Developing a short-form structured diagnostic interview for common mental disorders using signal detection theory

MATTHEW SUNDERLAND,¹ TIM SLADE² & GAVIN ANDREWS¹

¹ Clinical Research Unit for Anxiety and Depression, School of Psychiatry, 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

Key words

short forms, Composite International Diagnostic Interview, concordance, signal detection theory

Correspondence

M. Sunderland, NDARC, Building R1, University of New South Wales, Randwick Campus, Randwick, NSW 2052, Australia.
Telephone: +612 9385 0106
Email: matthews@unsw.edu.au

Received 20 May 2011;
revised 21 September 2011;
accepted 17 February 2012

Abstract

Diagnostic instruments must be relatively free from respondent burden and cost effective to administer whilst remaining faithful to the psychiatric nomenclature. It seems logical to develop short-form alternatives to rather lengthy and complicated diagnostic interviews to facilitate large scale data collection. The current study examines one method, signal detection theory, for developing a short-form interview based on the Composite International Diagnostic Interview version 3.0. The method was able to retain the smallest number of items to predict a lifetime and 30 day DSM-IV diagnosis for 10 disorders. Concordance analyses between the full-form and the short-form modules, demonstrated an excellent level of agreement in the whole sample and various subsamples of the Australian population as well as in an international comparison sample of the US population. The good concordance between the long form and the short form demonstrates the ability of signal detection theory to assist in the development of valid short forms, which could replace lengthy diagnostic interviews when the aim is to reduce respondent burden and overall research costs. Copyright © 2012 John Wiley & Sons, Ltd.

Introduction

Operationalized diagnostic criteria and structured interviewing has decreased the reliance on trained clinicians for interview administration and the resultant savings in cost has translated into the collection of diagnostic data from large general populations in excess of 10,000 respondents (see Andrews *et al.*, 2001; Grant *et al.*, 2005; Kessler *et al.*, 2004, 2005). Indeed, estimates of prevalence, incidence, and associated risk factors of the common disorders from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) and the International Classification of Diseases, 10th revision (ICD-10) observed in 28

nationally representative surveys, as part of the World Mental Health Survey initiative (Kessler and Üstün, 2008), have all been derived using one structured interview, the World Mental Health Composite International Diagnostic Interview, version 3.0 (CIDI 3.0; Kessler & Üstün, 2004). Despite the proven reliability and validity of the CIDI 3.0, the interview ultimately suffers from a long administration time in order to systematically assess each DSM-IV and ICD-10 criteria, which contributes to undesirable levels of respondent and interviewer burden.

The development of short-form diagnostic instruments has proliferated over the past 20 years in response to the need for more economical methods of psychological

assessment, not only to reduce respondent burden but also to reduce the overall costs of administration. This idea is particularly pertinent in routine clinical practice where the collection of standardized diagnostic information has diminished in recent years due to the added pressures of shorter consultations times and difficulty obtaining reimbursement for psychological assessment (Mullins-Sweatt and Widiger, 2009; Piotrowski, 1999). A short-form version has previously been derived using stepwise logistic regression in an earlier version of the CIDI (known as the University of Michigan CIDI) with independent studies demonstrating the short form is suitable for screening at the broad diagnostic level (Kessler *et al.*, 1998; Sunderland *et al.*, 2011). To date there is no short form available for the widely used CIDI 3.0 to facilitate comparisons between smaller scale studies and the larger World Mental Health surveys.

Signal detection theory (SDT) can be used to successfully identify items that predict a binary outcome and can assist in the development of a short-form structured diagnostic interview. SDT has been previously utilized to evaluate the performance of diagnostic tests in medical research but has since been applied with a form of recursive partitioning, an iterative non-parametric process, to produce a series of binary decision rules that identify homogenous subgroups of cases who are more or less likely to have the examined outcome (Kraemer, 1992). The procedure outputs an empirically derived decision tree that: (1) displays each subgroup identified by the procedure; (2) calculates the probability of the potential outcome associated with each subgroup; (3) identifies the interaction between the predictor variables required to predict subgroup membership. As a result, SDT can handle non-linear interactions between dichotomous variables that generate predicted probabilities for the dependent outcome. In a similar manner to logistic regression, the predicted probabilities for each subgroup can be combined to estimate the prevalence of a given disorder. SDT has been successfully applied in the literature to identify significant predictors associated with late life depression (Andreescu *et al.*, 2008), the prescription of atypical antipsychotics (Hoblyn *et al.*, 2006), retention in opioid agonist therapy (Villafranca *et al.*, 2006), hospital discharge following mechanical ventilation (Kim *et al.*, 2006), and the correlates and determinants of dietary supplement use (Davis *et al.*, 2008).

A study by James *et al.* (2005) compared the performance of SDT and stepwise logistic regression when identifying significant predictors of cognitive impairment. They demonstrated that both procedures performed equally well in terms of predictive accuracy. However, SDT offered several significant advantages over logistic regression that make it

ideal for the purpose of developing a short-form structured diagnostic interview. First, due to the non-parametric nature of recursive partitioning, SDT is able to systematically examine all possible interactions between predictor items, including higher order interactions such as three and four way interactions (Kraemer, 1992). In contrast, higher order interactions must be manually entered into the regression model after the inclusion of all lower order interactions and main effects, which can result in greater Type II error due to a lack of statistical power. In addition, individual researchers must decide what interactions and how many to include, something that can become increasingly impractical in the presence of multiple predictors and can influence the final variable selection (Hosmer and Lemeshow, 1989). It should be noted that the exploratory recursive partitioning method utilized by SDT has the potential to overfit the data, meaning that the predicted probabilities generated in one dataset may not adequately generalize to other independent samples. Thus, stringent stopping rules are required to ensure that the decision trees do not become overly complicated.

Based on the previous evidence, the current study aims to use methods from SDT to develop shortened CIDI 3.0 interview modules for the common DSM-IV mental disorders using data from the Australian 2007 National Survey of Mental Health and Well-being (NSMHWB). The predicted probability for each case generated by the best SDT model will be converted to a binary diagnostic decision to examine the overall performance of a short-form interview in comparison to the long form. In addition, data will be analysed to examine if the short form can adequately generalize to an independent and cross-national sample. Excellent concordance between the short-form and long-form modules will provide sufficient evidence that the method for developing short-form interviews outlined in the current study is effective.

Methods

Sample

The data for the current study were from the 2007 Australian NSMHWB, a cross-sectional household survey of the Australian general population (excluding very remote areas). Private households were randomly selected in each state and territory using a stratified, multistage area design. Of the eligible households selected, there were 8841 households that completed the interview (a response rate of 60%). Socio-demographic information of the final sample representing the Australian population aged 16 to 85 (a total estimated population count of 16,015,000) is presented in Table 1.

Table 1. Socio-demographic characteristics of the 2007 Australian National Survey of Mental Health and Well-being

	Sample size	Weighted percentage (SE)
Age		
16–24	1471	15.9 (0.04)
25–34	1290	17.6 (0.07)
35–44	1638	19.2 (0.10)
45–54	1264	17.9 (0.06)
55–64	1273	14.5 (0.06)
65–74	1104	9.0 (0.06)
75+	801	6.1 (0.03)
Sex		
Male	4021	49.7 (0.01)
Female	4814	50.4 (0.01)
Country of birth		
Australia	6530	72.9 (0.74)
English speaking country	1032	11.3 (0.41)
Other	1279	15.8 (0.70)
Marital status		
Never married	2894	32.5 (0.60)
Widowed	694	4.5 (0.18)
Divorced	915	7.5 (0.35)
Separated	336	2.5 (0.21)
Married	4002	53.0 (0.67)
Employment status		
Employed	5499	65.2 (0.20)
Unemployed	216	2.6 (0.06)
Not in labour force	3126	32.2 (0.18)

Note: weighted percentage is weighted for the age and sex distribution of the Australian general population; SE, standard error.

Measures

The CIDI 3.0 was used as the base instrument to derive psychiatric diagnoses. The CIDI 3.0 possesses sound psychometric properties described elsewhere (Kessler and Üstün, 2004). The instrument underwent minor wording alterations to make it suitable for the Australian context. Further alterations were made to post-traumatic stress disorder (PTSD) questions relating to the respondents' worst traumatic event and the substance use disorders section was altered to correct for sequencing issues. Finally, the Part I/II structure operationalized in the standard CIDI 3.0, where a subsample of respondents is administered additional demographic information, was removed. The diagnoses assessed by the CIDI 3.0 in the Australian survey included; depression, dysthymia, manic episode,

agoraphobia, social phobia, panic disorder, generalized anxiety disorder (GAD), obsessive compulsive disorder (OCD), PTSD, substance use, and substance dependence.

Short-form development

To enhance the validity of the results and protect against over-fitting the model, the total sample from the 2007 Australian NSMHWB was randomly split into two subsamples, labelled the calibration sample and the validation sample respectively (Daumer *et al.*, 2008). The calibration sample was used to conduct the signal detection analysis to identify significant predictors and their interactions separately for each diagnostic module whilst the validation sample was used to estimate the predicted probabilities for each disorder based on the best SDT model. Finally, the validation sample was also used to estimate the concordance rate between the two binary diagnostic variables generated by the short-form and long-form modules.

The CIDI 3.0 uses a stem/branch method of administration in order to reduce the overall length of administration, which means that one or two core items are administered and, depending on the answer, the respondent either skips the remaining items or continues on to the next items. As a result of the skip sequences, the statistical analyses must be run on the subsample of respondents who endorse the skip patterns for the corresponding module given that the respondents who "skip out" do not have complete data and are excluded from receiving a positive diagnosis. The inclusion of these particular items in the final short form is essential. Furthermore, items that were not directly related to a diagnosis were not included in the item pools for the SDT analyses since they could be considered redundant in light of the aim of the study (e.g. questions relating to specific treatment seeking behaviour, effectiveness of help or treatment seeking, hospitalization rates, specific age of onset and age of recency).

Signal detection analysis

For the current study, the outcome variable (the signal) was the lifetime DSM-IV diagnosis variables for each module under investigation. The predictor variables (the detection) were the set of variables that form the selected item pool. The procedure first identifies the optimal cut-point for each predictor variable before identifying subgroups at greater risk for the examined outcome. For binary items the cut-point is naturally defined, however for ordinal or continuous variables the procedure makes use of the receiver operating characteristic (ROC) curve to select the optimal cut-point across all increments of the variable in relation to the outcome. Once this is achieved the procedure then identifies the most predictive variable of a diagnosis

based on the weighted kappa coefficient and divides the sample using the levels of that predictor variable, consequently forming two “branches” of a decision tree.

After the most predictive item is chosen, the procedure continues by splitting the sub-samples identified in the previous step using the same method to further identify additional predictive items until a pre-determined stopping rule is met. Two stopping rules were implemented in the current study, the first rule required the number of cases in at least one cell of the 2×2 cross-tabulation to drop below 10, and second, the chi-square value associated with the 2×2 cross-tabulation fails to reach significance at the 1% level. If either of these two stopping rules were met the procedure ceased the splitting process and finalized the decision tree. Each final subgroup identified by the decision tree is assigned a predicted probability for a diagnosis based on the likelihood for that subgroup as a whole has the dependent outcome. For example, if the decision tree identifies a subgroup containing 200 cases who answered positively to three questions regarding symptoms of depression and of those 200 cases there are 100 cases with a confirmed diagnosis, then the predicted probability of a diagnosis of depression for any individual who answered positively to those three questions would be 0.50 or 50%. These predicted probabilities were assigned to each individual and ROC analysis was used to evaluate the predictive power for each module when estimating a full CIDI 3.0 diagnosis.

Scoring algorithms

Once the best model based on SDT was selected, the predicted probabilities generated by that best model were then converted into a binary diagnostic decision to examine the possibility of using the short-form modules as an interview that generates a positive or negative on each disorder for every individual rather than a series of predicted probabilities. For an initial rudimentary approach, it was decided that any individual below a 50% probability of caseness, identified by the best method, was classified as a non-case and any individual at or above a 50% probability of caseness was classified as a case for the disorder. This resulted in variables that estimated the lifetime short-form prevalence rates that were used for comparison with the full CIDI 3.0 variables, which is described further in the following section.

To estimate 30 day prevalence, the full CIDI 3.0 algorithms code anybody who received a lifetime diagnosis and indicated that they experienced symptoms in the past 30 days as positive for 30 day prevalence. The same technique was applied to the short-form modules to produce 30 day short-form prevalence rates. This requires that one additional question be included in each short-form

module addressing whether the respondent has experienced symptoms in the past 30 days.

Concordance analysis

Australian data

To measure concordance between the short-form modules and the full CIDI 3.0 in the validation sample, simple percentages of agreement were calculated that included; sensitivity, specificity, positive predicted value (PPV), and negative predicted value (NPV). The prevalence estimates and the percentages of agreement were weighted for the Australian general population and standard errors were corrected for any bias arising from the complex sampling design using jack-knife replicate weights (Wolter, 1985).

Weighted 2×2 kappa coefficients, as described by Gilchrist (2009), were calculated to adjust for agreements that occur purely due to chance. High $\kappa+$ values indicate the short-form results in the detection of few false positives whereas high $\kappa-$ values indicate the short-form results in the detection of few false negatives. The area under the ROC curve (AUC) values were calculated for binary variables by summing the sensitivity with the specificity and dividing by two. Since AUC values are not unduly influenced by the initial base prevalence rate, these values were used to clarify the results found by kappa. To interpret the kappa coefficients the current study utilized the specific rules of thumb prescribed by Landis and Koch (1977), who state that excellent agreement is evidenced by a $\kappa \geq 0.8$. The AUC values were interpreted in a similar fashion by using the rules of thumb prescribed by Swets (1988), who state that excellent agreement is evidenced by an $AUC \geq 0.9$. Due to the stringent requirements of the current study that the short form identify virtually the same disordered cases as the CIDI 3.0, it was decided that each module must display kappa and AUC values within the excellent range to be considered as a useful substitute.

To investigate whether the short form could generalize across multiple socio-demographic groups, the AUC values for selected subpopulations were calculated and compared to each other. The population was split according to age (above and below the median age of 43), sex, and country of birth (Australia, other English speaking country, and other non-English speaking country). Furthermore, to examine if the short form classifies respondents differently depending on their overall level of psychological distress, the population was split according to whether they received a high (≥ 30) or not high (< 30) score on the Kessler 10 psychological distress scale (K10; Kessler *et al.*, 2002). Pair-wise comparisons between the examined groups proceeded

by generating the individual AUC values, their associated standard errors, and 95% confidence intervals using a non-parametric method described by Hanley and McNeil (1982). For each comparison the *z*-statistic (i.e. subtracting the two AUC values before dividing by the square root of the sum of the two standard errors) and associated statistical significance level were then calculated (Motulsky, 2007).

American data

As a final test of the generalizability of the short-form modules, the concordance analysis was re-run in an independent and cross-national sample in an attempt to replicate the Australian rates. Data were from the National Comorbidity Survey – Replication (NCS-R), a multistage stratified nationally representative household survey of the US general population, freely accessible in the public domain (for further information on the NCS-R sampling design and field procedures see Kessler *et al.*, 2004). The NCS-R sample contained data from 9282 English speaking adults aged 18 years or over, representing a response rate of 70.9%. For the purposes of the current study all respondents that were administered Part 1 of the survey were utilized in the analysis.

Using the NCS-R survey it was possible to score the data using the short-form modules for the majority of diagnoses, including: major depressive episode (MDE), mania, GAD, social phobia, agoraphobia, panic disorder, alcohol abuse and alcohol dependence. However, due to content changes, it was not possible to score the NCS-R PTSD and dysthymia modules. The prevalence estimates were weighted for the US general population and standard errors were corrected for bias arising from the complex sampling design using the Taylor series linearization method (Wolter, 1985).

Results

Mini-CIDI 3.0 modules

Short-form diagnostic modules were produced for 10 common mental disorders, representing nearly all modules assessed in the Australia 2007 NSMHWB (except the OCD module). The reduction in items from the CIDI 3.0 to the short form for all modules is presented in Table 2 and demonstrates the ability of the signal detection analysis to significantly reduce the number of items required to identify homogenous groups of disordered cases. The total number of items for the CIDI 3.0 was calculated by summing the items required for a diagnosis including the questions required for the skip sequences within the module. For the majority of diagnostic modules, the number of items was reduced by around two to four times that of the original modules. The predicted probabilities for each case generated

Table 2. Comparison of total items administered in the short-form and long-form CIDI 3.0 to generate a diagnosis (Lifetime version)

Disorder	Long form	Short form	Difference
Depression/dysthymia	107	31	3.45
Mania	66	21	3.14
GAD	74	14	5.29
Social phobia	56	20	2.80
Agoraphobia	60	14	4.29
Panic disorder	68	26	2.62
PTSD	99	38	2.61
Alcohol use disorders	45	14	3.21
Total	575	178	3.23

Note: Difference is calculated by dividing number of items from the CIDI 3.0 with number of items from the short-form CIDI 3.0

by the SDT analysis were calculated and used to compare with the full CIDI 3.0 diagnosis in a ROC analysis. The results are presented in Table 3 and indicate that all modules were universally excellent at detecting full CIDI 3.0 disorders (AUC values > 0.9).

Concordance analysis

Australian data

The initial concordance rates between the lifetime and 30 day CIDI 3.0 and the short forms generated by SDT for each diagnostic module in the Australian NSMHWB are presented in Tables 4 and 5, respectively. The scoring

Table 3. Predictive power of short-form modules when estimating the prevalence of full CIDI 3.0 diagnosed disorders

Disorder	AUC	SE	CI
Depression	0.9929	0.0009	0.9911–0.9947
GAD	0.9963	0.0007	0.9950–0.9976
Dysthymia	0.9990	0.0007	0.9976–1.0000
Social phobia	0.9817	0.0035	0.9747–0.9886
Agoraphobia	0.9910	0.0063	0.9783–1.0000
Panic disorder	0.9926	0.0048	0.9832–1.0000
PTSD	0.9812	0.0047	0.9720–0.9904
Manic episode	0.9892	0.0029	0.9836–0.9948
Alcohol abuse	0.9824	0.0022	0.9780–0.9868
Alcohol dependence	0.9448	0.0086	0.9279–0.9616

Note: AUC, area under the receiver operating characteristic curve; SE, standard error; CI, confidence interval.

	Short form	Long form											
	Percentage (SE)	Percentage (SE)	Sensitivity (SE)	Specificity (SE)	PPV (SE)	NPV (SE)	K +	K -	AUC				
	14.0 (0.7)	14.2 (0.7)	0.89 (1.8)	0.98 (0.3)	0.90 (1.7)	0.98 (0.3)	0.87	0.89	0.94				
MDME													
Mania	0.7 (0.2)	0.6 (0.1)	0.63 (10.2)	0.99 (0.1)	0.57 (10.8)	0.99 (0.1)	0.62	0.57	0.81				
Dysthymia	2.8 (0.4)	2.7 (0.4)	1.00 (0.0)	0.99 (0.1)	0.98 (1.5)	1.00 (0.0)	1.00	0.98	0.99				
GAD	7.5 (0.5)	7.4 (0.5)	0.94 (1.6)	0.99 (0.2)	0.92 (2.3)	0.99 (0.2)	0.93	0.91	0.97				
Social phobia	7.9 (0.5)	7.7 (0.4)	0.84 (2.7)	0.98 (0.3)	0.82 (2.9)	0.99 (0.3)	0.82	0.80	0.91				
Agoraphobia	2.4 (0.3)	2.4 (0.3)	1.00 (0.0)	1.00 (0.0)	1.00 (0.0)	1.00 (0.0)	1.00	1.00	1.00				
Panic disorder	3.7 (0.4)	3.6 (0.4)	0.98 (1.4)	0.99 (0.1)	0.96 (1.8)	0.99 (0.1)	0.98	0.96	0.99				
PTSD	7.4 (0.6)	7.3 (0.5)	0.89 (2.1)	0.99 (0.2)	0.88 (1.7)	0.99 (0.2)	0.88	0.87	0.94				
Alcohol abuse	20.2 (0.8)	22.0 (0.9)	0.92 (1.7)	1.00 (0.0)	1.00 (0.0)	0.98 (0.5)	0.90	1.00	0.96				
Alcohol dependence	4.6 (0.5)	4.1 (0.4)	0.71 (4.4)	0.98 (0.3)	0.63 (5.4)	0.99 (0.3)	0.70	0.61	0.85				

Table 5. Concordance rates between modules of the short-form and long-form CIDI 3.0 in a random half of the 2007 Australian National Survey of Mental Health and Well-being (30 day diagnosis).

Note: Prevalence and concordance rates were estimated using weighted data. PPV, positive predicted value; NPV, negative predicted value; AUC, area under the receiver operating characteristic curve; $\kappa+$, weighted 2×2 kappa coefficient with emphasis on detecting false positives; $\kappa-$, weighted 2×2 kappa coefficient with emphasis on detecting false negatives (see Gilchrist, 2009).

algorithms used by both interviews produced similar estimates of lifetime prevalence with a 0.2% difference occurring between the majority of diagnoses and the greatest difference of 1.8% occurring between the alcohol abuse modules. The simple percentages of agreement were uniformly high with scores ≥ 0.8 for all disorders, except for mania and alcohol dependence. The specificity and NPV estimates were considered in the excellent to perfect range indicating that the short-form modules correctly classified almost all cases that were classified as not disordered by the CIDI 3.0. The sensitivity and PPV scores were somewhat lower but still considered within the excellent range indicating that the short form correctly classified a substantial proportion of cases that were classified as disordered by the CIDI 3.0. Controlling for the number of agreements that occurred purely by chance, the weighted kappa's provided further evidence that the majority of diagnostic modules were considered within the excellent range.

Conversely, the performance of the mania and alcohol dependence modules exhibited kappa coefficients ranging between 0.57 and 0.70, despite being considered within the substantial range of agreement as indicated by Landis and Koch (1977), they did not meet the stringent criteria

requiring an excellent level of agreement that was set out at the beginning of the study. The AUC values confirmed the findings of the kappa coefficients with all diagnostic modules within the excellent range of ≥ 0.9 , except for mania and alcohol dependence. The data from the 30 day prevalence estimates provide a similar conclusion.

The AUC values and associated 95% confidence intervals for each lifetime disorder stratified by age, sex, country of birth, and psychological distress are presenting in graphical format in Figure 1. For all disorders no significant differences emerged between AUC values as evidenced by non-significant z-tests and overlapping 95% confidence intervals. The greatest differences between the AUC values were observed in the mania module. Whilst the results suggest there is no significant difference, there is some concern that the small sample of cases with manic episode resulted in insufficient statistical power as evidenced by the large confidence intervals around the AUC values, particularly when stratified by country of birth. Despite these concerns, the lack of any significant differences in the remaining disorder modules indicates that the short form does not disproportionately classify cases depending on group membership in certain socio-demographic features. The

Figure 1 Graphs representing lifetime diagnostic module AUC values and 95% confidence intervals stratified by sex (a), age (b), country of birth (c), and psychological distress (d). Note: MDD, major depressive disorder; SOC, social phobia; GAD, generalized anxiety disorder; DYS, dysthymia; MAN, manic episode; PAN, panic disorder; PTSD, post-traumatic stress disorder; ALA, alcohol abuse; ALD, alcohol dependence.

Table 6. Concordance rates between modules of the short-form and long-form CIDI 3.0 in Part 1 of the US National Comorbidity Survey – Replication (Lifetime diagnosis)

	Short form	Long form								
	Percentage (SE)	Percentage (SE)	Sensitivity (SE)	Specificity (SE)	PPV (SE)	NPV (SE)	$\kappa +$	$\kappa -$	AUC	
MDE	19.1 (0.5)	19.2 (0.5)	0.90 (0.8)	0.98 (0.2)	0.90 (0.8)	0.98 (0.2)	0.88	0.88	0.94	
Mania	1.2 (0.1)	3.5 (0.2)	0.25 (2.2)	0.99 (0.1)	0.74 (4.7)	0.97 (0.2)	0.24	0.73	0.62	
GAD	8.3 (0.3)	7.8 (0.3)	0.90 (1.4)	0.99 (0.1)	0.84 (1.4)	0.99 (0.1)	0.89	0.83	0.94	
Social phobia	13.1 (0.4)	12.1 (0.4)	0.89 (1.0)	0.97 (0.2)	0.82 (1.3)	0.98 (0.2)	0.87	0.80	0.93	
Agoraphobia	2.4 (0.2)	2.4 (0.2)	0.95 (1.8)	0.99 (0.0)	0.95 (0.2)	0.99 (0.0)	0.95	0.95	0.98	
Panic disorder	5.2 (0.2)	4.7 (0.2)	0.97 (0.8)	0.99 (0.1)	0.88 (1.5)	0.99 (0.0)	0.97	0.88	0.98	
Alcohol abuse	11.5 (0.5)	11.7 (0.5)	0.98 (0.4)	1.00 (0.0)	1.00 (0.0)	0.99 (0.1)	0.98	1.00	0.99	
Alcohol dependence	5.3 (0.3)	5.1 (0.3)	0.81 (1.8)	0.99 (0.1)	0.77 (2.3)	0.99 (0.1)	0.80	0.76	0.90	

Note: Prevalence and concordance rates were estimated using weighted data. PPV, positive predicted value; NPV, negative predicted value; AUC, area under the receiver operating characteristic curve; $\kappa +$, weighted 2×2 kappa coefficient with emphasis on detecting false positives, $\kappa -$, weighted 2×2 kappa coefficient with emphasis on detecting false negatives (see Gilchrist, 2009).

agoraphobia module was not included in the analyses because the diagnostic agreement was found to be perfect and therefore all agoraphobic cases identified by the CIDI 3.0, regardless of any socio-demographic feature, were identified by the short forms.

International comparison data (NCS-R)

The concordance rates for selected disorders, generated by running the lifetime short-form diagnostic algorithms over the NCS-R sample, are presented in Table 6. The percentages of agreement were again well within the excellent range for the majority of disorders, except for mania and alcohol dependence. The weighted kappa coefficients and AUC values support this finding with the majority generating values ≥ 0.8 and ≥ 0.9 , respectively. Interestingly, the concordance rates exhibited by the alcohol dependence module in this sample were greater than the rates generated in the Australia sample, with an AUC value of 0.90 compared to an AUC of 0.85. The mania module, however, once again displayed undesirable results with weighted kappa coefficients below 0.8 and AUC values below 0.9, more importantly the mania module appeared to perform worse in the US sample in comparison to the Australian sample. In general, the results from the US data provided further evidence that the short forms could accurately identify disordered and non-disordered cases in comparison to the CIDI 3.0. Furthermore, the results indicate that the short-form scoring algorithms are not specific to the Australian general population and provide evidence that the findings can easily generalize to another independent and cross-national sample.

Discussion

Using data from the 2007 Australian NSMHWB, our results were able to demonstrate that applying SDT could substantially reduce the number of items required to accurately generate a diagnosis, representing approximately a three-fold decrease in the number of items. Re-scoring the data using the reduced item set combined with the converted predicted probability to produce a binary diagnosis and assuming the CIDI 3.0 as the “gold” standard, revealed that a large proportion of cases were correctly classified by the short form as either present or absent for the disorder in the total Australian population, in different socio-demographic subpopulations, and in an independent sample of the US general population. However, the current study utilized a rather rudimentary approach to convert the predicted probabilities to a diagnostic decision to make it suitable for quick and easy scoring. Other more sophisticated approaches to convert the predicted probabilities into an accurate binary decision (i.e. approaches that weigh up the costs and benefits

of false positive and false negative diagnostic decisions) will likely result in greater concordance since the results using the raw predicted probabilities to predict a diagnosis were universally excellent for almost all disorders ($AUC \geq 0.98$). This indicates that the short form and long form could be used interchangeably to identify significantly disordered cases, which justifies the use of SDT as an effective method for constructing short-form diagnostic interviews.

The results produced by the signal detection analysis indicate not only a reduction in items but also a reduction in respondent burden and overall length of administration. Hypothetically, let us consider that the length of administration for the CIDI 3.0 was on average 90 minutes, then using the total number of items displayed in Table 2, this represents on average 9.4 seconds per item. Extrapolating this ratio to the reduced set of items indicates that the average saving in time will be around an hour, with the short form requiring on average 28 minutes to administer, a substantial saving in time. However, there is a need for additional research to empirically compare the length of administration required for both interviews, administered independently from each other, before a firm conclusion can be drawn.

The reduction in items from the full CIDI 3.0 to produce the short-form interview does require a cautionary note relating to clinical relevance. The modules were developed under the assumption that the majority of cases in the population identified by the DSM criteria are homogenous in nature and therefore some items could be removed or not administered once sufficient evidence was found to make a diagnosis. However, in practice, cases of disorder tend to be heterogeneous and many items that were removed from the short-form modules may be useful when distinguishing between qualitatively different cases of the same disorder. As a result, it is recommended that the short-form modules be used to estimate broader disorder prevalence and incidence rates in various populations rather than as the sole source to guide clinical treatments and practice. The results from the short-form modules as well as further examination once the individual has received a short-form diagnosis is required in order to generate a complete clinical picture.

One methodological limitation of the current study that must be taken in account involves the possibility that the concordance rates between the CIDI 3.0 and the short form may have been over-estimated due to the analysis of *part-whole* relationships. Similar to the limitation highlighted in the Kessler *et al.* (1998) study, the current study displays concordance rates between two diagnostic modules generated via one interview administration. Any context bias or systematic errors that may occur in a sample during independent administration have not been taken into account (Coste *et al.*, 1997). For example, it has been shown

that certain items can influence the pattern of responding to other items that follow; if these items are removed then it could potentially alter the pattern of responding to items that remain (Tourangeau and Rasinski, 1988; Tourangeau *et al.*, 2000). Without administering the short form independently from the CIDI 3.0 there is no way of knowing how much this limitation has influenced the current results.

Smith *et al.* (2000) warn that part-whole relationships should be avoided when attempting to make a final conclusion regarding the short-form validity. Instead, they argue that part-whole relationships may be better suited to providing an initial indication of short-form performance during the pilot stage of development. The primary goal of the current study was to demonstrate the ability of SDT to select a reduced set of items for each module and clarify that the concordance rates were sufficiently high to justify their selection. A stringent cutoff criterion requiring that each module exhibit an excellent level of agreement was set as one attempt to take into account the possibility that the results were over-estimated due to the part-whole relationships. Furthermore, the analysis was conducted using a split-half validation method and the concordance estimates were run across different subpopulations and across an independent cross-national sample in an attempt to provide a further justification for the reduced item selection. These features are considered as significant strengths of the current study particularly since the current study aimed to examine the performance of the short-form development method and not the actual validity of the short-form modules.

It should be noted that the current study implemented only one procedure amongst a range of alternative procedures that are suitable for data mining and predictive modelling. The aim of this paper was to demonstrate the applicability of these methods in order to generate a short-form diagnostic interview that has strong concordance with the full version but the added benefit of possessing a significant reduction in administration time. SDT was selected as one method to demonstrate this aim based on previous comparisons in the literature with logistic regression. In contrast, there are other methods such as Random Forest analysis and computerized adaptive testing that may show similar benefits when developing short-form interviews and further investigation of these techniques is required to demonstrate their effectiveness.

The promising initial results regarding the performance of the short forms provide a tentative conclusion indicating that it is feasible to design a short-form diagnostic interview using SDT. The method presented in this study could be replicated to design additional short-form diagnostic interviews as well as contributing to the

development of short forms of revisions made to the full-form CIDI, this is rather pertinent given the latest revision of the diagnostic manuals is approaching and new instruments that operationalize the revised criteria will need to be developed. Nevertheless, the earlier limitations have drawn our attention to the need for additional research to examine if the short-form modules continue to exhibit the good validity and reliability shown by the long forms. Predominantly, a detailed examination of the short-form item content, followed by subsequent revisions, is called for to address any misinterpretations of particular items when the short-form interview is run

independently from the full form. Likewise, additional independent psychometric testing of the short forms is needed to establish an accurate representation of the reliability, validity, and utility.

Acknowledgements

The research was funded in part by an Australian Postgraduate Awards scholarship.

Declaration of interest statement

The authors have no competing interests.

References

- Andresescu C., Mulsant B.H., Houck P.R., Whyte E.M., Mazumdar S., Dombrovski A.Y., *et al.* (2008) Empirically derived decision trees for the treatment of late-life depression. *The American Journal of Psychiatry*, **165**(7), 855–862, DOI: 10.1176/appi.ajp.2008.07081340
- Andrews G., Henderson S., Hall W. (2001) Prevalence, comorbidity, disability and service utilisation: overview of the Australian National Mental Health Survey. *The British Journal of Psychiatry*, **178**(2), 145–153.
- Coste J., Guillemin F., Pouchot J., Fermanian J. (1997) Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, **50**(3), 247–252, DOI: 10.1016/S0895-4356(96)00363-0.
- Daumer M., Held U., Ickstadt K., Heinz M., Schach S., Ebers G. (2008) Reducing the probability of false positive research findings by pre-publication validation – experience with a large multiple sclerosis database. *BMC Medical Research Methodology*, **8**, 18, DOI:10.1186/1471-2288-8-18.
- Davis R.E., Resnicow K., Atienza A.A., Peterson K. E., Domas A., Hunt A., *et al.* (2008) Use of signal detection methodology to identify subgroups of dietary supplement use in diverse populations. *Journal of Nutrition*, **138**(1), 205S–211S.
- Gilchrist J.M. (2009) Weighted 2 × 2 kappa coefficients: recommended indices of diagnostic accuracy for evidence-based practice. *Journal of Clinical Epidemiology*, **62**(10), 1045–1053.
- Grant B.F., Hasin D.S., Stinson F.S., Dawson D.A., Ruan W.J., Goldstein R.B., *et al.* (2005) Prevalence, correlates, co-morbidity, and comparative disability of DSM-IV generalised anxiety disorder in the USA: results from the National Epidemiologic Survey of Alcohol and Related Conditions. *Psychology Medicine*, **35**(12), 1747–1759.
- Hanley J.A., McNeil B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1), 29–36.
- Hoblyn J., Noda A., Yesavage J.A., Brooks III J.O., Sheikh J., Lee T., *et al.* (2006) Factors in choosing atypical antipsychotics: toward understanding the bases of physicians' prescribing decisions. *Journal of Psychiatric Research*, **40**(2), 160–166, DOI: 10.1016/j.jpsychires.2005.06.004
- Hosmer D.W., Lemeshow S. (1989) *Applied Logistic Regression*. New York, John Wiley & Sons.
- James K.E., White R.F., Kraemer H.C. (2005) Repeated split sample validation to assess logistic regression and recursive partitioning: an application to the prediction of cognitive impairment. *Statistics in Medicine*, **24**(19), 3019–3035, DOI: 10.1002/sim.2154
- Kessler R.C., Andrews G., Colpe L.J., Hiripi E., Mroczek D.K., Normand S.L., *et al.* (2002) Short screening scales to monitor population prevalence and trends in non-specific psychological distress. *Psychology in Medicine*, **32**(6), 959–976, DOI: 10.1017/S0033291702006074.
- Kessler R.C., Andrews G., Mroczek D.K., Üstün B., Wittchen H. (1998) The World Health Organization Composite International Diagnostic Interview Short Form. *International Journal of Methods in Psychiatric Research*, **7**(4), 171–185, DOI: 10.1002/mpi.47
- Kessler R.C., Berglund P., Chiu W.T., Demler O., Heeringa S., Hiripi E., *et al.* (2004) The US National Comorbidity Survey Replication (NCS-R): design and field procedures. *International Journal of Methods in Psychiatric Research*, **13**(2), 69–92, DOI: 10.1002/mpi.167
- Kessler R.C., Chiu W.T., Demler O., Walters E.E. (2005) Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**(6), 617–627.
- 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 in Psychiatric Research*, **13**(2), 93–121, DOI: 10.1002/mpi.168
- Kessler R.C., Üstün T.B. (2008) *The WHO World Mental Health Survey: Global Perspectives on the Epidemiology of Mental Disorders*. New York, Cambridge University Press.
- Kim Y., Hoffman L.A., Choi J., Miller T.H., Kobayashi K., Donahoe M.P. (2006) Characteristics associated with discharge to home following prolonged mechanical ventilation: a signal detection analysis. *Research in Nursing and Health*, **29**(6), 510–520, DOI: 10.1002/nur.20150
- Kraemer H. (1992) *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Newbury Park, CA, Sage Publications.
- Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.
- Motulsky H. (2007) *Prism 5 Statistical Guide*. San Diego, CA, Graphpad Software Inc.
- Mullins-Sweatt S.N., Widiger T.A. (2009) Clinical utility and DSM-V. *Psychological Assessment*, **21**(3), 302–312.
- Piotrowski C. (1999) Assessment practices in the era of managed care: current status and future directions. *Journal of Clinical Psychology*, **55**(7), 787–796, DOI: 10.1002/(SICI)1097-4679(199907)
- Smith G.T., McCarthy D.M., Anderson K.G. (2000) On the sins of short-form development. *Psychological Assessment*, **12**(1), 102–111.

- Sunderland M., Andrews G., Slade T., Peters L. (2011) Measuring the level of diagnostic concordance and discordance between modules of the CIDI-Short Form and the CIDI-Auto 2.1. *Social Psychiatry and Psychiatric Epidemiology*, **46**(8), 775–785.
- Swets J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**(4857), 1285–1293.
- Tourangeau R., Rasinski K.A. (1988) Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, **103**(3), 299–314, DOI: 10.1037/0033-2909.103.3.299
- Tourangeau R., Rips L.J., Rasinski K.A. (2000) *The Psychological of Survey Response*. New York, Cambridge University Press.
- Villafranca S.W., McKellar J.D., Trafton J.A., Humphreys K. (2006) Predictors of retention in methadone programs: a signal detection analysis. *Drug and Alcohol Dependence*, **83**(3), 218–224, DOI: 10.1016/j.drugalcdep.2005.11.020
- Wolter K. (1985) *Introduction to Variance Estimation*. New York, Springer-Verlag.