUC = 0.6–0.7, κ = 0.2–0.4), moderate (AUC = 0.7–0.8, κ = 0.4–0.6), substantial (AUC = 0.8–0.9, κ = 0.6–0.8) and almost perfect (AUC $\geqslant$ 0.9, $\kappa \geqslant$ 0.8) (Landis & Koch, 1977). We also report total classification accuracy (TCA), the proportion of all respondents whose CIDI-SC and SCID classifications are consistent.

In addition, we report disaggregated measures of operating characteristics, including SN and SP, positive predictive value (PPV; proportion of screened positives confirmed by the SCID), negative predictive value (NPV; proportion of screened negatives confirmed as non-cases by the SCID), LR positive [(LR+); SN/(1 – SP)] and LR negative [LR–; (1 – SN)/SP)]. LR+ and LR− assess the relative proportions of screened positives *versus* screened negatives confirmed as cases (LR+) or non-cases (LR–). LR+ values $\geqslant$5 and LR− values $\leqslant$0.2 are generally considered useful, whereas LR+ values $\geqslant$10 and LR− values $\leqslant$0.1 are considered sufficient to rule in/out diagnoses (Haynes *et al.* 2006).

Comparison with other widely used screening scales

To compare CIDI-SC operating characteristics with other screening scales, a 1990–2012 Medline search of screening scale validity studies was carried out using search terms 'screening', 'validity', 'sensitivity', 'specificity', 'case finding' and 'AUC' crossed with 'depression', 'bipolar disorder', 'manic depression', 'generalized anxiety disorder' and 'panic disorder'. We focused on studies where screening scales were compared to blinded clinical reappraisal interviews in samples of patients, community residents or internet users. Only key studies were considered.

Results

Stepwise logistic regression

Separate stepwise logistic regression analyses were used to predict each SCID disorder from the corresponding CIDI screening items.

MDE

Three CIDI questions were entered stepwise to predict 30-day dysphoria (sad–depressed, down–discouraged) and anhedonia (little–no interest in day-to-day activities) in the total sample. Among respondents with dysphoria and/or anhedonia, five additional questions were entered to screen for other DSM-IV Criterion A symptoms of MDE or the Criterion C requirement of clinically significant distress or impairment. The AUC for the continuous CIDI-SC scale with these eight questions was 0.93.

GAD

Two CIDI questions were entered to screen for 30-day DSM-IV GAD Criterion A (excessive anxiety–worry about multiple events–activities) in the total sample. Among Criterion A screened positives, five additional questions were entered to screen for Criteria B (difficulty controlling worry), C (restless, difficulty relaxing) and E (clinically significant distress or impairment). The AUC for the continuous CIDI-SC scale with these seven questions was 0.88.

PD

Two CIDI questions were entered to screen for having lifetime attacks of intense fear or discomfort that came on very suddenly in the total sample. Among

https://doi.org/10.1017/S0033291712002334 Published online by Cambridge University Press

Table 1. *Consistency of DSM-IV diagnoses based on the CIDI screening scales (CIDI-SC) at their optimal (to estimate prevalence) thresholds and based on the SCID (n = 206)*

	CIDI-SC		SCID								
	%	(S.E.)	%	(S.E.)	McNemar[a] χ^2_1	AUC	κ	TCA	SN	(S.E.)	SP	(S.E.)
MDE	23.8	(3.5)	19.7	(3.3)	2.9	0.85	0.65	88.2	80.2	(7.6)	90.1	(2.5)
GAD	10.8	(1.7)	7.7	(1.7)	2.5	0.81	0.52	91.9	68.0	(10.0)	93.9	(1.2)
PD	13.7	(2.6)	10.7	(2.5)	2.4	0.85	0.62	91.9	76.4	(10.3)	93.8	(1.4)
BPD	4.4	(1.0)	4.4	(1.2)	0.0	0.86	0.73	97.7	74.0	(13.8)	98.8	(0.5)

CIDI, Composite International Diagnostic Interview; SCID, Structured Clinical Interview for DSM-IV; AUC, area under the receiver operating characteristic curve; TCA, total classification accuracy; SN, sensitivity of the screening scale at the designated threshold; SP, specificity of the screening scale at the designated threshold; MDE, major depressive episode; GAD, generalized anxiety disorder; PD, panic disorder; BPD, bipolar disorder; S.E., standard error.

[a] Prevalence estimates based on the CIDI-SC do not differ significantly from those based on the SCID for any of the diagnoses ($p = 0.09$–0.98).

respondents with such attacks, seven follow-up questions were entered about psycho-physiological symptoms. Among patients with such symptoms, an additional question asked about symptoms reaching a peak within 10 min and two question asked about the Criterion A1 DSM-IV PD requirement that attacks be recurrent and unexpected. Four questions asked about Criterion A2 that attacks be followed by a month of persistent concern about another attack, worry about implications or significant change in behavior. Final questions then asked about 30-day prevalence. The AUC for the continuous CIDI-SC scale with these 15 symptom questions crossed with reports of 30-day recency was 0.90.

BPD

Two CIDI questions were entered to screen for lifetime DSM-IV mania-hypomania Criterion A (distinct periods of abnormally persistently elevated, expansive or irritable mood) in the total sample. Among respondents who endorsed at least one such question, four additional questions were entered to screen for Criterion B (more talkative than usual, racing thoughts, psychomotor agitation, excessive involvement in activities having high potential for painful consequences) and two for Criterion D (mania)/E (hypomania) involving presence–absence of marked impairment or hospitalization. A final question then asked about 30-day prevalence. The AUC for the continuous CIDI-SC scale with these eight questions for lifetime or 30-day mania–hypomania crossed with the CIDI-SC screen for 30-day MDE to define 30-day BPD was 0.97.

The three CIDI-SC diagnostic stem questions for MDE combined with two for GAD, two for PD and two for BPD create a set of only nine items that screen out the majority of primary care patients in less than 3 min. The maximum number of items (40) can be completed in no more than 8 min.

Concordance of DSM-IV CIDI-SC and SCID diagnoses

CIDI-SC *versus* SCID prevalence estimate differences are insignificant for all disorders at optimal (for estimating prevalence) CIDI-SC thresholds ($\chi^2_1 = 0.0$–2.9, $p = 0.09$–0.98) (Table 1). Aggregate diagnostic concordance at these thresholds is substantial for all disorders (AUC = 0.81–0.86), with proportions of SCID cases detected (SN) of 68.0–80.2%. The proportions of SCID non-cases classified correctly (SP) are 90.1–98.8%. Lower SN than SP is expected for thresholds designed to estimate prevalence without bias when only a minority of patients has a disorder, in which case LR+ is more informative than SN. LR+ is >10 for three of the four CIDI-SC at the optimal thresholds, indicating that screened positives are much more likely than screened negatives to be confirmed as SCID cases. LR+ is 8.1 for MDE, an informative but not definitive value. The proportions of screened positives at the optimal thresholds confirmed as SCID cases (PPV) are in the range 48.2% (GAD) to 73.7% (BPD) (Table 2).

The screen for MDE, the only one where LR+ is <10, can be made more conservative by raising the threshold (LR+ = 24.5, PPV = 85.9%), but at the cost of reducing SN from 80.2% to 46.5%. All four CIDI-SC can be made less conservative by lowering their thresholds, increasing SN to between 94.8% (BPD) and 100% (GAD and PD), but at the cost of increasing the estimated prevalence and decreasing LR+ and PPV. The only disorder where this conservative change is efficient is BPD, with an estimated prevalence

https://doi.org/10.1017/S0033291712002334 Published online by Cambridge University Press

Table 2. *CIDI screening scale (CIDI-SC) classification of DSM-IV/SCID cases and non-cases at different thresholds on the CIDI-SC (n = 206)*[a]

	p	(S.E.)	SN	(S.E.)	LR+	PPV	(S.E.)	SP	(S.E.)	LR−	NPV	(S.E.)
I. MDE												
Conservative	10.7	(1.9)	46.5	(8.3)	24.5	85.9	(4.6)	98.1	(0.6)	0.5	88.2	(3.2)
Optimal	23.8	(3.5)	80.2	(7.6)	8.1	66.5	(7.0)	90.1	(2.5)	0.2	94.9	(2.2)
Anti-conservative	45.9	(4.7)	99.2	(0.8)	3.0	42.6	(6.1)	67.2	(5.1)	0.0	99.7	(0.3)
II. GAD												
Optimal	10.8	(1.7)	68.0	(10.0)	11.1	48.2	(6.7)	93.9	(1.2)	0.3	97.2	(1.4)
Anti-conservative	44.3	(4.9)	100	(–)	2.5	17.3	(3.7)	60.3	(5.2)	0.0	100	(–)
III. PD												
Optimal	13.7	(2.6)	76.4	(10.3)	12.3	59.5	(8.6)	93.8	(1.4)	0.2	97.1	(1.4)
Anti-conservative	62.1	(4.9)	100	(–)	1.7	17.2	(4.0)	42.5	(5.2)	0.0	100	(–)
IV. BPD												
Optimal	4.4	(1.0)	74.0	(13.8)	61.7	73.7	(9.0)	98.8	(0.5)	0.3	98.8	(0.8)
Anti-conservative	7.0	(1.8)	94.8	(3.4)	31.6	59.1	(13.4)	97.0	(1.4)	0.1	99.8	(0.2)

CIDI, Composite International Diagnostic Interview; SCID, Structured Clinical Interview for DSM-IV; *p*, proportion of patients who screened positive on the CIDI-SC at the designated threshold; SN, sensitivity of the CIDI-SC at the designated threshold; LR+, likelihood ratio positive of the CIDI-SC at the designated threshold; PPV, positive predictive value of the CIDI-SC at the designated threshold; SP, specificity of the CIDI-SC at the designated threshold; LR−, likelihood ratio negative of the CIDI-SC at the designated threshold; NPV, negative predictive value of the CIDI-SC at the designated threshold; MDE, major depressive episode; GAD, generalized anxiety disorder; PD, panic disorder; BPD, bipolar disorder; S.E., standard error.

increasing from 4.4% to 7.0%, LR+ and PPV both remaining high (31.6, 59.1%) and SN increasing from 74.0% to 94.8%.

Although SP is above 90% for all disorders, this is not a definite rule-out when only a small minority of respondents has the disorder, in which case LR− is more informative. LR− is ⩽0.2 for only two diagnoses at the optimal threshold (MDE and PD) whereas LR− is never below 0.2, meaning that none of the diagnoses can be ruled out confidently with CIDI-SC scores below the optimal diagnostic threshold. However, thresholds can be lowered to produce LR− values less than 0.1 for all disorders, although at the cost of reducing SN. For MDE, 54.1% of patients can be ruled out [i.e. at a threshold where 45.9% (100%–45.9% = 54.1%) of patients screen positive] with LR− 0.1 (NPV = 99.7%). For GAD, 55.7% of patients can be ruled out with LR− 0.0 (NPV = 100%). For PD, 86.3% of patients can be ruled out with LR− 0.2 and 37.9% with LR− 0.0. None of the PD screened negatives at the lower thresholds and 2.8% at the next lowest threshold had a SCID diagnosis. For BPD, 93% of patients can be ruled out with LR− 0.1 (NPV = 99.8%).

Severe and non-severe cases

SN is higher for severe than non-severe cases of all four diagnoses at the optimal threshold (Table 3). SN is 85.4–92.9% for severe MDE, PD and BPD *versus* 68.9–69.6% for non-severe cases and 70.8% *versus* 59.6% for severe and non-severe GAD. However, none of the severe *versus* non-severe SN differences are statistically significant because of the small numbers of cases ($\chi^2_1 = 0.2$–2.4, $p = 0.12$–0.68).

Comparisons with other screening scales

MDE

The nine-item Patient Health Questionnaire (PHQ-9; Spitzer *et al.* 1999) is the most widely used major depression screening scale. Reviews of many PHQ-9 primary care validity studies (Gilbody *et al.* 2007; Wittkampf *et al.* 2007; Kroenke *et al.* 2010; Manea *et al.* 2012) show a central tendency of the AUC to be 0.85–0.88, which does not differ meaningfully from the CIDI-SC MDE AUC of 0.85. (See online Appendix Tables 1–4 for detailed results.) CIDI-SC SN and SP (0.80, 0.90) are also in the middle of the PHQ-9 ranges (0.77–0.88, 0.88–0.94). The AUC of other MDE screening scales is generally lower (0.72–0.84) and LR+ uninformative (Zigmond & Snaith, 1983; Broadhead *et al.* 1995; Farvolden *et al.* 2003; Hunter *et al.* 2005; Donker *et al.* 2009; Gaynes *et al.* 2010; Houston *et al.* 2011). One exception was found in a community survey of the Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977), with an AUC of 0.94 (Beekman *et al.* 1997), but other CES-D validity studies found considerably lower AUC, at 0.76–0.82 (Schulberg *et al.* 1985; Klinkman *et al.* 1997; Thomas *et al.* 2001).

https://doi.org/10.1017/S0033291712002334 Published online by Cambridge University Press

Table 3. *CIDI screening scale (CIDI-SC) sensitivity (SN) and likelihood ratio positive (LR+) for detecting severe and non-severe DSM-IV/SCID cases (n = 206)*

	Severe			Non-severe			Total			Severe *v.* non-severe[a]
	SN	(S.E.)	LR+	SN	(S.E.)	LR+	SN	(S.E.)	LR+	χ^2_1
MDE	92.9	(4.1)	9.4	69.6	(12.6)	7.0	80.2	(7.6)	8.1	2.1
GAD	70.8	(15.2)	11.6	59.6	(14.7)	9.8	68.0	(11.7)	11.1	0.2
PD	90.8	(7.5)	14.6	68.9	(14.7)	11.1	76.4	(10.3)	12.3	2.3
BPD	85.4	(10.0)	71.2	69.0	(18.6)	57.5	74.0	(13.8)	61.7	2.4

CIDI, Composite International Diagnostic Interview; SCID, Structured Clinical Interview for DSM-IV; SN, sensitivity of the CIDI-SC at the designated threshold; LR+, likelihood ratio positive of the CIDI-SC at the designated threshold; MDE, major depressive episode; GAD, generalized anxiety disorder; PD, panic disorder; BPD, bipolar disorder; S.E., standard error.

[a] Although SN is consistently higher for severe than non-severe cases, none of these differences is statistically significant ($p = 0.12$–0.68).

GAD

The CIDI-SC GAD AUC (0.81) is in the middle of the range for the GAD screening scales reviewed (0.74–0.85) (Broadhead *et al.* 1995; Farvolden *et al.* 2003; Kroenke *et al.* 2007; Donker *et al.* 2009, 2011; Houston *et al.* 2011). However, CIDI-SC SN and SP (0.68, 0.94) are closest to those of one specialty treatment screening scale, the Web-Based Depression and Anxiety Test (WB-DAT; Farvolden *et al.* 2003). Other screening scales have higher SN (0.83–0.93) but much lower SP (0.45–0.82). CIDI-SC and WB-DAT consequently have much higher LR+ (11.1, 10.5) than other scales (1.7–4.9), indicating higher confirmation of screened positives. This can be illustrated using Bayes' theorem to calculate post-test probability of SCID GAD for screened positives (Altman & Bland, 1994), which shows that for a true GAD prevalence of 5–15%, confirmation of screened positives would be only 21–46% for screening scales but much higher for WB-DAT (36–65%) and CIDI-SC (37–66%). Caution is needed in interpreting the WB-DAT results, however, as they were obtained in a specialty treatment setting.

PD

The CIDI-SC PD AUC (0.85) is at the upper end of the PD screening scales reviewed (0.69–0.88) (Broadhead *et al.* 1995; Stein *et al.* 1999; Farvolden *et al.* 2003; Lowe *et al.* 2003; Hunter *et al.* 2005; Bunevicius *et al.* 2007; Kroenke *et al.* 2007; Donker *et al.* 2009). CIDI-SC has among the highest LR+ (12.3) along with WB-DAT (12.5) and one of two GAD-7 (19.0) validity studies (Lowe *et al.* 2003). The high LR+ in that GAD-7 study, however, is offset by a much lower LR+ (3.9) in a second much larger GAD-7 study (Kroenke *et al.* 2007). The scales with high LR+ are much more distinct for their high SP (0.94–0.96) than high SN. If we assume that the true PD prevalence is in the range 5–15% in primary care and SN–SP estimates are accurate, confirmation of screened positives would be 35–65% for CIDI-SC and WB-DAT, 17–77% for the GAD-7, and no higher than 20–45% for other screening scales.

BPD

Although the Mood Disorder Questionnaire (MDQ; Hirschfeld *et al.* 2000) is by far the most widely used BPD screening scale, the vast majority of MDQ studies focus on patients in treatment for depression and investigate whether those with BPD can be distinguished from non-bipolar depressives (Hirschfeld *et al.* 2000, 2005; Miller *et al.* 2004, 2011; Weber Rouget *et al.* 2005; Parker *et al.* 2008, 2012; Twiss *et al.* 2008; Zimmerman *et al.* 2009). This focus reflects the fact that the MDQ was developed to address the underdetection of BPD among depressed patients (Hirschfeld & Vornik, 2004). We are aware of only two MDQ validity studies that evaluated ability to distinguish patients with BPD from all other patients (including those without depression) in settings other than a specialty clinic (Hirschfeld *et al.* 2003; Dodd *et al.* 2009). These studies were both carried out in community samples. The MDQ AUC was fairly low in both studies (AUC = 0.62) compared to much higher AUCs (0.86–0.96) for the CIDI-SC BPD scale at its two informative thresholds.

Only one other BPD screening scale, the Mood Swings Questionnaire (MSQ; Parker *et al.* 2006), had an AUC as high as the CIDI-SC, but this was in a study in a mental health specialty clinic among patients presenting for treatment of depression. Two subsequent studies in that same clinic produced lower MSQ AUC estimates (0.73–0.81; Parker *et al.* 2008, 2012). Other BPD screening scales reviewed had lower AUC (0.66–0.81; Hunter *et al.* 2005; Gaynes *et al.* 2010).

https://doi.org/10.1017/S0033291712002334 Published online by Cambridge University Press

The advantage of CIDI-SC over these other scales can be traced to high CIDI-SC SN at its anti-conservative threshold (0.95). Although, as noted earlier, high SN is often accompanied by low LN+, this is not the case with CIDI-SC BPD, where LR+ is 31.6 at the anti-conservative threshold and 62–85% of screened positives would be confirmed as SCID cases if the true BPD prevalence was in the range 5–15%.

Discussion

CIDI-SC operating characteristics are equivalent to the best alternative screening scales for MDE and GAD and superior to other screening scales for PD and BPD. CIDI-SC results can be compared directly to general population epidemiological CIDI surveys because CIDI-SC items all come from the CIDI. Such nested screening scales can be useful in targeting and streamlining CIDI follow-up interviews by 'preloading' CIDI-SC responses into the CIDI computerized interview program to guide interview question skip logic. Such an integrated computerized CIDI interviewing system is currently in development and includes options for self-administering CIDI-SC on tablet computers in primary care waiting rooms, web-based CIDI-SC self-administration to track treatment response, and interviewer-based CIDI interview administration using pre-loaded CIDI-SC responses.

The fact that AUCs of continuous CIDI-SC scales in ROC analyses (0.88–0.97) are considerably higher than AUCs of dichotomized CIDI-SC scales at their unbiased thresholds (0.81–0.86) means that meaningful variation in SCID prevalence exists throughout the CIDI-SC scale ranges. One implication, as shown in the comparative analyses of LR+ and LR− at multiple thresholds, is that different thresholds can be useful for screening in than screening out cases. Importantly, the CIDI-SC has excellent LR+ and LR− at multiple informative thresholds. Furthermore, continuous CIDI-SC scores can be converted into predicted probabilities of clinical diagnoses in epidemiological studies to yield more accurate estimates of prevalence than by dichotomizing scores and classifying each respondent as either a definite case or a non-case. This predicted probability approach is discussed in more detail elsewhere (Kessler *et al.* 2010*b*).

Despite these positive findings, several study limitations are noteworthy. First, CIDI-SC SN is lower for GAD than other diagnoses. Disaggregation shows that this is because CIDI-SC have difficulty operationalizing the DSM-IV requirement that worries be excessive. CIDI-SC questions for this requirement have a higher threshold than the SCID. A similar result was found in an earlier study of the full CIDI (Wittchen *et al.* 1995). Concerns exist about clinician ability to determine when worries are excessive (Ruscio *et al.* 2005), leading to the suggestion that more concrete guidance be given in DSM-5 about defining excessiveness (Andrews *et al.* 2010). Although such guidance does not appear in currently proposed DSM-5 criteria (www.dsm5.org/ProposedRevisions), new behavioral requirements (proposed DSM-5 Criterion D) of marked avoidance, time–effort, procrastination or seeking reassurance might help to establish a threshold for excessiveness that could be the basis for improving revised CIDI-SC GAD SN.

Second, although we want to use CIDI-SC results to create a cross-walk between general population CIDI epidemiological surveys and primary care CIDI-SC screening studies, no guarantee exists that CIDI-SC operating characteristics will be similar in community epidemiological surveys and primary care samples. It is consequently important to include CIDI-SC in future CIDI community surveys and validate their operating characteristics relative to diagnoses based on the full CIDI and SCID. Such methodological studies are currently underway in new CIDI surveys in the WHO WMH Survey Initiative (Kessler & Üstün, 2008*a*).

Third, our clinical reappraisal sample was relatively small because of funding limitations, precluding cross-validation, subgroup analysis, or analysis of information values across the range of continuous CIDI-SC scores to evaluate sensitivity to change. These limitations make it especially important to replicate the current study in independent primary care samples, to investigate the stability of the encouraging results reported here and to carry out analyses of the clinical sensitivity of variation in continuous CIDI-SC scores to assess the severity of anxiety and depression. Larger replication studies could also help to establish an empirical foundation for determining whether even shorter versions of CIDI-SC might be developed based on computerized adaptive testing (Gibbons *et al.* 2011).

Supplementary material

For supplementary material accompanying this paper visit: www.hcp.med.harvard.edu/wmhcidi/resources.php.

Acknowledgments

Support for this study was provided by AstraZeneca. The sponsor had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; and preparation, review or approval of the manuscript.

Declaration of Interest

R. C. Kessler has consulted for AstraZeneca, Analysis Group, Bristol–Myers Squibb, Cerner-Galt Associates, Eli Lilly and Company, GlaxoSmithKline Inc., HealthCore Inc., Health Dialog, Integrated Benefits Institute, John Snow Inc., Kaiser Permanente, Matria Inc., Mensante, Merck & Co. Inc., Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Primary Care Network, Research Triangle Institute, Sanofi-Aventis Groupe, Shire US Inc., SRA International Inc., Takeda Global Research & Development, Transcept Pharmaceuticals Inc., and Wyeth-Ayerst; has served on advisory boards for Appliance Computing II, Eli Lilly and Company, Mindsite, Ortho-McNeil Janssen Scientific Affairs, Plus One Health Management and Wyeth-Ayerst; and has had research support for his epidemiological studies from Analysis Group Inc., Bristol–Myers Squibb, Eli Lilly and Company, EPI-Q, GlaxoSmithKline, Johnson & Johnson Pharmaceuticals, Ortho-McNeil Janssen Scientific Affairs., Pfizer Inc., Sanofi-Aventis Groupe, and Shire US Inc. **J. R. Calabrese** has received research grant support from Abbott, AstraZeneca, Bristol-Myers Squibb, Cephalon, Eli Lilly and Company, GlaxoSmithKline, Janssen, Repligen, Sunovion/DSPA, Takeda and Wyeth; has consulted to or served on advisory boards of AstraZeneca, Bristol–Myers Squibb, Cephalon, Sunovion/DSPA, Forest, GlaxoSmithKline, Janssen, Johnson and Johnson, Lundbeck, Neurosearch, OrthoMcNeil, Otsuka, Pfizer, Repligen, Schering-Plough, Servier, Solvay, Supernus, Synosia, and Wyeth; and has provided CME lectures supported by Abbott, AstraZeneca, Bristol–Myers Squibb, GlaxoSmithKline, Janssen, Johnson and Johnson, Lundbeck, Merck, Sanofi Aventis, Schering-Plough, Pfizer, Solvay, and Wyeth. **P. A. Farley, M. A. Jewell and A. C. Shillington** are employees of EPI-Q, the organization that implemented the primary care screening and recruitment for the SCID clinical re-appraisal interviews in this study. In addition, they carry out contract research for AstraZeneca, Cephalon, Merck & Co. Inc., Sanofi-Aventis, GlaxoSmithKline, Genentech, Biogen, Roche, Transcept Pharmaceuticals Inc., Lundbeck, Shire US Inc., Takeda, Novartis, Pfizer, Abbott, and Adolor. Their compensation related to these activities is limited to their salary. They also own stock in EPI-Q. **P. E. Keck** is employed by the University of Cincinnati College of Medicine and University of Cincinnati Physicians, the organization that carried out the clinical reappraisal interviews in this study. He is presently or has been in the past year a principal or co-investigator on research studies sponsored by Alkermes, AstraZeneca, Cephalon, GlaxoSmithKline, Eli Lilly and Company, Marriott Foundation, National Institute of Mental Health (NIMH), Orexigen, Pfizer Inc., and Shire. He has been reimbursed for consulting to, in the past 2 years: 2011: Pamlab, 2012: Bristol–Myers Squibb. Patents: Dr Keck is a co-inventor on United States Patent No. 6,387,956: Shapira NA, Goldsmith TD, Keck PE Jr. (University of Cincinnati) Methods of treating obsessive-compulsive spectrum disorder comprises the step of administering an effective amount of tramadol to an individual. Filed 25 March 1999; approved 14 May 2002. Dr Keck has received no financial gain from this patent. **A. A. Nierenberg** has consulted for the American Psychiatric Association, Appliance Computing Inc. (Mindsite), Basliea, Brain Cells Inc., Brandeis University, Bristol–Myers Squibb, Dey Pharmaceuticals, Dainippon Sumitomo, Eli Lilly and Company, EpiQ, L.P./Mylan Inc., Novartis, PGx Health, Shire, Schering-Plough, Takeda Pharmaceuticals, and Targacept; consulted through the MGH Clinical Trials Network and Institute (CTNI) for AstraZeneca, Brain Cells Inc., Dianippon Sumitomo/Sepracor, Johnson and Johnson, Labopharm, Merck, Methylation Science, Novartis, PGx Health, Shire, Schering-Plough, Targacept and Takeda/Lundbeck Pharmaceuticals; received grant/research support from NIMH, PamLabs, Pfizer Pharmaceuticals, and Shire; received honoraria from Belvoir Publishing, University of Texas Southwestern Dallas, Hillside Hospital, American Drug Utilization Review, American Society for Clinical Psychopharmacology, Baystate Medical Center, Columbia University, CRICO, Dartmouth Medical School, IMEDEX, Israel Society for Biological Psychiatry, Johns Hopkins University, MJ Consulting, New York State, Medscape, MBL Publishing, National Association of Continuing Education, Physicians Postgraduate Press, SUNY Buffalo, University of Wisconsin, University of Pisa, University of Michigan, University of Miami, APSARD, ISBD, SciMed, Slack Publishing and Wolters Kluwer Publishing; was currently or formerly on the advisory boards of Appliance Computing Inc., Brain Cells Inc., Eli Lilly and Company, Johnson and Johnson, Takeda/Lundbeck, Targacept, and InfoMedic; owns stock options in Appliance Computing Inc. and Brain Cells Inc.; has copyrights to the Clinical Positive Affect Scale and the MGH Structured Clinical Interview for the Montgomery Asberg Depression Scale exclusively licensed to the MGH Clinical Trials Network and Institute (CTNI); and has a patent extension application for the combination of buspirone, bupropion, and melatonin for the treatment of depression. **M. B. Stein** has a financial interest/arrangement or affiliation with one or more organizations that could be perceived as a real or apparent conflict of interest in

the context of the subject of this presentation. He receives or has in the past 3 years received Research Support from: Hoffmann-La Roche; is currently or in the past 3 years has been a Consultant for Care Management Technologies; and receives payment for editorial work from Depression and Anxiety (Publisher: Wiley-Blackwell). **M. E. Thase** has served as an advisory/consultant for Alkermes, AstraZeneca, Bristol–Myers Squibb Company, Eli Lilly and Company, Dey Pharma, L.P., Forest Laboratories, Gerson Lehman Group, GlaxoSmithKline (ended 2008), Guidepoint Global, H. Lundbeck A/S, MedAvante Inc., Merck and Co. Inc. (formerly Schering Plough and Organon), Neuronetics Inc., Novartis (ended 2008), Otsuka, Ortho-McNeil Pharmaceuticals (Johnson & Johnson), Pamlab, L.L.C., Pfizer (formerly Wyeth Ayerst Pharmaceuticals), PGx Inc., Shire US Inc., Sunovion Pharmaceuticals Inc., Supernus Pharmaceuticals, Takeda, Transcept Pharmaceuticals; has received grant support from Agency for Healthcare Research and Quality, Eli Lilly and Company, Forest Pharmaceuticals, GlaxoSmithKline (ended July 2010), National Institute of Mental Health, Otsuka Pharmaceuticals, Sepracor Inc. (ended January 2009); is or has been on the Speakers bureaus for AstraZeneca (ended June 2010), Bristol–Myers Squibb Company, Dey Pharmaceutical, Eli Lilly and Company (ended June 2009), Merck and Co. Inc., Pfizer (formerly Wyeth Ayerst Pharmaceuticals); holds equity in MedAvante Inc.; receives royalties from American Psychiatric Foundation, Guilford Publications, Herald House and W.W. Norton & Company Inc.; and has a spouse employed by Embryon (Formerly Advogent; Embryon does business with BMS and Pfizer/Wyeth). **H.-U. Wittchen** has consulted for and received research grants from Lundbek, Pfizer Inc., Novartis, Abbott, Sanofi-Aventis, Eli Lilly and Company, Merck and Co., and GlaxoSmithKline.

References

Altman DG, Bland JM (1994). Diagnostic tests 2: Predictive values. *British Medical Journal* **309**, 102.

Andrews G, Hobbs MJ, Borkovec TD, Beesdo K, Craske MG, Heimberg RG, Rapee RM, Ruscio AM, Stanley MA (2010). Generalized worry disorder: a review of DSM-IV generalized anxiety disorder and options for DSM-V. *Depression and Anxiety* **27**, 134–147.

Barkow K, Maier W, Üstün TB, Gansicke M, Wittchen HU, Heun R (2003). Risk factors for depression at 12-month follow-up in adult primary health care patients with major depression: an international prospective study. *Journal of Affective Disorders* **76**, 157–169.

Barrett B, Byford S, Knapp M (2005). Evidence of cost-effective treatments for depression: a systematic review. *Journal of Affective Disorders* **84**, 1–13.

Beekman AT, Deeg DJ, Van Limbeek J, Braam AW, De Vries MZ, Van Tilburg W (1997). Criterion validity of the Center for Epidemiologic Studies Depression scale (CES-D): results from a community-based sample of older subjects in The Netherlands. *Psychological Medicine* **27**, 231–235.

Broadhead WE, Leon AC, Weissman MM, Barrett JE, Blacklow RS, Gilbert TT, Keller MB, Olfson M, Higgins ES (1995). Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Archives of Family Medicine* **4**, 211–219.

Bruce SE, Yonkers KA, Otto MW, Eisen JL, Weisberg RB, Pagano M, Shea MT, Keller MB (2005). Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. *American Journal of Psychiatry* **162**, 1179–1187.

Bunevicius A, Peceliuniene J, Mickuviene N, Valius L, Bunevicius R (2007). Screening for depression and anxiety disorders in primary care patients. *Depression and Anxiety* **24**, 455–460.

Cohen J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Committee on Integration to Improve Population Health (2012). *Primary Care and Public Health: Exploring Integration to Improve Population Health*. National Academy Press: Washington, DC.

Converse J, Presser S (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Sage: Thousand Oaks, CA.

Cook RJ (1998). Kappa and its dependence on marginal rates. In *The Encyclopedia of Biostatistics* (ed. P. Armitage and T. Colton), pp. 2166–2168. Wiley: New York, NY.

Dodd S, Williams LJ, Jacka FN, Pasco JA, Bjerkeset O, Berk M (2009). Reliability of the Mood Disorder Questionnaire: comparison with the Structured Clinical Interview for the DSM-IV-TR in a population sample. *Australian and New Zealand Journal of Psychiatry* **43**, 526–530.

Donker T, van Straten A, Marks I, Cuijpers P (2009). A brief Web-based screening questionnaire for common mental disorders: development and validation. *Journal of Medical Internet Research* **11**, e19.

Donker T, van Straten A, Marks I, Cuijpers P (2011). Quick and easy self-rating of Generalized Anxiety Disorder: validity of the Dutch web-based GAD-7, GAD-2 and GAD-SI. *Psychiatry Research* **188**, 58–64.

Farvolden P, McBride C, Bagby RM, Ravitz P (2003). A Web-based screening instrument for depression and anxiety disorders in primary care. *Journal of Medical Internet Research* **5**, e23.

First MB, Spitzer RL, Gibbon M, Williams JBW (2002). *Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-Patient Edition* (*SCID-I/NP*). Biometrics Research, New York State Psychiatric Institute: New York, NY.

Gaynes BN, DeVeaugh-Geiss J, Weir S, Gu H, MacPherson C, Schulberg HC, Culpepper L, Rubinow DR (2010). Feasibility and diagnostic validity of

the M-3 checklist: a brief, self-rated screen for depressive, bipolar, anxiety, and post-traumatic stress disorders in primary care. *Annals of Family Medicine* **8**, 160–169.

Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, Schumacher J, Saag M, Kitahata MM, Crane PK (2011). Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research* **20**, 1349–1357.

Gilbody S, Richards D, Brealey S, Hewitt C (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of General Internal Medicine* **22**, 1596–1602.

Gili M, Luciano JV, Serrano MJ, Jimenez R, Bauza N, Roca M (2011). Mental disorders among frequent attenders in primary care: a comparison with routine attenders. *Journal of Nervous and Mental Disease* **199**, 744–749.

Green JG, Avenevoli S, Finkelman M, Gruber MJ, Kessler RC, Merikangas KR, Sampson NA, Zaslavsky AM (2011). Validation of the diagnoses of panic disorder and phobic disorders in the US National Comorbidity Survey Replication Adolescent (NCS-A) supplement. *International Journal of Methods in Psychiatric Research* **20**, 105–115.

Groves RM, Fowler Jr. FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2009). *Survey Methodology*, 2nd edn. Wiley: New York, NY.

Haynes RB, Sackett DL, Guyatt GH, Tugwell P (2006). *Clinical Epidemiology: How to Do Clinical Practice Research*, 3rd edn. Lippincott Williams & Wilkins: Philadelphia, PA.

Hirschfeld RM, Cass AR, Holt DC, Carlson CA (2005). Screening for bipolar disorder in patients treated for depression in a family medicine clinic. *Journal of the American Board of Family Practice* **18**, 233–239.

Hirschfeld RM, Holzer C, Calabrese JR, Weissman M, Reed M, Davies M, Frye MA, Keck P, McElroy S, Lewis L, Tierce J, Wagner KD, Hazard E (2003). Validity of the Mood Disorder Questionnaire: a general population study. *American Journal of Psychiatry* **160**, 178–180.

Hirschfeld RM, Vornik LA (2004). Recognition and diagnosis of bipolar disorder. *Journal of Clinical Psychiatry* **65** (Suppl. 15), 5–9.

Hirschfeld RM, Williams JB, Spitzer RL, Calabrese JR, Flynn L, Keck Jr. PE, Lewis L, McElroy SL, Post RM, Rapport DJ, Russell JM, Sachs GS, Zajecka J (2000). Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *American Journal of Psychiatry* **157**, 1873–1875.

Houston JP, Kroenke K, Faries DE, Doebbeling CC, Adler LA, Ahl J, Swindle R, Trzepacz PT (2011). A provisional screening instrument for four common mental disorders in adult primary care patients. *Psychosomatics* **52**, 48–55.

Hunter EE, Penick EC, Powell BJ, Othmer E, Nickel EJ, Desouza C (2005). Development of scales to screen for eight common psychiatric disorders. *Journal of Nervous and Mental Disease* **193**, 131–135.

Katz MM, Secunda SK, Hirschfeld RM, Koslow SH (1979). NIMH clinical research branch collaborative program on the psychobiology of depression. *Archives of General Psychiatry* **36**, 765–771.

Kessler RC, Adler LA, Gruber MJ, Sarawate CA, Spencer T, Van Brunt DL (2007). Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members. *International Journal of Methods in Psychiatric Research* **16**, 52–65.

Kessler RC, Akiskal HS, Angst J, Guyer M, Hirschfeld RM, Merikangas KR, Stang PE (2006). Validity of the assessment of bipolar spectrum disorders in the WHO CIDI 3.0. *Journal of Affective Disorders* **96**, 259–269.

Kessler RC, Coulouvrat C, Hajak G, Lakoma MD, Roth T, Sampson N, Shahly V, Shillington A, Stephenson JJ, Walsh JK, Zammit GK (2010*a*). Reliability and validity of the Brief Insomnia Questionnaire in the America Insomnia Survey. *Sleep* **33**, 1539–1549.

Kessler RC, Demler O, Frank RG, Olfson M, Pincus HA, Walters EE, Wang P, Wells KB, Zaslavsky AM (2005). Prevalence and treatment of mental disorders, 1990 to 2003. *New England Journal of Medicine* **352**, 2515–2523.

Kessler RC, Green JG, Gruber MJ, Sampson NA, Bromet E, Cuitan M, Furukawa TA, Gureje O, Hinkov H, Hu CY, Lara C, Lee S, Mneimneh Z, Myer L, Oakley-Browne M, Posada-Villa J, Sagar R, Viana MC, Zaslavsky AM (2010*b*). Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) Survey Initiative. *International Journal of Methods in Psychiatric Research* **19** (Suppl. 1), 4–22.

Kessler RC, Üstün TB (2004). The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research* **13**, 93–121.

Kessler RC, Üstün TB (2008*a*). Overview and future directions for the World Mental Health Survey Initiative. In *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders* (ed. R. C. Kessler and T. B. Üstün), pp. 555–568. Cambridge University Press: New York, NY.

Kessler RC, Üstün TB (eds) (2008*b*). *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. Cambridge University Press: New York, NY.

Kessler RC, Wittchen H-U, Abelson JM, McGonagle KA, Schwarz N, Kendler KS, Knäuper B, Zhao S (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *International Journal of Methods in Psychiatric Research* **7**, 33–55.

Kisely S, Scott A, Denney J, Simon G (2006). Duration of untreated symptoms in common mental disorders: association with outcomes: international study. *British Journal of Psychiatry* **189**, 79–80.

Kisely S, Simon G (2005). An international study of the effect of physical ill health on psychiatric recovery in primary care. *Psychosomatic Medicine* **67**, 116–122.

Klinkman MS, Coyne JC, Gallo S, Schwenk TL (1997). Can case-finding instruments be used to improve

physician detection of depression in primary care? *Archives of Family Medicine* **6**, 567–573.

Konnopka A, Leichsenring F, Leibing E, Konig HH (2009). Cost-of-illness studies and cost-effectiveness analyses in anxiety disorders: a systematic review. *Journal of Affective Disorders* **114**, 14–31.

Kraemer HC (1992). *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Sage: Newbury Park, CA.

Kroenke K, Spitzer RL, Williams JB, Lowe B (2010). The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry* **32**, 345–359.

Kroenke K, Spitzer RL, Williams JB, Monahan PO, Lowe B (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine* **146**, 317–325.

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lowe B, Grafe K, Zipfel S, Spitzer RL, Herrmann-Lingen C, Witte S, Herzog W (2003). Detecting panic disorder in medical and psychosomatic outpatients: comparative validation of the Hospital Anxiety and Depression Scale, the Patient Health Questionnaire, a screening question, and physicians' diagnosis. *Journal of Psychosomatic Research* **55**, 515–519.

Lowe B, Spitzer RL, Williams JB, Mussell M, Schellberg D, Kroenke K (2008). Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *General Hospital Psychiatry* **30**, 191–199.

Manea L, Gilbody S, McMillan D (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal* **184**, E191–E196.

Margolis DJ, Bilker W, Boston R, Localio R, Berlin JA (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology* **55**, 518–524.

Miller CJ, Johnson SL, Kwapil TR, Carver CS (2011). Three studies on self-report scales to detect bipolar disorder. *Journal of Affective Disorders* **128**, 199–210.

Miller CJ, Klugman J, Berv DA, Rosenquist KJ, Ghaemi SN (2004). Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. *Journal of Affective Disorders* **81**, 167–171.

Parker G, Fletcher K, Barrett M, Synnott H, Breakspear M, Hyett M, Hadzi-Pavlovic D (2008). Screening for bipolar disorder: the utility and comparative properties of the MSS and MDQ measures. *Journal of Affective Disorders* **109**, 83–89.

Parker G, Graham R, Hadzi-Pavlovic D, Fletcher K, Hong M, Futeran S (2012). Further examination of the utility and comparative properties of the MSQ and MDQ bipolar screening measures. *Journal of Affective Disorders* **138**, 104–109.

Parker G, Hadzi-Pavlovic D, Tully L (2006). Distinguishing bipolar and unipolar disorders: an isomer model. *Journal of Affective Disorders* **96**, 67–73.

Pepe MS (2003). *Statistical Analysis of Medical Tests for Classification and Prediction*. Oxford University Press: New York, NY.

Presser S, Blair J (1994). Survey pretesting: do different methods yield different results? *Sociological Methodology* **24**, 73–104.

Presser S, Couper MP, Lessler JT, Martin E, Martin J, Rothgeb JM, Singer E (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly* **68**, 109–130.

Radloff LS (1977). The CES-D Scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385–401.

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA, Sartorius N, Towle L (1988). The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

Ruscio AM, Lane M, Roy-Byrne P, Stang PE, Stein DJ, Wittchen HU, Kessler RC (2005). Should excessive worry be required for a diagnosis of generalized anxiety disorder? Results from the US National Comorbidity Survey Replication. *Psychological Medicine* **35**, 1761–1772.

Schulberg HC, Saul M, McClelland M, Ganguli M, Christy W, Frank R (1985). Assessing depression in primary medical and psychiatric practices. *Archives of General Psychiatry* **42**, 1164–1170.

Spitzer RL, Kroenke K, Williams JB (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *Journal of the American Medical Association* **282**, 1737–1744.

Stein MB, Roy-Byrne PP, McQuaid JR, Laffaye C, Russo J, McCahill ME, Katon W, Craske M, Bystritsky A, Sherbourne CD (1999). Development of a brief diagnostic screen for panic disorder in primary care. *Psychosomatic Medicine* **61**, 359–364.

Susser E, Schwartz S, Morabia A, Bromet EJ (2006). *Psychiatric Epidemiology: Searching for the Causes of Mental Disorders*. Oxford Universiy Press: New York, NY.

Thomas JL, Jones GN, Scarinci IC, Mehan DJ, Brantley PJ (2001). The utility of the CES-D as a depression screening measure among low-income women attending primary care clinics. The Center for Epidemiologic Studies-Depression. *International Journal of Psychiatry in Medicine* **31**, 25–40.

Tsuang MT, Tohen M, Jones P (eds) (2011). *Textbook of Psychiatric Epidemiology*. Wiley: New York, NY.

Twiss J, Jones S, Anderson I (2008). Validation of the Mood Disorder Questionnaire for screening for bipolar disorder in a UK sample. *Journal of Affective Disorders* **110**, 180–184.

Üstün TB, Sartorius N (eds) (1995). *Mental Illness in General Health Care: An International Study*. Wiley: New York, NY.

Weber Rouget B, Gervasoni N, Dubuis V, Gex-Fabry M, Bondolfi G, Aubry JM (2005). Screening for bipolar disorders using a French version of the Mood Disorder Questionnaire (MDQ). *Journal of Affective Disorders* **88**, 103–108.

Willis GB (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage: Thousand Oaks, CA.

Wittchen HU (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.

Wittchen HU, Kessler RC, Beesdo K, Krause P, Hofler M, Hoyer J (2002). Generalized anxiety and depression in primary care: prevalence, recognition, and management. *Journal of Clinical Psychiatry* **63** (Suppl. 8), 24–34.

Wittchen HU, Kessler RC, Zhao S, Abelson J (1995). Reliability and clinical validity of UM-CIDI DSM-III-R generalized anxiety disorder. *Journal of Psychiatric Research* **29**, 95–110.

Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *General Hospital Psychiatry* **29**, 388–395.

Wolter KM (1985). *Introduction to Variance Estimation*. Springer-Verlag: New York, NY.

Zigmond AS, Snaith RP (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica* **67**, 361–370.

Zimmerman M, Galione JN, Ruggero CJ, Chelminski I, McGlinchey JB, Dalrymple K, Young D (2009). Performance of the mood disorders questionnaire in a psychiatric outpatient setting. *Bipolar Disorders* **11**, 759–765.