International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 15(4): 167–180 (2006)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.196

Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys

JOSEP MARIA HARO,[1] SAENA ARBABZADEH-BOUCHEZ,[2] TRAOLACH S. BRUGHA,[3] GIOVANNI DE GIROLAMO,[4] MARGARET E. GUYER,[5] ROBERT JIN,[6] JEAN PIERRE LEPINE,[2] FAUSTO MAZZI,[7] BLANCA RENESES,[8] GEMMA VILAGUT,[9] NANCY A. SAMPSON,[6] RONALD C. KESSLER[6]

1 Fundació Sant Joan de Déu per la Recerca i la Docència, Barcelona, Spain
2 Hôpital Fernand Widal, Paris, France
3 University of Leicester, UK
4 AUSL, Città de Bologna, Italy
5 Massachusetts Mental Health Center, Boston, USA
6 Harvard Medical School, Department of Health Care Policy, Boston, USA
7 Universita degli Studi di Modena e Regio, Emilia, Italy
8 Hospital Clinico San Carlos, Department of Psychiatry, Madrid, Spain
9 Health Services Research Unit, Institut Municipal d´Investigació Mèdica, Barcelona, Spain

Abstract

The DSM-IV diagnoses generated by the fully structured lay-administered Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) in the WHO World Mental Health (WMH) surveys were compared to diagnoses based on follow-up interviews with the clinician-administered non-patient edition of the Structured Clinical Interview for DSM-IV (SCID) in probability subsamples of the WMH surveys in France, Italy, Spain, and the US. CIDI cases were oversampled. The clinical reappraisal samples were weighted to adjust for this oversampling. Separate samples were assessed for lifetime and 12-month prevalence. Moderate to good individual-level CIDI-SCID concordance was found for lifetime prevalence estimates of most disorders. The area under the ROC curve (AUC, a measure of classification accuracy that is not influenced by disorder prevalence) was 0.76 for the dichotomous classification of having any of the lifetime DSM-IV anxiety, mood and substance disorders assessed in the surveys and in the range 0.62–0.93 for individual disorders, with an inter-quartile range (IQR) of 0.71–0.86. Concordance increased when CIDI symptom-level data were added to predict SCID diagnoses in logistic regression equations. AUC for individual disorders in these equations was in the range 0.74–0.99, with an IQR of 0.87–0.96. CIDI lifetime prevalence estimates were generally conservative relative to SCID estimates. CIDI-SCID concordance for 12-month prevalence estimates could be studied powerfully only for two disorder classes, any anxiety disorder (AUC = 0.88) and any mood disorder (AUC = 0.83). As with lifetime prevalence, 12-month concordance improved when CIDI symptom-level data were added to predict SCID diagnoses. CIDI 12-month prevalence estimates were unbiased relative to SCID estimates. The validity of the CIDI is likely to be under-estimated in these comparisons due to the fact that the reliability of the SCID diagnoses, which is presumably less than perfect, sets a ceiling on maximum CIDI-SCID concordance. Copyright © 2006 John Wiley & Sons, Ltd.

Key words: Composite International Diagnostic Interview (CIDI), validity, WHO World Mental Health (WMH) Survey Initiative

This paper presents the results of a clinical reappraisal study carried out in conjunction with the WHO World Mental Health (WMH) Survey Initiative. The purpose of the study was to estimate the concordance of diagnoses based on the instrument used in the WMH surveys, the WHO Composite International Diagnostic Interview (CIDI) Version 3.0 (CIDI 3.0) (Kessler and Ustun, 2004) with diagnoses based on followup clinical interviews. The clinical interview schedule used for this purpose was the Axis I research version, non-patient edition of the Structured Clinical Interview for DSM-IV (SCID) (First et al. 2002).

Previous clinical reappraisal studies showed that earlier versions of the CIDI, which were based on DSM-III-R criteria, generated DSM diagnoses generally consistent with those obtained in SCID clinical reappraisal interviews in community surveys (Wittchen, 1994; Wittchen et al., 1995; Wittchen et al., 1996; Kessler et al., 1998). In a variety of the settings, the results of CIDI clinical reappraisal studies that generated diagnoses based on ICD-10 criteria using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1990) as the clinical gold standard were more variable, with some studies showing the CIDI diagnoses to have poor agreement with SCAN diagnoses (Brugha et al., 2001) in a community sample and others showing agreement to be either good (Andrews et al., 1995) in a patient sample or excellent (Jordanova et al., 2004) in a primary care-provider sample.

The WMH surveys are the first to administer CIDI 3.0, which operationalizes both ICD-10 and DSM-IV criteria and was developed to improve the validity of the CIDI with the benefit of insights gained from the CIDI clinical reappraisal studies cited in the previous paragraph. Methodological studies are needed to evaluate the consistency of CIDI diagnoses based on these new criteria with clinical reinterviews, as it is not legitimate to assume that the results of the earlier methodological studies hold for this new version of CIDI. The results reported in the current paper are based on methodological studies of this sort that were carried out in probability subsamples of WMH samples in four countries: France, Italy, Spain, and the US. CIDI cases were oversampled. The data were weighted to adjust for this oversampling. Separate examinations were carried out of lifetime prevalence and 12-month prevalence.

The analysis of the clinical reappraisal data had three phases, the third of which went beyond earlier studies. The first phase considered aggregate CIDI-SCID consistency of prevalence estimates. The second phase considered CIDI-SCID consistency of individual-level diagnostic classifications. The third phase considered whether CIDI-SCID concordance would be improved significantly by developing prediction equations in which CIDI item-level data were used along with CIDI diagnostic data to predict SCID diagnoses. As discussed in more detail previously in this journal (Kessler et al., 2004a) and illustrated in a series of recent disorder-specific reports (Kessler et al., 2005; Kessler et al., 2006; Lenzenweger et al. in press), these predicted probabilities can either be treated as outcomes in substantive analyses or can be used as input to more complex analyses that use the method of multiple imputation (MI) (Rubin, 1987) to make estimates of the prevalence and correlates of clinical diagnoses from CIDI data. Comparison with parallel estimates of the prevalence and correlates of CIDI diagnoses allows much more fine-grained consideration of consistency with clinical diagnoses than conventional analyses of diagnostic concordance.

Methods

The main samples

As noted above, clinical reappraisal studies were carried out in WMH surveys in France, Italy, Spain, and the US. The first three of these four surveys are part of the European Study of the Epidemiology of Mental Disorders (ESEMeD) regional consortium within the WMH Survey Initiative. The ESEMeD surveys were administered face-to-face to a combined sample of 21,425 respondents (from a high of 5,473 in Spain to a low of 2,372 in the Netherlands) in 2001–2 (Belgium, France, Italy, Spain) and 2002–3 (Germany, Netherlands). The sample designs were distinct in each country but all featured stratified multistage clustered samples of the household population. The surveys focused on household residents ages 18 and older who could speak the official language(s) of the countries. Verbal or written consent was obtained prior to data collection. The combined response rate was 61.2% (from a high of 78.6% in Spain to a low of 45.9% in France). The use of respondent incentives varied across countries. The use of population registries in some countries allowed direct selection of individuals, avoiding the need for a within-household probability of selection weight. Non-response adjustment weights were used along with more conventional within-household and post-stratification

DOI: 10.1002/mpr

weights to create a composite weight in each country. A more detailed discussion of ESEMeD sampling and weighting is presented elsewhere (Alonso et al., 2004).

The US survey was the National Comorbidity Survey Replication (NCS-R) (Kessler and Merikangas, 2004), a face-to-face survey of 9,282 adult respondents (ages 18+) carried out in 2001–3. The sample was based on a multistage clustered area probability design described in more detail previously in this journal (Kessler et al., 2004b). The response rate was 70.9%. Respondents were given a $50 incentive for participation. A probability subsample of hard-to-recruit predesignated respondents was selected for a brief telephone non-respondent survey. Non-respondent survey participants were given a $100 incentive. The results of the non-respondent survey were used to create a non-response adjustment weight that was added to more conventional within-household probability of selection and post-stratification weights to create a composite NCS-R weight.

The clinical reappraisal samples

The clinical reappraisal samples oversampled CIDI cases but also included non-cases. The methods of subsampling differed in ESEMeD and the NCS-R, but was based on probability procedures in both cases so that the clinical reappraisal sample could be weighted back to be representative of the total original sample. This weighting took into consideration the sample design of the original surveys, including differential probability of selection in households depending on sample size and post-stratification, so that significance tests could be made using appropriate design-based methods.

The ESEMeD clinical reappraisal study was based on a probability subsample of respondents who lived in targeted geographical areas in France, Italy and Spain. One hundred per cent of the main survey respondents with any 12-month DSM-IV/CIDI diagnosis in these targeted regions were selected for participation in the clinical reappraisal study along with a random 10% subsample of other respondents in the same regions, for a sample of 428 completed clinical reappraisal interviews (87 in France, 194 in Italy, and 137 in Spain). As the focus was on consistency of assessing 12-month prevalence, the ideal design would have been one that administered reappraisal interviews within a short time of the initial interviews. However, logistical complications led to delays, with many of the reappraisal interviews carried out as much as six months after the initial CIDI interviews. This means that the overlap in the recall period of 12-month prevalence estimates varied across respondents and that the number of interviews completed according to protocol (within two months of the CIDI interview) is much smaller than the total number of clinical interviews (n = 143). As described below in the results section, this variation was taken into account in analysis.

The NCS-R clinical reappraisal study was based on a probability subsample of all respondents in telephone households who participated in the NCS-R throughout the entire US. The clinical reappraisal sample oversampled respondents who met DSM-IV/CIDI lifetime criteria for one or more relatively uncommon disorders (for example, agoraphobia, panic disorder, generalized anxiety disorder, substance abuse with dependence) at a higher rate than respondents in a second sampling stratum who met criteria only for more common disorders (for example, specific phobia, major depression, substance abuse with or without dependence), with the lowest sampling fraction for a third stratum made up of respondents who failed to meet criteria for any lifetime DSM-IV/CIDI disorder. Selection was made proportional to respondent weights in the main sample so as partially to cancel out the within-household probability of selection weight. A total of 325 clinical reappraisal interviews were completed according to protocol. In addition to the main clinical reappraisal sample, separate, more focused clinical reappraisal samples were selected to validate the CIDI diagnoses of bipolar disorder (n = 40) and adult attention-deficit/hyperactivity disorder (ADHD) (n = 154).

All the SCID clinical reappraisal interviews in the NCS-R clinical reappraisal study and approximately 40% of those in the ESEMeD clinical reappraisal sample were administered over the telephone. The remaining ESEMeD clinical reappraisal interviews were administered face-to-face. Telephone administration is now widely accepted in clinical reappraisal studies based on evidence of comparable validity to in-person administration (Kendler et al., 1992, Rohde et al., 1997, Sobin et al., 1993). A great advantage of telephone administration is that a centralized and closely supervised clinical interview staff can carry out the interviews throughout the entire sample area without the geographic restriction that is typically required for face-to-face clinical assessment. A disadvantage is that the small part of the population without telephones cannot be included in clinical calibration studies when interviews are done by telephone. This limitation was removed in the ESEMeD study because respondents in non-phone households were interviewed face-to-face. It is a limitation in the NCS-R study, though, that the roughly 5% of US households without phones were excluded from the clinical reappraisal sample.

The clinical reappraisal study design

The clinical reappraisal study was designed to determine the extent to which the diagnostic classifications made on the basis of the CIDI would have been different if the surveys had been carried out entirely by carefully trained clinical interviewers using the SCID rather than by trained lay interviewers with the CIDI. As the entry questions (the diagnostic stem questions) in the CIDI and SCID are very similar, the distinction between the two types of interview hinges on two CIDI-SCID differences: differences in the ability to elicit endorsement of diagnostic stem questions based on CIDI yes-no questions versus the more flexible open-ended probing in the SCID; and differences in symptom assessments among respondents who endorse diagnostic stem questions based on fully structured CIDI questions versus the more conversational probes in the SCID. We had no doubt but that the SCID procedures were superior to the CIDI procedures in eliciting clear information about symptom characteristics. It was less clear to us, though, whether the SCID was superior to the CIDI in eliciting endorsement of diagnostic stem questions, as our previous work carrying out CIDI-SCID comparisons documented some cases in which respondents were more comfortable admitting embarrassing feelings and behaviours to lay interviewers than to clinical interviewers (Kessler et al., 1998).

A major impediment to making accurate CIDI-SCID comparisons of the sort described in the last paragraph is that respondents are inconsistent in their reports over time. Indeed, our own previous experience and that of other researchers shows consistently that respondents in community surveys tend to report less and less as they are interviewed more and more due to respondent fatigue (Bromet et al., 1986). Part of this pattern is a tendency for respondents to endorse a smaller number of diagnostic stem questions in follow-up interviews than in initial interviews (Kessler et al., 1998), leading to the biased perception that initial structured interviews overestimate prevalence compared to second clinical interviews. Based on this observation, we modified the conventional blinded clinical reinterview design in two important ways in the WMH clinical reappraisal study. First, we unblinded the clinical interviewers to whether the respondents endorsed diagnostic stem questions in the CIDI but not to the final CIDI diagnoses. Second, we encouraged respondents to endorse diagnostic stem questions in the clinical reappraisal interviews by reminding respondents who endorsed the CIDI stem question in their initial interview of this fact. This partial unblinding of interviewers might be seen as introducing a bias, but that turns out not to be the case due to the fact that the majority of community survey respondents who endorse CIDI stem questions do not go on to meet full CIDI criteria for the associated disorder.

The stem question reminder process, in comparison, had a substantial effect on the completeness of respondent reports in clinical re-interviews. Respondents were told at the beginning of their clinical reinterview that they will be asked some of the same questions as in their earlier interview. They were also told that this was being done to test the interview and not to test their memory, so they should answer without trying to remember what they said to the earlier interviewer. Respondents were then taken through the clinical interview in the usual fashion, with the exception that the sections of the clinical reinterview in which they endorsed a diagnostic stem question in the CIDI were started with the introduction: 'During the first interview, you said [presentation of the stem question endorsed in the NCS interview]. Has that happened in the past 12 months?' Reinterview respondents could still deny that they reported a diagnostic stem question in the initial interview, although this was uncommon. In cases where the respondent had not endorsed the CIDI stem question in the original interview, the SCID probing for a diagnostic stem endorsement was carried out in the conventional fashion so as to discover false negative responses in the CIDI. The clinical interviewers also had complete flexibility to go back to a diagnostic section that was previously skipped if any information subsequently surfaced in the interview to suggest a positive response to the diagnostic stem question.

Clinical interviewer training and supervision

Clinical interviewers were carefully trained by the same SCID training team in the NCS-R and ESEMeD studies and were closely supervised during the course

DOI: 10.1002/mpr

of fieldwork using the same quality assurance protocol. As noted above, the version of the SCID used was a modified version of the Axis I research version, non-patient edition (First et al., 2002). An expanded version of the model training programme created by the developers of the SCID (Gibbon et al., 1981) was used for interviewer training. This programme featured

- the use of the standard SCID training tapes and manuals, which take an average of approximately 30 hours of self-study, followed by
- 40 hours of in-person group training by experienced SCID trainers, and
- ongoing quality control monitoring throughout the field period.

Textual extracts from the DSM-IV manual were supplied to interviewers that described where available specific study diagnostic criteria in order to maximize the reliability of clinical ratings.

Quality control monitoring included clinical supervisor review of all hard copy completed SCID interviews, recontact of respondents whenever the clinical supervisor felt that more information was needed to make a rating, periodic consultation with diagnostic experts who served as consultants for complex cases, consultant review of a random subsample of interview audiotapes, and biweekly interviewer-supervisor meetings to prevent drift. As training materials were all in English, the clinical interviewers in Europe had to be bilingual and were trained in what was to them a foreign language. Because of this, special care was taken to expand the second phase of training in Europe and to have the clinical supervisors in these studies specially trained by the US trainers. In addition, the US trainers provided ongoing telephone and email consultation to clinical supervisors in the ESEMeD countries throughout the field period in order to maintain comparability between the NCS-R and ESEMeD reappraisal exercises.

Analysis methods

After weighting the clinical reappraisal sample data to be representative of the main samples, we investigated whether CIDI prevalence estimates are biased in comparison to SCID prevalence estimates using McNemar χ^2 tests to evaluate the statistical significance of differences in the proportions of respondents who were false positives versus false negatives. As with all our significance tests, McNemar tests were carried out using design-based estimation methods that adjusted for the effects of weighting and clustering and oversampling of CIDI cases (Kish and Frankel, 1974; Wolter, 1985).

Individual-level CIDI-SCID diagnostic concordance was next evaluated using two different descriptive measures: the area under the receiver operator characteristic curve (AUC) (Hanley and McNeil, 1982) and Cohen's κ (Cohen, 1960). Although κ is the most widely used measure of concordance in validity studies of psychiatric disorders, it has been criticized because it is dependent on prevalence and consequently is often low in situations where there appears to be high agreement between low-prevalence measures (Byrt et al., 1993; Cook, 1998; Kraemer et al., 2003). An important implication is that κ varies across populations that differ in prevalence even when the populations do not differ in sensitivity (SN) (the percentage of true cases correctly classified by the CIDI) or specificity (SP) (the percent of true non-cases correctly classified). As sensitivity and specificity are considered to be fundamental parameters, this means that the comparison of κ across different populations cannot be used to evaluate the cross-population performance of a test.

Critics of κ prefer to assess concordance with measures that are a function of SN and SP. The odds ratio (OR) meets this requirement, as OR is equal to $[SN \times SP]/[(1 - SN) \times (1 - SP)]$ (Agresti, 1996). However, the upper end of the OR is unbounded, making it difficult to use the OR to evaluate the extent to which CIDI diagnoses are consistent with clinical diagnoses. Yules Q has been proposed as an alternative measure to resolve this problem (Spitznagel and Helzer, 1985), as Q is a bounded transformation of OR $[Q = (OR - 1)/(OR + 1)]$ that ranges between −1 and +1. Q can be interpreted as the difference in the probabilities of a randomly selected clinical case and a randomly selected clinical non-case that differ in their classification on the CIDI being correctly versus incorrectly classified by the CIDI. The difficulty with Q is that 'tied pairs' (clinical cases and non-cases that have the same CIDI classification) are excluded, which means that Q does not tell us about actual prediction accuracy.

The AUC is a measure that resolves this problem, as AUC can be interpreted as the probability that a randomly selected clinical case will score higher on the CIDI than a randomly selected non-case. Although

DOI: 10.1002/mpr

developed to study the association between a continuous predictor and a dichotomous outcome, the AUC can be used in the special case where the predictor is a dichotomy, in which case AUC equals (SN + SP)/2. As a result of this useful interpretation, we focus on AUC in our evaluation of CIDI-SCID diagnostic concordance. We also report SN and SP, the key components of AUC in the dichotomous case, as well as positive predictive value (PPV; the proportion of CIDI cases that are confirmed by the SCID), negative predictive value (NPV; the proportion of CIDI non-cases that are confirmed as non-cases by the SCID), and κ.

The third phase of analysis involved estimation of a series of stepwise logistic regression equations in which SCID diagnoses were treated as dichotomous outcomes and CIDI symptom variables were included along with CIDI diagnoses as predictors in order to determine whether CIDI symptom-level data could significantly improve the prediction of SCID diagnoses compared to prediction from CIDI diagnoses. As discussed in more detail elsewhere (Kessler et al., 2004a), significant improvement of this sort could be used to generate predicted probabilities of SCID diagnoses for each survey respondent who was not in the clinical reappraisal sample. Diagnostic imputations based on these predicted probabilities could then be used to make estimates of the prevalence and correlates of clinical diagnoses in the full sample so as to incorporate the analysis of validity into substantive investigations. For example, it would be possible in this way to carry out parallel analyses of the extent to which the correlates of predicted SCID diagnoses differ from the correlates of CIDI diagnoses.

The AUC was the descriptive statistic used to describe these improvements. As noted above, the AUC is typically used with a dimensional predictor and a dichotomous outcome. As a result, it is a simple matter to think of the AUC as the association between a predicted probability of a dichotomous outcome, in our case based either on prediction from the dichotomous CIDI case classification or from a logistic regression equation containing both CIDI diagnoses and symptom measures as predictors, and the observed classifications on the outcome. This makes it possible to evaluate the extent to which AUC increases as more complex predictors are added to an equation over and above the initial CIDI dichotomous diagnostic classification.

Results

Lifetime aggregate concordance

Separate disorder-specific CIDI-SCID comparisons of lifetime prevalence were made in the NCS-R for panic disorder, phobias, PTSD, major depression, bipolar disorder, and alcohol or drug abuse with or without dependence. (Table 1) McNemar tests of CIDI versus SCID differences in estimated lifetime prevalence are insignificant for panic disorder, agoraphobia (with or without panic), specific phobia, and bipolar I–II disorder, but are significant for all other disorders. As shown by the fact that PPV is consistently higher than SN, these differences are due to the CIDI lifetime prevalence estimates being conservative relative to the SCID estimates. SCID prevalence estimates are 34% higher than CIDI prevalence estimates for any anxiety disorder, 33% higher for major depression, 53% higher for alcohol or drug abuse or dependence and 42% higher for any of the above disorders.

Lifetime individual-level concordance

Using descriptors modelled on those used for roughly comparable values of κ (Landis and Koch, 1977), individual-level CIDI-SCID lifetime prevalence concordance is moderate (AUC in the range 0.7–0.8) for the majority of diagnoses assessed in the NCS-R clinical reappraisal study (panic disorder, any phobia, panic disorder or any phobia, any anxiety disorder, major depression, alcohol dependence, drug abuse, and any disorder) (Table 1). Concordance is almost perfect (AUC greater than or equal to 0.9), in comparison, for bipolar disorder, substantial (AUC in the range 0.8–0.9) for agoraphobia and alcohol abuse, and fair (AUC in the range 0.6–0.7) for the remaining disorders (specific phobia, social phobia, PTSD, drug dependence). The majority of SCID cases are detected by the CIDI (SN) for anxiety disorders (54.4%; 38.3–62.6%), major depression (55.3%), bipolar disorder (86.8%), substance dependence (73.6%) and any disorder (62.8%). The vast majority of CIDI cases, in comparison, are confirmed by the SCID (PPV), including 74.5% (43.9%–86.1%) with anxiety disorder, 58.3%–73.7% with mood disorder, 82.0%–98.7% with substance disorder and 84.3% with any disorder.

Lifetime concordance using CIDI symptom-level data

Stepwise logistic regression analysis was used to select CIDI symptom questions for each diagnosis that signifi-

DOI: 10.1002/mpr

Table 1. Consistency of lifetime DSM-IV CIDI and SCID diagnoses in the NCS-R clinical reappraisal sample (n = 325)[1]

	McNemar χ^2_1 Test	AUC	κ	(se)	OR	(95% CI)	SN	(se)	SP	(se)	PPV	(se)	NPV	(se)
I. Anxiety disorders														
Panic disorder	0.0	0.72	0.45	(0.15)	56.3*	(15.5–204.6)	45.8	(13.0)	98.5	(0.5)	48.4	(14.1)	98.4	(0.5)
Agoraphobia	0.0	0.81	0.61	(0.15)	174.8*	(39.1–780.6)	62.6	(12.9)	99.1	(0.5)	62.0	(15.7)	99.1	(0.3)
Specific phobia	0.0	0.67	0.33	(0.07)	6.3*	(2.7–14.6)	45.2	(8.7)	88.5	(2.3)	43.9	(7.6)	89.0	(2.8)
Social phobia	5.7*	0.65	0.35	(0.07)	8.4*	(3.9–18.2)	36.6	(7.0)	93.6	(1.4)	53.9	(7.6)	87.8	(2.6)
Any phobia	7.5*	0.71	0.45	(0.06)	9.8*	(5.0–19.4)	51.7	(5.7)	90.2	(1.9)	68.1	(6.0)	82.1	(3.0)
Panic or any phobia	7.4*	0.71	0.46	(0.06)	9.9*	(5.1–19.5)	52.6	(5.6)	90.0	(2.0)	68.7	(5.8)	81.9	(3.1)
Post-traumatic stress disorder	11.4*	0.69	0.49	(0.10)	64.9*	(14.9–281.9)	38.3	(11.8)	99.1	(0.5)	86.1	(7.7)	91.3	(3.0)
Any anxiety disorder	12.1*	0.73	0.48	(0.05)	11.6*	(6.0–22.4)	54.4	(5.3)	90.7	(1.8)	74.5	(5.0)	80.0	(3.2)
II. Mood disorders														
Major depressive disorder	7.2*	0.75	0.54	(0.06)	18.4*	(7.9–42.9)	55.3	(6.8)	93.7	(1.9)	73.7	(7.0)	86.8	(2.7)
Bipolar I/II[2]	0.6	0.93	0.69	(0.3)	582.6*	(72.6–4674)	86.8	(10.5)	98.9	(0.5)	58.3	(14.5)	99.8	(0.2)
III. Substance disorders														
Alcohol abuse[3]	7.3*	0.81	0.70	(0.06)	93.3*	(28.0–311.3)	64.1	(7.4)	98.1	(1.0)	88.1	(5.6)	92.7	(2.0)
Drug abuse[3]	8.9*	0.76	0.63	(0.08)	111.8*	(26.3–476.3)	53.7	(12.7)	99.0	(0.5)	88.2	(6.0)	93.8	(2.7)
Alcohol dependence with abuse [3]	18.5*	0.72	0.56	(0.09)	877.0*	(105.8–7266.2)	43.1	(9.3)	99.9	(0.1)	98.7	(1.3)	91.9	(1.7)
Drug dependence with abuse [3]	11.3*	0.62	0.36	(0.12)	74.0*	(9.2–625.0)	25.0	(10.6)	99.6	(0.4)	82.0	(13.9)	94.2	(2.2)
Revised alcohol dependence[4]	0.1	0.879	0.77	(0.06)	129.1*	(38.9–428.0)	78.6	(7.7)	97.2	(1.1)	81.5	(6.6)	96.7	(1.3)
Revised drug dependence[4]	0.0	0.798	0.59	(0.10)	52.7*	(10.0–278.2)	62.7	(16.9)	96.9	(1.2)	62.3	(11.9)	97.0	(2.0)
Revised alcohol or drug dependence[4]	2.6	0.856	0.76	(0.06)	113.7*	(32.8–394.8)	73.6	(9.1)	97.6	(1.1)	86.6	(5.5)	94.6	(2.3)

Table 1. Continued

	McNemar χ^2_1 Test	AUC	κ	(se)	OR	(95% CI)	SN	(se)	SP	(se)	PPV	(se)	NPV	(se)
IV. Other disorders														
Adult ADHD[5]	3.5	0.861	0.49	(0.16)	60.2	(14.7–246.6)	77.7	(6.9)	94.5	(3.1)	39.6	(9.2)	98.9	(0.6)
V. Any disorder[5]														
Any disorder	21.1*	0.76	0.52	(0.05)	13.6*	(7.3–25.4)	62.8	(4.4)	89.0	(2.1)	84.3	(3.2)	71.7	(4.3)

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

[1] OR: odds ratio; AUC: area under the ROC curve; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

[2] Results for bipolar disorder are based on a separate clinical reappraisal sample of 40 respondents (Kessler et al., under review).

[3] Substance abuse was diagnosed in both the CIDI and SCID with or without dependence. The CIDI assessment of substance dependence was made only among respondents who met criteria for abuse based on the finding in an early study that the prevalence of dependence without abuse is very uncommon (Kessler et al., 1998). However, this result has recently been called into question (Hasin and Grant, 2004), leading subsequent WMH surveys to remove this restriction and to assess substance dependence even among respondents who failed to report a history of abuse.

[4] In light of the underestimation of prevalence of alcohol and drug dependence in the CIDI compared to the SCID, a revised coding scheme was used in which CIDI diagnoses of abuse were used to predict SCID diagnoses of dependence with abuse. As shown in the body of the table, AUC increased meaningfully with this revised scoring approach.

[5] Results for adult ADHD are based on a separate clinical reappraisal sample of 154 respondents (Kessler et al., 2006).

[6] The disorders considered here are only the ones included in the main clinical reappraisal sample of 325 respondents. Bipolar disorder and adult ADHD are not included.

cantly predicted the parallel SCID diagnosis after including the CIDI diagnosis in the prediction equation. Each respondent in the clinical reappraisal sample was then assigned a predicted probability of each SCID diagnosis based on the resulting logistic regression equation as a way to summarize the predictive information contained in the CIDI diagnostic and symptom data. The AUC for the predicted probability is consistently higher than the AUC for the dichotomous CIDI diagnostic classification in predicting each of the SCID diagnoses. (Table 2) This improved prediction is due to the fact that the CIDI collects a substantial amount of information from respondents who endorse diagnostic stem questions but fail to meet full diagnostic criteria for DSM disorders. This information was used in the prediction equations to adjust for the consistently higher diagnostic thresholds in the CIDI than the SCID. When the CIDI data are transformed through these equations into predicted probabilities of SCID diagnoses, CIDI-SCID concordance changes from largely moderate (AUC in the range 0.7–0.8) to largely substantial (AUC in the range 0.8–0.9) or almost perfect (AUC in the range 0.9–1.0). Bias in prevalence estimates is also removed when predicted probabilities of SCID diagnoses are used instead of dichotomous CIDI disorder classifications.

Table 2. Area under the ROC curve (AUC) for dichotomous (DICH) DSM-IV CIDI diagnostic classifications and continuous (CONT) CIDI-based predicted probabilities in predicting lifetime DSM-IV/SCID diagnoses in the NCS-R clinical reappraisal sample (n = 325)

	DICH[1]	CONT[1]
I. Anxiety disorders		
Panic disorder	0.72	0.93
Agoraphobia	0.81	0.96
Specific phobia	0.67	0.84
Social phobia	0.65	0.74
Post-traumatic stress disorder	0.69	0.88
II. Mood disorders		
Major depressive disorder	0.75	0.87
Bipolar I/II[2]	0.93	0.97
III. Substance disorders		
Revised alcohol dependence[3]	0.72	0.99
Revised drug dependence[3]	0.62	0.95
IV. Other disorders		
Adult ADHD[4]	0.86	0.99

[1] DICH = AUC values for the dichotomous (DICH) DSM-IV CIDI diagnostic classifications; CONT = AUC values for continuous CIDI-based predicted probabilities of SCID diagnoses derived from logistic regression equations.

[2] Results for bipolar disorder are based on a separate clinical reappraisal sample of 40 respondents (Kessler et al., under review).

[3] In light of the under-estimation of prevalence of alcohol and drug dependence in the CIDI compared to the SCID, a revised coding scheme was used in which CIDI diagnoses of abuse were used to predict SCID diagnoses of dependence with abuse. As shown in the body of the table, AUC increased meaningfully with this revised scoring approach.

[4] Results for adult ADHD are based on a separate clinical reappraisal sample of 154 respondents (Kessler et al., 2006).

Twelve-month aggregate concordance

Disorder-specific CIDI-SCID comparisons of 12-month prevalence were made in ESEMeD for most of the same disorders assessed in the NCS-R. The one exception in the case of anxiety disorders was that generalized anxiety disorder (GAD) was assessed in the ESEMeD but not the NCS-R reappraisal interviews (due to the fact that the SCID includes GAD only as a current diagnosis, not a lifetime diagnosis). The two exceptions in the case of mood disorders were that dysthymic disorder was assessed in ESEMeD but not the NCS-R and that bipolar disorder was assessed in the NCS-R but not ESEMeD. In the case of substance disorders, ESEMed assessed only alcohol abuse dependence, not illicit drug abuse-dependence. Finally, the separate assessments of adult ADHD and bipolar disorder in the NCS-R was not repeated in the ESEMeD assessments.

Because of the narrower time frame of assessment in the ESEMeD (12-months) than NCS-R (lifetime) reappraisal interviews, the number of respondents with individual CIDI disorders was much smaller in the ESEMeD than NCS-R reappraisal samples. As a result, CIDI-SCID 12-month diagnostic consistency was assessed by focusing on summary measures of any anxiety disorder, any mood disorder, and any overall disorder. Any alcohol disorder could not be assessed separately because of too few cases (n = 3). As noted in the section on the clinical reappraisal samples, logistical complications led to many of the ESEMeD clinical reappraisal interviews being carried out as much as six months after the initial CIDI interviews, resulting in the overlap in the recall period of 12-month prevalence estimates varying substantially across respondents. This variation was taken into account by carrying out the analyses separately for respondents with a time lag between CIDI and SCID interviews less than two months, between two and four months, and more than four months. Concordance was found to increase monotonically across these three subsamples as time between the two interviews decreased. As a result, we focus here on results for the subsample with a length of time between interviews less than two months (Table 3). Focusing first on aggregate concordance, McNemar tests for CIDI versus SCID differences in estimated 12-month prevalence are insignificant for all four summary measures in this subsample.

Twelve-month individual-level concordance

Individual-level CIDI-SCID concordance for 12-month prevalence is substantial (AUC in the range 0.8–0.9) for any mood disorder, any anxiety disorder and any overall disorder. Within-country estimates of AUC (results not shown in table, but available on request) have substantial consistency for any anxiety disorder (0.83–0.94) and any mood disorder (0.83–0.84), but more variability for any disorder (0.78–0.93).

The majority of 12-month SCID cases were detected by the CIDI (SN) for any anxiety (83.7%), any mood (69.1%) and any overall disorder (77.9%). Lower proportions of CIDI cases were confirmed by the SCID (PPV) for 12-month estimates in the ESEMeD surveys than for lifetime estimates in the NCS-R, including 31.3% with anxiety disorder, 49.6% with mood disorder and 41.5% with any disorder.

Twelve-month concordance using CIDI symptom-level data

The AUC was found to increase substantially in the case of mood disorders when CIDI symptom data were added to equations that included the dichotomous CIDI diagnosis to predict the 12 month SCID diagnosis (from 0.83 to 0.93). The AUC increased more modestly in predicting 12-month SCID anxiety disorders (from 0.88 to 0.91).

Summary

Several limitations of the current report should be taken into account when interpreting the results. First, most of the SCID reinterviews were carried out over the telephone, while initial CIDI interviews were face-to-face. It has been shown that telephone interviews

Table 3. Consistency of twelve-month DSM-IV CIDI and SCID diagnoses in the ESEMeD clinical reappraisal sample with time between CIDI and SCID interviews less than 60 days (n = 143)

	McNemar χ^2_1 Test	AUC	κ	(se)	OR	(95% CI)	SN	(se)	SP	(se)	PPV	(se)	NPV	(se)
Any anxiety disorder	6.9	0.88	0.42	(0.1)	66.7	(18.2–244.2)	83.7	(8.3)	92.9	(1.7)	31.3	(7.8)	99.3	(0.4)
Any mood disorder	0.8	0.83	0.56	(0.2)	76.7	(21.4–274.9)	69.1	(11.8)	97.2	(0.9)	49.6	(10.5)	98.7	(0.6)
Any disorder	6.0	0.84	0.49	(0.1)	34.1	(6.6–176.8)	77.9	(13.2)	90.6	(2.7)	41.5	(8.5)	98.0	(1.5)

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

[1] OR: odds ratio; AUC: area under the ROC curve; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

constitute a valid mode of clinical assessment (Kendler et al., 1992; Sobin et al., 1993; Rohde et al., 1997) but we do not know what would have happened if the same mode of administration were to have been consistently employed in both cases. Second, assessment of lifetime and 12-month reliability was conducted in two different countries, making it difficult to compare the two sets of results. Third, the investigation of 12-month concordance found strong inverse associations of CIDI-SCID concordance with time between interviews, leading us to focus on the subset of respondents in which the two interviews were carried out less than two months apart. This restriction of the sample reduced statistical power, requiring us to examine 12-month concordance only for classes of disorder rather than for individual disorders. It is conceivable that a larger sample would have shown CIDI-SCID concordance to be even higher among respondents who completed both interviews within a period of only a few days or week.

Although not a limitation, it should be noted in interpreting the results reported here that the evaluation of clinically relevant information in epidemiological studies includes more than the simple investigation of prevalence (Brugha, 2002). This is true in two ways. First, given that the population prevalence of mental disorders far outstrips available treatment resources, mental health policy decision makers have proposed several more restrictive definitions based on severity and impact that can be used to narrow the number of people qualifying for treatment (Regier, 2000). Categorical measures of this sort are included in the WMH surveys (Demyttenaere et al., 2004) but were not considered in the current report. Second, dimensional measures of clinical severity are widely used in treatment studies, and need to be included as well in epidemiological studies if we want to make the results of the latter relevant to clinicians (Kessler and Ustun, 2004). Fully structured versions of standard clinical severity scales are included in the WMH surveys, such as the Inventory of Depressive Symptomatology (Rush et al., 1996) and the Panic Disorder Severity Scale (Shear et al., 1997) to assess the severity of individual disorders. In addition, the WHO Disability Assessment Schedule (WHO-DAS) (Rehm et al., 1999) is included in the WMH surveys to assess the severity of overall psychopathology. These dimensional measures were not considered in the current report.

Based on the additional measures described in the last paragraph, the assessment of clinical significance in the WMH surveys does not hinge entirely on concordance between the categorical DSM-IV diagnoses based on the CIDI and those based on the SCID. Nonetheless, information on CIDI-SCID diagnostic concordance is useful in determining whether the diagnostic thresholds and DSM-IV diagnostic criteria disorders are defined in a consistent way in the CIDI and SCID. We have seen that the CIDI diagnostic thresholds for lifetime prevalence of DSM-IV disorders are generally somewhat more conservative than those of the SCID, at least in the US, whereas diagnostic thresholds for 12-month prevalence are generally unbiased, at least in the three Western European countries considered here. We also saw that individual-level diagnostic concordance is generally good when we use the CIDI to make categorical diagnostic classifications and very good when we develop CIDI-based dimensional probability-of-disorder measures.

Although the word *validation* is often used to characterize the kind of investigation described in the last two paragraphs, this is not an entirely accurate term due to the fact that the SCID diagnoses cannot be taken as perfect representations of DSM diagnoses. This is true both because the test-retest reliability of the SCID is far from perfect (Segal et al., 1994), especially in community samples (Williams et al., 1992) and because some respondents in community surveys consciously hide information about their mental or substance problems from clinical interviewers (Kranzler et al., 1997). Based on these considerations, the estimates of CIDI-SCID concordance should be considered lower bound estimates of CIDI validity. A good illustration can be found in the work of Booth et al. (1998), who compared lifetime diagnoses of major depression based on an earlier version of CIDI with diagnoses based on SCID clinical reappraisal interviews, where κ was 0.53. However, when the CIDI was compared with more accurate LEAD standard diagnoses (Spitzer, 1983), which used not only the SCID but also all the clinical information available to arrive at an improved estimate of clinical diagnoses, κ increased to 0.67.

The difference in aggregate concordance between CIDI and SCID prevalence estimates is another finding that warrants comment. As noted above, CIDI lifetime prevalence estimates are generally conservative compared to SCID estimates, whereas CIDI 12-month prevalence estimates are unbiased compared to SCID estimates. It is noteworthy in this regard that the CIDI first assesses lifetime prevalence of each disorder by asking the respondent to concentrate on the worst lifetime episode of the disorder. Lifetime assessment is based on that worst episode. Twelve-month prevalence is then assessed in the CIDI with a single question that asks about the last time the individual experienced an episode similar to the worst one. The 12-month SCID, in comparison, carries out a detailed assessment of all symptoms present in the past 12 months. The most plausible interpretation of the discrepancy between CIDI versus SCID lifetime and 12-month prevalence estimates based on these instrument differences is that the CIDI probably underestimates lifetime prevalence because its diagnostic thresholds are too high: while it overestimated 12-month prevalence among the lifetime cases it does not assess all required 12-month symptoms for the diagnosis. The global consequence is that 12-month CIDI prevalence estimates appear to be unbiased because the downward bias in lifetime prevalence estimates and downward bias in condition 12-month prevalence estimates among lifetime cases cancel each other out.

As mentioned earlier in the paper, in order to overcome the CIDI limitations described in the last paragraph, Version 3.0 of the CIDI includes a number of questions about 12-month clinical severity that were absent from earlier versions. These clinical questions assess disorder severity not only for respondents who meet full CIDI lifetime criteria for the disorder but also for subthreshold cases with 12-month symptoms, allowing correction of prevalence estimates in both time frames by decreasing diagnostic thresholds for lifetime prevalence and increasing clinical severity requirements for 12-month prevalence. A rough sense of the extent to which these recalibration exercises, which are only now beginning to be carried out in the WMH clinical reappraisal samples, might be able to improve CIDI diagnostic validity is provided by examining the increases in AUC associated with using regression-based CIDI symptom scoring rather than categorical diagnostic scoring to predict SCID diagnoses. This new work aims to refine CIDI diagnoses to correct the problem of lifetime thresholds being too high and 12-month thresholds among lifetime cases being too low.

In conclusion, the WMH clinical reappraisal studies have shown that CIDI-SCID agreement in DSM-IV diagnoses is generally good, that CIDI lifetime prevalence estimates are generally conservative relative to SCID estimates, and that CIDI 12-month prevalence estimates are generally unbiased relative to SCID estimates. The estimates of concordance probably underestimate true CIDI validity due to the fact that SCID diagnoses, which were implicitly taken as a gold standard, are in fact known to be imperfect. Finally, the inclusion of subthreshold assessments and 12-month scales of clinical severity provide enough information to improve CIDI-SCID concordance based on the results of currently ongoing methodological studies.

Acknowledgements

The ESEMeD project is funded by the European Commission (Contracts QLG5-1999-01042; SANCO 2004123), the Piedmont Region (Italy), Fondo de Investigación Sanitaria, Instituto de Salud Carlos III, Spain (FIS 00/0028), Ministerio de Ciencia y Tecnología, Spain (SAF 2000-158-CE), Departament de Salut, Generalitat de Catalunya, Spain, and other local agencies and by an unrestricted educational grant from GlaxoSmithKline. The US National Comorbidity Survey Replication (NCS-R) is supported by the National Institute of Mental Health (NIMH; U01-MH60220) with supplemental support from the National Institute of Drug Abuse (NIDA), the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044780) and the John W. Alden Trust. These surveys were carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We thank the WMH staff for assistance with instrumentation, fieldwork and data analysis. These activities were supported by the US National Institute of Mental Health (R01MH070884), 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 & Company, Ortho-McNeil Pharmaceutical, Inc., GlaxoSmithKline, and Bristol-Myers Squibb. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/.

References

Agresti A (1996) An Introduction to Categorical Data Analysis. New York: John Wiley & Sons.

Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha TS, Bryson H, de Girolamo G, Graaf R, Demyttenaere K, Gasquet I, Haro JM, Katz SJ, Kessler RC, Kovess V, Lepine JP, Ormel J, Polidori G, Russo LJ, Vilagut G, Almansa J, Arbabzadeh-Bouchez S, Autonell J, Bernal M, Buist-Bouwman MA, Codony M, Domingo-Salvany A, Ferrer M, Joo SS, Martinez-Alonso M, Matschinger H, Mazzi F, Morgan Z, Morosini P, Palacin C, Romera B,

DOI: 10.1002/mpr

Taub N, Vollebergh WA (2004) Sampling and methods of the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatr Scand Suppl 420: 8–20.

Andrews G, Peters L, Guzman AM, Bird K (1995) A comparison of two structured diagnostic interviews: CIDI and SCAN. Aust N Z J Psychiatry 29: 124–32.

Booth BM, Kirchner JE, Hamilton G, Harrell R, Smith GR (1998) Diagnosing depression in the medically ill: validity of a lay-administered structured diagnostic interview. J Psychiatr Res 32: 353–60.

Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC (1986) Long-term reliability of diagnosing lifetime major depression in a community sample. Arch Gen Psychiatry 43: 435–40.

Brugha TS (2002) The end of the beginning: a requiem for the categorization of mental disorder? Psychol Med 32: 1149–54.

Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE (2001) A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Psychol Med 31: 1001–13.

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. J Clin Epidemiol 46: 423–9.

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

Cook RJ (1998) Kappa and its dependence on marginal rates. In Armitage P, Colton T (eds) The Encyclopedia of Biostatistics. New York: Wiley, pp. 2166–8.

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 (2004) Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 291: 2581–90.

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) New York: Biometrics Research, New York State Psychiatric Institute.

Gibbon M, McDonald-Scott P, Endicott J (1981) Mastering the art of research interviewing. A model training procedure for diagnostic evaluation. Arch Gen Psychiatry 38: 1259–62.

Hasin DS, Grant BF (2004) The co-occurrence of DSM – IV alcohol dependence: results of the National Epidemiologic survey on Alcohol and Related Conditions on heterogeneity that differ by population subgroup Arch Gen Psychiatry 61: 891–96.

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

Jordanova V, Wickramesinghe C, Gerada C, Prince M (2004) Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. Psychol Med 34: 1013–24.

Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ (1992) A population-based twin study of major depression in women. The impact of varying definitions of illness. Arch Gen Psychiatry 49: 257–66.

Kessler RC, Abelson J, Demler O, Escobar JI, Gibbon M, Guyer ME, Howes MJ, Jin R, Vega WA, Walters EE, Wang P, Zaslavsky A, Zheng H (2004a) Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI) Int J Methods Psychiatr Res 13: 122–39.

Kessler RC, Akiskal HS, Angst J, Guyer M, Hirschfeld RM, Merikangas KR, Stang PE (in press). Validlity of the assessment of bipolar spectrum disorders in the WHO CIDI 30 J Affect Disord.

Kessler RC, Adler L, Barkley R, Biederman J, Conners CK, Demler O, Faraone SV, Greenhill LL, 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 163: 716–23.

Kessler RC, Berglund P, Chiu WT, Demler O, Heeringa S, Hiripi E, Jin R, Pennell BE, Walters EE, Zaslavsky A, Zheng H (2004b) The US National Comorbidity Survey Replication (NCS-R): design and field procedures. Int J Methods Psychiatr Res 13: 69–92.

Kessler RC, Birnbaum H, Demler O, Falloon IR, Gagnon E, Guyer M, Howes MJ, Kendler KS, Shi L, Walters E, Wu EQ (2005) The prevalence and correlates of nonaffective psychosis in the National Comorbidity Survey Replication (NCS-R). Biol Psychiatry 58: 668–76.

Kessler RC, Merikangas KR (2004) The National Comorbidity Survey Replication (NCS-R): background and aims. Int J Methods Psychiatr Res 13: 60–8.

Kessler RC, Ustun TB (2004) The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 13: 93–121.

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. Int J Methods Psychiatr Res 7: 33–55.

DOI: 10.1002/mpr

Kish L, Frankel MR (1974) Inferences from complex samples. J Roy Stat Soc 36: 1–37.

Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ (2003) Measures of clinical significance. J Am Acad Child Adolesc Psychiatry 42: 1524–9.

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

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

Lenzenweger MF, Lane MC, Loranger AW, Kessler RC (in press) DSM-IV personality disorders in the National Comorbidity Survey Replication. Biol Psychiatry.

Regier DA (2000) Community diagnosis counts. Arch Gen Psychiatry 57: 223–4.

Rehm J, Ustun TB, Saxena S, Nelson CB, Chatterji S, Ivis F, Adlaf E (1999) On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. Int J Methods Psychiatr Res 8: 110–23.

Rohde P, Lewinsohn PM, Seeley JR (1997) Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. Am J Psychiatry 154: 1593–8.

Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

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

Segal DL, Hersen M, Van Hasselt VB (1994) Reliability of the Structured Clinical Interview for DSM-III-R: an evaluative review. Compr Psychiatry 35: 316–27.

Shear MK, Brown TA, Barlow DH, Money R, Sholomskas DE, Woods SW, Gorman JM, Papp LA (1997) Multicenter collaborative panic disorder severity scale. Am J Psychiatry 154: 1571–5.

Sobin C, Weissman MM, Goldstein RB, Adams P, Wickramaratne PJ, Warner V, Lisch JD (1993) Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. Psychiatr Genet 3: 227–34.

Spitzer RL (1983) Psychiatric diagnosis: are clinicians still necessary? Compr Psychiatry 24: 399–411.

Spitznagel EL, Helzer JE (1985) A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 42: 725–28.

Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HG, Jr., Rounsaville B, Wittcher HU. [(1992) The Structured Clinical Interview for DSM-III-R (SCID) II. Multisite test-retest reliability. Arch Gen Psychiatry 49: 630–6.

Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablenski A, Regier D, Sartorius N (1990) SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch Gen Psychiatry 47: 589–93.

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

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

Wittchen HU, Zhao S, Abelson JM, Abelson JL, Kessler RC (1996) Reliability and procedural validity of UM-CIDI DSM-III-R phobic disorders. Psychol Med 26: 1169–77.

Wolter K (1985) Introduction to Variance Estimation. New York: Springer-Verlag.

Correspondence: JM Haro, Fundació Sant Joan de Déu per la Recerca i la Docència
Carrer Santa Rosa, 39-57, 08950 – Esplugues de Llobregat (Barcelona), Spain.
Telephone (+34) 93 600 97 51.
Fax (+34) 93 600 97 71.
Email: jmharo@fsjd.org.

DOI: 10.1002/mpr