International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1398

Clinical reappraisal of the Composite International Diagnostic Interview Screening Scales (CIDI-SC) in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

RONALD C. KESSLER,[1] PATCHO N. SANTIAGO,[2] LISA J. COLPE,[3] CATHERINE L. DEMPSEY,[2] MICHAEL B. FIRST,[4,5] STEVEN G. HEERINGA,[6] MURRAY B. STEIN,[7,8] CAROL S. FULLERTON,[2] MICHAEL J. GRUBER,[1] JAMES A. NAIFEH,[2] MATTHEW K. NOCK,[9] NANCY A. SAMPSON,[1] MICHAEL SCHOENBAUM,[3] ALAN M. ZASLAVSKY[1] & ROBERT J. URSANO[2]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
3 National Institute of Mental Health, Bethesda, MD, USA
4 Department of Psychiatry, Columbia University, New York, USA
5 New York State Psychiatric Institute, New York, USA
6 University of Michigan, Institute for Social Research, Ann Arbor, MI, USA
7 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
8 VA San Diego Healthcare System, San Diego, CA, USA
9 Department of Psychology, Harvard University, Cambridge, MA, USA

Key words
Composite International Diagnostic Interview (CIDI), CIDI Screening Scales (CIDI-SC), diagnostic concordance, PTSD checklist (PCL), screening scales, validity

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

Abstract

A clinical reappraisal study was carried out in conjunction with the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) All-Army Study (AAS) to evaluate concordance of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnoses based on the Composite International Diagnostic Interview Screening Scales (CIDI-SC) and post-traumatic stress disorder (PTSD) checklist (PCL) with diagnoses based on independent clinical reappraisal interviews (Structured Clinical Interview for DSM-IV [SCID]). Diagnoses included: lifetime mania/hypomania, panic disorder, and intermittent explosive disorder; six-month adult attention-deficit/hyperactivity disorder; and 30-day major depressive episode, generalized anxiety disorder, PTSD, and substance (alcohol or drug) use disorder (abuse or dependence). The sample ($n = 460$) was weighted for over-sampling CIDI-SC/PCL screened positives. Diagnostic thresholds were set to equalize false positives and false negatives. Good individual-level concordance was found between CIDI-SC/PCL and SCID diagnoses at these thresholds (area under curve

Received 10 July 2013;
accepted 15 July 2013

[AUC] = 0.69–0.79). AUC was considerably higher for continuous than dichotomous screening scale scores (AUC = 0.80–0.90), arguing for substantive analyses using not only dichotomous case designations but also continuous measures of predicted probabilities of clinical diagnoses. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

As described in more detail earlier in this issue (Kessler *et al.*, 2013b) and elsewhere (Ursano *et al.*, submitted for publication), the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.armystarrs.org) is a multi-component epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates in the US Army. The literature on risk and resilience factors for suicidality makes it clear that mental disorders are powerful risk factors (Nock *et al.*, 2008; Nock *et al.*, 2013). As a result, a wide range of mental disorders were assessed in the Army STARRS surveys. However, due to the size and logistical complexities of these surveys, which are described earlier in this issue (Heeringa *et al.*, 2013), it was impossible to administer an in-depth psychiatric diagnostic interview to participants. Instead, mental disorders were assessed with short self-administered screening scales.

A number of screening scales exist to assess such disorders as attention-deficit hyperactivity disorder (ADHD; Kessler *et al.*, 2005a), bipolar disorder (BPD; Hirschfeld *et al.*, 2000), generalized anxiety disorder (GAD; Spitzer *et al.*, 2006), major depressive episode (MDE; Kroenke *et al.*, 2001), and post-traumatic stress disorder (PTSD; Breslau *et al.*, 1999). Although in some cases these scales were developed originally to assess symptom severity among patients in treatment, they subsequently have been adapted for use either as web-based tools for self-diagnosis (Donker *et al.*, 2009; Farvolden *et al.*, 2003) or as brief evaluations of mental disorders in primary care settings or community surveys (Broadhead *et al.*, 1995; Gaynes *et al.*, 2010; Hunter *et al.*, 2005; Kessler *et al.*, 2013a). Clinical reappraisal studies comparing scores on these screening scales with independent clinical diagnoses show that many of these screening scales have good concordance with clinical diagnoses (Kessler and Pennell, in press).

The screening scales that form the core diagnostic assessment in Army STARRS are the World Health Organization (WHO) Composite International Diagnostic Interview Screening Scales (CIDI-SC) (Kessler *et al.*, 2013a). These were selected largely because they are a coordinated set of short scales that cover a wide range of disorders and have good psychometric properties. However, another appeal of the CIDI-SC is that they are embedded in the WHO Composite International Diagnostic Interview (CIDI) (Kessler and Üstün, 2004), the research diagnostic interview used in most large-scale epidemiological surveys of psychiatric disorders throughout the world (Haro *et al.*, 2006). Use of the CIDI-SC in Army STARRS thereby creates a crosswalk to an in-depth diagnostic interview that might be used in more focused follow-up studies of Army STARRS high-risk subsamples. The exception is that we used the PTSD checklist (PCL) (Weathers *et al.*, 1993) to assess PTSD based on the widespread use of this screening scale in previous military studies of PTSD (Barnes *et al.*, 2013; Brown *et al.*, 2012; Jones *et al.*, 2013) coupled with strong evidence for the validity of the PCL in both military and civilian samples (Wilkins *et al.*, 2011).

Although good concordance of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994) diagnoses based on the CIDI-SC (Kessler *et al.*, 2005a; Kessler *et al.*, 2007; Kessler *et al.*, 2006a; Kessler *et al.*, 2013a) and PCL (Wilkins *et al.*, 2011) with diagnoses based on independent clinical reappraisal interviews has been reported in a number of studies, this does not guarantee that these screening scales will perform equally well among soldiers in the Army STARRS surveys. As a result, a new clinical reappraisal study (CRS) was carried out in conjunction with the Army STARRS All-Army Study (AAS) (Ursano *et al.*, submitted for publication) to examine the psychometric characteristics of the CIDI-SC and PCL in the context of the field conditions encountered in the Army STARRS surveys. Results of this CRS are presented in the current report.

Methods

The samples

The All-Army Study (AAS)

As described in more detail previously in this issue (Kessler *et al.*, 2013b), the AAS is a cross-sectional survey of active duty Army personnel exclusive of soldiers in basic

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

combat training administered in quarterly replicates to a total of nearly 50,000 soldiers during calendar years 2011–2012. Each quarterly AAS replicate consisted of a stratified (by Army Command-location and unit size) probability sample of Army units, excluding units of fewer than 30 soldiers (less than 2% of all Army personnel). All targeted personnel in these units were ordered to attend an informed consent presentation explaining study purposes, confidentiality procedures, and the voluntary nature of participation before requesting written informed consent for a group self-administered questionnaire (SAQ). Respondents were additionally asked for consent to link their Army and Department of Defense administrative records to their SAQ responses and to participate in future longitudinal follow-up data collections. Identifying information (name, birthday, Social Security number for record linkage; telephone number, email, secondary contact information for longitudinal follow-up) was collected from consenting respondents and kept in a separate secure file. These recruitment, consent, and data protection procedures were approved by the Human Subjects Committees of the Uniformed Services University of the Health Sciences for the Henry M. Jackson Foundation (the primary grantee), the Institute for Social Research at the University of Michigan (the organization implementing Army STARRS surveys), and all other collaborating organizations.

The CRS was carried out between March 2012 and November 2012. All quarterly AAS replicates over that time period were based on representative samples of soldiers stationed both in the continental United States and elsewhere in the world other than a combat theater, while the Q2–3 2012 replicates also included probability samples of soldiers stationed in Afghanistan who were surveyed in group-administered sessions while they were passing through Kuwait either leaving for or returning from their mid-tour leave. However, because of logistical issues requiring that the CRS interviews be administered within two weeks of the AAS survey, the CRS was implemented exclusively in the continental United States among Regular (active component) Army AAS respondents providing consent for administrative data linkage and completing the SAQ. Activated Army Reserve and National Guard respondents were excluded from the CRS due to small numbers.

Although, as noted earlier, all unit members in these replicates were ordered to report to the informed consent session, 19.4% of those in the replicates used for the CRS were absent due to conflicting duty assignments. The vast majority of those attending (99.6%) consented to the survey and 98.8% of consenters completed the survey. In addition, 71.4% of completers provided successful record linkage. Most incomplete surveys were due to logistical complications (e.g. units either arriving late to survey sessions or having to leave early), although some respondents needed more than the allotted 90 minutes to complete the survey. The survey completion-successful-linkage *cooperation* rate was 63.9% and the completion-successful-linkage *response* rate was 51.5% based on the American Association of Public Opinion Research COOP1 and RR1 calculation methods (American Association for Public Opinion Research, 2009).

The clinical reappraisal study (CRS) sample

In order to evaluate the concordance of diagnoses based on the CIDI-SC and PCL in the AAS with independent clinical diagnoses, a sample of AAS respondents was selected to participate in clinical follow-up interviews within two weeks of completing the AAS in selected AAS sessions. As soon as the AAS survey was completed in these sessions, each AAS respondent was classified as *threshold*, *subthreshold* or *no* on each of the eight screening scales considered here. A probability subsample of AAS respondents from the session was then invited to participate in a confidential clinical reappraisal interview with the goal of obtaining a total (i.e. over the entire nine-month interview recruitment period) of 30 CRS interviews with respondents selected at random from those classified as threshold cases on each diagnosis, 10 from among those classified as subthreshold on each diagnosis, and 40 respondents selected at random from those classified as meeting neither threshold nor subthreshold criteria for any diagnosis. CRS respondents with each diagnosis were selected *with replacement* (i.e. the same respondent could be selected for more than one diagnosis). The initial sampling fractions varied across disorders due to differences in prevalence among the disorders. These sampling fractions were then modified over sessions in order to achieve a roughly equal distribution of cases within each diagnosis across sessions while meeting the sample quotas. The 460 clinical interviews completed by the end of the CRS is more than the 360 needed (i.e. 30 interviews with threshold CIDI-SC/PCL cases for each of eight disorders plus 10 interviews with CIDI-SC/PCL subthreshold cases for each of these disorders plus 40 respondents screening negative on all eight CIDI-SC/PCL scales) because it was necessary to recruit additional respondents in the later replicates to fill the sample quotas for the least common disorders.

Invitations to participate in the CRS were made through unit points of contact who scheduled two-hour time blocks during which respondents were relieved of

their usual duty assignments in order to report to the Army STARRS office on the installation. Once at the study office, an Army STARRS data collection specialist explained the content and purposes of the CRS and obtained written informed consent to participate. Consenting respondents were then assigned to a private room where they were administered the CRS interview telephonically by one of the CRS clinical interviewers, all of whom were located at the Uniformed Services University of the Health Sciences (USUHS) in Bethesda, Maryland. The CRS clinical supervisor (CLD), also located at USUHS, coordinated with Army STARRS data collection specialists at the local AAS installations to schedule these remote CRS telephone interviewers.

An overview of screening scale content

Screening scales were included in the AAS for eight DSM-IV disorders that have been found in previous general population studies to be significant predictors of suicidality (Nock *et al.*, 2008; Nock *et al.*, 2013; Nock *et al.*, 2009). These include two mood disorders (MDE, mania/hypomania [MHM]), three anxiety disorders (panic disorder with or without agoraphobia, GAD, PTSD), and three externalizing disorders (adult ADHD, intermittent explosive disorder, substance use disorder [SUD]).

Symptom questions in most CIDI-SC ask respondents about the frequency of particular symptoms over the 30 days before interview using the response options *all or almost all of the time*, *most of the time*, *some of the time*, *a little of the time*, and *none of the time*. Each CIDI-SC has an embedded skip logic whereby all respondents are administered one or more entry questions and then either skipped if they fail to endorse these questions or continue to a series of follow-up questions if they endorse the entry question(s). This approach was designed to reduce overall scale administration time and respondent burden while minimizing the number of true positives incorrectly skipped out by the entry questions. Respondents who fail to endorse any of the entry questions are asked a total of 46 questions across all eight scales combined, while respondents who endorse every single question are asked an additional 82 questions.

The CIDI-SC MDE scale begins with four entry questions that ask about *being sad*, *depressed*, *or discouraged*, *having little or no interest or pleasure in things*, and *feeling down on yourself*, *no good*, *or worthless* (Kessler *et al.*, 2013a). Respondents who report that at least one of these symptoms occurred at least *some of the time* in the past 30 days are administered 10 additional questions to assess the inclusion criteria of MDE. The *some of the time* threshold, while low for a DSM-IV diagnosis of MDE (which requires depressive symptoms to last most of the day nearly every day for two weeks or longer), was chosen because we wanted to collect information not only on threshold cases but also on subthreshold manifestions of MDE. A similar attempt to collect information about subthreshold symptoms was made in selecting stem question skip rules for each of the other screening scales.

The CIDI-SC MHM scale focuses on subthreshold hypomania as well as mania and hypomania based on evidence that subthreshold hypomania can be highly impairing (Merikangas *et al.*, 2007). In addition, the questions focus on lifetime rather than 30-day prevalence due to the fact that recent BPD can manifest as either MHM or as MDE. As described in more detail elsewhere (Kessler *et al.*, 2006a; Kessler *et al.*, 2013a), the single MHM entry question begins with a vignette describing a hypomanic episode and then asks respondents if they ever had an episode of this sort at any time in their life. A positive response is followed by four questions about the frequency of core MHM symptoms during *a typical intense episode of this sort*. These symptoms include being *much higher*, *happier*, *or optimistic than usual*; *much more irritable than usual*; *so hyper or wound up that you felt out of control*; *having thoughts race through your mind so fast you could hardly keep track of them*. Respondents who report that at least one of these symptoms occurrs at least *some of the time* during a typical intense episode are then administered six additional questions about the inclusion criteria of MHM and are then asked about episode recency to assess 30-day prevalence of MHM. Lifetime rather than 30-day MHM is evaluated here due to the rarity of 30-day MHM in the AAS sample.

The CIDI-SC panic disorder (PD) scale includes two entry questions about lifetime atacks of *panic*, *anxiety*, *or strong fear that came on very suddenly and made you feel very frightened or uneasy*; and *attacks of heart pounding or chest pain that came on very suddenly and made you feel very frightened or uneasy* (Kessler *et al.*, 2013a). A positive response to either entry question is followed by one additional question on how often these attacks are triggered (i.e. occur in situations where the respondent has a strong fear – like a fear of snakes or heights – or where the respondent is in real danger – like a car accident) versus untriggered (i.e. occur without provocation "out of the blue"). Respondents who report ever having untriggered attacks are then administered 13 additional questions to assess the remaining DSM-IV inclusion criteria of PD. Lifetime rather than 30-day PD is evaluated here due to the rarity of 30-day PD in the AAS sample.

The CIDI-SC GAD scale includes five entry questions about 30-day frequency of being *anxious or nervous*;

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

worried about a number of different things; *more anxious or worried than other people in your same situation*; *worried about things most other people don't worry about*; and *having trouble controlling your worry or anxiety* (Kessler *et al.*, 2013a). Respondents who report any of these symptoms at least *some of the time* are administered an additional nine questions to assess the remaining DSM-IV inclusion criteria of GAD along with a final question to assess persistence of symptoms. As a minimum duration of six months is required to meet DSM-IV criteria of GAD, the CIDI-SC assesses duration of symptoms, although the concordance data reported here are for symptoms in the 30-days before interview.

As noted earlier, PTSD is assessed in the AAS with the PCL. The PCL Civilian version (Weathers *et al.*, 1993) was used in Army STARRS because we covered traumatic experiences both in and out of the line of duty. This is a 17-question scale that assesses the 17 DSM-IV Criterion B–D symptoms of PTSD. Although there are no entry questions in the PCL, AAS respondents are first asked 15 questions about traumatic experiences (TEs) that might have happened to them during deployments and 15 additional questions about TEs that might have happened to them at any other time in life. Only respondents who report at least one of these 30 TEs are administered the PCL. The PCL questions ask *how much* respondents were *bothered* in the past 30 days by symptoms associated with any of the TEs they ever experienced. Response categories are *extremely*, *quite a bit*, *moderately*, *a little bit*, and *not at all*.

The CIDI-SC adult ADHD scale includes four entry questions found in previous research to provide an optimal short inclusion screen for ADHD in the adult general population (Kessler *et al.*, 2010a). Respondents who report at least two of these symptoms at least *some of the time* in the past six months then receive an additional eight questions shown in a number of previous studies to detect adult ADHD with good accuracy (Kessler *et al.*, 2007; Kessler *et al.*, 2010a; Kessler *et al.*, 2009).

The CIDI-SC intermittent explosive disorder (IED) scale includes one entry question about lifetime attacks of anger when the respondent *all of a sudden … lost control and either broke or smashed something worth more than a few dollars*, *hit or tried to hurt someone*, *or threatened someone* (Kessler *et al.*, 2006b). A positive response is followed by six additional questions that assess the remaining DSM-IV inclusion criteria of IED. As the assessment of IED followed the same logic as the assessment of PD, lifetime rather than 30-day IED is evaluated here in parallel with the evaluation of PD.

The CIDI-SC assessment of SUD, finally, begins with 12 entry questions about quantity-frequency of alcohol use, illicit drug use, and prescription drug misuse, where the latter is defined as use *either without a doctor's prescription*, *more than prescribed*, *or to get high*, *buzzed*, *or numbed out*. Prescription drug misuse is included in the assessment based on evidence that it is considerably more common than illicit drug use in the Army (Bray *et al.*, 2010). Respondents who report any of these types of substance use are then administered the four CIDI-SC questions about DSM-IV substance abuse in the 30 days before interview and eight additional questions to screen for substance dependence in the 30 days before interview including five from the Severity of Dependence Scale (Gossop *et al.*, 1995) and three additional CIDI-SC questions. SUDs (i.e. either abuse or dependence) are assessed only once for alcohol and/or drugs combined.

Scoring the screening scales

Each screening scale was initially scored continuously by summing values across all items in the scale, assigning respondents who were skipped out after screening questions the lowest possible scores on the remaining items. Receiver operating characteristic (ROC) curve analysis (Margolis *et al.*, 2002) was then used to estimate area under the ROC curve (AUC) for the entire continuous scale and to dichotomize the scale at a point that optimized aggregate concordance between the prevalence estimate based on the Structured Clinical Interview for DSM-IV (SCID) and the prevalence estimate based on the CIDI-SC at the designated threshold. This threshold also makes the number of false positives equal the number of false negatives. It is noteworthy, though, that other criteria exist to select diagnostic thresholds and that decisions about which threshold to choose can vary depending on the criterion used. For example, if we had wanted to use the screening scales in a primary care setting to select patients for more in-depth evaluation, we might have lowered the threshold to the point where the vast majority of SCID cases were detected. Or if we were using the screening scales to select patients for a clinical intervention, we might have raised the threshold to the point where the vast majority of screened positives consisted of SCID cases. If the relative importance of minimizing false positives and minimizing false negatives can be specified based on the considerations of such competing criteria, it is possible to minimize this weighted sum of errors in a formal way (Kraemer, 1992). Based on these considerations, a number of alternative thresholds are examined later.

The clinical reappraisal interview

The clinical reappraisal interview was a modified Research Version, Non-Patient Edition of the Structured Clinical Interview for DSM-IV (SCID-I) (First *et al.*, 2002) focused on the eight syndromes under study with the variations in recall periods noted earlier to match the recall periods used in the screening scales. As noted earlier, these interviews were administered by telephone. 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 without the geographic restrictions required for face-to-face clinical assessment. A disadvantage is that people without telephones cannot be included in the assessment. As noted later, though, this difficulty was resolved in the Army STARRS CRS by having pre-designated respondents report to the central Army STARRS research office on their installations, where they were placed in a private room and interviewed remotely by telephone.

A major impediment to making accurate evaluations of concordance between screening scales and clinical diagnoses is the fact 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 fully-structured assessments overestimate prevalence compared to clinical reappraisal interviews. Consistent with the approach used in a number of other clinical reappraisal studies (Haro *et al.*, 2006; Kessler *et al.*, 2005b; Kessler *et al.*, 1998), we modified the conventional blinded clinical re-interview design in three important ways to address this problem.

First, we unblinded the clinical interviewers to whether respondents endorsed diagnostic stem questions in the CIDI-SC. Importantly, though, we did not unblind clinical interviewers to whether the respondents who endorsed CIDI-SC diagnostic stem questions went on to meet full diagnostic criteria.

Second, we rephrased entry questions in the clinical reappraisal interviews to acknowledge prior endorsement of diagnostic stem questions in the CIDI-SC/PCL in order to minimize the problem of false negative diagnostic stem responses in the SCID. For example, rather than repeating a question about presence-absence of 30-day depressed mood in the SCID to respondents who reported 30-day depressed mood in the CIDI-SC, SCID began the assessment of major depression with a declarative sentence: "In your earlier survey you reported feeling sad or depressed most of the time over the past 30 days. The next questions ask more about those feelings."

Third, in order to guarantee that this partial unblinding did not bias clinical interviewers in the direction of rating all stem-positive respondents as cases, we enriched the clinical reappraisal sample to include a higher proportion of respondents than in the sample who endorsed CIDI-SC/PCL diagnostic stem questions but did *not* meet full CIDI-SC/PCL diagnostic criteria. This third feature of the design actually makes the interviewer task more difficult than it would be in a standard CRS in which there is an over-sample of respondents classified as meeting full diagnostic criteria but not of respondents meeting partial criteria.

Clinical interviewer training and quality control

The SCID were administered by 14 trained clinical interviewers. These included four doctoral-level psychologists, seven MA-level psychologists, and three MSW-level clinical social workers. Half of the interviewers had a decade or more of clinical experience (10–21 years), while the other half had 3–9 years of clinical experience (two with three years of experience and one each with five, six, seven, eight, and nine years of experience). The 32-hour SCID interviewer training program began with a 16-hour centralized group training session taking place over a full weekend that was taught by one of the developers of the SCID (MBF) with the assistance of an experienced SCID supervisor (CLD). Training then continued with biweekly individual and group training sessions with homework assignments totaling 32 hours. The training was carried out at USUHS using a modification of the standard SCID training protocol tailored to the diagnoses assessed by the screening scales. In addition to completing this training, each clinical interviewer was required to pass a proficiency test before they began production interviewing based on trainer and supervisor ratings of three practice interviews using a modified version of the SCID Interviewing Skills Evaluation Form created specifically for this study.

All SCID interviews were audio-recorded with permission of respondents and responses recorded on a hard copy interview. The supervisor reviewed the tape recordings of the first five interviews carried out by each interviewer and a minimum of 10% of all subsequent

interviews carried out by each interviewer. The supervisor also reviewed all hard copy interviews completed by all interviewers and reviewed tape recordings of all interviews in which concerns were raised by the hard copy reviews. The symptom-level hard copy clinical ratings were double-entered into a computerized data file after supervisor review and approval. Each interviewer had a weekly one-on-one feedback meeting with the supervisor and participated in a biweekly group calibration meeting with the supervisor and trainer to prevent rater drift. Diagnoses were made without diagnostic hierarchy rules but with organic exclusions.

Analysis methods

Weighting

The CRS sample was weighted to adjust for over-sampling respondents screened as threshold or subthreshold using a weighting method that adjusted for the fact that sampling was made with replacement. This is important because a number of the statistics used to describe scale characteristics are biased when differential selection of screened positives and negatives is not taken into account.

Analysis of screening scale operating characteristics

As noted earlier in the description of screening scale scoring, a summary continuous screening scale score was created for each diagnosis by summing scores across the screening scale items. ROC curve analysis (Margolis *et al.*, 2002) was then used to estimate AUC for the entire scale. Each continuous screening scale was then dichotomized at a threshold that equalized the (weighted) number of false positives and false negatives, thereby maximizing concordance between prevalence estimates based on the SCID and the screening scales. The McNemar x^2 test was used to evaluate the significance of differences between screening scale and SCID prevalence estimates at this threshold. A range of other thresholds was then selected so that SCID prevalence estimates increased monotonically across screening scale strata but did not differ significantly within strata using the logic of stratum-specific likelihood ratio analysis (Pepe, 2003).

Screening scale operating characteristics were then evaluated for each of these thresholds. Individual-level concordance was evaluated using AUC and Cohen's κ (Cohen, 1960). Although κ is the traditional measure used in psychiatric research, κ is not emphasized here because it varies across populations that differ in prevalence even when sensitivity (SN; the percent of true cases correctly classified) and specificity (SP; the percent of true non-cases correctly classified) are constant (Cook, 1998). AUC, in comparison, is a function of SN and SP, which are considered the fundamental parameters of agreement (Kraemer, 1992). AUC equals (SN + SP)/2 when the screen is dichotomous. AUC scores between 0.5 and 1.0 are often interpreted in parallel with κ as *slight* (AUC = 0.50–0.59; κ = 0.0–0.19), *fair* (AUC = 0.6–0.69; κ = 0.2–0.39), *moderate* (AUC = 0.7–0.79; κ = 0.4–0.59), *substantial* (AUC = 0.8–0.89; κ = 0.6–0.79), and *almost perfect* (AUC = 0.9+; κ = 0.8+) (Landis and Koch, 1977). We also report total classification accuracy (TCA), the proportion of all respondents whose CIDI-SC and SCID classifications are consistent.

In addition, we report disaggregated measures of operating characteristics, including SN and SP, positive predictive value (PPV; the proportion of screened positives confirmed by the SCID), negative predictive value (NPV; the proportion of screened negatives confirmed as non-cases by the SCID), likelihood ratio positive (LR+; [SN/(100 − SP)]), and likelihood ratio negative (LR−; [(100 − SN)/SP)]). LR + and LR − assess *relative* proportions of screened positives versus screened negatives confirmed as cases (LR+) or non-cases (LR−). LR + values greater than or equal to five and LR − values less than or equal to 0.2 are generally considered useful, while LR + values greater than or equal to 10 and LR − values less than or equal to 0.1 are considered sufficient to rule in/out diagnoses (Haynes *et al.*, 2006). Significance tests were based on Taylor series design-based standard errors to adjust for data weighting (Wolter, 1985).

Multiple imputation of predicted probabilities of DSM-IV/SCID diagnoses

As noted earlier in the subsection on scoring the screening scales, each screening scale was originally scored continuously and then dichotomized. However, it is not necessary to dichotomize screening scales to make them useful. This is true even in clinical applications, where simple dichotomous scoring rules can be refined by using polychotomous rules that collapse screening scale scores into strata based on analysis of data in a CRS such that the observed prevalence of the clinical outcome differs significantly across strata but not within strata (Guyatt and Rennie, 2001). Designations of patients into multiple risk strata can be useful for clinical purposes when no sharp distinction between cases and non-cases exists in the screening scale (e.g. borderline hypertension).

An extension of this approach can be used in epidemiological surveys to classify respondents into multiple risk

strata based on screening scale scores and to assign predicted probabilities of clinical diagnoses to respondents in each stratum based on the results of a clinical reappraisal survey. It is also possible to ignore the construction of strata in this approach when a monotonic association exists throughout the scale range between a screening scale and probability of a diagnosis, in which case regression analysis can be used to generate predicted probabilities of clinical diagnoses for each respondent in a large sample based on regression coefficients estimated in a smaller clinical reappraisal subsample. These predicted probabilities can then be used either as continuous variables or as the basis for making dichotomous distinctions using any of several different methods discussed elsewhere (Kessler *et al.*, 2010b; Kessler and Pennell, in press).

The creation of continuous scores of this sort is only useful, though, when significant monotonic associations exist between screening scale scores and probabilities of having the clinical diagnosis. We demonstrate later that such associations exist between screening scale scores and diagnoses based on the SCID in the Army STARRS data by comparing AUC for the continuous versions of the screening scales with AUC based on various dichotomous versions of the scales. Given that these monotonic associations exist, we used the method of multiple imputation (MI) (Rubin, 1987) to assign predicted probabilities of SCID diagnoses based on screening scale scores to all respondents in the Army STARRS surveys. MI is a two-phase method designed to impute missing values of particular variables to respondents who have information on variables strongly related to the variable(s) with missing values in such a way as to maximize the use of all available data in examining multivariate associations.

The first phase of MI develops prediction equations based on any of several different complex search methods (Schafer, 2003; White *et al.*, 2011) to estimate multivariate associations of predictors with the variables to be imputed in the subset of respondents with complete data and to use those equations to generate predicted values (*imputations*) for the missing variables in the remainder of the sample. In order to address the fact that imputed values are less precise than observed values, this first phase uses pseudo-replication (i.e. estimation of a new set of coefficients based on the same model from pseudo-samples selected with replacement from the actual sample of people with complete data) to generate multiple imputations for each missing value. The second phase of MI, in which the multiple imputations are used in substantive analysis, then uses each set of imputed values to carry out the substantive analysis separately and then combines the coefficient values across these replications to adjust standard errors of estimates for the fact that some of the data used in the analyses were imputed rather than observed.

Importantly, the first phase of MI allows the inclusion not only of a screening scale (in this case, the CIDI-SC or PCL) designed to provide a proxy measure for the unmeasured variable of interest (in this case, DSM-IV/SCID diagnoses), but also other variables that might be used in second-phase analyses as predictors or consequences of the imputed variable. This is important because the use of only the CIDI-SC or PCL to impute clinical diagnoses would lead to under-estimation of the associations of predictors and consequences of clinical diagnoses with the components of the clinical diagnoses that are not predicted by the CIDI-SC or PCL scores (Collins *et al.*, 2001). As a result, the multiply-imputed predicted probabilities of DSM-IV/SCID diagnoses in Army STARRS were based on complex multivariate equations that included the complete set of CIDI-SC/PCL scores to impute each clinical diagnosis (to adjust for comorbidities among clinical disorders) along with a wide range of substantive correlates included in the AAS and Army/Department of Defense administrative data systems. We produced 20 imputations for each respondent in Army STARRS, a number at the high end of the number recommended in applying MI (Graham *et al.*, 2007).

Results

Concordance of screening scale scores with DSM-IV/SCID diagnoses

Differences in prevalence estimates based on the dichotomized screening scales and SCID are insignificant for all disorders at optimal screening scale thresholds for estimating prevalence ($x_1^2 = 0.0–0.6$, $p = 0.89–0.43$). (Table 1) This is not surprising, of course, as the thresholds were selected to make CIDI-SC prevalence as similar as possible to SCID prevalence. But this is no guarantee of good concordance at the individual level. Individual-level diagnostic concordance at these thresholds is *moderate* for seven diagnoses (AUC = 0.70–0.79) and *fair* for the other diagnosis (ADHD; AUC = 0.69). Total classification accuracy is in the range 86.0–95.9%. The screening scale estimate of 30-day prevalence of any of the seven disorders assessed for 30-day prevalence (the exception being MHM, which was only assessed over the entire lifetime), like most of the individual disorders, has moderate concordance with the estimate based on the SCID (AUC = 0.78).

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Table 1. Aggregate (McNemar x^2) and individual-level (AUC, κ, TCA) consistency of DSM-IV diagnoses based on the CIDI screening scales (CIDI-SC) at their optimal (to estimate prevalence) thresholds and on blinded SCID clinical reappraisal interviews ($n = 460$)[a]

	Aggregate concordance[b]					Individual-level concordance[c]		
	Prevalence estimates							
	CIDI-SC		SCID		McNemar			
	Percent	(SE)	Percent	(SE)	x_1^2	AUC	κ	TCA
I. Mood disorders								
Major depressive episode	6.8	(1.0)	6.7	(1.0)	0.0	0.78	0.55	94.3
Mania/hypomania	4.9	(1.0)	5.2	(0.9)	0.1	0.70	0.42	94.4
II. Anxiety disorders								
Panic disorder	5.0	(0.9)	5.1	(0.7)	0.0	0.78	0.57	95.9
Generalized anxiety disorder	6.6	(0.9)	6.8	(1.0)	0.0	0.70	0.41	92.6
Post-traumatic stress disorder	6.7	(1.0)	6.4	(0.8)	0.1	0.75	0.49	93.7
III. Externalizing disorders								
Adult attention-deficit/hyperactivity disorder	8.2	(1.1)	7.1	(1.1)	0.6	0.69	0.35	90.8
Intermittent explosive disorder	20.8	(2.3)	20.4	(2.0)	0.1	0.79	0.57	86.0
Substance use disorder	4.9	(0.4)	5.4	(0.8)	0.1	0.73	0.47	94.8
IV. Any disorder[d]	18.9	(1.6)	20.3	(1.9)	0.6	0.78	0.58	86.6

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are set at the thresholds designed to maximize concordance with prevalence estimates based on the blinded SCID clinical reappraisal interviews. The McNemar x^2 tests evaluate concordance of these two prevalence estimates.
[c]AUC = area under the receiver operating characteristic curve; κ = Cohen's κ; TCA = total classification accuracy. See the text for definitions of these statistics, all three of which provide information about the overall individual-level concordance between diagnoses based on the CIDI-SC and the blinded SCID clinical reappraisal interviews.
[d]Any of the seven disorders other than mania/hypomania, as mania/hypomania were assessed only over the entire lifetime.

Operating characteristics of the tests

The proportions of SCID cases detected (SN) at the optimal screening scale diagnostic thresholds for estimating SCID prevalence are in the range 42.8–66.8% and the proportions of screening scale cases confirmed by the SCID (PPV) at these thresholds are in the range 37.1–65.3% (68.3% for any 30-day disorder) (Table 2). The proportions of SCID non-cases classified correctly (SP) are 90.9–97.9% and the proportions of screening scale non-cases confirmed as non-cases by the SCID (NPV) are 91.5–97.8%. Lower SN and PPV than SP and NPV are expected for thresholds designed to estimate prevalence without bias when only a minority of respondents has a disorder. LR + is generally considered more informative than SN in such cases (Haynes *et al.*, 2006). LR + is in the *definitive* range (i.e. greater than 10.0) at these thresholds for six of the eight disorders and in the *informative* range (i.e. greater than 5.0) for the others (7.3 for IED; 7.8 for ADHD) and for any 30-day disorder (8.5), indicating that screened positives at these thresholds are much more likely than screened negatives to be confirmed as cases in the clinical reappraisal interviews. LR − values, in comparison, are in a range that would not be considered useful in screening out true non-cases (0.4–0.6).

The implications of modifying diagnostic thresholds

The proportions of screened positives confirmed as SCID cases (PPV) could be increased by raising the screening scale diagnostic thresholds beyond the optimal for estimating prevalence. However, this increase in PPV would be obtained at the expense of decreasing SN and creating downwardly biased (conservative) prevalence estimates. The value of making such a change in threshold while still attempting to approximate clinical prevalence can be

Table 2. CIDI screening scale (CIDI-SC) operating characteristics at optimal thresholds for estimating DSM-IV/SCID prevalence ($n = 460$)[a]

	Positive operating characteristics[b]					Negative operating characteristics[c]				
	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
I. Mood disorders										
Major depressive episode	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
Mania/hypomania	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
II. Anxiety disorders										
Panic disorder	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
Generalized anxiety disorder	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
Post-traumatic stress disorder	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
III. Externalizing disorders										
Adult attention-deficit/hyperactivity disorder	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
Intermittent explosive disorder	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
Substance use disorder	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
IV. Any disorder[d]	63.7	(4.3)	68.3	(4.0)	8.5	92.5	(1.2)	90.9	(1.6)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[c]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[d]Any of the seven disorders other than mania/hypomania, as mania/hypomania were assessed only over the entire lifetime.

evaluated by examining relative changes in PPV versus SN associated with modest increases in screening scale thresholds around the optimal thresholds for estimating SCID prevalence. When we make these small increases in threshold we see that the increases in PPV are much less than the decreases in SN for four disorders (MDE, GAD, ADHD, SUD) (proportional screening scales decreases of 20%, 7%, 25%, and 31%, respectively; proportional PPV increases of 2%, 0%, 18%, and 4%, respectively) (Table 3). In addition, PPV actually *decreases* slightly for the other four disorders due to respondents with CIDI-SC scores just above the optimal thresholds for estimating SCID prevalence of these disorders having high SCID prevalence. These results argue against small changes to increase the screening scale thresholds in the service of making diagnoses more conservative while still maintaining estimates that approximate the SCID prevalence estimates.

We also examined the implications of making small changes in the thresholds in the other direction to increase the proportions of clinical cases screening positive by lowering the screening scale thresholds. Such changes increase SN by definition. This is desirable for purposes of guaranteeing comprehensive detection in treatment samples when PPV does not decrease more than SN increases. However, such anticonservative changes can lead to upward bias in prevalence estimates as well as to reductions in LR + when the proportional increases in SN are lower than the proportional decreases in SP. An analysis of these changes associated with modest decreases in screening scale thresholds shows that LR + consistently decreases when modest changes are made to decrease thresholds (Table 3). These results argue against making the screening scale thresholds less conservative while still maintaining estimates that approximate SCID prevalence.

Selecting alternative optimization rules in selecting screening scale diagnostic thresholds

As noted earlier in the section on analysis methods, the most useful thresholds for screening scales differ depending on the uses to which the screening scales are put. As Army STARRS is an epidemiological study rather than a clinical study, we place a premium on accurate estimation of SCID prevalence. But in a clinical study, where screening scales

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Table 3. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from the optimal for estimating prevalence to either more conservative or more anticonservative thresholds ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
Major depressive episode												
Conservative	6.0	(0.8)	49.1	(8.8)	54.9	(8.5)	16.9	97.1	(0.7)	96.4	(1.0)	0.5
Optimal	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
Anticonservative	7.5	(1.2)	62.5	(9.2)	55.6	(6.3)	17.4	96.4	(0.9)	97.3	(0.9)	0.4
Mania/hypomania												
Conservative	2.7	(0.5)	20.9	(5.9)	39.5	(8.3)	12.3	98.3	(0.4)	95.8	(1.0)	0.8
Optimal	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
Anticonservative	11.6	(1.4)	72.6	(9.1)	32.3	(6.6)	8.7	91.7	(1.4)	98.4	(0.6)	0.3
Panic disorder												
Conservative	3.4	(0.7)	37.1	(9.9)	54.8	(8.9)	23.2	98.4	(0.4)	96.7	(0.7)	0.6
Optimal	5.0	(0.9)	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
Anticonservative	6.4	(0.9)	71.4	(10.4)	57.1	(6.9)	24.6	97.1	(0.5)	98.4	(0.5)	0.3
Generalized anxiety disorder												
Conservative	6.0	(0.8)	40.8	(4.6)	46.7	(7.7)	10.9	96.6	(0.8)	95.7	(0.7)	0.5
Optimal	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
Anticonservative	7.1	(1.0)	44.2	(4.8)	42.6	(7.8)	10.0	95.6	(1.0)	95.9	(0.7)	0.6
Post-traumatic stress disorder												
Conservative	6.2	(1.0)	46.0	(7.0)	47.5	(7.2)	13.1	96.5	(0.9)	96.3	(0.7)	0.6
Optimal	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
Anticonservative	7.7	(1.1)	56.8	(7.2)	47.2	(6.3)	13.2	95.7	(0.9)	97.0	(0.6)	0.4
Adult attention-deficit/hyperactivity disorder												
Conservative	6.8	(1.1)	31.8	(6.2)	33.4	(6.8)	6.5	95.1	(1.1)	94.8	(1.0)	0.7
Optimal	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
Anticonservative	8.8	(1.1)	44.1	(7.9)	35.5	(6.5)	7.2	93.9	(1.1)	95.7	(1.0)	0.5
Intermittent explosive disorder												
Conservative	16.8	(1.4)	47.3	(4.5)	57.2	(5.5)	5.2	90.9	(1.6)	87.1	(2.0)	0.6
Optimal	20.8	(2.3)	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
Anticonservative	26.7	(3.3)	73.5	(7.4)	56.0	(5.7)	5.0	85.3	(2.7)	92.6	(1.9)	0.9
Substance use disorder												
Conservative	4.1	(0.5)	38.1	(8.3)	49.5	(8.4)	17.3	97.8	(0.4)	96.5	(0.8)	0.6
Optimal	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
Anticonservative	6.7	(0.6)	57.3	(9.1)	45.6	(5.6)	14.7	96.1	(0.5)	97.5	(0.8)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to values both above (conservative) and below (anticonservative) the thresholds designed to maximize concordance with prevalence estimates based on the blinded SCID clinical reappraisal interviews.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).

might be used for case-finding to select people for additional assessment and treatment, it might make more sense to lower the threshold to capture as large a proportion of clinical cases as feasible within the constraints of the cost-benefit ratio of screening and treatment. To investigate the implications of using such a rule in setting

screening scale thresholds, we compared scale operating characteristics when the threshold was selected to detect 80% of DSM-IV/SCID cases (i.e. SN = 80.0%).

This change leads to a lowering of screening scale thresholds for all disorders because SN is consistently lower than 80% at the optimal threshold for estimating SCID prevalence. And this, in turn, leads to substantial increases in screening scale prevalence (2.5–7.0 times the prevalence estimates based on the optimal threshold for estimating SCID prevalence) for all disorders other than PD and IED (where CIDI-SC prevalence estimates increase to 1.2–1.3 times the optimal for estimating SCID prevalence) and to correspondingly large reductions in PPV (Table 4). While PPV at the optimal threshold for estimating SCID prevalence averages 51.6% (i.e. 51.6% of screened positives are true clinical cases, with a range 37.1–65.3%), average PPV drops to 30.0% (range: 11.9–57.1%) when thresholds are selected so that SN exceed 80%. This

Table 4. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from (i) the optimal for estimating prevalence to (ii) having high SN (i.e. detecting at least 80% of DSM-IV/SCID cases) ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR−
Major depressive episode												
Optimal for prevalence	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
High SN	25.2	(2.1)	80.2	(11.8)	21.3	(3.1)	3.8	78.7	(2.1)	98.2	(1.2)	0.3
Mania/hypomania												
Optimal for prevalence	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
High SN	20.5	(2.2)	82.7	(8.4)	20.8	(4.1)	4.8	82.9	(2.1)	98.9	(0.6)	0.2
Panic disorder												
Optimal for prevalence	5.0	(0.9)	58.5	(11.2)	59.5	(7.7)	27.9	97.9	(0.4)	97.8	(0.6)	0.4
High SN[e]	6.4	(0.9)	71.4	(10.4)	57.1	(6.9)	24.6	97.1	(0.5)	98.4	(0.5)	0.3
Generalized anxiety disorder												
Optimal for prevalence	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
High SN	19.1	(2.1)	80.6	(5.2)	28.9	(5.3)	5.5	85.4	(2.3)	98.4	(0.5)	0.2
Post-traumatic stress disorder												
Optimal for prevalence	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
High SN	43.5	(4.0)	81.2	(6.6)	11.9	(2.0)	2.0	59.0	(4.2)	97.9	(0.8)	0.3
Adult attention-deficit/hyperactivity disorder												
Optimal for prevalence	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
High SN	40.3	(2.9)	84.3	(7.5)	14.9	(2.5)	2.3	63.1	(3.1)	98.1	(1.0)	0.2
Intermittent explosive disorder												
Optimal for prevalence	20.8	(2.3)	66.8	(6.2)	65.3	(5.2)	7.3	90.9	(1.6)	91.5	(1.6)	0.4
High SN[e]	26.7	(3.3)	73.5	(7.4)	56.0	(5.7)	5.0	85.3	(2.7)	92.6	(1.9)	0.3
Substance use disorder												
Optimal for prevalence	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
High SN[e]	12.4	(1.6)	66.8	(9.,4)	28.8	(5.4)	7.1	90.6	(1.7)	98.0	(0.8)	0.4

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to have a minimum SN of 80.0% based on the blinded SCID clinical reappraisal interviews.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR+ = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[e]As none of the CIDI-SC thresholds for this disorder had SN as high as 80%, the threshold with the highest SN is reported.

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

means that it would require an average of about three SCID interviews to detect each clinical case among the screened positives at the lower threshold compared to roughly two at the higher threshold. Clinical intervention cost-effectiveness calculations would be needed to determine whether this additional expense of case-finding could be justified based on the human costs (i.e. quality of life, morbidity, mortality) of an untreated case, the costs of treatment, and the likely effectiveness of treatment in reducing human costs. From the perspective of epidemiological research, lowering the thresholds below the optimal for estimating prevalence might still be desirable even though such an anticonservative change introduces upward bias in prevalence estimates, as it is possible that lowering thresholds will lead to greater proportional increases in SN than in (100 − SP), in which case LR + will increase. However, LR + decreases consistently when the screening scale thresholds are lowered, arguing against making these thresholds less conservative for purposes of epidemiological analysis of the Army STARRS data.

Another goal of screening might be to select screening scale thresholds to have a minimum proportion of screened positives confirmed in clinical interviews (i.e. high PPV). For example, minimum PPV might be set at 50% to guarantee that the majority of screened positives are true clinical cases or at 80% to guarantee that the vast majority of screened positives are true clinical cases. However, this will lead to a reduction in SN that might make the true cases detected unrepresentative of all true cases. If minimum PPV is set at 50%, the thresholds selected to maximize estimation of SCID prevalence meet the PPV criterion in five of eight cases, the exceptions being MHM (PPV = 45.8%), GAD (PPV = 45.6%), and ADHD (PPV = 37.1%). In the case of MHM, the threshold can be raised to increase PPV to 71.1%, but this leads to a dramatic reduction in estimated prevalence (from 4.9% to 0.7%) and in SN (from 43.5% to 9.7%) (Table 5). While more than two-thirds of the small fraction of respondents defined as positive for MHM in the CIDI-SC are SCID cases, the exclusion of the vast majority of SCID cases of MHM from this small fraction (100 − SN = 90.3% of SCID cases not detected) means that the proportion of SCID cases among the screened negatives is nearly as high as the proportion among screened negatives (LR − = 0.9), arguing against making the screening scale thresholds this conservative for purposes of epidemiological analysis of the Army STARRS data.

In the case of GAD, raising the CIDI-SC threshold to make PPV exceed 50% leads to halving both estimated prevalence (from 6.6% to 3.2%) and SN (from 43.9% to 23.6%) in the service of only a relatively modest increase in PPV (from 45.6% to 50.2%) compared to when the threshold is set to maximize estimation of SCID prevalence. It is difficult to argue for a threshold that decreases SN so dramatically for such a modest increase in PPV. The situation is similar but less dramatic for ADHD, where a change in the CIDI-SC threshold that increased PPV by roughly 50% (from 37.1% to 55.9%) decreased estimated prevalence by 70% (from 8.2% to 2.4%) and SN by 55% (from 42.8% to 19.3%). Selecting thresholds to have even higher PPV (a minimum of 80%) for disorders where screening scale PPV is greater than 50% at the optimal threshold for estimating SCID prevalence consistently has the same negative effects in that the proportional increases in PPV (in the range 47–59%) are much less than the proportional decreases in prevalence (84–91%), resulting in extremely low levels of SN (7.3–13.4%). These results argue against using such restrictive thresholds for purposes of epidemiological analysis of the Army STARRS data.

Continuous versus dichotomous diagnostic classification

As noted earlier in the section on analysis methods, we calculated ROC curves for the entire screening scale distributions (Figure 1). AUC was calculated for each of these curves and compared to the AUC of the dichotomous version of the same screening scale. AUC was found to be substantially higher for the continuous than dichotomous scoring rule for each of the eight screening scales (Range: 0.80–0.90 continuous versus 0.69–0.79 dichotomous; inter-quartile range: 0.85–0.87 continuous versus 0.70–0.78 dichotomous) (Table 6). This suggests that meaningful variation in SCID prevalence exists at other places on the screening scale ranges than the optimal diagnostic threshold for estimating SCID prevalence. The important implication of this finding for our purposes is that continuous screening scale scores defining predicted probabilities of clinical diagnoses might be more useful than dichotomous diagnostic classifications based on the screening scales for purposes of epidemiological analysis. We consequently calculated both continuous (predicted probability of having a DSM-IV/SCID diagnosis) and dichotomous versions of each screening scale for use in analysis of the Army STARRS data. The continuous versions were produced using the MI method. Importantly, not only the screening scale scores but also a wide range of other significant correlates of the DSM-IV/SCID diagnoses were used in the first-phase MI analysis in order to minimize bias in subsequent substantive analyses that

Table 5. Variation in CIDI screening scale (CIDI-SC) operating characteristics when diagnostic thresholds are changed from (i) the optimal for estimating prevalence to (ii) having high PPV (i.e. at least 80% of screened positives having a DSM-IV/SCID diagnosis) ($n = 460$)[a]

	CIDI-SC prevalence estimate[b]		Positive operating characteristics[c]					Negative operating characteristics[d]				
	Percent	(SE)	SN	(SE)	PPV	(SE)	LR+	SP	(SE)	NPV	(SE)	LR–
Major depressive episode												
Optimal for prevalence	6.8	(1.0)	58.8	(9.0)	57.5	(6.6)	19.0	96.9	(0.8)	97.0	(0.9)	0.4
High PPV	0.6	(0.3)	7.3	(4.1)	84.7	(11.5)	73.0	99.9	(0.1)	93.8	(1.0)	0.9
Mania/hypomania												
Optimal for prevalence	4.9	(1.0)	43.5	(8.9)	45.8	(6.3)	15.5	97.2	(0.6)	96.9	(0.8)	0.6
High PPV[e]	0.7	(0.2)	9.7	(3.0)	71.1	(12.5)	48.5	99.8	(0.1)	95.3	(1.0)	0.9
Generalized anxiety disorder												
Optimal for prevalence	6.6	(0.9)	43.9	(4.7)	45.6	(7.4)	11.5	96.2	(0.8)	95.9	(0.7)	0.6
PPV GT 50%	3.2	(0.6)	23.6	(3.9)	50.2	(8.5)	13.9	98.3	(0.5)	94.6	(0.9)	0.8
High PPV[e]	1.0	(0.3)	10.4	(3.5)	69.1	(12.2)	34.7	99.7	(0.2)	93.8	(0.9)	0.9
Post-traumatic stress disorder												
Optimal for prevalence	6.7	(1.0)	53.7	(6.9)	50.8	(7.0)	15.3	96.5	(0.8)	96.8	(0.6)	0.5
High PPV[e]	1.1	(0.3)	13.4	(3.3)	79.5	(10.5)	67.0	79.5	(10.5)	99.8	(0.1)	0.9
Adult attention-deficit/hyperactivity disorder												
Optimal for prevalence	8.2	(1.1)	42.8	(7.8)	37.1	(7.0)	7.8	94.5	(1.1)	95.6	(1.0)	0.6
PPV GT 50%	2.4	(0.5)	19.3	(3.8)	55.9	(12.0)	16.1	98.8	(0.4)	94.1	(1.0)	0.8
High PPV[e]	2.0	(0.4)	17.8	(3.7)	63.1	(12.7)	22.3	99.2	(0.3)	94.1	(0.9)	0.8
Substance use disorder												
Optimal for prevalence	4.9	(0.4)	47.6	(8.5)	51.6	(8.0)	19.0	97.5	(0.5)	97.0	(0.8)	0.5
High PPV	0.6	(0.2)	8.8	(3.6)	81.7	(13.3)	88.0	99.9	(0.1)	95.1	(0.8)	0.9

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.
[b]The CIDI-SC prevalence estimates are varied by changing the threshold to have a minimum PPV of 80.0% based on the blinded SCID clinical reappraisal interviews. Results are not reported for PD or IED because optimal thresholds for predicting SCID prevalence of these disorders also had the highest values of PPV.
[c]SN = sensitivity (the percent of SCID cases detected by the CIDI-SC); PPV = positive predictive value (the percent of CIDI-SC cases confirmed by the SCID); LR + = likelihood ratio positive (the relative proportions of SCID cases among CIDI-SC cases versus non-cases).
[d]SP = specificity (the percent of SCID non-cases classified as non-cases by the CIDI-SC); NPV = negative predictive value (the percent of CIDI-SC non-cases confirmed as non-cases by the SCID); LR − = likelihood ratio negative (the relative proportions of SCID non-cases among CIDI-SC cases versus non-cases).
[e]As none of the CIDI-SC thresholds for this disorder had PPV as high as 80%, the threshold with the highest PPV is reported.

will use these variables as correlates of predicted probabilities of DSM-IV/SCID disorders.

Discussion

Previous research has shown that CIDI-SC operating characteristics are equivalent to or better than those of alternative screening scales in samples of the general population (Kessler *et al.*, 2005a; Kessler *et al.*, 2006a; Kessler *et al.*, 2013a) and that the PCL has very good concordance with clinical diagnoses of PTSD in samples of both the military and the general population (Wilkins *et al.*, 2011). We nonetheless carried out an independent CRS of these screening scales in Army STARRS due to the fact that the operating characteristics of the same screening scale can differ substantially across surveys depending on such fundamental survey conditions as auspices, level of confidentiality (e.g. complete anonymity versus de-identification), mode of data collection, and situational factors, such as constraint on the amount of time available to complete the survey (Kessler and Pennell, in press).

It is not surprising in light of the challenging survey conditions in Army STARRS – including group-administration in settings with suboptimal physical facilities (e.g. sitting on

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

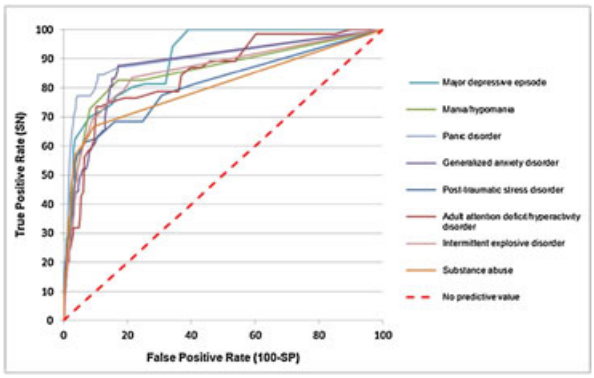

Figure 1. ROC curves for the associations between continuous screening scales and DSM-IV/SCID diagnoses ($n = 460$) ROC = receiver operating characteristic; SN = sensitivity; SP = specificity; DSM-IV = Diagnostic and Statistical Manuel of Mental Disorders, Fourth Edition; SCID = Structured Clinical Interview for DSM-IV.

Table 6. Comparison of area under the receiver operating characteristic curve (AUC) based on the dichotomous versions of the CIDI-SC scales as the optimal thresholds for estimating DSM-IV/SCID prevalence and based on the continuous versions of the CIDI-SC scales ($n = 460$)[a]

	Area under the curve (AUC)	
	Dichotomous	Continuous
I. Mood disorders		
Major depressive episode	0.78	0.90
Mania/hypomania	0.70	0.86
II. Anxiety disorders		
Panic disorder	0.78	0.90
Generalized anxiety disorder	0.70	0.87
Post-traumatic stress disorder	0.75	0.81
III. Externalizing disorders		
Adult attention deficit/ hyperactivity disorder	0.69	0.85
Intermittent explosive disorder	0.79	0.86
Substance use disorder	0.73	0.80

[a]Analyses are based on weighted data to adjust for the over-sampling of respondents screening positive on the CIDI-SC scales.

folding chairs in full field gear in temporary data collection locations) – that we found that the CIDI-SC and PCL AUCs are somewhat lower than in previous psychometric studies of these scales. Individual-level concordance of diagnoses based on the CIDI-SC and PCL with diagnoses based on independent SCID clinical reappraisal interviews in the AAS is for most part *moderate* (AUC = 0.70–0.79; $\kappa = 0.4$–0.6), whereas most previous evaluations found concordance of the CIDI-SC and the PCL with SCID diagnoses to be *substantial* (AUC = 0.80–0.89; $\kappa = 0.6$–0.8). However, the administrative conditions of the screening scales in most previous studies that carried out clinical reappraisals were much better than in Army STARRS, including self-administration in primary care waiting rooms (Kessler *et al.*, 2013a), face-to-face interviewer administration in household surveys (Kessler *et al.*, 2006a), and interviewer administration over the telephone with health plan subscribers (Kessler *et al.*, 2005a).

Perhaps the more striking result in light of the challenging Army STARRS field conditions is that the positive CIDI-SC/PCL operating characteristics for dichotomous versions of the scales designed to optimize aggregate concordance with SCID prevalence estimates are generally quite good. LR + values for six of the eight disorders are in the range 11.5–27.9, all of which are well above the 10.0 value generally considered sufficient to rule in diagnoses (Haynes *et al.*, 2006), while the 7.3–7.8 LR + values for the other two diagnoses and the 8.5 LR + value for any 30-day disorder are well above the 5.0 value considered useful in ruling in diagnoses. However, these good LR + values are accompanied by LR − values generally considered not to be useful in screening out true negatives (0.4–0.6); that is, to contain proportions of true negative that are not strikingly different from the proportions found among screened positives.

As discussed in more detail elsewhere (Kessler *et al.*, 2013a), the definitions of screened positives and screened negatives could be purified for clinical purposes by selecting thresholds at the tails of the distributions that have operating characteristics deemed useful for clinical purposes. For example, an upper threshold of a screening scale could be selected to have a minimum PPV of 0.5 in order to make sure that at least 50% of screened positives are SCID cases. As we saw, though, this desirable feature of that threshold would generally mean that a substantial proportion of SCID cases are missed. Alternatively, the upper threshold of a screening scale could be set at a minimum SN of 0.80 to make sure that the vast majority of SCID cases are picked up by the screen, but this desirable feature of that threshold would mean that only a small proportion of screened positives have SCID diagnoses. In a similar way, a lower threshold of a screening could be purified by requiring NPV to be, say, at least $1 - p/5$, where p = SCID prevalence of the disorder, thereby guaranteeing that the proportion of SCID cases among patients screening negative is no more than 20% as high

as the prevalence of the disorder in the sample, but this desirable feature of that threshold might mean that a substantial proportion of true non-cases are excluded from this ruled-out group.

It is also possible to select multiple thresholds at upper and lower tails both to maximize the positives (i.e. definitive screen-ins and/or screen-outs) and minimize the negatives (i.e. minimizing the numbers of false positives and/or false negatives) and leave one or more intermediate strata that define those with high-but-not-definitively-high scores, low-but-not-definitively-low scores, and uninformative intermediate scores. We noted earlier that such polychotomous scoring rules are fairly common in screening scales developed for clnical practice (Guyatt and Rennie, 2001). Indeed, CIDI-SC polychotomous thresholds have been developed for exactly this reason to facilitate the use of these scales in primary care screening (Kessler *et al.*, 2013a).

However, a more useful approach for purposes of epidemiological analysis of the screening scales considered here is likely to be retention of the entire screening scale range given that AUCs of continuous versions of the screening scales are higher than AUCs of dichotomized versions of the scales at their unbiased thresholds. Based on this observation, we are using MI to assign predicted probabilities of DSM-IV/SCID diagnoses to all Army STARRS respondents who completed the screening scales. We are addressing the uncertainty of inference from prediction equations using imputed rather than observed values by estimating 20 MI estimates of the predicted probability of having each clinical diagnosis for each respondent. The practical use of this approach is illustrated in a more detailed methodological exposition published previously in this journal (Kessler and Üstün, 2004) as well as in a number of subsequent substantive reports that used this approach to estimate the prevalence and correlates of several different DSM-IV/SCID disorders in other psychiatric epidemiological studies (Fayyad *et al.*, 2007; Huang *et al.*, 2009; Kessler *et al.*, 2005a). However, second-phase of MI analysis can be computationally intensive even after the first-phase multiple imputations, as each model has to be estimated 20 separate times rather than once and the coefficients in these 20 replicates then need to be combined to calculate adjusted standard errors. As a result, we also plan to work with dichotomously-scored screening scale measures at the optimal diagnostic thresholds and to investigate the extent to which substantive results differ depending on whether this dichotomous approach is used instead of MI. Dichotomous screening scale scoring will be used in cases where results are relatively insensitive to the more refined estimates using MI.

Acknowledgements

On behalf of the Army STARRS Collaborators

Funding/Support: Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors: As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Additional Contributions: The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine L. Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences);

Int. J. Methods Psychiatr. Res. 22(4): 303–321 (2013). DOI: 10.1002/mpr

Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho N. Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School).

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. First received consultation fees from the Henry M. Jackson Foundation for the Advancement of Military Medicine, the sponsor of the study. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report nothing to disclose.

References

American Association for Public Opinion Research. (2009) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, Deerfield, IL, American Association for Public Opinion Research.

American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, Fourth Edition, Washington, DC, American Psychiatric Association.

Barnes J.B., Nickerson A., Adler A.B., Litz B.T. (2013) Perceived military organizational support and peacekeeper distress: A longitudinal investigation. *Psychological Services*, **10**(2), 177–185, DOI: 10.1037/a0032607

Bray R.M., Pemberton M.R., Lane M.E., Hourani L.L., Mattiko M.J., Babeu L.A. (2010) Substance use and mental health trends among U.S. military active duty personnel: key findings from the 2008 DoD Health Behavior Survey. *Military Medicine*, **175**(6), 390–399.

Breslau N., Peterson E.L., Kessler R.C., Schultz L.R. (1999) Short screening scale for DSM-IV posttraumatic stress disorder. *American Journal of Psychiatry*, **156**(6), 908–911.

Broadhead W.E., Leon A.C., Weissman M.M., Barrett J.E., Blacklow R.S., Gilbert T.T., Keller M.B., Olfson M., Higgins E.S. (1995) Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Archives of Family Medicine*, **4**(3), 211–219.

Bromet E.J., Dunn L.O., Connell M.M., Dew M.A., Schulberg H.C. (1986) Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry*, **43**(5), 435–440.

Brown J.M., Williams J., Bray R.M., Hourani L. (2012) Postdeployment alcohol use, aggression, and post-traumatic stress disorder. *Military Medicine*, **177**(10), 1184–1190.

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46, DOI: 10.1177/001316446002000104

Collins L.M., Schafer J.L., Kam C.M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**(4), 330–351.

Cook R.J. (1998) Kappa and its dependence on marginal rates. In Armitage P., Colton T. (eds) *The Encyclopedia of Biostatistics*, p. 2166–2168, New York, John Wiley & Sons.

Donker T., van Straten A., Marks I., Cuijpers P. (2009) A brief Web-based screening questionnaire for common mental disorders: development and validation. *Journal of Medical Internet Research*, **11**(3), e19, DOI: 10.2196/jmir.1134

Farvolden P., McBride C., Bagby R.M., Ravitz P. (2003) A Web-based screening instrument for depression and anxiety disorders in primary care. *Journal of Medical Internet Research*, **5**(3), e23, DOI: 10.2196/jmir.5.3.e23

Fayyad J., De Graaf R., Kessler R., Alonso J., Angermeyer M., Demyttenaere K., De Girolamo G., Haro J.M., Karam E.G., Lara C., Lepine J.P., Ormel J., Posada-Villa J., Zaslavsky A.M., Jin R. (2007) Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, **190**, 402–409, DOI: 10.1192/bjp.bp.106.034389

First M.B., Spitzer R.L., Gibbon M., Williams J.B.W. (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.

Gaynes B.N., DeVeaugh-Geiss J., Weir S., Gu H., MacPherson C., Schulberg H.C., Culpepper L., Rubinow D.R. (2010) Feasibility and diagnostic validity of the M-3 checklist: a brief, self-rated screen for depressive, bipolar, anxiety, and post-traumatic stress disorders in primary care. *Annals of Family Medicine*, **8**(2), 160–169, DOI: 10.1370/afm.1092

Gossop M., Darke S., Griffiths P., Hando J., Powis B., Hall W., Strang J. (1995) The Severity of Dependence Scale (SDS): psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. *Addiction*, **90**(5), 607–614.

Graham J.W., Olchowski A.E., Gilreath T.D. (2007) How many imputations are really

needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, **8**(3), 206–213, DOI: 10.1007/s11121-007-0070-9

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

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., de Girolamo G., Guyer M.E., Jin R., Lepine J.P., Mazzi F., Reneses B., Vilagut G., Sampson N. A., Kessler R.C. (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *International Journal of Methods in Psychiatric Research*, **15**(4), 167–180, DOI: 10.1002/mpr.196

Haynes R.B., Sackett D.L., Guyatt G.H., Tugwell P. (2006) *Clinical Epidemiology: How to Do Clinical Practice Research*, Third Edition, Philadelphia, PA, Lippincott Williams & Wilkins.

Heeringa S.G., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Kessler R.C. (2013) Field procedures in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 276–287.

Hirschfeld R.M., Williams J.B., Spitzer R.L., Calabrese J.R., Flynn L., Keck P.E., Jr., Lewis L., McElroy S.L., Post R.M., Rapport D.J., Russell J.M., Sachs G.S., Zajecka J. (2000) Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *American Journal of Psychiatry*, **157**(11), 1873–1875, DOI: 10.1176/appi.ajp.157.11.1873

Huang Y., Kotov R., de Girolamo G., Preti A., Angermeyer M., Benjet C., Demyttenaere K., de Graaf R., Gureje O., Karam A.N., Lee S., Lepine J.P., Matschinger H., Posada-Villa J., Suliman S., Vilagut G., Kessler R.C. (2009) DSM-IV personality disorders in the WHO World Mental Health Surveys. *British Journal of Psychiatry*, **195**(1), 46–53, DOI: 10.1192/bjp.bp.108.058552

Hunter E.E., Penick E.C., Powell B.J., Othmer E., Nickel E.J., Desouza C. (2005) Development of scales to screen for eight common psychiatric disorders. *Journal of Nervous and Mental Disease*, **193**(2), 131–135, DOI: 10.1097/01.nmd.0000152786.61048.a1

Jones M., Sundin J., Goodwin L., Hull L., Fear N.T., Wessely S., Rona R.J. (2013) What explains post-traumatic stress disorder (PTSD) in UK service personnel: deployment or something else? *Psychological Medicine*, **43**(8), 1703–1712, DOI: 10.1017/S0033291712002619

Kendler K.S., Neale M.C., Kessler R.C., Heath A.C., Eaves L.J. (1992) A population-based twin study of major depression in women. *The impact of varying definitions of illness. Archives of General Psychiatry*, **49**(4), 257–266.

Kessler R.C., Adler L., Ames M., Demler O., Faraone S., Hiripi E., Howes M.J., Jin R., Secnik K., Spencer T., Ustun T.B., Walters E. E. (2005a) The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. *Psychological Medicine*, **35**(2), 245–256, DOI: 10.1017/S0033291704002892

Kessler R.C., Adler L.A., Gruber M.J., Sarawate C. A., Spencer T., Van Brunt D.L. (2007) Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members. *International Journal of Methods and Psychiatric Research*, **16**(2), 52–65, DOI: 10.1002/mpr.208

Kessler R.C., Akiskal H.S., Angst J., Guyer M., Hirschfeld R.M., Merikangas K.R., Stang P.E. (2006a) Validity of the assessment of bipolar spectrum disorders in the WHO CIDI 3.0. *Journal of Affective Disorders*, **96**(3), 259–269, DOI: 10.1016/j.jad.2006.08.018

Kessler R.C., Berglund P., Demler O., Jin R., Merikangas K.R., Walters E.E. (2005b) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**(6), 593–602, DOI: 10.1001/archpsyc.62.6.593

Kessler R.C., Calabrese J.R., Farley P.A., Gruber M. J., Jewell M.A., Katon W., Keck P.E., Nierenberg A.A., Sampson N.A., Shear M.K., Shillington A.C., Stein M.B., Thase M.E., Wittchen H.U. (2013a) Composite International Diagnostic Interview screening scales for DSM-IV anxiety and mood disorders. *Psychological Medicine*, **43**(8), 1625–1637, DOI: 10.1017/S0033291712002334

Kessler R.C., Coccaro E.F., Fava M., Jaeger S., Jin R., Walters E. (2006b) The prevalence and correlates of DSM-IV intermittent explosive disorder in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **63**(6), 669–678, DOI: 10.1001/archpsyc.63.6.669

Kessler R.C., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Heeringa S.G. (2013b) Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 267–275.

Kessler R.C., Green J.G., Adler L.A., Barkley R.A., Chatterji S., Faraone S.V., Finkelman M., Greenhill L.L., Gruber M.J., Jewell M., Russo L.J., Sampson N.A., Van Brunt D.L. (2010a) Structure and diagnosis of adult attention-deficit/hyperactivity disorder: analysis of expanded symptom criteria from the Adult ADHD Clinical Diagnostic Scale. *Archives of General Psychiatry*, **67**(11), 1168–1178, DOI: 10.1001/archgenpsychiatry.2010.146

Kessler R.C., Green J.G., Gruber M.J., Sampson N. A., Bromet E., Cuitan M., Furukawa T.A., Gureje O., Hinkov H., Hu C.Y., Lara C., Lee S., Mneimneh Z., Myer L., Oakley-Browne M., Posada-Villa J., Sagar R., Viana M.C., Zaslavsky A.M. (2010b) Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. *International Journal of Methods and Psychiatric Research*, **19**(Suppl 1), 4–22, DOI: 10.1002/mpr.310

Kessler R.C., Lane M., Stang P.E., Van Brunt D.L. (2009) The prevalence and workplace costs of adult attention deficit hyperactivity disorder in a large manufacturing firm. *Psychological Medicine*, **39**(1), 137–147, DOI: 10.1017/S0033291708003309

Kessler R.C., Pennell B.-E. (in press) Developing and selecting mental health measures. In T.P. Johnson (ed.) *Handbook of Health Survey Methods*, New York, John Wiley & Sons.

Kessler R.C., Üstün T.B. (2004) The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods and Psychiatric Research*, **13**(2), 93–121, DOI: 10.1002/mpr.168

Kessler R.C., Wittchen H.-U., Abelson J.M., McGonagle K.A., Schwarz N., Kendler K.S., Knäuper B., Zhao S. (1998) Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *International Journal of Methods in Psychiatric Research*, **7**(1), 33–55.

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

Kroenke K., Spitzer R.L., Williams J.B. (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**(9), 606–613, DOI: 10.1046/j.1525-1497.2001.016009606.x

Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174, DOI: 10.2307/2529310

Margolis D.J., Bilker W., Boston R., Localio R., Berlin J.A. (2002) Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology*, **55**(5), 518–524, DOI: S0895435601005121 [pii]

Merikangas K.R., Akiskal H.S., Angst J., Greenberg P.E., Hirschfeld R.M., Petukhova M., Kessler R.C. (2007) Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry*, **64**(5), 543–552, DOI: 10.1001/archpsyc.64.5.543

Nock M.K., Borges G., Bromet E.J., Cha C.B., Kessler R.C., Lee S. (2008) Suicide and suicidal behavior. *Epidemiologic Reviews*, **30**(1), 133–154, DOI: 10.1093/epirev/mxn002

Nock M.K., Deming C.A., Fullerton C.S., Gilman S.E., Goldenberg M., Kessler R.C., McCarroll J.E., McLaughlin K.A., Peterson C., Schoenbaum M., Stanley B., Ursano R.J. (2013) Suicide among Soldiers: a review of psychological risk and protective factors. *Psychiatry*, **76**(2), 97–125, DOI: 10.1521/psyc.2013.76.2.97

Nock M.K., Hwang I., Sampson N., Kessler R.C., Angermeyer M., Beautrais A., Borges G., Bromet E., Bruffaerts R., de Girolamo G., de Graaf R., Florescu S., Gureje O., Haro J.M., Hu C., Huang Y., Karam E.G., Kawakami N., Kovess V., Levinson D., Posada-Villa J., Sagar R., Tomov T., Viana M.C., Williams D.R. (2009) Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO World Mental Health Surveys. *PLoS Medicine*, **6**(8), e1000123, DOI: 10.1371/journal.pmed.1000123

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

Rohde P., Lewinsohn P.M., Seeley J.R. (1997) Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. *American Journal of Psychiatry*, **154**(11), 1593–1598.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley & Sons.

Schafer J.L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**(1), 19–35.

Sobin C., Weissman M.M., Goldstein R.B., Adams P., Wickramaratne P., Warner V., Lish J.D. (1993) Diagnostic interviewing for family studies: comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. *Psychiatric Genetics*, **3**(4), 227–233.

Spitzer R.L., Kroenke K., Williams J.B., Lowe B. (2006) A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, **166**(10), 1092–1097, DOI: 10.1001/archinte.166.10.1092

Ursano R.J., Heeringa S., Stein M.B., Kessler R.C. (submitted for publication) The Army Study to Assess Risk and Resilience in Servicemembers (STARRS).

Weathers F., Litz B., Herman D., Huska J., Keane T. (1993) The PTSD checklist (PCL): reliability, validity, and diagnostic utility. *Annual meeting of the International Society for Traumatic Stress Studies*, San Antonio, TX.

White I.R., Royston P., Wood A.M. (2011) Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, **30**(4), 377–399, DOI: 10.1002/sim.4067

Wilkins K.C., Lang A.J., Norman S.B. (2011) Synthesis of the psychometric properties of the PTSD checklist (PCL) military, civilian, and specific versions. *Depression and Anxiety*, **28**(7), 596–606, DOI: 10.1002/da.20837

Wolter K.M. (1985) *Introduction to Variance Estimation*, New York, Springer-Verlag.