Measuring the level of diagnostic concordance and discordance between modules of the CIDI-Short Form and the CIDI-Auto 2.1

Author/Contributor:
Sunderland, Matthew Martyn; Andrews, Gavin; Slade, Tim; Peters, Lorna

Publication details:
Social psychiatry and psychiatric epidemiology
v. 46
Chapter No. 8
pp. 775-785
1433-9285 (ISSN)

Publication Date:
2011

Publisher DOI:
http://dx.doi.org/10.1007/s00127-010-0247-6

Word Count: 5,263 (excluding abstract and references), 3 Tables.

Running Head: Concordance and discordance between CIDI-SF and CIDI-Auto

Measuring the level of diagnostic concordance and discordance between modules of the CIDI-Short Form and the CIDI-Auto 2.1

Matthew Sunderland[1]

Gavin Andrews[1]

Tim Slade[2]

Lorna Peters[3]

[1] Clinical Research Unit for Anxiety and Depression, University of New South Wales at St. Vincent's Hospital, Sydney, Australia.

[2] National Drug and Alcohol Research Centre, University of New South Wales, Sydney Australia.

[3] Centre for Emotional Health, Department of Psychology, Macquarie University, Sydney, Australia.

Address for Correspondence: Mr. Matthew Sunderland, Clinical Research Unit for Anxiety and Depression, 299 Forbes Street, Darlinghurst NSW 2010, Australia.

Email: matthews@student.unsw.edu.au. Ph: +612 8382 1722. Fax: +612 8382 1721.

ABSTRACT

Purpose: The Composite International Diagnostic Interview- Short Form (CIDI-SF) is a short disorder-specific diagnostic interview for the common mental disorders. Many researchers have been attracted to the CIDI-SF because of its brevity and cost-effectiveness. As a result the CIDI-SF has been used in multiple epidemiological studies and clinical trials. Despite the widespread use, a search of the literature has revealed relatively few validation studies. This investigation aims to provide estimates of concordance and discordance between the CIDI-SF disorder modules and the full CIDI as well as providing evidence regarding the potential screening utility of the CIDI-SF.

Methods: The sample comprised of 83 patients attending a tertiary referral clinic for the anxiety disorders. Patients were administered the CIDI-SF and the full CIDI-Auto and estimates of agreement between the two measures were calculated. Interview transcripts were examined for cases that disagreed on a diagnosis to elicit a likely reason for the lack of agreement between the two measures. Finally, the screening properties of the dimensionally scored CIDI-SF were calculated and compared with the Depression Anxiety Stress Scale.

Results: The CIDI-SF tended to over-estimate the rate of diagnoses as evidenced by a high degree of false positives. However, the CIDI-SF exhibited favorable screening properties (ruling out non-disordered cases).

Conclusions: These results suggest that caution must be taken when using the CIDI-SF as the sole diagnostic instrument in epidemiological research to estimate prevalence and incidence. The CIDI-SF may be more useful for screening out potential candidates in clinical research and psychopharmacological trials.

Key Words: Composite International Diagnostic Interview, Concordance, Discordance, Validity, Screening Utility.

INTRODUCTION

The Composite International Diagnostic Interview- Short Form (CIDI-SF) is a brief, disorder-specific, fully-structured diagnostic interview. It was developed out of the need for quick and cost effective psychiatric assessment in the annual US National Health Interview Survey [1]. The original CIDI-SF was developed using a stepwise regression technique with data from the National Co-morbidity Survey (NCS; [2]) that estimated DSM-III-R prevalence. The aim of the measure was to maintain a small subset of items from the full Composite International Diagnostic Interview (CIDI), which appeared to be highly predictive of a diagnosis, to allow a substantial saving in time while maximizing the overall diagnostic accuracy [3]. The CIDI-SF has since undergone specific alterations and revisions to assess a DSM-IV diagnosis.

Many researchers have been attracted to the CIDI-SF because of its brevity and cost effectiveness. Since its initial development, the CIDI-SF or variants of the original CIDI-SF have been used in a diverse range of settings for a variety of purposes. For instance, the Canadian National Population Health Survey has used the CIDI-SF to assess the prevalence of DSM-IV mood disorders in the general population [4]; the HIV Cost and Services Utilization Study administered the CIDI-SF in a longitudinal study of persons receiving medical care for HIV [5]; the Healthcare for Communities study used the CIDI-SF to rapidly assess the impact of health care delivery on psychiatric disorders [6]; and many independent researchers have implemented the CIDI-SF to assist with the selection of a suitable sample for pharmacological trials and clinical research [7, 8].

In epidemiological settings it is paramount that a diagnostic tool estimates prevalence and incidence rates with a high degree of accuracy. Over- or under-estimation of prevalence has critical implications for developing useful health care policy and assessing the population's need

for mental health prevention and treatment programs [9]. In clinical research a diagnostic tool may not require the same level of accuracy, however the tool should at least adequately screen potential subjects so that researchers are confident that the desired sample is targeted [10, 11]. A critical question is whether the CIDI-SF has the level of accuracy required for use in epidemiological and clinical studies. A search of the literature has uncovered relatively few validation studies of the CIDI-SF and a firm conclusion regarding the validity of the interview has yet to be established.

Kessler and colleagues [3] were the first to show that the disorder modules of the CIDI-SF exhibited an excellent level of diagnostic agreement with the University of Michigan's version of the CIDI (UM-CIDI). The total classification accuracy ranged from 93.2% for depression to 99.9% for agoraphobia. However, since the CIDI-SF and UM-CIDI responses were elicited in a single interview, the authors stated that the concordance estimates "represent part-whole associations between sets of responses to symptom questions collected in a single survey" (p. 182). It was concluded that further independent validation, where both forms of the CIDI are administered separately, is required.

Patten [12, 13] investigated the performance of the CIDI-SF depression module in comparison to the full depression section of the CIDI version 2.1 (CIDI-Auto), administered independently from each other. In both community and clinical samples they found the CIDI-SF tended to over-estimate the true prevalence rate and produced a high number of false positives. This suggests the CIDI-SF is a highly sensitive instrument but not very specific. The majority of false positives were due to the inadequacy of the CIDI-SF to assess the clinical significance of symptoms and their relationship to organic causes (i.e., due to a chronic physical condition, medication use, or bereavement). Furthermore, it appeared that the CIDI-SF did not identify

DSM-IV major depression in the strictest sense but identified a broader category of depressive disorder. These findings have since been replicated in a sample of young adults using the Schedules for Clinical Assessment in Neuropsychiatry as the comparison interview [14].

Finally, four modules of the CIDI-SF (depression, panic disorder, generalized anxiety disorder, dysthymia) were validated against the UM-CIDI in a representative sample of persons in care for HIV [15]. The results provided further evidence that the CIDI-SF was highly sensitive at detecting DSM-III-R diagnoses but not very specific. The authors concluded that the CIDI-SF may be better suited for use as a valid disorder screener instead of a disorder specific diagnostic tool. Burnam and colleagues [16] support this conclusion and showed that the CIDI-SF scales were highly correlated to another screening measure of emotional health when predicting health service use. However, the high degree of physical co-morbidity observed in the sample of persons in care for HIV and the failure to validate the full range of CIDI-SF disorder modules against DSM-IV diagnoses has left a gap in the literature that requires further examination.

The current study aims to investigate the validity and diagnostic performance of all ten CIDI-SF modules in a clinical sample of adults presenting to a tertiary referral anxiety disorders clinic. According to the previous literature, it was hypothesized that the CIDI-SF would identify more false positives in comparison to the full CIDI, predominantly due to a lack of items addressing the clinical significance of symptoms and their relationship with organic causes. Furthermore, the previous literature has suggested that the CIDI-SF may be useful as a general screening measure for the mood and anxiety disorders. Consequently, an additional aim was to directly compare the screening utility of the CIDI-SF with a well-developed general screening measure for mood and anxiety, the Depression Anxiety Stress Scale [17]. It was hypothesized

that the CIDI-SF modules, used as dimensionally scored screening instruments, would possess similar screening properties to the DASS.

METHOD

Sample and setting

The sample comprised of patients referred for treatment to a Sydney Metropolitan Anxiety Disorders clinic between May 1999 and January 2000. During the eight month period assigned for this study, a total of 325 patients were assessed at the clinic. Data for this study were collected from patients across two phases of assessment at the Anxiety Disorders clinic. CIDI-SF data were collected from 305 patients during their first day assessment. The first day assessment is designed to screen patients and determine whether the treatment programs offered at the clinic are suitable for each patient. Those offered treatment were asked to return for a second day assessment approximately three weeks later where a comprehensive diagnosis is determined. It was at this point that CIDI-Auto data were collected on the 103 patients returning for their second day assessment. Due to some missing data on the CIDI-SF, matching CIDI-SF and CIDI-Auto data were limited to a subsample of 83 patients. All subsequent analyses were conducted on this subsample of patients. Post-hoc comparisons were conducted on a range of demographic features and the type of CIDI-SF diagnoses received to confirm that this sample did not significantly differ from the remaining sample of patients that attended the clinic during this period. These comparisons revealed that, on average, the patients in the subsample were approximately four years younger than patients who were not included in this study ($t = 2.64$, $p = 0.009$). Nonetheless, this subsample did not differ in terms of sex ($\chi^2 = 1.06, p = 0.30$), marital status ($\chi^2 = 3.24, p = 0.66$), education ($\chi^2 = 0.03, p = 0.87$), prevalence of any anxiety disorder

($\chi^2 = 1.57, p = 0.21$), affective disorder ($\chi^2 = 0.71, p = 0.40$), and substance disorder ($\chi^2 = 0.55, p = 0.46$).

Measures

CIDI-SF

Modifications and revisions were made to the original CIDI-SF [3] to make it suitable for assessing a probable DSM-IV 12-month diagnosis of the common mental disorders. Social phobia, agoraphobia, panic attack, generalized anxiety disorder (GAD), major depressive episode (MDE), dysthymia, obsessive compulsive disorder (OCD), post-traumatic stress disorder (PTSD), alcohol dependence, and drug dependence were assessed. To reduce respondent burden and total administration time further the CIDI-SF utilizes a stem and branch method of screening patients. If the patient fails to endorse the screening questions they skip from the remaining questions associated with that particular disorder and proceed to the next module.

A probable diagnosis for each disorder is calculated by summing the number of symptom questions endorsed by the patient and applying empirically derived cut-offs, which were designed to maximize concordance with the full DSM-III-R diagnostic criteria (see [3] for a detailed description of how the empirically derived cut-off scores were estimated). The CIDI-SF assesses a probable diagnosis and not a definite diagnosis since the complete diagnostic criteria, such as the clinical significance and associated physical exclusion criteria, are not fully operationalised.

CIDI-Auto

The characteristics of the CIDI-Auto have been well described elsewhere [18]. Briefly, it is a fully-structured self-report diagnostic interview administered via a computer to derive a 12 month diagnosis of the common DSM-IV and ICD-10 mental disorders. In contrast to the CIDI-

SF, the CIDI-Auto extensively assesses each diagnostic criterion, including the symptom criteria and the associated levels of distress, severity, and impairment as well as the relevant exclusion criteria. A stem and branch system is implemented within the CIDI-Auto to reduce respondent burden. There are separate modules to determine the presence of MDE, dysthymia, social phobia, agoraphobia, panic disorder, GAD, OCD, PTSD, alcohol dependence, and drug dependence, in the past 12 months.

It has been shown that the CIDI-Auto possesses adequate validity in comparison to fully-structured and unstructured clinical interviews and excellent test-retest reliability across multiple settings [19-21]. The CIDI-Auto has been used extensively to estimate prevalence in nationally representative epidemiological surveys, for example the 1997 Australian National Survey of Mental Health and Well-being [22]. For these reasons the CIDI-Auto was chosen as the "gold" diagnostic standard in the current study when estimating the diagnostic validity of the CIDI-SF.

DASS

The Depression Anxiety Stress Scale (DASS) was developed to measure the full range of symptoms associated with depression and anxiety. It consists of 42 self-report items answered on a 4 point scale where 0 indicates that the item did not apply to the respondent at all while 3 indicates that the item applies very much to the respondent. Factor analyses have consistently identifying three factors that translate into separate sub-scales representing depression, anxiety, and, stress, which is characteristic of symptoms similar to those observed in GAD [17]. Each of the three DASS sub-scales comprise of 14 items. The responses to each of the 14 items are summed and a total score ranging from 0 to 42 is calculated for all three subscales.

The DASS has excellent internal consistency and favourable test-retest reliability in non-clinical and clinical samples [23]. Furthermore, the DASS is able to significantly differentiate

between patients with panic disorder, GAD, and the mood disorders from other diagnostic groups [24]. The DASS is used here as the "gold" standard when examining the screening utility of the CIDI-SF. As the DASS was not intended to be a measure of the alcohol and substance use disorders the alcohol and drug dependence modules were not included when investigating the screening utility of the CIDI-SF.

Procedure

During the first day assessment, patients were given a written consent form to read and sign stating that any information collected during the diagnostic assessment could be used for research purposes and that publication of results would be in de-identified format. Patients then completed a routine diagnostic assessment that included the DASS, CIDI-SF, the Medical Outcomes Study Short Form Health Survey (SF-12; [25]), and a clinical interview with an experienced consultant psychiatrist. Patients who were offered treatment at the clinic for either anxiety (social phobia, panic disorder, agoraphobia, generalized anxiety disorder) or depression returned for a second, comprehensive diagnostic assessment, approximately three weeks later. Returning patients then completed the CIDI-Auto via self-report computer administration. A research assistant was available if any patient required assistance with the CIDI-Auto.

Statistical Analysis

Concordance Analysis

To generate concordance estimates between the CIDI-SF and CIDI-Auto each module was scored using algorithms that result in 12 month diagnostic variables. Each module was analyzed separately using 2x2 cross tabulations in SPSS version 17. Measures of diagnostic agreement including sensitivity, specificity, positive predictive value, negative predictive value, overall percentage of agreement, and likelihood ratios were calculated to estimate the

independent effect of both false positives and false negatives. However, despite the above indices providing valuable information regarding the level of agreement between the two measures, they do not take into account agreements that occur purely by chance.

To control for the level of chance agreements while maintaining information regarding the relative importance of both false positives and false negatives, weighted kappa's (κ- and κ+) were also calculated [26]. Both indices produce scores ranging from 0 to 1. A low score on κ- indicates a high rate of false negatives while a high score indicates a low rate of false negatives. Therefore, the diagnostic test will possess favourable properties for ruling out the disorder when κ- approaches 1. On the other hand, a low score on κ+ indicates a high rate of false positives while a high score indicates a low rate of false positives. Therefore, the diagnostic test will possess favourable properties for ruling in the disorder when κ+ approaches 1 [27]. To accurately interpret the magnitude of each kappa coefficient, Landis and Koch [28] suggest that there is excellent agreement when kappa exceeds 0.8, substantial agreement between 0.6 and 0.8, moderate agreement between 0.4 and 0.6, and poor agreement when kappa drops below 0.4.

Studies have shown that in some circumstances kappa can be unduly influenced by the initial base prevalence rate of the disorder and therefore comparing two coefficients from separate disorders can be problematic [29, 30]. Furthermore, simulations studies have revealed that small kappa coefficients are possible when the percentages of agreement are high therefore creating disparate results [31-33]. To overcome these issues alternative indices should be generated that are not influenced by the base prevalence rate to clarify the results of kappa. Area under the Receiver Operating Characteristic curve (AUC) was therefore calculated by summing the sensitivity with specificity and dividing by two. The AUC is interpreted as a score between 0.5 and 1.0 with larger values indicating a greater level of agreement between the two measures.

According to Swets [34], an AUC ranging from 0.5 to 0.7 indicates the measure has low accuracy, 0.7 to 0.9 indicates a reasonably useful measure, and an AUC above 0.9 indicates a highly accurate measure.

Discordance Analysis

One of the additional aims of the current study was to quantify the likely reasons for discordant results. Therefore, several possible reasons were specified for why a discrepancy might occur between the two interviews prior to the analysis and each discrepant case was classified accordingly. For each disorder, the CIDI-Auto interview transcripts for false positives and false negatives were compared to the matching CIDI-SF questionnaires to isolate the likely reasons for a discordant result.

The predetermined classification groupings for the discordance analysis are as follows. The two diagnostic measures could totally disagree on a diagnosis as evident by a patient meeting the screening requirements for a disorder on one instrument but not on the other. This category was labeled as total misclassification due to screening. A diagnostic disagreement could also occur when the majority of criteria are endorsed but the patient fails to meet full criteria due to one or two key questions. A diagnostic disagreement like this could be attributed to the patient's pattern of responding, for example the patient may fail to meet criteria for a disorder using the CIDI-Auto but meets criteria in the CIDI-SF due to providing different responses on the same question included in both measures, or the diagnostic disagreement could be attributed to certain differences in the diagnostic instruments, for example the patient fails to meet criteria for a disorder using the CIDI-Auto but meets criteria using the CIDI-SF due to a question being included in the CIDI-Auto but excluded in the CIDI-SF. These two categories were labeled as

partial agreement but patient response disagreement and partial agreement but measurement response disagreement.

Finally, investigating cases with partial agreement allows the discrepancy between the two instruments to be isolated to a specific reason or component of the DSM-IV criteria. Therefore, three specific reasons for disagreement were included in cases with partial agreement, these reasons include; not meeting the required symptom criteria and/or diagnostic thresholds, insufficient levels of severity/distress/interference related to the disorder, and finally not meeting the common exclusion criteria (i.e. physical/medical reasons attributed to symptom, or a co-occurring diagnosis that better explains the symptoms).

Screening Utility

As a means of estimating the screening utility of the CIDI-SF the screening properties of the CIDI-SF were compared with the DASS when predicting a CIDI-Auto diagnosis. First, the CIDI-SF was re-scored similar to a screening instrument by using the sum of the total symptoms endorsed for each disorder rather than generating diagnostic variables using empirically derived cut-points. Each disorder module is scored using a continuous scale that ranges from 0 to the maximum number of questions in each module so that higher scores indicate a higher probability of having the disorder.

Second, Receiver Operating Characteristic (ROC) curves were fitted to each CIDI-Auto disorder separately using the CIDI-SF and the DASS as predictor variables. The ROC curve plots the sensitivity and 1 minus the specificity for every possible cut-off point on the CIDI-SF and DASS. The AUC values associated with each ROC curve are able to estimate the screening tool's ability to accurately predict a diagnosis [35]. For each disorder the two AUC values and 95% confidence intervals were estimated using nonparametric methods and compared using a

technique described by Delong, Delong, and Clarke-Pearson [36]. This technique results in a chi-square (χ^2) difference test so that rejecting the null hypothesis at the nominal 0.05 level indicates that one screening tool has significantly greater predictive power than the other.

RESULTS

Sample Characteristics

This study assessed 83 patients with a mean age of 32 and a range from 18 to 54. 70% of the patients were female. The sample was, as expected, highly disordered with approximately 88% diagnosed by the CIDI-Auto with an anxiety disorder, 42% with a mood disorder, and 20% with one form of substance dependence present in the past 12 months. The rate of co-morbid diagnoses was high, with two diagnoses observed as the median and a range from 0 to 7. The average level of disability attributed to the patient's mental health was 34, as measured by the SF-12, a score that is consistent with treatment seeking clinical populations [37]. Due to missing data some patients were excluded from analyses of specific modules.

Concordance Analysis

The results of the concordance analysis are presented in Table 1. The CIDI-SF consistently diagnosed patients at a higher rate than the CIDI-Auto with the largest discrepancies occurring between the agoraphobia (47.5% vs. 3.8%) and OCD modules (60.2% vs. 13.3%). The weighted kappa coefficients reveal that for all disorders, apart from OCD, the κ- values are within the moderate to excellent range indicating that the CIDI-SF produces relatively few false negatives and possesses favourable properties for ruling out disorders. On the other hand, the κ+ values are within the poor range for all disorders indicating a large number of false positives and that the CIDI-SF possesses undesirable properties for ruling in disorders. The AUC values support the kappa coefficients by providing evidence that dysthymia, OCD, and social phobia

modules have poor diagnostic agreement with the CIDI-Auto but the remaining modules can be considered reasonably useful screening measures. No disorder module scored an AUC value within the excellent range. This is most likely due to the large number of false positives.

Discordance Analysis

The results of the discordance analysis are presented in Table 2. The primary reason for false positive and false negative results differed depending on the disorder under investigation. The few false negatives examined in the discordance analysis were predominantly due to total misclassification between the screening questions. For example, of the 7 false negative cases observed in Panic Disorder, five (71%) were due to failing the CIDI-SF screening question but they met full criteria on the CIDI-Auto and two (29%) partially agreed on a diagnosis yet the patients disagreed on symptom questions contained in both the CIDI-SF and CIDI-Auto.

It is of interest to note that for false positive results, apart from OCD and dysthymia, the CIDI-SF and CIDI-Auto partially agreed on most cases (>50%). This indicates that the CIDI-SF is possibly detecting cases that have some degree of disorder but are considered sub-threshold by the CIDI-Auto. The only disorder to show a large proportion of false positive cases due to the exclusion criteria was agoraphobia with 66% of cases. These cases tended to fail a CIDI-Auto diagnosis because their symptoms were better explained by panic disorder whilst this requirement was not incorporated into the diagnostic algorithms for the CIDI-SF.

All of false positives identified in the dysthymia module were due to total misclassification between the screening items. This may be due to the inability of the CIDI-SF dysthymia screening questions to adequately screen out non-disordered patients in comparison to the screening questions for the CIDI-Auto. The false positives in the remaining disorder modules were due to a mix of failing the symptom and/or threshold criteria, or the symptoms endorsed did

not meet the clinical significance and distress requirements of the CIDI-Auto. In cases with partial agreement, the cause of the disagreement was split between differing patient responses and differences in the two instruments, except for the depression module, where all of the partial agreement cases (65% of all false positive cases) were due to differences between the two instruments.

Screening Utility

The AUC values, 95% confidence intervals, and χ^2 difference tests comparing the CIDI-SF disorder modules and the DASS subscales predicting CIDI-Auto diagnoses are presented in Table 3. The AUC values for the majority of disorders are within the acceptable range except for the CIDI-SF module predicting dysthymia (AUC=0.63), the DASS-Anxiety scale predicting social phobia (AUC=0.50), and the DASS-Anxiety scale predicting OCD (AUC=0.54). The data provide no evidence that the CIDI-SF modules differ in terms of predictive power to the DASS subscales when identifying MDE, dysthymia, GAD, panic disorder, PTSD, and agoraphobia. However, for a diagnosis of social phobia or OCD it appears that the CIDI-SF has a significantly higher predictive power in comparison to the DASS-Anxiety scale, with χ^2=8.09, p=0.005 and χ^2=22.27, p<0.001 respectively.

DISCUSSION

The primary aim of the current study was to investigate the validity and level of diagnostic concordance between the CIDI-SF and the full CIDI-Auto in a sample of patients attending a tertiary referral anxiety disorders clinic. Secondary aims included quantifying the level of discordance and investigating the screening properties of the CIDI-SF. The results of the study confirmed the hypothesis that the CIDI-SF would generate an undesirable proportion of false positives when compared with the CIDI-Auto. This finding provides additional support to

the previous literature [4, 12, 14, 15]. However, the suggestion that false positives were primarily due to the lack of questions assessing the influence of medication and co-morbid physical conditions is unfounded. This is perhaps not surprising given that a previous study has suggested the physical exclusion criteria has little influence on estimating overall prevalence rates when they were removed from the CIDI-Auto algorithms [38]. Instead, the false positive cases were better explained by a variety of reasons that included insufficient screening questions, inaccurate symptom thresholds, and a failure to assess clinically significant levels of distress. Finally, the results of the current study confirmed the hypothesis that the CIDI-SF possesses good screening properties in comparison to a valid general screening measure for anxiety and depression. In fact, the CIDI-SF exhibited significantly better screening properties for social phobia and OCD in comparison to the DASS.

The findings of the current study raise several important points for discussion. It is of some value to understand why the results presented in the current study are different from the results previously found by Kessler and colleagues [3]. There are two possible explanations for the different results. First, it is a well-established fact that clinical respondents often differ in their pattern and degree of responding to diagnostic instruments in comparison to community samples [39]. It is therefore difficult to accurately compare and generalize the findings of this study to epidemiological samples. The high degree of false positives demonstrated here may be isolated to a highly-disordered treatment-seeking clinical population where the prevalence rates of these disorders are higher than those observed in the general population. Fortunately, we can use the likelihood ratio statistic to model the performance of the CIDI-SF using the base prevalence rates found in the 1997 Australian National Survey of Mental Health and Well-being [40]. For instance, the 12 month prevalence for major depressive episode in the Australian

survey was 6.3%, using the likelihood ratio generated from the current sample we would expect the CIDI-SF prevalence to be in the range of 55%, a dramatic over-estimation. Similar results were observed for the remaining disorders suggesting that the clinical sample used in the current study may have exaggerated the level of discordance between the two instruments. On the other hand, evidence from Patten and colleagues [4] suggests that the CIDI-SF depression module likewise identified a high degree of false positives in a community sample of the Canadian general population, although not as high as those observed in the current study.

The second, and more likely explanation, involves the use of different methodologies. Kessler and colleagues [3] calculated the degree of concordance by comparing responses to the CIDI-SF that were embedded within the full CIDI, thus the two interviews were not administered independently. This has the serious implication of artificially inflating the true level of concordance as it does not take into account the influence that prior questioning has on priming responses and establishing the correct context. Survey questionnaires undergo comprehensive cognitive testing during development to address any respondent misunderstandings and to confirm that the questions assess the desired construct using the proper context and format [41]. Shifting the order or removing certain questions can influence the respondents understanding of the questionnaire as a whole and may cause differential response patterns [42].

The next critical issue that requires discussion is how the CIDI-SF can be altered or calibrated to increase the level of accuracy in relation to the full CIDI. Epidemiological studies that quote prevalence estimates derived from the CIDI-SF run the risk of over-estimating the true prevalence and identifying cases with a broader degree of psychopathology. Fortunately, methods to correct for this over-estimation have been developed by adjusting the prevalence post-hoc and have proven to be effective [43]. Alternatively, for studies that are yet to use the

CIDI-SF it may be advantageous to include additional questions addressing problematic areas identified by the current study. For example, the CIDI-SF dysthymia module had 23 false positive cases all due to the failure of the CIDI-SF screening questions to adequately rule out non-disordered cases. After examination of the CIDI-Auto transcripts it was concluded that if the CIDI-SF included one additional question from the CIDI-Auto, designed to screen out individuals if the prolonged feelings of sadness and depression lasted 2 years or more without an interruption of 2 months feeling okay, then the level of false positives may decrease and the concordance between the dysthymia modules could substantially improve. However, this is assuming that the majority of cases would respond to questions in the CIDI-SF and CIDI-Auto in the same manner therefore re-classifying all 23 cases is considered the best case scenario.

The final point for discussion involves the capacity of the CIDI-SF to act as a general screening measure for the anxiety and depressive disorders in a clinical setting, particularly when using the CIDI-SF to screen patients for potential disorder specific CBT treatments and psychopharmacological trials. The findings of the current study demonstrate an interesting phenomenon that the majority of false positive cases (>50%) have some degree of disorder (i.e. the presence of some symptoms, mild levels of distress, etc.), however they fail to meet full diagnostic criteria due to one or two specific reasons (i.e. fail to meet the full symptom threshold, lack clinically significant levels of distress). In a clinical sense this has some key advantages, for one it is important that true positive cases are not left un-detected as greater weight is given to reducing false negatives instead of false positives, and two it has been shown that detecting and treating sub-threshold cases may be an important first step at preventing a full episode from developing [44, 45]. Given that the CIDI-SF has adequate screening properties the current study provides evidence that the CIDI-SF may be better suited as a short screening tool in clinical

settings than for use in large scale community epidemiological studies to accurately estimate prevalence. Indeed, Orlando and colleagues [43] came to a similar conclusion when examining the performance of the CIDI-SF to screen for respondents at high risk for a psychiatric disorder in a population of HIV positive adults.

The findings of the current study must be considered in relation to some of the relevant limitations. The moderate sample size used in the study is of some concern particularly when comparing AUC values. The wide 95% confidence intervals around some of the AUC values indicate the possibility of inadequate statistical power and could have increased the risk of Type II error. However, as the study used a clinical sample to generate the concordance estimates we are confident that there are sufficient cases with and without diagnoses of the examined disorders to make meaningful conclusions. As mentioned above, the use of a clinical sample does carry further limitations when attempting to generalize the findings to a community sample. Therefore, additional validation of the full range of CIDI-SF modules is required to confirm that the findings found here are not specific to a subset of highly disordered treatment seeking individuals.

Another limitation of the current study was that the order of administration of the CIDI-SF and CIDI-Auto was not randomized and the CIDI-SF was administered prior to the CIDI-Auto in all cases. As mentioned above, self-report questionnaires are susceptible to order effects and the results of the CIDI-Auto may be unduly influenced by the outcome of the CIDI-SF as patients attempt to remain consistent. Although, this effect may be somewhat reduced due to the gap between interview administrations of approximately three weeks and therefore some patients may not recall their specific responses to the CIDI-SF. The gap in administration can likewise have an effect on the results due to changes in pathology over time; however, once again this

effect may be reduced due to the 12 month time frame of the CIDI-SF and CIDI-Auto questions. The gap in administration time would have a greater effect if the interviews assessed the disorders as present in the prior 30 days.

Conclusions

The current study is the first to provide DSM-IV diagnostic agreement data between the full range of CIDI-SF disorder modules and the associated disorder sections of the CIDI-Auto in a clinical population. The results highlight the importance of independent validation studies for a new diagnostic instrument and the potential risks that can occur when researchers rely on concordance estimates that are derived using the same sample from which it was constructed. The results indicate that the CIDI-SF possesses favorable properties for ruling out disorders but undesirable properties for ruling in disorders. Consequently, the CIDI-SF is at risk of over-estimating the true prevalence of mental disorders and therefore caution should be taken when the CIDI-SF is used as the sole diagnostic instrument in epidemiological studies. On the other hand, the CIDI-SF does appear to be a valid diagnostic screening instrument that can detect mental disorders at the threshold and to some degree sub-threshold level. Therefore, for clinical research and psychopharmacological trials the CIDI-SF is a valid tool for screening potential candidates.

REFERENCES

1. Adams PF, Marano MA (1995) Current estimates from the National Health Interview Survey, 1994. Vital and health statistics. Series 10, Data from the National Health Survey, 1-260.
2. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS (1994) Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. Arch Gen Psychiatry 51:8-19.
3. Kessler RC, Andrews G, Mroczek DK, Ustun B, Wittchen H (1998) The World Health Organization Composite International Diagnostic Interview Short Form. Int J Meth Psychiat Res 7:171-185. DOI: 10.1002/mpr.47.
4. Patten SB, Brandon-Christie J, Devji J, Sedmak B (2000) Performance of the composite international diagnostic interview short form for major depression in a community sample. Chronic Dis Can 21:68-72.
5. Bozzette SA, Berry SH, Duan N, Frankel MR, Leibowitz AA, Leftkowitz D, Emmons CA, Senterfitt JW, Berk ML, Morton SC, Shapiro MF (1998) The care of HIV-infected adults in the United States: HIV Cost and Services Utilization Study Consortium. New Engl J Med 339:1897-1904.
6. Sturm R, Gresenz CR, Sherboume CD, Minnium K, Klap R, Bhattacharya J, Farley D, Young AS, Burnam MA, Wells KB (1999) The design of Healthcare for Communities: a study of health care delivery for alcohol, drug abuse, and mental health conditions. Inquiry 36:221-233.

7. Calbring P, Forslin P, Ljungstrand P, Willebrand M, Strandlund C, Ekselius L, Anderson G (2002) Is the internet-administered CIDI-SF equivalent to a clinician-administered SCID interview? Cog Behav Therapy 31:183-189. DOI: 10.1080/165060702321138573.
8. Andersson G, Carlbring P, Kaldo V, Strom L (2004) Screening of psychiatric disorders via the Internet: A pilot study with tinnitus patients. Nord J Psychiatry 58:287-291. DOI: 10.1080/08039480410005792.
9. Jenkins R (2001) Making psychiatric epidemiology useful: The contribution of epidemiology to government policy. Acta Psychiat Scand 103:2-14. DOI: 10.1034/j.1600-0447.2001.00004.x.
10. Sayre SL, Evans M, Hokanson PS, Schmitz JM, Stotts AL, Averill P, Grabowski J (2004) "Who gets in?": Recruitment and screening processes of outpatient substance abuse trials. Addict Behav 29:389-398. DOI: 10.1016/j.addbeh.2003.08.010.
11. Adami H, Elliott A, Zetlmeisl M, McMahon R, Thaker G (2002) Use of telephone screens improves efficiency of healthy subject recruitment. Psychiatry Res 113:295-301. DOI: 10.1016/S0165-1781(02)00265-2.
12. Patten SB (1997) Performance of the Composite International Diagnostic Interview Short Form for major depression in community and clinical samples. Chronic Dis Can 18:109-112.
13. Patten SB (2001) Long-term medical conditions and major depression in a Canadian population study at waves 1 and 2. J Affect Disorders 63:35-41. DOI: 10.1016/S0165-0327(00)00186-5.
14. Aalto-Setala T, Haarasilta L, Marttunen M, Tuulio-Henriksson A, Poikolainen K, Aro H, Lonnqvist J (2002) Major depressive episode among young adults: CIDI-SF versus SCAN consensus diagnoses. Psychol Med 32:1309-1314. DOI: 10.1017/S0033291702005810.

15. Orlando M, Burnam MA, Sherboume CD, Morton SC, London AS, Hays RD, Bing EG (2001) Brief screening of psychiatric disorders among a national sample of HIV-positive adults: Concordance between the Composite International Diagnostic Interview (CIDI) and the CIDI Short form (CIDI-SF). Int J Meth Psychiat Res 10:97-107. DOI: 10.1002/mpr.104.
16. Burnam MA, Bing EG, Morton SC, Sherbourne C, Fleishman JA, London AS, Vitiello B, Stein M, Bozzette SA, Shapiro MF (2001) Use of mental health and substance abuse treatment services among adults with HIV in the United States. Arch Gen Psychiatry 58:729-736.
17. Lovibond PF, Lovibond SH (1995) The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 33:335-342. DOI: 10.1016/0005-7967(94)00075-U.
18. World Health Organization (1997) CIDI-Auto 2.1: Administrator's guide and reference. World Health Organization: Sydney.
19. Peters L, Andrews G (1995) Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorder. Psychol Med 25:1269-1280.
20. Peters L, Clark D, Carroll F (1998) Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders. Psychol Med 28:893-901.
21. Andrews G, Peters L (1998) The psychometric properties of the Composite International Diagnostic Interview. Soc Psychiatry Psychiatr Epidemiol 33:80-88. DOI: 10.1007/s001270050026.

22. Andrews G, Henderson S, Hall W (2001) Prevalence, comorbidity, disability and service utilization: Overview of the Australian National Mental Health Survey. Br J Psychiatry 178:145-153. DOI: 10.1192/bjp.178.2.145.
23. Antony MM, Cox BJ, Enns MW, Bieling PJ, Swinson RP (1998) Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. Psychol Assess 10:176-181. DOI: 10.1037/1040-3590.10.2.176.
24. Brown TA, Chorpita BF, Korotitsch W, Barlow DH (1997) Psychometric properties of the depression anxiety stress scales (DASS) in clinical samples. Behav Res Ther 35:79-89. DOI: 10.1016/S0005-7967(96)00068-X.
25. Ware JE, Kosinski M, Keller SD (1996) A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 34:220-233.
26. Kraemer, H. C., Periyakoil, V. S., Noda, A. (2002). Kappa coefficients in medical research. Stat Med 21:2109-2129. DOI: 10.1002/sim.1180.
27. Gilchrist JM (2009) Weighted 2x2 kappa coefficients: recommended indices of diagnostic accuracy for evidence-based practice. J Clin Epidemiol in press.
28. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159-174.
29. Kraemer HC (1979) Ramifications of a population model for Kappa as a coefficient of reliability. Psychometrika 44:461-472. DOI: 10.1007/BF02296208.
30. Streiner DL (2003) Diagnostic tests: Using and misusing diagnostic and screening tests. J Pers Assess 81:209-219.
31. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: 1. The problems of two paradoxes. J Clin Epidemiol 43:543-549. DOI: 10.1016/0895-4356(90)90158-L.

32. Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. J Clin Epidemiol 46:423-429. DOI: 10.1016/0895-4356(93)90018-V.
33. Hoehler FK (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol 53:499-503. DOI: 10.1016/S0895-4356(99)00174-2.
34. Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285-1293.
35. Streiner DL, Cairney J (2007) What's under the ROC? An introduction to receiver operating characteristics curves. Can J Psychiatry 52:121-128.
36. Delong ER, Delong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837-845.
37. Andrews G (2002) A brief integer scorer for the SF-12: validity of the brief scorer in Australian community and clinic settings. Aust NZ J Publ Heal 26: 508-510.
38. Sunderland M, Slade T, Anderson TM, Peters L (2008) Impact of substance-induced and general medical condition exclusion criteria on the prevalence of common mental disorders as defined by the CIDI. Aust NZ J Psychiatry 42:898-904. DOI: 10.1080/00048670802345508.
39. Wittchen HU, Ustun TB, Kessler RC (1999) Diagnosing mental disorders in the community: A difference that matters? Psychol Med 29:1021-1027. DOI: 10.1017/S0033291799008831.
40. Streiner D, Geddes J (1998) Some useful concepts and terms used in articles about diagnosis. Evid Based Mental Health 1:6-8. DOI: 10.1136/ebmh.1.1.6.
41. Collins D (2003) Pretesting survey instruments: An overview of cognitive methods. Qual Life Res 12:229-238. DOI: 10.1023/A:1023254226592.

42. Smith TW (1991) Context effects in the general social survey. In: Beimer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S (eds) Measurement Errors in Surveys. Wiley-Interscience. pp 57-72.
43. Orlando M, Burnam MA, Beckman R, Morton SC, London AS, Bing EG, Fleishman JA (2002) Re-estimating the prevalence of psychiatric disorders in a nationally representative sample of persons receiving care for HIV: Results from the HIV cost and services utilization study. Int J Meth Psychiat Res 11:75-82. DOI: 10.1002/mpr.125.
44. Kessler RC, Merikangas KR, Berglund P, Eaton WW, Koretz DS, Walters EE (2003) Mild Disorders Should Not Be Eliminated from the DSM-V. Arch Gen Psychiatry 60:1117-1122.
45. Preisig M, Merikangas KR, Angst J (2001) Clinical significance and comorbidity of subthreshold depression and anxiety in the community. Acta Psychiat Scand 104:96-103. DOI: 10.1034/j.1600-0447.2001.00284.x.

ACKNOWLEDGEMENTS

This study was funded in part by an Australian Post-graduate Award doctorial scholarship and NHMRC project grant no. 510137. The authors wish to report no conflict of interest.

TABLES

Table 1: Concordance estimates between CIDI-SF and CIDI-Auto modules for the common mental disorders

		CIDI-SF	CIDI-Auto	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Overall Agreement	Likelihood Ratios		Weighted Kappa		Area Under the ROC Curve
Dx	N	**%**	**%**	Se	Sp	PPV	NPV	P	LR+	LR-	κ -	κ +	AUC
Social Phobia	82	73.2	50.0	0.90	0.44	0.62	0.81	0.67	1.61	0.23	0.636	0.233	0.671
Agoraphobia	80	47.5	3.8	1.00	0.55	0.08	1.00	0.56	2.22	0.00	1.000	0.043	0.773
Panic	81	65.4	55.6	0.84	0.58	0.72	0.75	0.73	2.00	0.28	0.550	0.363	0.714
GAD	81	63.0	46.9	0.89	0.60	0.66	0.86	0.74	2.23	0.18	0.716	0.372	0.750
MDE	82	68.3	39.0	0.94	0.48	0.53	0.92	0.65	1.81	0.13	0.803	0.239	0.709
Dysthymia	82	34.1	12.2	0.50	0.68	0.17	0.90	0.66	1.56	0.74	0.241	0.064	0.590
OCD	83	60.2	13.3	0.91	0.44	0.20	0.97	0.50	1.63	0.20	0.771	0.078	0.677
PTSD	78	12.8	6.4	0.60	0.90	0.30	0.97	0.88	6.00	0.44	0.541	0.252	0.752
Alcohol Dependence	83	39.8	14.5	1.00	0.70	0.36	1.00	0.75	3.33	0.00	1.000	0.256	0.852
Drug Dependence	81	28.4	6.2	1.00	0.76	0.21	1.00	0.77	4.12	0.00	1.000	0.166	0.882

Note: Cases with missing data on CIDI-SF disorder modules were excluded.

Table 2: Reasons for discordance between CIDI-SF and CIFI-Auto modules for the common mental disorders

		Social Phobia		**Agoraphobia**	
		False Positives (n=23)	False Negatives (n=4)	False Positives (n=35)	False Negatives (n=0)
Total Misclassification	Screening	9	4	6	0
Partial Agreement - Patient Response Disagreement	Symptoms/Threshold	0	0	1	0
	Severity/Distress/Interference	8	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
Partial Agreement - Measure Response Disagreement	Symptoms/Threshold	1	0	5	0
	Severity/Distress/Interference	4	0	0	0
	Physical/Med/Hierarchy Exclusions	1	0	23	0
		MDE		**Panic Disorder**	
		False Positives (n=26)	False Negatives (n=2)	False Positives (n=15)	False Negatives (n=7)
Total Misclassification	Screening	9	1	1	5
Partial Agreement - Patient Response Disagreement	Symptoms/Threshold	0	1	7	2
	Severity/Distress/Interference	0	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
Partial Agreement - Measure Response Disagreement	Symptoms/Threshold	12	0	4	0
	Severity/Distress/Interference	2	0	3	0
	Physical/Med/Hierarchy Exclusions	3	0	0	0

		OCD		PTSD	
		False Positives (n=40)	False Negatives (n=1)	False Positives (n=7)	False Negatives (n=2)
Total Misclassification	Screening	23	1	0	2
Partial Agreement - Patient Response Disagreement	Symptoms/Threshold	4	0	3	0
	Severity/Distress/Interference	0	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
Partial Agreement - Measure Response Disagreement	Symptoms/Threshold	13	0	3	0
	Severity/Distress/Interference	0	0	1	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0

		Drug Dependence		GAD	
		False Positives (n=18)	False Negatives (n=0)	False Positives (n=17)	False Negatives (n=4)
Total Misclassification	Screening	8	0	6	4
Partial Agreement - Patient Response Disagreement	Symptoms/Threshold	6	0	0	0
	Severity/Distress/Interference	0	0	2	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
Partial Agreement - Measure Response Disagreement	Symptoms/Threshold	4	0	8	0
	Severity/Distress/Interference	0	0	1	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0

		Alcohol Dependence False Positives (n=21)	False Negatives (n=0)	**Dysthymia** False Positives (n=23)	False Negatives (n=5)
Total Misclassification	Screening	1	0	23	5
Partial Agreement - Patient Response Disagreement	Symptoms/Threshold	16	0	0	0
	Severity/Distress/Interference	0	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
Partial Agreement - Measure Response Disagreement	Symptoms/Threshold	4	0	0	0
	Severity/Distress/Interference	0	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0
	Physical/Med/Hierarchy Exclusions	0	0	0	0

Table 3: AUC values and Chi-square difference tests comparing CIDI-SF disorder modules with DASS sub-scales when predicting a CIDI-Auto diagnosis

Diagnosis	Scale	AUC	95% CI	Difference (df=1)
MDE	CIDI-SF Depression Module	0.79	0.70-0.89	χ^2=1.17, p=0.279
	DASS-Mood Scale	0.73	0.62-0.84	
Dysthymia	CIDI-SF Dysthymia Module	0.63	0.46-0.80	χ^2=0.77, p=0.380
	DASS-Mood Scale	0.74	0.59-0.88	
GAD	CIDI-SF GAD Module	0.78	0.69-0.88	χ^2=0.06, p=0.804
	DASS-Stress Scale	0.80	0.70-0.89	
Panic Disorder	CIDI-SF Panic Module	0.75	0.65-0.86	χ^2=0.59, p=0.442
	DASS-Anxiety Scale	0.70	0.58-0.82	
Social Phobia	CIDI-SF Social Phobia Module	0.72	0.62-0.83	χ^2=8.09, p=0.005
	DASS-Anxiety Scale	0.50	0.37-0.63	
OCD	CIDI-SF OCD Module	0.86	0.72-1.00	χ^2=22.27, p<0.001
	DASS-Anxiety Scale	0.54	0.39-0.69	
PTSD	CIDI-SF PTSD Module	0.83	0.60-1.05	χ^2=0.32, p=0.574
	DASS-Anxiety Scale	0.76	0.60-0.91	
Agoraphobia	CIDI-SF Agoraphobia Module	0.84	0.79-0.89	χ^2=2.34, p=0.126
	DASS-Anxiety Scale	0.58	0.24-0.91	