Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMH-CIDI)

RONALD C. KESSLER,[1] JAMIE ABELSON,[2] OLGA DEMLER,[1] JAVIER I. ESCOBAR,[3] MIRIAM GIBBON,[4] MARGARET E. GUYER,[1] MARY J. HOWES,[1] ROBERT JIN,[1] WILLIAM A. VEGA,[3] ELLEN E. WALTERS,[1] PHILIP WANG,[1] ALAN ZASLAVSKY,[1] HUI ZHENG[1]

[1] Department of Health Care Policy, Harvard Medical School, Boston MA, USA
[2] Institute for Social Research, University of Michigan, Ann Arbor MI, USA
[3] University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, New Brunswick NJ, USA
[4] New York State Psychiatric Institute, New York City NY, USA

ABSTRACT *An overview is presented of the rationale, design, and analysis plan for the WMH-CIDI clinical calibration studies. As no clinical gold standard assessment is available for the DSM-IV disorders assessed in the WMH-CIDI, we adopted the goal of calibration rather than validation; that is, we asked whether WMH-CIDI diagnoses are 'consistent' with diagnoses based on a state-of-the-art clinical research diagnostic interview (SCID; Structured Clinical Interview for DSM-IV) rather than whether they are 'correct'. Consistency is evaluated both at the aggregate level (consistency of WMH-CIDI and SCID prevalence estimates) and at the individual level (consistency of WMH-CIDI and SCID diagnostic classifications). Although conventional statistics (sensitivity, specificity, Cohen's* κ*) are used to describe diagnostic consistency, an argument is made for considering the area under the receiver operator curve (AUC) to be a more useful general-purpose measure of consistency. In addition, more detailed analyses are used to evaluate consistency on a substantive level. These analyses begin by estimating prediction equations in a clinical calibration subsample, with WMH-CIDI symptom-level data used to predict SCID diagnoses, and using the coefficients from these equations to assign predicted probabilities of SCID diagnoses to each respondent in the remainder of the sample. Substantive analyses then investigate whether estimates of prevalence and associations when based on WMH-CIDI diagnoses are consistent with those based on predicted SCID diagnoses. Multiple imputation is used to adjust estimated standard errors for the imprecision introduced by SCID diagnoses being imputed under a model rather than measured directly. A brief illustration of this approach is presented in comparing the precision of SCID and predicted SCID estimates of prevalence and correlates under varying sample designs.*

Key words: clinical calibration, concordance, epidemiologic research design, reliability, validity

Introduction

This paper discusses the rationale, design, and analysis plan for the clinical calibration studies carried out in conjunction with the US National Comorbidity Survey Replication (NCS-R). Similar procedures are being used in clinical calibrations in two other US surveys that are being carried out in conjunction with the NCS-R – the National Latino and Asian American Survey (NLAAS) and the National Survey of American Life (NSAL) – as well

as in a number of surveys participating in the World Health Organization's (WHO) World Mental Health (WMH) Survey Initiative (Kessler, 1999; Kessler and Üstün, 2000). These studies are being conducted in order to increase the clinical relevance of the survey results by calibrating the American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnoses from the fully structured WMH-CIDI interviews to independent clinical diagnoses from semi-structured research diagnostic interviews based on the Structured Clinical Interview for DSM-IV (SCID) (First, Spitzer, Gibbon and Williams, 2002).

As detailed in the body of the paper, the calibration includes three phases. The first two phases compare WMH-CIDI diagnoses with independent clinical diagnoses. The first is an aggregate investigation of bias in prevalence estimates that includes an initial estimation aimed at characterizing the performance of the WMH-CIDI, followed by an evaluation aimed at determining whether this performance could be improved by substantively reasonable modifications of the WMH-CIDI diagnostic algorithm. The second phase is an individual-level investigation of concordance. The third phase consists of the development of prediction equations in the clinical calibration subsamples in order to generate predicted probabilities of clinical diagnoses from WMH-CIDI data for each respondent in each of the three surveys. As detailed in the body of the paper, these predicted probabilities can be treated directly as outcomes in substantive analyses or can be used as the 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 WMH-CIDI data. Comparison with parallel estimates of the prevalence and correlates of WMH-CIDI diagnoses allows direct evaluations of consistency that are much more meaningful than conventional analyses of diagnostic concordance.

The clinical relevance of prevalence estimates in epidemiological surveys

The prevalence estimates in the Epidemiologic Catchment Area (ECA) study (Robins and Regier, 1991) and subsequently in the National Comorbidity Survey (NCS) (Kessler, McGonagle, Zhao, Nelson, Hughes, Eshleman, Wittchen and Kendler, 1994) were considerably higher than expected by most health policy analysts. Approximately 30% of respondents in the 18–54 age range met criteria for one or more of the DSM-III (ECA) or DSM-III-R (NCS) disorders assessed in those surveys in the 12 months before interview (Regier, Kaelber, Rae, Farmer, Knauper, Kessler and Norquist, 1998). Furthermore, this is a lower bound estimate, as neither of the surveys included the full range of DSM disorders in their assessments and even this range of disorders is likely to be non-inclusive (Angold, Costello, Farmer, Burns and Erkanli, 1999; Pincus, Davis and McQueen, 1999).

The initial reaction to these results among health policy analysts was one of disbelief. The most obvious interpretation was that the lay administered diagnostic interviews – the Diagnostic Interview Schedule (DIS) in the ECA (Robins, Helzer, Croughan and Ratcliff, 1981) and a modified version of the Composite International Diagnostic Interview (CIDI) in the NCS (Robins, Wing, Wittchen, Helzer, Babor, Burke, Farmer, Jablenski, Pickens, Regier et al., 1988) – were upwardly biased. However, clinical calibration studies carried out in conjunction with both surveys showed that the DIS and CIDI prevalence estimates were not any higher than those found in blind semi-structured clinical interviews (Kessler, Wittchen, Abelson, McGonagle, Schwarz, Kendler, Knauper and Zhao, 1998; Eaton, Neufeld, Chen and Cai, 2000). This led critics to conclude that the DSM system itself is overly inclusive (Pincus, Zarin and First, 1998; Regier et al., 1998; Üstün, Chatterji and Rehm, 1998). This conclusion was instrumental in causing an APA task force to add a clinical significance criterion to many disorders in the DSM-IV in order to remind readers of the basic definition of a mental disorder in the introduction of the manual.

The NCS-R, NLAAS, and NSAL are the first major general population surveys in the US to administer a modified version of the CIDI that includes questions designed to operationalize the DSM-IV criteria. An earlier *post hoc* analysis of the ECA and NCS data attempted to produce provisional DSM-IV prevalence estimates by making use of questions in the ECA and NCS that were relevant to the new criteria (Narrow, Rae, Robins and Regier, 2002). However, as the ECA and NCS were not designed with DSM-IV criteria in mind, the new criteria were not well operationalized in those

surveys, raising doubts about the accuracy of the provisional prevalence estimates.

Given that the WMH-CIDI, the version of the CIDI used in the NCS-R, NLAAS, NSAL, and the WHO World Mental Health (WMH) surveys, is a new version of CIDI, it is important to check its consistency with diagnoses based on clinical reinterviews. Confirmation of consistency with clinical diagnoses is needed before accepting the new DSM-IV prevalence estimates as accurate. Based on this realization, a clinical calibration phase was built into all three US surveys as well as into WMH surveys in a number of other countries (China, France, India, Italy, Lebanon, Nigeria, South Africa, Spain). The methods used in these clinical calibration studies were co-ordinated across all surveys and countries in order to facilitate comparative analysis.

Before turning to a discussion of these co-ordinated methods, it should be noted that the discussion of clinical relevance has now advanced beyond a simple consideration of prevalence. Specifically, in recognition of the fact that the true population prevalence of mental disorders, when it is finally established, will almost certainly turn out to exceed the resources available to treat these disorders, mental health policy advocates have proposed several more restrictive definitions that can be used to narrow the number of people qualifying for treatment. The National Institute of Mental Health (NIMH) National Advisory Mental Health Council, for example, distinguished people with Severe and Persistent Mental Illness (SPMI) from other mentally ill people (National Advisory Mental Health Council, 1993), while the Alcohol, Drug Abuse, and Mental Health Administration Reorganization Act stipulated that state mental health block grant funds can be used only to treat people with serious mental illness (Substance Abuse and Mental Health Services Administration, 1993). Many health plans in the US have followed suit by restricting mental health coverage to a subset of DSM disorders that they consider to be 'biologically based'.

In terms of research, the suggestion has been made that epidemiological surveys should go beyond simple dichotomous diagnostic distinctions to include additional dimensional measures of clinical severity (Regier, 2000). As noted by Kessler and Üstün (2004), such measures are included in the WMH-CIDI. Fully structured versions of standard clinical severity scales, such as the Inventory of Depressive Symptomatology (Rush, Gullion, Basco, Jarrett and Trivedi, 1996) and the Panic Severity Test (Houck, Spiegel, Shear and Rucci, 2002), are embedded in the WMH-CIDI to assess the severity of individual disorders. The WHO Disability Assessment Schedule (WHO-DAS) (Rehm, Üstün, Saxena, Nelson, Chatterji, Ivis and Adlaf, 1999) is included in the WMH-CIDI to assess the severity of overall psychopathology. As a result, the assessment of clinical significance in the WMH-CIDI does not hinge entirely on concordance of the categorical DSM-IV diagnoses generated by the WMH-CIDI with independent clinical diagnoses. The current paper, however, focuses on only this one aspect of clinical significance.

The conventional clinical calibration design

There is considerable uncertainty regarding the best design to use in validating fully structured diagnostic interviews like the WMH-CIDI. The standard design is the double-blind test-retest design in which a probability subsample of respondents in a community survey that over-samples respondents with CIDI diagnoses is reinterviewed by trained clinical interviewers who administer a gold standard semi-structured research diagnostic interview such as the SCID (First et al., 2002). The clinical interviewers are blind to the CIDI diagnoses. Assessment of concordance between the CIDI and clinical diagnoses is then made by evaluating the sensitivity and specificity of the CIDI and by calculating Cohen's κ (Cohen, 1960) to measure overall concordance. However, as noted by Robins (1985), this design is flawed in a number of ways.

First, and perhaps most obviously, the semi-structured clinical research diagnostic interviews that are used as the gold standard against which to validate lay-administered interviews have their own errors. Although psychometric data on the DSM-IV version of the SCID have not been published, a short-term test-retest reliability study of the DSM-III-R version of the instrument in two non-patient samples found quite low concordance over time, with a mean κ of 0.37 for current diagnoses and 0.51 for lifetime diagnoses (Williams, Gibbon, First, Spitzer, Davies, Borus, Howes, Kane, Harrison, Rounsaville and Wittchen, 1992). If classical test theory assump-

tions hold (Crocker and Algina, 1986), then the concordance between WMH-CIDI and SCID diagnoses can be no higher than these reliabilities. The low reliability of the SCID is more pronounced in community than clinical samples (Williams et al., 1992). This could be due to patients reporting more consistently over time than community cases, to clinical interviewers being more consistent in rating the presumably more severe symptoms of patients than community cases, or, depending on whether a measure of reliability is being used that is sensitive to prevalence, on different prevalence in clinical than community samples. Some insight into the first two of these possibilities can be gleaned from studies that tape record clinical interviews and have the tapes reviewed by a second clinical interviewer who re-rates all the symptoms. This is the typical method used to confirm the adequacy of the training of clinical interviewers to carry out CIDI validation studies. This exercise consistently finds very high levels of agreement among clinical interviewers in their ratings of a single tape (Wittchen, 1994). A confounding factor here is that the audiotape only includes responses to questions solicited by the first interviewer. The second interviewer does not have the opportunity to ask additional questions. Nonetheless, to the extent that the second interviewer feels comfortable making a diagnostic rating based on the questions asked by the first interviewer, the high concordance found in studies of this type implies that the low test-retest reliability of clinical interviews in community samples is due more to inconsistency over time on the part of respondents than to inconsistent ratings on the part of clinical interviewers.

Second, even though the low test-retest reliability of clinical interviews is a serious problem, there is an even greater problem in using semi-structured clinical interviews as a validation standard. The formal way of stating this problem is to say that the classical test theory assumption of inter-temporally uncorrelated measurement errors is unlikely to hold for long, emotionally draining interviews such as the CIDI or SCID. The reason is that respondents with complex histories of mental disorder are aware after the first interview that they can skip out of large portions of the reinterview simply by denying diagnostic stem questions. Indeed, debriefing of pilot test respondents in conjunction with the NCS showed that respondents with complex histories of mental disorder usually pick up on the stem-branch logic of the interview part way through and deny stem questions that they know they should endorse in latter sections in a conscious effort to shorten the interview (Kessler, Wittchen, Abelson and Zhao, 2000). This means that estimates of test-retest reliability will be biased. Furthermore, it means that second interviews will underestimate disorder prevalence significantly. Bromet et al. (1986) clearly demonstrated this last finding in a community survey that interviewed 391 women twice over an 18-month time interval with a semi-structured clinical research diagnostic interview that was a predecessor to the SCID. Test-retest reliability was poor and the prevalence estimate was significantly lower in the second than the first wave of data collection.

Third, moving beyond the issue of reliability to the deeper issue of validity, questions can be raised about the validity of semi-structured clinical interviews in operationalizing DSM criteria. The problem is that many DSM criteria are not sufficiently precise to be operationalized unequivocally even with exhaustive clinical assessments. For example, criterion A.1.(e) in the DSM-IV diagnosis of Attention-Deficit/Hyperactivity Disorder is that the patient 'often has difficulty organizing tasks and activities'. But how often is 'often'? What kinds and severities of organizational problems constitute 'difficulty'? Answers to these questions need to be more clearly specified in order to allow clinical interviewers to make accurate assessments. As the DSM system provides little guidance in this regard, the developers of semi-structured clinical interviews have created their own operational standards. Although this is done to improve consistency of clinical rating (reliability), it also has the potential to reduce validity to the extent that the developers of these interviews have different ideas than the developers of the DSM about operational characteristics. To the extent that discrepancies between CIDI and SCID diagnoses are due to operational decisions that go beyond the DSM (for example, the decision that the word 'often' means 'at least several times a week' or the decision that the word 'difficulty' means 'serious enough to be noticed as a problem by a teacher or a work supervisor'), the discrepancies have to be seen as due to lack of validity in the DSM system rather than in the WMH-CIDI.

Fourth, questions have been raised about the standard interpretation of the statistics most often used to describe concordance between CIDI diagnoses and blind clinical diagnoses. Cohen's κ is the main source of controversy (Byrt, Bishop and Carlin, 1993; Cook, 1998; Kraemer, Morgan, Leech, Gliner, Vaske and Harmon, 2003), as κ is dependent on prevalence and κ consequently is often low in situations where there appears to be high agreement between low-prevalence measures (Feinstein and Cicchetti, 1990). Furthermore, κ varies across populations that differ in prevalence even when the populations do not differ in sensitivity (the percentage of true cases correctly classified by the CIDI) or specificity (the percent of true non-cases correctly classified). As sensitivity and specificity are considered by many substantive researchers 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.

Alternatives to the conventional design

A number of approaches have been proposed to go beyond the conventional clinical calibration design in ways that address one or more of the problems described in the last section. None of them is perfect, but it is nonetheless useful to consider them here. The most simple of these approaches is to correct estimates of concordance based on the assumption that semi-structured clinical interviews are merely unreliable, but not invalid, and that measurement error is random. If this assumption is made, simple psychometric methods based on latent class models can be used both to correct estimates of overall concordance between CIDI diagnoses and clinical diagnoses and to estimate true prevalence (Garrett, Eaton and Zeger, 2002). However, the models on which these methods are based require the researcher to make identifying assumptions that are likely to be implausible for reasons described in the previous section of this paper (Faraone and Tsuang, 1994).

Another approach is to collect additional data to improve the accuracy of clinical assessments in order to generate diagnoses that might reasonably be taken as both reliable and valid. This is sometimes done by carrying out a second clinical interview aimed at reconciling discrepancies between diagnoses in the CIDI and in the initial clinical interview (Mannuzza, Fyer, Martin, Gallops, Endicott, Gorman, Liebowitz and Klein, 1989; Williams et al., 1992). It is also possible to augment the information collected in such reconciliation interviews with medical records, informant reports, and reviews of all available information by independent expert clinicians in order to arrive at the most accurate possible clinical classification (Fennig, Craig, Lavelle, Kovasznay and Bromet, 1994; Ramirez Basco, Bostic, Davies, Rush, Witte, Hendrickse and Barnett, 2000). The difficulty with this approach is the one mentioned in the last section in conjunction with the research of Bromet et al. (1986) – that community respondents tend to report less and less as we interview them more and more, leading to the biased perception that initial structured interviews overestimate prevalence compared to second clinical interviews. Added to this is the consistent finding that respondents in reconciliation interviews usually resolve discrepancies in favour of their most recent interview. Both of these findings argue against the likelihood of more extensive data collection efforts resulting in totally unbiased clinical classifications.

An interesting alternative approach is what might be called the mixed single-interview design. In this approach the researcher investigates the extent to which diagnostic decisions in a single interview differ depending on whether the information is obtained entirely from structured responses to CIDI questions or from open-ended responses to semi-structured clinical questions. Janca et al. (1992) carried out a study of this sort with an early version of the CIDI. Clinical interviewers either administered a CIDI or sat in the room with a lay interviewer who was administering a CIDI. The clinical interviewers filled out a DSM-III-R clinical checklist simultaneously with the CIDI administration. They could ask additional semi-structured questions whenever the CIDI response provided insufficient information to code the relevant symptom on the clinical checklist. Analysis consisted of comparing diagnoses based entirely on CIDI responses with those based on responses to the clinical checklist. Diagnostic concordance was quite good, averaging $\kappa = 0.78$ across diagnoses, with no systematic difference in prevalence estimates between diagnoses based on the CIDI versus on the clinical checklist. This result is consistent with the interpretation that the much greater discrepancy between CIDI diagnoses and clinical diagnoses, when the latter are

based on a second interview rsther than on a mixed single-interview, are due to community respondents becoming more reluctant to admit disorders in second interviews.

The ideal design to confirm this interpretation would be one in which random half-samples of respondents were assigned either to mixed single-interview or test-retest interview assessments. We are not aware of any study that has used this design. Nonetheless, the mixed single-interview approach is very useful even without this additional design element in shifting the underlying research question in a way that finesses the problem of not having a gold standard. In the conventional clinical calibration design, the research question is 'How valid is the CIDI?' In order to answer this question, the researcher is required to create a perfectly reliable and valid gold standard either explicitly (by carrying out a clinical assessment that he is willing to accept as perfectly reliable and valid) or implicitly (by using psychometric methods to estimate a latent variable model, which is assumed to create an unbiased representation of the disorder). In the mixed single-interview design, in comparison, the research question changes to: 'What difference does it make to survey results that lay interviewers administered fully-structured diagnostic interviews rather than clinical interviewers administered semi-structured diagnostic interviews?' This is a more modest question, of course, than the question asked in the conventional design. It is a more practical question too, as clinical decisions are made on the basis of clinical assessments that are not always perfectly reliable and valid. Furthermore, it is an answerable question, unlike the question asked in the conventional design, because it can be answered without having to create a perfectly reliable and valid gold standard.

A design that is indirectly related to the mixed single-interview design is the predictive validity design. In this approach, baseline diagnoses from CIDI interviews and clinical reinterviews are compared in relative predictive power for relevant outcome variables (for example, psychiatric hospitalization, onset of work disability for emotional problems, suicide attempt). As in the mixed single-interview design, the implicit question is: 'What difference does it make that diagnoses are based on fully structured versus semi-structured diagnostic interviews?' In this case, though, the question concerns the difference it makes to conclusions about effects of the disorders rather than about prevalence of the disorders. The limited amount of methodological research that has been carried out using the predictive validity design suggests that the associations of fully structured interviews with a range of clinical outcomes are nearly as strong as the associations of clinician assessments with the same outcomes (Helzer, Spitznagel and McEvoy, 1987; Robins, 1989). Unlike the mixed single-interview design, the predictive validity design can also be used to make some inferences about comparative reliability. Specifically, the relative predictive power of two types of diagnoses can be assumed equivalent to the relative strength of association between true unmeasured disorder and the two types of diagnoses if (1) true unmeasured disorder is the only systematic determinant of diagnoses in both types of diagnoses and (2) the effects of these diagnoses on the outcomes are totally due to their links with true unmeasured disorders. Needless to say, though, the plausibility of these assumptions can be called into question.

Alternative measures of concordance

As noted above, the fact that κ varies with prevalence even when sensitivity (SENS) and specificity (SPEC) are constant has been a concern to critics. These critics prefer to assess overall strength of association with measures that are a function of SENS and SPEC. The odds-ratio (OR) meets this requirement, as $OR = [SENS \times SPEC]/[(1 - SENS) \times (1 - SPEC)]$ (Agresti, 1996). However, the upper end of OR is unbounded, making it difficult to use 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. This means that Q does not tell us about actual prediction

accuracy. The area under the receiver operator characteristic curve (AUC) is a measure that resolves this problem. Although developed for quite a different purpose – to study the association between a continuous predictor and a dichotomous outcome – the AUC can also be used in the special case where the predictor is a dichotomy. In this special case, AUC is equal to (SENS + SPEC)/2. The 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 (Hanley and McNeil, 1982). As a result of this useful interpretation, we focus on AUC in our evaluation of diagnostic concordance between the WMH-CIDI and the SCID.

The WMH-CIDI clinical calibration design

Based on our evaluation of the alternatives to the conventional design reviewed in the last section, we were attracted to the mixed single-interview design as we planned the WMH-CIDI clinical calibration studies. However, further consideration showed that this design has four practical implementation problems in the context of a major community epidemiological survey that made us reject it as the design for the WMH-CIDI clinical calibration studies. First, as the clinical interviewer administers both the CIDI and the clinical interview together in the mixed single-interview design, the blindness that exists in the conventional design is broken and opportunities arise for bias in evaluation. It is possible to reduce this problem considerably by requiring the clinical interview to probe aggressively, but this then creates a second problem: that the interview becomes a good deal longer and much more conversational than the standard CIDI, possibly leading to changes in the response style of respondents due to differences in rapport and burden. A third problem is that full implementation of the mixed single-interview design would require a random subsample of survey respondents to be interviewed by clinical interviewers. There are daunting logistical and financial barriers to this in large-scale community epidemiological surveys, especially when the surveys are nationally representative. Fourth, the mixed single-interview design does not include any test of the possibility that the lay interviewers who administer the CIDI are performing less than optimally either in their use of the complex probes and feedback rules required for accurate administration of the interview or in their accuracy of data entry. In light of these four problems, we decided against the mixed single-interview design in evaluating the WMH-CIDI.

Despite this decision, we came away from our review of the conventional design and its alternatives with two conclusions: (1) that we would not be able to create a gold standard measure we could plausibly consider completely reliable and valid in a conventional double-blind test-retest design; and (2) that because of this first conclusion we should focus on the kinds of methodological questions that are addressed in the mixed single-interview and predictive validity designs. We wanted to do this, though, without actually using the mixed single-interview design. We decided to do this by using a modified version of the double-blind test-retest design based on an observation made in our earlier methodological studies of the CIDI in the NCS (Kessler et al., 1998), which was subsequently confirmed in a different context by other investigators (Lucas, Fisher, Piacentini, Zhang, Jensen, Shaffer, Dulcan, Schwab-Stone, Regier and Canino, 1999): that the drop-off in diagnostic reports in clinical reinterviews compared to initial fully structured diagnostic interviews is concentrated largely in diagnostic stem questions. This means that a method that reduces the underreporting of diagnostic stem questions in clinical reinterviews would resolve a large part of the drop-off problem. The resolution of this problem, in conjunction with a short enough time interval in a test-retest that clinical status could be assumed to remain unchanged, would make it possible to address the question in the mixed single-interview design: 'What difference does it make to survey results that we use lay interviewers to administer fully-structured diagnostic interviews rather than clinical interviewers to administer semi-structured diagnostic interviews?'

The modified double-blind test-retest method made two changes to clinical reinterviews. First, we unblinded the interviewers to whether the respondents endorsed diagnostic stem questions in the WMH-CIDI, but not to the final WMH-CIDI diagnoses. Second, we 'forced' respondents to endorse these stem questions in the clinical reinterviews. The partial unblinding of interviewers might be seen as introducing a bias, but it turns out that this is not

the case due to the fact that the majority of community survey respondents who endorse WMH-CIDI stem questions do not go on to meet full WMH-CIDI criteria for the associated disorder. The 'forcing' of stem question endorsement, in comparison, has a substantial effect on the completeness of respondent reports in clinical reinterviews. This is achieved by telling respondents at the beginning of their clinical reinterview that they will be asked some of the same questions as in their earlier interview. They are also told that this is 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. They are 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 WMH-CIDI are 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?' This introduction is substituted for the conventional stem-branch structure of the sections, in which diagnostic stem questions are asked and subsequent branch questions are asked only if the respondent endorses the stem questions. As is typical in clinical interviews, our clinical interviewers also have great flexibility in going back to a diagnostic section that was previously skipped if any information subsequently surfaces in the interview to suggest a positive response to the diagnostic stem question. Reinterview respondents can still deny that they reported a diagnostic stem question in the initial interview, but this is uncommon. We also rotate the order of administration of sections in the clinical interview to see if the order bias found in previous methodological studies of research diagnostic interviews (Jensen, Watanabe and Richters, 1999) can still be observed when we use this stem-forcing approach.

Clinical instrumentation, training and supervision

As noted in the introduction, the clinical interview used in the WMH-CIDI clinical calibration studies is the SCID (First et al., 2002). The version of the SCID used is a modified version of the 12-month Axis I research version, non-patient edition. An expanded version of the model programme created by the developers of the SCID (Gibbon, McDonald-Scott and Endicott, 1981) was used for interviewer training. The lifetime version is also used in the NCS-R. This programme featured the following three phases: (1) the use of the standard SCID training tapes and manuals, which take an average of approximately 30 hours of self-study; (2) 40 hours of in-person group training by experienced SCID trainers; and (3) ongoing quality-control monitoring throughout the field period. Quality control monitoring included clinical supervisor (JA, MG, FE) review of all hard copy completed SCID interviews, re-contact 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 (JE, MG, MH, PW), consultant review of a random subsample of interview audiotapes, and biweekly interviewer-supervisor meetings to prevent drift.

The first of the three training phases was constant across all the countries that participated in the WMH-CIDI clinical calibration study. However, training materials were all in English, which meant that clinical interviewers in non-English speaking countries 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 countries other than the US and to have the clinical supervisors in these studies specially trained by the US trainers (MH, MG). In addition, the US trainers provided an ongoing telephone and e-mail consultation service to clinical supervisors in these other countries throughout the field period in conjunction with their quality control monitoring.

In the US studies, the SCID interviews were administered by telephone and were audiotaped for future methodological study after obtaining permission from respondents. Telephone administration of SCID interviews is now widely accepted based on evidence of comparable validity to in-person administration (Kendler, Neale, Kessler, Heath and Eaves, 1992; Sobin, Weissman, Goldstein, Adams, Wickramaratne, Warner and Lish, 1993; Rohde and Seeley, 1997). A great advantage of telephone administration is that a centralized and closely supervised clinical interview staff can carry out the interviews throughout the country. A disadvantage is that the roughly 5% of people in the household population of the US without telephones cannot be

included in clinical calibration studies when interviews are done by telephone. We were willing to make this trade-off in the US because the proportion of respondents without a telephone was small. In countries with lower telephone penetration, though, the WMH researchers usually decided to carry out SCID reinterviews in person rather than by telephone in order to avoid sample bias. In these countries, the SCID subsamples were concentrated in selected cities or regions of the country in order to avoid the logistical complications of administering in-person clinical reinterviews throughout the country.

Sampling

The 12-month clinical calibration subsamples oversampled 12-month WMH-CIDI cases, but also included non-cases, while the lifetime clinical calibration subsample (carried out only in the NCS-R) oversampled lifetime WMH-CIDI cases. The exact methods of subsampling differed across surveys, but in each case disproportionate sampling was based on probability procedures so that the clinical calibration subsample could be weighted back to be representative of the total original sample. This reweighting took into consideration the sample design of the original surveys, including differential probability of selection in households depending on sample size and post-stratification. The clinical subsample sizes in the US surveys for 12-month clinical reappraisal were 360 in the NCS-R, 195 in the NLAAS, and 677 in the NSAL. Each of these clinical reappraisal studies evaluated DSM-IV Major Depressive Episode, Dysthymia, Generalized Anxiety Disorder, Panic Disorder, Phobia, Post-Traumatic Stress Disorder, Alcohol Abuse-Dependence, and Drug Abuse-Dependence. In addition, a separate lifetime clinical reappraisal study of the same diagnoses was carried out in the NCS-R (n = 328). The NCS-R also included separate clinical reappraisal studies of non-affective psychosis (n = 73) and adult attention-deficit/hyperactivity disorder (n = 154) as well as an ongoing study of bipolar spectrum disorder.

Investigating aggregate concordance

After weighting the data to be representative of the entire sample, the first question investigated in analysis of the clinical calibration data is whether the WMH-CIDI prevalence estimates are comparable to the SCID prevalence estimates. The McNemar test is used here to compare the significance of the difference between the number of +– and –+ cases in the 2 × 2 cross-classification of WMH-CIDI and SCID diagnoses. As with all significance tests in the clinical calibration subsamples, the McNemar tests are carried out using design-based estimation methods that adjust for the effects of weighting and clustering of the initial survey data as well as for the over-sampling of WMH-CIDI cases (Kish and Frankel, 1974; Wolter, 1985). Our previous work in the NCS, which used the same design, found that CIDI-SCID differences in DSM-III-R prevalence estimates were not statistically significant for the vast majority of the disorders assessed in the survey (Kessler et al., 1998). We do not yet know, though, whether the same result will hold for the DSM-IV diagnoses in the WMH-CIDI. As noted in the introduction, disaggregated analyses aimed at pinpointing symptoms responsible for diagnostic discrepancies are being carried out in cases where WMH-CIDI prevalence estimates differ significantly from SCID prevalence estimates. As the main difference between DSM-III-R and DSM-IV diagnostic criteria for most disorders is the inclusion of a new clinical significance criterion, we are especially interested in investigating the extent to which CIDI-SCID prevalence differences can be reduced by modifying the thresholds on the WMH-CIDI scales of clinically significant distress or impairment. If comparable WMH-CIDI modifications across disorders are found to reduce CIDI-SCID diagnostic discrepancies, these modifications will be implemented in the WMH-CIDI diagnostic algorithms.

Investigating individual-level concordance

Individual-level diagnostic concordance between the WMH-CIDI and the SCID is being evaluated with 2 × 2 cross-classifications of diagnoses based on the two interviews. Overall statistical significance is evaluated with design-based χ^2 tests. We do not aspire to estimate validity due to our belief, discussed in earlier sections of the paper, that the SCID does not represent a totally reliable and valid gold standard. Our main concern, instead, is with the extent to which WMH-CIDI diagnoses reproduce SCID aggregate prevalence estimates and individual-level diagnostic classifications. As a result, a number of descriptive measures of overall concordance are

being computed. As noted above, our preferred measure of overall classification accuracy is the AUC. However, we also compute Cohen's κ, OR, and Q for purposes of comparison to other studies. We take into consideration the fact that the upper bound of AUC is less than 1.0 because the SCID diagnoses are not perfectly reliable. More disaggregated descriptive measures are also computed from the perspective of the SCID as the gold standard, including sensitivity (the proportion of SCID cases who are detected in the WMH-CIDI), sensitivity (the proportion of SCID non-cases who are classified as non-cases in the WMH-CIDI), positive predictive value (PPV) the proportion of WMH-CIDI cases who are confirmed by the SCID), and negative predictive value (NPV) (the proportion of WMH-CIDI non-cases that are confirmed as non-cases by the SCID).

Calibration of WMH-CIDI data to SCID diagnoses

We are also estimating logistic regression equations in which SCID diagnoses are treated as dichotomous outcomes and WMH-CIDI symptom variables (not merely dichotomous WMH-CIDI diagnoses) are predictors. The goal is to see if WMH-CIDI symptom-level data can significantly improve the prediction of SCID diagnoses compared to an equation in which dichotomous WMH-CIDI diagnoses are the only predictors. Note that these predicted probabilities are extensions of the measures of PPV used in standard 2 × 2 tables of diagnostic concordance. The AUC is the descriptive statistic used to evaluate these improvements. As noted earlier in this paper, 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 on a logistic regression equation, 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 an initial dichotomous predictor (the WMH-CIDI dichotomous diagnostic classification).

Results from the NCS-R clinical calibration study show consistently that AUC increases significantly when WMH-CIDI symptom data are added to equations that include dichotomous WMH-CIDI diagnostic classifications. For example, concordance (AUC) between WMH-CIDI and SCID diagnoses of lifetime major depressive episode (MDE) in the NCS-R clinical calibration subsample increases from 0.77 to 0.86 when WMH-CIDI depression symptom measures are added as predictors to an equation that includes the dichotomous WMH-CIDI MDE diagnostic classification as the only predictor. This finding documents that the WMH-CIDI diagnostic algorithm is statistically suboptimal in using all the data in the WMH-CIDI to classify SCID cases.

Once these equations are estimated, the coefficients are used to impute predicted probabilities of SCID diagnoses for each survey respondent who is not in the clinical calibration subsample. This is actually done 10 different times by generating that number of pseudo replications of the clinical subsample. The reason for generating 10 separate estimates will be discussed shortly. For now, though, it is sufficient to note that a prevalence estimate of each SCID diagnosis in the total sample can be generated from these predicted probabilities. These prevalence estimates were unbiased so long as the clinical calibration subsample was weighted to adjust for the over-sampling of WMH-CIDI cases before estimating the imputation equations and so long as the imputation equations were saturated.

The standard error (SE) of these prevalence estimates can be obtained using conventional methods of double sampling (Shrout and Newman, 1989). This approach could be extended to the investigation of correlates of SCID disorders by calculating estimates of SCID prevalence and SEs in subgroups. However, such data could not be obtained for every subgroup of interest because of limitations in the size of the clinical calibration subsample. An alternative is to expand the initial imputation prediction equations to investigate whether coefficients are similar in subgroups (for example, whether WMH-CIDI symptom questions interact significantly with subgroup measures such as age or sex to predict SCID diagnoses). If no differences are found, or if such differences are built into the prediction equations, then significant differences in the predicted probabilities imply differences in the same direction in the SCID prevalence (although the predicted probabilities do not necessarily translate directly into estimates of the magnitude of the SCID prevalence differences).

This approach of expanding the imputation equations to adjust for all possible subgroup differences in interactions with WMH-CIDI measures will almost certainly be imperfect, especially because of the usual model specification strategy of excluding (setting to zero) model coefficients that cannot be estimated with adequate power from the data in the clinical calibration subsample. As a result, some bias will exist in using the predicted probabilities for subgroup comparisons of SCID prevalence. However, the greater the predictive power of the WMH-CIDI questions, the less the predicted probabilities will rely on the model and the less danger there will be that bias will affect important results. Furthermore, the assumption that non-significant or non-estimable coefficients are zero is often a scientifically plausible basis for making comparisons among subgroups and is quite efficient relative to an approach that requires estimating separate prediction models in each subgroup.

Several methods exist to analyse the correlates of SCID predicted probabilities. One is to treat the mean predicted probability as the variable of interest in linear or restricted linear (for example, Tobit) prediction equations. Another possibility, which is useful when a single SCID diagnosis is the focus of attention, is to reproduce the observational record for each respondent so that the sample is treated as having two times as many observations as it actually has. The pair of records for each respondent is then coded so that one is defined as having the SCID diagnosis and the other as not. The weights for these two records are then defined as the original sample weight multiplied by (p_s) for the SCID-positive record and ($1 - p_s$) for the SCID-negative record, where p_s is the mean predicted probability of the SCID diagnosis from the respondent's imputation equations. If the research question deals with comorbidity among a number of SCID diagnoses, though, this approach becomes quite cumbersome, as it requires assigning a weight to every possible combination of positive and negative diagnoses, thereby vastly expanding the dataset and rendering the variance estimation more problematical.

A more parsimonious alternative is to assign each respondent either to have or not have a given SCID diagnosis by using a random number generator from a binomial distribution that is defined by his mean predicted probability from the imputation equations. Multiple diagnoses can be imputed in the same way from the joint probability distribution. This approach generates unbiased and realistic representations of the predictions, although it also incorporates some noise due to the randomness of the imputations. This sort of single stochastic imputation can be very useful in allowing comparison of diverse survey results based on the use of either WMH-CIDI diagnoses or predicted SCID diagnoses. For example, along the same lines as the predictive validity studies mentioned earlier, we could investigate what difference it makes to use CIDI diagnoses rather than predicted SCID diagnoses as predictors in an equation aimed at evaluating the effects of mental disorders on role functioning. Alternatively, we could include both WMH-CIDI and predicted SCID diagnoses in the same prediction equation and formally evaluate their comparative effects in predicting role impairment (Murphy, Monson, Laird, Sobol and Leighton, 2000; Daskalakis, Laird and Murphy, 2002). It is also possible to create difference scores so that we include both measures of C_I and ($S_I - C_I$) as predictors in the same equation, where C_I is a dichotomous (0,1) of whether respondent i carries a particular WMH-CIDI diagnosis and S_I is a continuous measure (0 – 1) of the predicted probability of a SCID diagnosis for respondent i. This means that ($S_i - C_i$) is a continuous difference score ranging between –1 and 1 that defines the extent to which information about the respondent's predicted SCID diagnosis is not captured in his WMH-CIDI diagnosis. It is also possible to use ($S_i - C_i$) as an outcome variable to investigate whether there are any systematic determinants of differences in WMH-CIDI and predicted SCID diagnoses.

All the estimation methods described in the last three paragraphs are subject to the criticism that they do not readily facilitate proper estimation of the uncertainty of inference from the imputed data, as conventional analyses treat the imputed data as known rather than predicted from a model. The method of multiple imputation (MI) (Rubin, 1987) is designed to overcome this limitation by making use of the within-person variation in imputed values from the 10 separate imputations generated for each respondent for each SCID diagnosis. Specifically, MI combines information on between-person variance in mean imputations with information on within-person variance in individual imputations to

calculate SEs that account for the fact that the imputations are predicted rather than known. The operational implication is that we have to estimate each equation 10 separate times whenever we use MI. An MI aggregation method is then used to compute the best estimate of each coefficient by combining information about the mean of the coefficient across the 10 imputations with information about the standard deviation of the coefficient estimates across the replications.

In the extreme case where the WMH-CIDI is totally unrelated to a particular SCID diagnosis (AUC = 0.5), the only systematic information in the multiply imputed dataset will be the consistent 0.0 and 1.0 values in the subsample of respondents who were in the clinical calibration subsample. The expected value of predicted disorder prevalence for each respondent who was not in the clinical calibration subsample will be the SCID prevalence in the clinical calibration subsample. In a case of this sort, the MI predicted SCID prevalence estimate will be unbiased and the SE of the estimate will be equivalent to the design-based SE in the clinical calibration sub-sample. At the other extreme, where the WMH-CIDI perfectly predicts a particular SCID diagnosis (AUC = 1.0), the MI SE of the estimated SCID prevalence estimate will be equivalent to the design-based SE in the total sample. In more realistic cases in which AUC lies between 0.5 and 1.0, the MI SE will take into consideration both the size of the clinical calibration subsample and the strength of the association between the WMH-CIDI and the SCID. The situation is similar for higher-order statistics, with the exception that measures of association will be biased towards zero by lack of concordance between predicted and true SCID diagnoses.

An illustration: multiple imputation of SCID MDE in the NCS-R

An informative way to evaluate the MI approach is to compare the SEs of estimated SCID diagnoses to the expected SEs of diagnoses based on a hypothetical SCID survey. Differences in prevalence are not at issue here as we assume that the MI approach produces largely unbiased estimates. The issue, rather, is the precision with which prevalence is estimated in a CIDI survey with a SCID clinical calibration component that uses MI to impute predicted SCID diagnoses compared to a hypothetical survey that abandons the WMH-CIDI altogether and administers the SCID to all respondents. In carrying out this simulation we have to adjust for the fact that a widely dispersed nationally representative SCID survey would be much more expensive than a WMH-CIDI survey of the same size. Included here would be higher costs of hiring (needing to advertise for, screen, and hire a national clinical interview staff rather than use the existing lay interview staff of the Survey Research Center at the University of Michigan that carried out the US surveys), training (at least 2 weeks of clinical interviewer training compared to 1 week of lay interviewer training), salary (NCS-R clinical interviewers are paid $50/hour plus fringe benefits compared to an average of $13/hour with no fringe benefits for lay interviewers), supervision (both more intensive supervision of clinical than lay interviewers and much higher salaries of clinical supervisors than lay supervisors), and post-processing (laptop computer direct data entry of structured responses to lay interviews versus clinical supervisor review, coding, and key-punching of open-ended responses to clinical interviews). This means that, at a fixed cost, a WMH-CIDI survey would have a considerably larger sample size than a SCID survey. For purposes of the following simulation, we assumed that we could carry out a nationally representative SCID survey of approximately 3,000 cases for the same overall cost as a WMH-CIDI survey of approximately 10,000 cases (roughly the size of the NCS-R).

Size of SE in the MI approach depends on six variables: true prevalence, strength of association between the WMH-CIDI and SCID, the extent to which true cases are over-sampled in the clinical calibration subsample, the size of the total sample, the size of the clinical calibration subsample, and whether the parameter of interest is estimated in the total sample or in subsamples. We fixed the first two of these six variables in our simulation by focusing on the MI SCID lifetime prevalence estimation of MDE and assuming the actual lifetime prevalence (24.2%) and the actual strength of association (AUC = 0.86) found in the NCS-R (Kessler, Berglund, Demler, Jin, Koretz, Merikangas, Rush, Walters and Wang, 2003). We also fixed the proportional representation of true cases in the clinical calibration subsample to be twice as high as in the total sample. We yoked the next two variables by assuming that we could either

have a sample of 10,000 WMH-CIDI interviews in which 300 respondents were reinterviewed with the SCID (roughly equivalent to the actual NCS-R design) or successively smaller WMH-CIDI samples in increments of 1,000 cases in which the cost saving of decreasing the number of WMH-CIDI interviews was accompanied by an increase in the number of SCID reinterviews in the ratio 1:2 (for example, 10,000 WMH-CIDI plus 300 SCID, 9,000 WMH-CIDI plus 800 SCID, 8,000 WMH-CIDI plus 1,300 SCID, 7,000 WMH-CIDI plus 1,800 SCID). The 1:2 ratio is higher than the assumed ratio for total costs of a WMH-CIDI survey compared to a SCID survey because we assumed that the marginal cost of carrying out a SCID interview over the telephone after a WMH-CIDI interview is already done is lower than the marginal cost of an additional face-to-face SCID interview in the absence of an earlier WMH-CIDI interview. Finally, we varied the estimation to be carried out in the total sample or in subsamples of between 2% and 90% of the total sample.

The inclusion of subsamples in the simulation is critical because complex substantive analyses almost always work with subsamples. Even something as simple as examining sex differences in prevalence requires separate estimation among men and women, while studies of race-ethnic differences sometimes require comparisons in subsamples that are as small as 2% to 10% of the sample. The simulation uses the estimates of PPV in the total sample to generate predicted probabilities in subsamples; that is, it assumes that PPV and NPV are constant across subsamples. This assumption will be incorrect when SENS and SPEC are constant and prevalence differs. As a result of this problem, we looked for interactions between predicted probabilities and important subsampling variables (for example, age, sex, education, race-ethnicity) in expansions of the basic prediction equation before carrying out the simulation. When interactions were statistically significant, they were built into the final prediction equations to modify the assumptions of constant PPV and NPV in a way that reflected the actual structure in the data. These specifications were constrained by the limited statistical power to detect meaningful interactions in the clinical reappraisal subsample but were nonetheless useful in moving beyond the assumption of completely constant PPV and NPV.

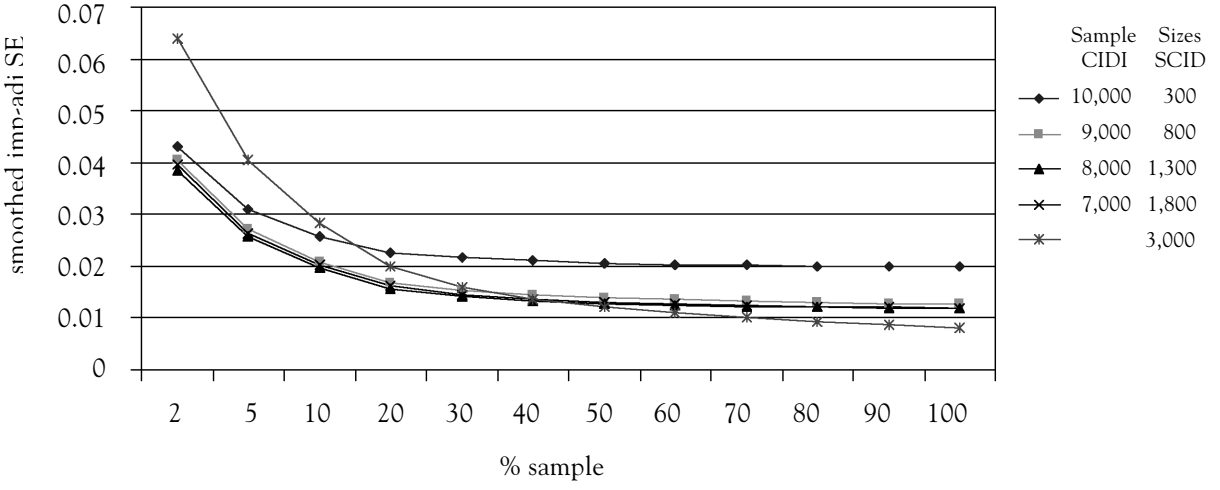
Before turning to the substantive results, one seemingly odd implication of the assumption regarding the relative costs of WMH-CIDI and SCID interviews is worthy of comment: the fact that the smallest of the hypothetical WMH-CIDI samples (7,000 with 1,800 SCID reinterviews) has only 1,200 fewer SCID reinterviews than the SCID-only sample. The reader might be puzzled that we would think these two designs are comparable in cost. Could it possibly be the case that a reduction in 1,200 SCID interviews from the SCID-only design could pay for 7,000 WMH-CIDI interviews? That seems implausible on the face of it. The answer is that we are assuming that the SCID interviews in the hypothetical WMH-CIDI sample designs would be carried out by telephone (an investment of only about 3 interviewer hours per completed interview) after the WMH-CIDI interviewer has already recruited the household and carried out the face-to-face WMH-CIDI interview (an investment of close to 12 interviewer hours). In the SCID-only sample, in comparison, the interviews would be carried out face-to-face. This means that the SCID interviewers would have to travel to the homes of respondents rather than working over the telephone from a centralized location. We assumed that the average number of interviewer hours per completed face-to-face interview would be higher in the SCID-only design than the WMH-CIDI designs because it would be more difficult to find competent clinical interviewers than lay interviewers in all the many different parts of the country where the NCS-R was carried out (over 170 counties in 34 different states). When competent interviewers cannot be found in a sample area, we have to fly interviewers from other areas of the country into the sample area to complete the interviews, substantially raising the number of hours the interviewers have to be paid per completed interview.

With these design considerations as a background, results of the simulation for estimating the lifetime prevalence of SCID MDE are presented in Figure 1. Each of the five curves in the figure represents a different hypothetical sample design. Four of the five are WMH-CIDI samples that range in size between 7,000 and 10,000, with the number of SCID reinterviews varying from a low of 300 in the WMH-CIDI 10,000 sample design to a high of 1,800 in the WMH-CIDI 7,000 sample design. The fifth is the SCID-only sample. The x-axis of the figure repre-

sents the proportion of the sample in which prevalence is being estimated, whereas the y-axis represents the size of the SE of the prevalence estimate. Two broad patterns in the figure are noteworthy. First, the SE, as one would expect, increases as the proportion of the sample in which prevalence is being estimated decreases. Second, this increase has a different functional form in the WMH-CIDI samples than in the SCID-only sample because of the assumption mentioned two paragraphs above of constant PPV and NPV in subsamples. This assumption allows us to use the precision in the total sample to estimate prevalence in WMH-CIDI subsamples. As a result of this two-phase sample advantage, the precision of estimates in the WMH-CIDI samples increases relative to the SCID-only sample as we move to smaller and smaller subsamples. Three of the four WMH-CIDI sample designs, the exception being WMH-CIDI 10,000, outperform the SCID-only design in estimating prevalence in the lower part of the sample proportion distribution (subsamples of 30% or less). The SCID-only design, in comparison, outperforms all the WMH-CIDI designs in the upper part of the distribution (subsamples of 50% or more), but these differences are very small. Taken together, these results lead to the conclusion that a WMH-CIDI sample design of 7,000–9,000 respondents with SCID reinterviews of 800–1,800 cases (oversampling WMH-CIDI MDE cases at a 2:1 rate compared to non-cases) would be superior to a SCID-only design of the same cost in maximizing the precision of lifetime SCID MDE prevalence estimates if the assumption of constant PPV and NPV holds. In the total sample or large subsamples, SEs are small for all these designs, which means that the modest advantage of the SCID-only design is of no substantive significance. In smaller subsamples, the WMH-CIDI designs have a clear advantage. It is important to remember that SCID subsamples in WMH-CIDI sample designs would not oversample only on MDE but on all disorders and that the 2:1 over-sampling into the SCID sub-sample assumed here for MDE would vary across disorders.

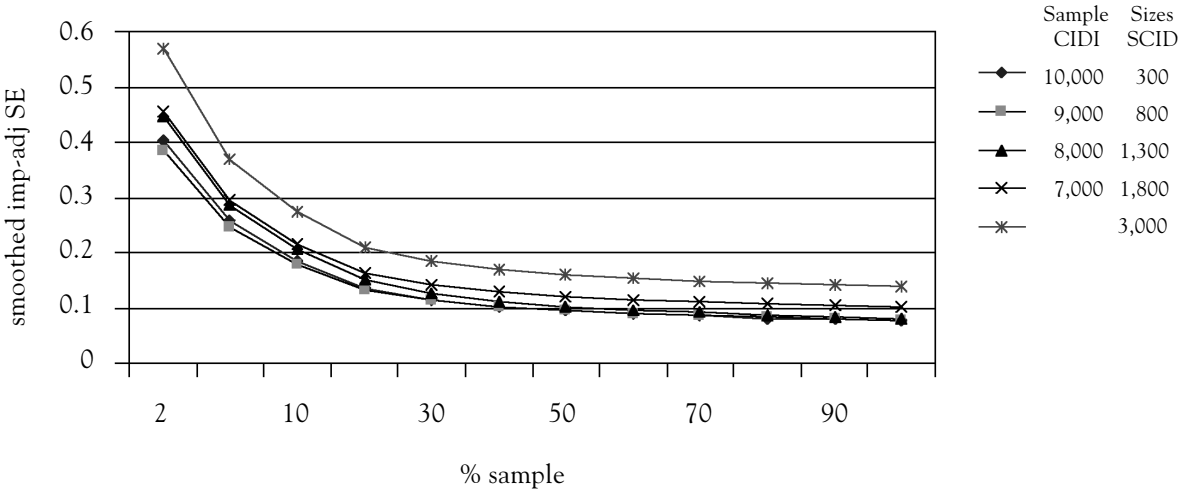
The conclusion that the WMH-CIDI designs are superior to the SCID-only design depends on the parameters being estimated. To illustrate this point, Figure 2 presents the results of a simulation for sex differences in the lifetime prevalence of SCID MDE, again assuming constant PPV. As shown there, all of the WMH-CIDI sample designs, including WMH-CIDI 10,000, outperform the SCID-only design

Figure 1. Multiple imputed standard errors of SCID DSM-IV MDE lifetime prevalence estimates for five hypothetical CIDI-SCID sample designs.[1]



[1]All designs assume the actual SCID DSM-IV Lifetime Prevalence estimate of MDE (24.2%) as in the NCS-R (Kessler et al., 2003), the actual association of CIDI symptom-level data in predicting SCID diagnoses as in the NCS-R (AUC = 0.86), and a 2:1 proportional representation of SCID cases in the clinical calibration subsample compared to the total sample.

Figure 2. Estimated standard errors of sex differences in lifetime SCID DSM-IV Major Depressive Episode prevalence estimates in total samples and random subsamples for five hypothetical CIDI-SCID sample designs.[1]



[1]All designs assume the actual SCID DSM-IV lifetime prevalence estimates of MDE among women (23.9%) and men (24.6%) as in the NCS-R (Kessler et al., 2003), the actual association of CIDI symptom-level data in predicting SCID diagnoses as in the NCS-R (AUC = 0.86), consistency of this association among women and men, and a 2:1 proportional representation of SCID cases in the clinical calibration subsample compared to the total sample.

throughout the entire range of the sample proportion distribution. All else being equal, the relative performance of the WMH-CIDI designs compared to the SCID-only design will increase as statistics become more complex and true prevalence becomes smaller because of the sample-size advantages over the SCID-only design. However, this increased performance is achieved at the expense of having to make increasingly more complex implicit assumptions about the consistency of the coefficients in the MI prediction equations across subsamples. It is always possible to evaluate the sensitivity of results to variations in these and other assumptions in the simulations to deal with these uncertainties. We do not pursue this line of investigation here, though, as our purpose is to be illustrative rather than to present a detailed investigation of the empirical results of any one survey.

Discussion

We presented an overview of the rationale, design, and analysis plan for the WMH-CIDI clinical calibration studies. These studies are being conducted in order to increase the clinical relevance of WMH-CIDI community epidemiological surveys. Previous methodological studies of fully structured diagnostic interviews like the WMH-CIDI have attempted to validate these instruments against a clinical gold standard. However, our review of this literature led us to conclude that validation of this sort is impossible because no highly reliable and valid clinical gold standard is available. As a result, we shifted the goal of the WMH-CIDI methodological studies to calibration rather than validation. In other words, rather than ask whether the diagnostic classifications in the WMH-CIDI are 'correct', we asked whether they are 'consistent' with diagnoses obtained from a state-of-the-art semi-structured clinician-administered research diagnostic interview (the SCID). Consistency is being evaluated both at the aggregate level (consistency of prevalence estimates) and at the individual level (concordance of prevalence estimates). Conventional statistics are being calculated to describe WMH-CIDI and SCID concordance (sensitivity, specificity, Cohen's κ) for purposes of comparison with previous diagnostic validity studies. We are also calculating several descriptive statistics that are a function of sensitivity and specificity

(odds-ratio, Yule's Q, area under the ROC curve) that, unlike κ, are not sensitive to variation in prevalence when sensitivity and specificity are constant. However, the main focus of the consistency analyses is to embed comparisons of consistency into substantive investigations. This involves creating prediction equations in clinical calibration subsamples that are used to assign predicted probabilities of SCID diagnoses to each respondent who completes a WMH-CIDI interview. Analyses of these predicted probabilities allow us to investigate the extent to which estimates of prevalence and correlates are similar or different when based on WMH-CIDI diagnoses versus predicted SCID diagnoses. Multiple imputation is used to adjust estimates of se for the SCID diagnoses being based on a model rather than on observation. When the substantive results are similar there is an advantage to focusing on WMH-CIDI results, as they are more precise than predicted SCID results because the WMH-CIDI is measured for each respondent while the SCID is based on a prediction model. When results differ, we are able to bound our uncertainty by comparing WMH-CIDI and SCID estimates.

Acknowledgements

This paper is a report of the National Comorbidity Survey Replication (NCS-R). The 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 044708), and the John W. Alden Trust. Collaborating investigators include Ronald C. Kessler (Principal Investigator, Harvard Medical School), Kathleen Merikangas (Co-Principal Investigator, NIMH), James Anthony (Michigan State University), William Eaton (The Johns Hopkins University), Meyer Glantz (NIDA), Doreen Koretz (Harvard University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Greg Simon (Group Health Cooperative), Michael Von Korff (Group Health Cooperative), Philip Wang (Harvard Medical School), Kenneth Wells (UCLA), Elaine Wethington (Cornell University) and Hans-Ulrich Wittchen (Institute of Clinical Psychology, Technical University Dresden and Max Planck Institute of Psychiatry). The authors appreciate the helpful comments of Kathleen Merikangas and Hans-Ulrich Wittchen on earlier drafts. A complete list of NCS publications and the full text of all NCS-R instruments can be found at http://www.hcp.med.harvard.edu/ncs. Send correspondence to NCS@hcp.med.harvard.edu.

References

Agresti A. An Introduction to Categorical Data Analysis. New York: Wiley, 1996.

Angold A, Costello EJ, Farmer EM, Burns BJ, Erkanli A. Impaired but undiagnosed. J Am Acad Child Adolesc Psychiatry 1999; 38(2): 129–37.

Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC. Long-term reliability of diagnosing lifetime major depression in a community sample. Arch Gen Psychiatry 1986; 43(5): 435–40.

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993; 46(5): 423–9.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20: 37–46.

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

Crocker L, Algina J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart & Winston, 1986.

Daskalakis C, Laird NM and Murphy JM. Regression analysis of multiple-source longitudinal outcomes: a 'Stirling County' depression study. Am J Epidemiol 2002; 155(1): 88–94.

Eaton WW, Neufeld K, Chen LS, Cai G. A comparison of self-report and clinical diagnostic interviews for depression: diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the Baltimore epidemiologic catchment area follow-up. Arch Gen Psychiatry 2000; 57(3): 217–22.

Faraone SV, Tsuang MT. Measuring diagnostic accuracy in the absence of a 'gold standard.' Am J Psychiatry 1994; 151(5): 650–7.

Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43(6): 543–9.

Fennig S, Craig T, Lavelle J, Kovasznay B, Bromet EJ. Best-estimate versus structured interview-based diagnosis in first-admission psychosis. Compr Psychiatry 1994; 35(5): 341–8.

First MB, Spitzer RL, Gibbon M, Williams JBW. 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, 2002.

Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat Med 2002; 21 (9): 1289–307.

Gibbon M, McDonald-Scott P, Endicott J. Mastering the

art of research interviewing. A model training procedure for diagnostic evaluation. Arch Gen Psychiatry 1981; 38(11): 1259–62.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1): 29–36.

Helzer JE, Spitznagel EL, McEvoy L. The predictive validity of lay diagnostic interview schedule diagnoses in the general population: a comparison with physician examiners. Arch Gen Psychiatry 1987; 44(12): 1069–77.

Houck PR, Spiegel DA, Shear MK, Rucci P. Reliability of the self-report version of the Panic Disorder Serverity Scale. Depress Anxiety 2002; 15(4): 183–5.

Janca A, Robins LN, Bucholz KK, Early TS, Shayka JJ. Comparison of the Composite International Diagnostic Interview and clinical DSM-III-R criteria checklist diagnoses. Acta Psychiatr Scand 1992; 85: 440–3.

Jensen PS, Watanabe HK, Richters JE. Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. J Abnorm Child Psychol 1999; 27(6): 439–45.

Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. A population-based twin study of major depression in women. The impact of varying definitions of illness. Arch Gen Psychiatry 1992; 49(4): 257–66.

Kessler R. The World Health Organization International Consortium in Psychiatric Epidemiology (ICPE): Initial work and future directions – the NAPE lecture 1998. Acta Psychiatr Scand 1999; 99: 2–9.

Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003; 289(23): 3095–105.

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. Arch Gen Psychiatry 1994; 51(1): 8–19.

Kessler RC, and Üstün TB. The World Health Organization World Mental Health 2000 Initiative. Hospital Management International 2000: 195–6.

Kessler RC and Üstün TB. The World Mental Health (WMH) survey initiative version of the World Health Organization Composite International Diagnostic Interview (CIDI). International Journal of Methods in Psychiatric Research 2004; (this issue).

Kessler RC, Wittchen H-U, Abelson JM, McGonagle K, Schwarz N, Kendler KS, Knauper B, Zhao S. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. International Journal of Methods in Psychiatric Research 1998; 7(1): 33–55.

Kessler RC, Wittchen HU, Abelson JM and Zhao S. Methodological issues in assessing psychiatric disorder with self-reports. In AA Stone, JS Turrkan, CA Bachrach, JB Jobe, HS Kurtzman, VS Cain eds. The Science of Self-Report: Implications for Research and Practice. Mahwah NJ: Erlbaum, 2000, 229–25.

Kish L, Frankel MR. Inferences from complex samples. Journal of the Royal Statistical Society 1974; 36 (Series B): 1–37.

Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. J Am Acad Child Adolesc Psychiatry 2003; 42(12): 1524–9.

Lucas CP, Fisher P, Piacentini J, Zhang H, Jensen PS, Shaffer D, Dulcan M, Schwab-Stone M, Regier D, Canino G. Features of interviews questions associated with attenuation of symptom reports. J Abnorm Child Psychol 1999; 27(6): 429–37.

Mannuzza S, Fyer AJ, Martin LY, Gallops MS, Endicott J, Gorman J, Liebowitz MR, Klein DF. Reliability of anxiety assessment. I. Diagnostic agreement. Arch Gen Psychiatry 1989; 46(12): 1093–101.

Murphy JM, Monson RR, Laird NM, Sobol AM, Leighton AH. A comparison of diagnostic interviews for depression in the Stirling County study: challenges for psychiatric epidemiology. Arch Gen Psychiatry 2000; 57(3): 230–6.

Narrow WE, Rae DS, Robins LN, Regier DA. Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile two surveys' estimates. Arch Gen Psychiatry 2002; 59(2): 115–23.

National Advisory Mental Health Council. Health care reform for Americans with severe mental illnesses: report of the National Advisory Mental Health Council. Am J Psychiatry 1993; 150: 1447–65.

Pincus HA, Davis WW, McQueen LE. 'Subthreshold' mental disorders. A review and synthesis of studies on minor depression and other 'brand names'. Br J Psychiatry 1999; 174: 288–96.

Pincus HA, Zarin DA, First M. 'Clinical significance' and DSM-IV. Arch Gen Psychiatry 1998; 55(12): 1145; author reply 1147–8.

Ramirez Basco M, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, Barnett V. Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 2000; 157(10): 1599–605.

Regier DA. Community diagnosis counts [Commentary]. Arch Gen Psychiatry 2000; 57: 223–4.

Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, Norquist GS. Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. Arch Gen Psychiatry 1998; 55(2): 109–15.

Rehm J, Üstün TB, Saxena S, Nelson CB, Chatterji S, Ivis F, Adlaf E. On the development and psychometric testing of the WHO screening instrument to assess

disablement in the general population. International Journal of Methods in Psychiatric Research 1999; 8: 110–23.

Robins LN. Epidemiology: reflections on testing the validity of psychiatric interviews. Arch Gen Psychiatry 1985; 42(9): 918–24.

Robins LN. Diagnostic grammar and assessment: translating criteria into questions. Psychol Med 1989; 19(1): 57–68.

Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry 1981; 38(4): 381–9.

Robins LN, Regier DA eds. Psychiatric Disorders in America: The Epidemiologic Catchment Area Study. New York: The Free Press, 1991.

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA, Sartorius N, Towle LH. The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry 1988; 45(12): 1069–77.

Rohde PL, Seeley JR. Comparability of telephone and face-to-face interviews in assesing axis I and II disorders. Am J Psychiatry 1997: 1593–8.

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

Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996; 26(3): 477–86.

Shear MK, Brown TA, Barlow DH, Money R, Sholomskas DE, Woods SW, Gorman JM, Papp LA. Multicenter collaborative panic disorder severity scale. Am J Psychiatry 1997; 154(11): 1571–5.

Shrout PE, Newman SC. Design of two-phase prevalence surveys of rare disorders. Biometrics 1989; 45: 549–55.

Sobin C, Weissman MM, Goldstein RB, Adams P, Wickramaratne P, Warner V, Lish JD. Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. Psychiatr Genet 1993; 3: 227–33.

Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985; 42(7): 725–8.

Substance Abuse and Mental Health Services Administration. Final notice establishing definitions for (1) children with a serious emotional disturbance, and (2) adults with a serious mental illness. Fed Regist 1993; 58: 29422–5.

Üstün TB, Chatterji S, Rehm J. Limitations of diagnostic paradigm: it doesn't explain 'need'. Arch Gen Psychiatry 1998; 55 (12): 1145–6; author reply 1147–8.

Williams JBW, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Harrison GP, Jr., Rounsaville B and Wittchen H-U. The structured clinical interview for DSM-III-R (SCID) II: Multisite test-retest reliability. Arch Gen Psychiatry 1992; 49: 630–6.

Wittchen H-U. Reliability and validity studies of the WHO – Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res 1994; 28(1): 57–84.

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

Correspondence: RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA, USA 02115.
Telephone (+1) 617-432-3587.
Fax (+1) 617-432-3588.
Email: kessler@hcp.med.harvard.edu.