Title

PROMIS depression measures perform similarly to legacy measures relative to a structured diagnostic interview for depression in cancer patients.

Authors

Kerrie Clover, BScHons, PhD, MPsychClin, MAPS [a b]

Sylvie D. Lambert, RN, PhD [c]

Dr Christopher Oldmeadow, BMath(Hons) PhD[d]

Benjamin Britton, DPsych [a]

Madeleine T. King, BSc Hons PhD [e]

Alex J Mitchell, MBBS; MSc; MD; MRCPsych [f]

Gregory Carter MBBS, FRANZCP, Cert Child Psych, PhD [a b]

a Psycho-Oncology Service, Calvary Mater Newcastle, New South Wales, Australia 2298.

b Centre for Translational Neuroscience and Mental Health, Faculty of Health and Medicine University of Newcastle, New South Wales, Australia 2308

c Ingram School of Nursing, McGill University & St. Marys' Research Centre, Wilson Hall, 3506 Sherbrooke Street, Montreal, Quebec, Canada, H3A 2A7

[d] Clinical Research Design, Information Technology and Statistical Support (CReDITSS) Hunter Medical Research Institute, NSW Australia,

[e] School of Psychology and Sydney Medical School, Level 6 North, Chris O'Brien Lifehouse (C39Z), University of Sydney NSW 2006 Australia

[f] University of Leicester, Cancer & Molecular Medicine, University Road, Leicester, United Kingdom LE1 5WW

Acknowledgments

We would like to thank our Psychologist interviewers and all of the participants who gave so generously of their time. Thanks also to Jessica Searston for assistance with manuscript preparation.

Abstract

Purpose: To assess the convergent validity of the Patient Reported Outcomes Measurement Information System (PROMIS) depression measures relative to legacy measures and criterion validity against a structured diagnostic interview for depression in an oncology sample.

Methods: 132 oncology/haematology outpatients completed the PROMIS Depression Computer Adaptive Test (PROMIS-D-CAT) and PROMIS Depression Short Form (PROMIS-D-SF) along with seven legacy measures: Beck Depression Inventory (BDI); Centre for Epidemiological Studies-Depression (CES-D); Depression, Anxiety and Stress Scale; Hospital Anxiety and Depression Scale; Patient Health Questionnaire; Distress Thermometer and PSYCH-6. Correlations, area under the curve (AUC) and diagnostic accuracy statistics were calculated with Structured Clinical Interview as the gold standard.

Results: Both PROMIS measures correlated with all legacy measures at $p<.001$ (rho=.589-.810) and all AUCs (>.800) were comparable. At the cut-off points for mild depression of 53 the PROMIS measures had sensitivity (.83 for PROMIS-D-CAT and .80 for PROMIS-D-SF) similar to or better than 6/7 legacy measures with high negative predictive value (> 90%). At cut-off points of 60 for moderate depression PROMIS measures had specificity >90%, similar to or better than all legacy measures and positive predictive value >=.50 (similar to 5/7 legacy measures).

Conclusions: The convergent and criterion validity of the PROMIS depression measures in cancer populations was confirmed, although the optimal cut-off points are not established. PROMIS measures were briefer than BDI-II and CES-D but do not offer any advance in terms of diagnostic accuracy, reduced response burden or cost over other legacy measures of depression in oncology patients.

INTRODUCTION

The importance of attending to cancer patients' emotional needs is well recognised [1]. As emotional distress often goes unrecognised and untreated [2] the use of validated measures to identify emotional concerns is recommended [3, 4]. Depression is an important aspect of emotional distress in cancer, because it is common, treatable and has serious negative effects if left untreated [1]. Numerous measures of depression symptomatology are available with acceptable psychometric properties; however, no measure has obtained optimal (perfect) performance in terms of diagnostic accuracy with minimal response burden and there is scope to develop improved measures [5, 6, 7].

An approach with considerable potential is the PROMIS initiative, a multi-centre collaboration aiming "to revolutionise the assessment of patient-reported outcomes for use in clinical research and healthcare delivery" through developing new measures [8]. Two types of PROMIS measures are available for depression: a "short form" (PROMIS-D-SF) measure in the traditional (static) format and a measure using computer adaptive technology (CAT)(PROMIS-D-CAT). CAT offers an advance in administering measures because it tailors the items to the patient's responses, using iterative logic, plus pre-determined calibrations of item pools for specific constructs such as depression. CAT measures reduce response burden by enabling the smallest possible number of items to be used [9].

While PROMIS measures show considerable potential, their validity in the emotional domain has not been widely established in oncology. Of 15 publications involving PROMIS depression or anxiety measures in oncology, only Baum et al [10] compared the PROMIS measures against another depression measure (convergent validity). They compared the Brief Symptom Inventory (BSI) depression subscale with the PROMIS-D-CAT in prostate cancer patients (n=136) with non-metastatic disease. They found a significant correlation (.85)

between the measures and an area under the curve of .966 against the BSI depression subscale. Thus, validation of the PROMIS instruments against a wider range of measures in oncology is needed.

The criterion validity of the PROMIS depression measures in cancer has also not been well established. Cella et al [11] used a clinical vignette study to propose clinical severity ratings for depression scores. However, we could not locate any published studies that compared the PROMIS depression measures to a structured diagnostic interview in an oncology population, which would provide a more robust method for determining criterion validity.

The increasing adoption of screening for emotional distress among cancer patients as part of routine care [12] drives the imperative for instruments with excellent psychometric performance, as the costs of false-positive and false-negative results are amplified by large scale implementation. Four statistics are commonly calculated to determine the (diagnostic) performance of measures: sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity is the ability of a measure to correctly detect the presence of disorder (that is, the percentage of people with a disorder that have the disorder detected by the measure). Specificity is the ability of a measure to detect the disorder of interest and only that disorder (the percentage of people without a disorder that the measure also classifies as not having a disorder). Positive predictive value reflects how likely an individual with a positive result on a measure is to have the disorder (the percentage of people that the measure classifies as having a disorder that truly have the disorder) and negative predictive value (NPV) reflects how likely an individual who tests negative is to not have the disease (the percentage of people that the measure classifies as not having a disorder that are truly do not have the disorder). Instruments with high sensitivity and high negative predictive validity (NPV) are required to ‘rule-out’ the likelihood of depression with minimal chance of false negatives [13]. Conversely, instruments with high specificity and high

positive predictive value (PPV) are needed to efficiently flag possible cases of depression for follow-up [13].

Brevity is also important for a depression screening tool in oncology since time constraints are a major barrier to integration in clinical practice [14]. The cumulative time saving associated with even modest reductions in length can be considerable [15]. CAT instruments may be ideally suited for screening, since they tailor the measure to ask the minimum number of items [9].

However, the functioning of the PROMIS-D-CAT in oncology patients has not been widely described. Only three studies, all in the USA, report using the PROMIS-D-CAT among oncology populations [10, 16, 17]. PROMIS depression and anxiety CAT measures have been used for screening [16] with a T-score in the severe range triggering automatic electronic notification to health care team members. Stone et al [16] compared a daily diary of symptoms to a weekly CAT assessment in a longitudinal study of 86 women with breast cancer undergoing chemotherapy. The study by Baum et al [10] is described above. Only one study reported the number of items administered by CAT [10].

AIM

Our primary aim was to examine the convergent and criterion validity of the PROMIS-D-SF and PROMIS-D-CAT among oncology outpatients. Convergent validity was examined using correlation with seven 'legacy' measures of depression and emotional distress. Criterion validity was examined in two ways. One assessed the two PROMIS measures against the gold standard of a structured interview diagnosis of depression, using area under the receiver-operator characteristic (ROC) curves. The other assessed diagnostic accuracy (sensitivity,

specificity, NPV, PPV) using proposed severity ratings [8]. A secondary aim was to describe the functioning of the PROMIS-D-CAT (score distribution, number of items administered), since the use of CAT measures is still relatively new to oncology settings.

METHODS

Data were from a larger study where we conducted a Rasch analysis of five depression measures in 162 cancer patients [6]. We report on the subset of patients (n=132) who additionally completed the PROMIS items.

Patients and Setting

A convenience sample of 132 oncology outpatients from a regional Australian cancer centre. Exclusion criteria: first clinic visit; insufficient English language skills or too unwell. Ethics Committee approval was obtained.

Procedure

Patients were approached in the waiting room and consenting participants returned to complete the study. Study materials clearly outlined the potential time commitment involved. A Registered Psychologist conducted the structured diagnostic interview. A different staff member then administered the legacy and PROMIS measures via computer. Potential participants were advised that the purpose of the study was to compare different questionnaires and that there would be overlap between the questions, so that they were prepared for, and not confused or frustrated by, the repetition.

Measures

PROMIS measures

Both PROMIS measures asked participants to indicate the frequency of symptoms over the past seven days on a five point response scale (0-4).

PROMIS Depression CAT

The PROMIS-D-CAT selects items from a 28-item bank [6]. The PROMIS default settings for standard error (0.3) and maximum number of items (12) were used and the minimum was changed to one item (from the default of four) to allow the lowest possible number of items to be asked. The Assessment Centre website used to administer CAT transforms raw, summed scale scores into T-Scores using norms to give a mean of 50 and a standard deviation (SD) of 10 [8]. T-scores from the PROMIS-D-CAT were categorised into severity level using cut-off points suggested for the PROMIS total item bank [11] as follows: normal 0-54; mild 55-64; moderately symptomatic 65-74 and 75+ severely symptomatic.

PROMIS Short Form

The PROMIS Depression 8b Short Form is an eight item measure focussing on emotional aspects of depression symptomatology. Item scores were summed to obtain the total raw score which was then converted to a T-score (mean 50, SD10) (http://www.healthmeasures.net accessed Nov 2017). A minimum of seven completed items was required and in this case, the total score was imputed following instructions in the PROMIS scoring guide (http://nihpromis.org/measures, accessed Nov 2013). To our knowledge, no cut-off points have been established for the PROMIS-D-SF, so the cut-off points that yielded sensitivity and specificity estimates for interview-detected MDE closest to those of the PROMIS-D-CAT were determined and used as a starting point for later analysis.

The eight PROMIS-SF items are a subset of the PROMIS-CAT 28-item bank. For any particular individual, if any PROMIS-D-SF items had been included in the PROMIS-D-CAT, they were not asked again in the PROMIS-D-SF.

Legacy depression and distress measures.

Five measures of depression and two emotional distress measures with established psychometric properties [6] were administered. Higher scores indicated more severe symptoms on all measures. The Beck Depression Inventory-2 (*BDI-II*) [19] is a 21-item scale where respondents rate their symptoms over the past two weeks. Scores of 0-13 indicate minimal depression; 14-19 mild; 20-28 moderate and 29-63 severe depression symptoms. The Centre for Epidemiological Studies-Depression (*CES-D*) consists of 20 items rated over the past week. Scores of 16-26 suggest mild depression and 27 and over suggest major depression. The Patient Health Questionnaire (*PHQ-9)* encapsulates the nine criteria for depression from the Diagnostic and Statistical Manual of the American Psychiatric Association, Fourth Edition, Text Revision (DSM-IV-TR). Items are rated over the past two weeks. Depression symptoms are categorised as mild (0-4); moderate (10-14); moderately severe (15-19) or severe (20 and over).

The Depression Anxiety and Stress Scale-Depression scale (*DASS-D*) is a seven item sub-scale of the 21 item DASS scale, rated over the past week. Depression symptoms are categorised as mild (5-6); moderate (7-10); severe (11-13) and extremely severe (14 and over). The Hospital Anxiety and Depression Scale (*HADS*) is a 14 item measure including a seven item depression subscale (HADS-D) that has been widely used in oncology [20, 21]. Possible depression is indicated by a score of 8-10 and a score of 11 or more indicating probable depression [22]. The DT is a one-item rating of distress over the past week on a 0-10 scale with an established cut-off score of four or more [23]. The PSYCH-6 is the six-item

psychological symptom subscale of the Somatic Psychological Health Report (SPHERE) [24]. Respondents indicate how often items have troubled them over the past few weeks. We have previously established a cut-off score of three or more for oncology outpatients [15].

Structured Diagnostic Interview

As the gold standard we used the Major Depressive Episode (MDE) section of the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID) [25] which assesses whether respondents met criteria for MDE in the past month. A diagnosis of MDE does not cover the full range of possible depressive symptomatology which may be experienced by cancer patients. However, it provides a defined statement of the likely need for intervention and an evidence base for the type of treatment that would be effective.

Anxiety modules were also administered, but are not reported here. The interview was conducted by two trained Registered Psychologists with experience in adult mental health disorders. Training was conducted according to SCID recommendations with private study, videotape training, role play and inter-rater comparisons of a video tape supervised by a Clinical Psychologist (BB) with extensive experience in psycho-oncology. Inter-rater reliability was not formally calculated. The SCID interviewers were blind to participants' responses on the self-report measures.

After the interview, legacy measures were administered first, in a fixed order using the QUICATOUCH platform [18]. The PROMIS-D-CAT and PROMIS-D-SF were then presented using the Assessment Centre platform run by the PROMIS initiative [8]. The order was: DT, HADS-D, PHQ-9, PSYCH-6, CES-D, BDI-II, DASS-D, PROMIS-D-CAT, PROMIS-D-SF. The total number of items administered was 80-91, depending on how many CAT items were answered.

Statistical Analysis

Analyses used Stata V14 (Statacorp, College Station, TX) and Statistical Package for the Social Sciences 22.0.

PROMIS-D-CAT Functioning

We reported the range of scores and the mean scores by SCID diagnosis (depressed vs not). We also reported the distribution and mean number of items administered by the PROMIS-D-CAT for the total sample, by SCID diagnosis (depressed vs not) and by proposed severity level [11].

Convergent Validity

Two series of bivariate (Spearman) correlations were performed, one to compare the PROMIS-D-CAT T-Score with all other depression and distress measures and the other to compare the PROMIS-D-SF to the same measures. The extent of correlation was classified as weak (<.4), moderate (.4 to < .7) and strong (.7 – 1) [26] Spearman was chosen because the data were not symmetric by nature, since they are diagnostic measures with potential bi-modalities in the data.

Criterion Validity

Two Logistic Regression analyses were performed, firstly with the PROMIS-D-CAT T-score as the predictor variable and secondly with the PROMIS-D-SF score. In each analysis, the SCID diagnosis of MDE in the past month was the dependent variable and the measure score (continuous) was the independent variable. Discrimination was assessed according to the area under the ROC curves (AUC) from these models. The same analysis was performed for each of the legacy measures to provide a context for interpreting the PROMIS measures. We have

previously reported on the relative performance of the legacy measures in a larger sample [6]. AUC were classified as: >.7= 'useful'; >.8= 'good' and >.9= 'very good' [27].

Measures of Diagnostic Accuracy

For the PROMIS-D-CAT, we calculated sensitivity, specificity, PPV and NPV at the cut-off points derived by Cella et al [11]. Values for adjacent T-scores were also examined to determine whether the proposed cut-off point was optimal. We anticipated that the mild cut off score would have higher sensitivity and lower specificity than the moderate and severe categories, making the mild category more suitable for a screening measure, to "rule-out" cases of depression, and the moderate / severe category more suitable for "ruling-in" cases of depression.

RESULTS

Sample

Of 322 eligible people indicating initial interest, 168 returned to commence the study. Of these, 18 missed PROMIS measures due to technical difficulties at our hospital, nine missed them due to intermittent problems with Assessment Centre and two did not have time to complete the PROMIS measures. This left 139 who commenced the PROMIS measures and 132 contributed to this analysis with at least one score from each source: the SCID MDE module, the PROMIS measures and the legacy measures. The median time to complete the legacy measures was 29 minutes, (inter-quartile range (IQR) 18-53); the median time to do the SCID interview was 28 minutes (IQR 20-35) and the median time to complete the PROMIS measures was 13 minutes (IQR 11-17).

The sample was 69% (n=91) female; 68% (n=90) married/living as married; 28% (n=37) widowed/separated/divorced and 4% (n=5) never married. Time since diagnosis ranged from five weeks to 21years (median=74 weeks), with 37% being within 12 months of diagnosis and a further 23% being within two years of diagnosis. Cancer stage was reported as early for 14% (n=19) of participants; 23% (n=30) reported Stage 2 or 3; 17% (n=23) reported advanced cancer, whereas stage could not be determined by self-report for 30% (n=40) of patients. Almost half (45%, n=59) of the sample had breast cancer, with 14% (n=18) reporting haematological cancers; 12% (n=16) colorectal cancer; 10% (n=13) lung cancer and 20% (n=26) reporting other cancers.

SCID Diagnosis

Among the 132 patients who completed the SCID, 14% (n=18) met criteria for MDE in the past month.

PROMIS-D-CAT functioning

T-scores on the PROMIS-D-CAT ranged from 34-69 with mean=49.77, median=50.00 and standard deviation (SD)=9.33. Patients with MDE had a higher mean T-score on the PROMIS-D-CAT (59.63, SD=7.46, median=61.1) than those without (48.18, SD= 8.62, median=48.1, t= -5.33, df=130, p<.001).

Using proposed severity levels for the PROMIS-D-CAT [5] most patients scored in the normal (non-depressed) range (70%, n=92), with 25% (n=33) in the mild and 5% (n=7) in the moderate range. No participant scored in the severe range.

The mean number of items answered on the PROMIS-D-CAT was 5.71 (range 2-12). Many (38%, n=50) participants only answered two or three items on the PROMIS-D-CAT and it terminated for about three quarters of participants with seven or fewer items (Table 1).

Participants with higher levels of depression symptomatology answered fewer items (Table 1).

Convergent Validity

Correlation with legacy measures

All correlations between PROMIS scores and scores on legacy measures were statistically significant ($p<.001$) (Table 2). Correlations with the PROMIS-D-CAT were generally higher (by .03-.04) than with the PROMIS-D-SF. Strong correlations were observed between the PROMIS-D-CAT and CES-D, DASS-D and BDI-II. Moderate correlations were observed with the PHQ-9; HADS-D, PSYCH-6 and DT.

Criterion Validity

AUC

The AUC obtained for the PROMIS-D-CAT was .844 (95%CI=.754-.934) and for the PROMIS-D-SF was .827 (95%CI=.707-.947) placing them in the 'good' range (Table 3). This was in the mid-range of AUC for legacy measures, with 6/7 measures in the good range (.809 to .856).

Diagnostic Accuracy Statistics

The mild (55) cut-off point on the PROMIS-D-CAT had sensitivity of .72 and specificity of .76 (Table 4). The moderate cut-off point (65) had a sensitivity of .22 and specificity of .97 (Table 4).

The cut-off scores for the PROMIS-D-SF that obtained comparable results to the PROMIS-D-CAT in terms of sensitivity and specificity were 55.3 (17 total raw score) for the mild

range (sensitivity=.73, specificity=.79) and 67.4 (30 total raw score) for the moderate range (sensitivity=.20, specificity=.98) (Table 4).

The ROC curves for all measures are provided as a context for interpreting the performance of the PROMIS measures (Figure 1). At the cut-off point for mild depression symptoms two measures (BDI-II and DT) had sensitivity similar to the PROMIS measures (.72); one had lower sensitivity (.61 HADS-D) and the remaining four measures had higher sensitivity (>.80 CES-D, DASS-D, PHQ-9 and PSYCH-6). All measures had a NPV in excess of 90%. Thus ability of the PROMIS measures to "rule-out" depression at the proposed cut off points was similar to two legacy measures, better than one measure and lower than four of the seven legacy measures.

Better performance as a measure to "rule-out" depression for the PROMIS-D-CAT was obtained at a slightly lower cut-off point of 53, with sensitivity of .83 and NPV of .96, which was similar to or better than all legacy measures except the PSYCH-6. The corresponding cut-off point on the PROMIS-D-SF was 14 with sensitivity of .80 and NPV of .97.

At the proposed moderate cut-off point the PROMIS-D-CAT and PROMIS-D-SF had specificity >.95 which was higher than but comparable to 3/7 other measures (CES-D .93, DASS-D .92and HADS-D .94) and higher than the remaining 4/7 measures (.82-.89). The PPVs of the PROMIS measures (>=.60) were comparable to the DASS-D (.59) and higher than all other measures.

However, it is worth noting that the sensitivity of the PROMIS measures at the proposed moderate cut-off point was low (<.25), comparable to the HADS-D (.22) but substantially lower than 6/7 other legacy measures (.61-.88)

At an alternate cut-off point of 60 the PROMIS-D-CAT showed more comparable performance to other measures for "ruling-in" depression, with specificity of .90, similar to 3/7 legacy measures (CES-D, DASS-D BDI-II) and higher than the remaining 4/7 measures (Table 4). At this cut-off point the PPV was .50, which was the same as the BDI-II and better than 4/7 other measures (PHQ-9, PSYCH-6, DT and HADS-D). The sensitivity (.61) at this cut-off point was also more similar to other measures at their moderate cut off points. A cut-off point of 23 on the PROMIS-D-SF offered similar performance with specificity of .93, PPV of .53 and sensitivity of .60 (Table 4).

Discussion

This study demonstrated the convergent and criterion validity of two PROMIS depression measures, PROMIS-D-CAT and PROMIS-D-SF, by comparing them with five legacy depression measures and two legacy measures of general emotional distress, and against the gold standard SCID for MDE in oncology outpatients. It also described how the PROMIS depression CAT functioned in an oncology population in terms of number of items presented and range of scores for those with and without a diagnosis of MDE.

Functioning of the CAT

T scores

The mean PROMIS-D-CAT T-score for our sample (49.77, SD 9.33) matched the reference population norm. As 14% of the sample had a clinical diagnose of MDE we might have expected our mean to be higher, noting that we found a moderate separation of T-scores between MDE cases (mean =59.63, SD= 7.46) and non-cases (mean 48.18, SD= 8.62).

The other three studies that reported using the PROMIS depression CAT in oncology samples also reported mean T scores somewhat lower than the PROMIS reference group: with means of 45.9 (SD=8.9); 48.8 (SD=8.5) and approximately 47.5 reported by Baum et al [10], Wagner et al [16] and Stone et al [17]. It may be that these patient samples had lower depression levels than the PROMIS reference population. However, Baum et al [10] reported their sample was comparable to the non-clinical reference population on the BSI. The other two studies did not have comparison measures of depression. Another explanation may be that some aspect of the CAT such as item selection or the T-score algorithm may not be optimal for oncology populations.

Number of CAT items

A stated benefit of CAT is reduced response burden through a reduced number of items. Respondents answered an average of 5.71 items on the PROMIS-D-CAT confirming that the CAT can operate as a brief instrument [28]. We should note however that we allowed a minimum of one item (maintaining standard pre-specified precision levels) where the standard protocol is for a minimum of four items. This was to allow the CAT to use as few items as possible.

Baum et al [10] reported that the average number of items answered on the PROMIS-D-CAT was higher (9.6 +/- 6.5) than the BSI depression subscale. They did not report the maximum number of items allowed or other CAT stopping rules, so their results may not be directly comparable. We could not locate any other studies using the PROMIS-D-CAT in oncology that reported the number of items presented.

Respondents with higher depressive symptoms completed fewer items on the PROMIS-D-CAT. The clinical effect of this is not clear. Placing a higher response burden on those less likely to be depressed may lower acceptability as the PROMIS-D-CAT may seem irrelevant.

Conversely, since aspects of depression, such as feeling overwhelmed and difficulty with decision-making could increase the response burden felt by depressed patients, a reduced number of items may be beneficial.

Convergent validity

In the development of PROMIS measures, Pilkonis et al [28] found a strong correlation (.83) between the CES-D and the summed score of the full item bank for depression. As mentioned earlier, Baum et al [10] observed a strong (.85) correlation between the PROMIS-D-CAT and the BSI depression scale in patients with non-metastatic prostate cancer. This degree of correlation was apparent in our sample only for the BDI-II, CES-D and DASS-D with the PROMIS-D-CAT T score.

The high correlation with the CES-D is perhaps unsurprising since the CES-D was used as a validation tool in first wave of PROMIS measure development [8, 28]. Furthermore, the BDI-II was one of the measures on which the CES-D was based [30]. Informal comparison of the items showed that about half of the items on the BDI-II (11/21) and CES-D (12/20) were represented on the PROMIS item bank.

Correlations with the DT, HADS-D, PHQ-9 and PSYCH-6 were considerably lower, around .6. This is possibly due to less similarity in content than for the CES-D and BDI-II. However, it is unclear why our correlations for HADS-D and PHQ-9 with PROMIS-D-SF were markedly lower than those found by Fischer et al [29] in German heart failure patients (n=194); PROMIS-D-SF vs PHQ-9 (.96) and HADS-D (.98).

Criterion Validity

None of our sample was considered “severe” on the PROMIS-D-CAT, despite 18 people meeting criteria for MDE, suggesting the proposed cut-off point might be too high. Nine of

the 18 people who met criteria for MDE scored in the mild range with another four scoring in the moderate range and different labels might better reflect the extent of symptomatology experienced.

Desirable characteristics for a brief measure to rule-out cases of depression include high sensitivity and high NPV, to minimise false negatives [13]. This is generally achieved by setting cut-off scores at a lower level of symptom severity. Conversely, desirable characteristics for case-finding measures are high specificity and high PPV, to reduce the number of false positives [13]. This is usually achieved by setting a higher cut-off score for symptom severity.

At the proposed cut-off point for mild symptoms, both the PROMIS-D-CAT and the PROMIS-D-SF demonstrated lower (.72 & .73) sensitivity for clinical depression than four of the legacy measures, suggesting that they were less suitable for "ruling-out" cases of depression than the other measures. At a lower cut-off point (53) the PROMIS measures had sensitivity (.83 & .80) for clinical depression which was similar to or better than 6/7 legacy measures while maintaining high NPV (>90%). Thus at the lower cut off points suggested by our analysis the PROMIS measures would be suitable for use as a screening tool for ruling out depression. At the lower cut-off points the sensitivity and specificity of the PROMIS measures compared favourably with pooled values of sensitivity (78.4%) and specificity (66.8%) for ultra-short tools to detect depression reviewed by Mitchell et al [14].

Although the PROMIS measures had very high (>.95) specificity at the proposed moderate (65) cut-off point the sensitivity of the PROMIS measures was substantially lower than most (6/7) other measures, suggesting that a high false negative rate would result from using the proposed cut-off (65). At a lower cut-off point (60) the performance of the PROMIS-D-CAT

was comparable to the legacy measures, with specificity similar to or better than all legacy measures and PPV similar to or better than 5/7 legacy measures while also maintaining adequate and comparable sensitivity (.61).

The PPVs for all measures were modest and none had sufficiently high PPV to be considered an accurate substitute for a SCID or mental health clinician assessment when attempting to diagnose depression, if used as a single measure. It is important to recognise that PPV will be limited by the relatively low prevalence of MDE in oncology. Since the PPV of a measure increases considerably when applied as the second step in a two stage screening process [31], these measures may be useful for case-finding as the second step in a two stage process.

We note that we reduced the minimum number of PROMIS-D-CAT items to be completed by a respondent from the default of four to one item. This was done in order to maximize the efficiency of assessment with the PROMIS-D-CAT measure (to lower respondent burden), but this may have affected the psychometric and screening performance of the PROMIS-D-CAT. As participants with interview-detected depression tended to answer fewer items (no-one with depression answered more than six items) it was not possible to examine the performance of the CAT among those who answered four or more questions. Future studies using the default number of items might obtain different results.

Relative value of PROMIS instruments

While we have confirmed the criterion and convergent validity of the PROMIS depression measures it is not clear that they offer any advance on legacy measures for detecting depression in oncology patients. For example both of the PROMIS measures obtained similar AUC to the one-item DT and lower AUC than the six-item PSYCH-6. In our analysis the

CES-D, DASS-D and PHQ-9 had equivalent sensitivity and NPV to the PROMIS-D-CAT (at a cut-off point of 53 for mild depression), making them equally suitable for ruling out depression. At a moderate cut-off point of 60 most (5/7) measures had equally high specificity and NPV as the PROMIS measures. On a positive note the PROMIS measures offer similar performance to the CES-D and BDI-II with significantly reduced respondent burden supporting their superior utility in cancer populations.

PROMIS have recently introduced financial charges for the use of the CAT measure. From 2017 using the PROMIS Assessment Centre platform will incur a $5000 USD charge per study [32]. An alternative is to use PROMIS via an App downloaded from the iTunes store for a subscription fee of $500 USD per annum. Given that instruments such as the DASS, PHQ-9 and PSYCH-6 are brief, free to use and offer similar diagnostic accuracy we anticipate that this may limit the uptake of the PROMIS measures for widespread use in depression screening.

We acknowledge that the PROMIS measures were initially intended to measure depression symptomatology rather than detect cases of clinical depression and that we do not address the use of PROMIS for monitoring health outcomes. However, an important aspect of validating any measure is to compare it against a gold standard. Additionally, there appears to be some impetus towards using PROMIS measures to identify caseness, given the development of depression severity categories (11) and the use of these categories clinically to refer patients for psychosocial care (16). We have confirmed the ability of the PROMIS measures to detect interview-determined depression, although more research is needed to establish the optimum cut-off points for depression severity. Other aspects of the PROMIS initiative will be valued by users, such as measurement equivalence across a range of health conditions, the establishment of an item bank which can offer multiple short forms that can be scored on the same metric and a focus on measurement precision.

Strengths and Limitations

The strengths of this study include the wide range of established measures, comparison with the SCID and blinding of clinical interviewers to responses on the self-report measures. Another advantage is the use of the PROMIS-D-CAT software in real-time, as it would be used clinically. Some other studies have used simulated CATs to estimate performance [9]. Although the sample size was reasonable, the number of patients with clinical depression was small, reflecting the level expected in oncology populations [33]. Replication in a larger sample is warranted. Additionally, the study involved a convenience sample at one centre and the results may not generalise. The survey required a sizeable time commitment from participants, which may explain the relatively low consent rate (41%). It is possible that patients who were more physically well were more likely to participate than those who were less well, but it was not feasible to collect data to assess this. However, as there is no obvious reason why the relationship between the diagnostic interview and the self-report survey questions would differ between respondents and non-respondents, we feel this would not affect the observed relationships.

Unfortunately it was not possible to randomise the order in which the legacy measures were administered by the QUICATOUCH platform, nor to randomise order of legacy versus PROMIS measures as two platforms were used. As the measures were presented in a fixed order, it is possible that order effects may have influenced the results. This would happen only if the process of completing the earlier measures somehow influenced how respondents answered the later questions. We are not aware of any directly relevant and consistent evidence in the literature to suggest this would be the case with this set of measures. The completion rates for the measures were uniformly high (95-100%), and were not lower for the measures presented later in the assessment, suggesting that respondent fatigue was not induced by order. Formal measures of inter-rater reliability were not obtained; however, the

SCID interview was administered by experienced psychologists, trained according to the SCID protocol and the structured nature of the interview reduces variation.

Conclusion

This study demonstrated the convergent and criterion validity of the PROMIS-D-CAT and PROMIS-D-SF measures in an oncology sample. More work is needed to determine the optimum cut-off points on the PROMIS measures.The PROMIS measures offer advantages in terms of brevity over the BDI-II and CES-D. However, they had no particular advantage in diagnostic accuracy, brevity or cost over the other legacy measures used in this study (DASS-D, DT, HADS-D, PHQ-9, PSYCH-6).

Compliance with Ethical Standards:

This study was funded by Calvary Mater Newcastle (grant number 11-09) and the Centre for Translational Neuroscience and Mental Health of the University of Newcastle (Australia) provided funding for statistical analysis. Professor King is supported by the Australian Government through Cancer Australia. Dr Lambert was initially supported by a National Health and Medical Research Council Research Fellowship (APP1012869) during data collection and by an FRQS Junior 1 Research Scholar Award subsequently.

Conflict of Interest: Kerrie Clover declares that she has no conflict of interest. Sylvie D. Lambert declares that she has no conflict of interest. Christopher Oldmeadow declares that he has no conflict of interest. Madeleine T. King declares that she has no conflict of interest Benjamin Britton declares that he has no conflict of interest. Alex J Mitchell declares that he has no conflict of interest. Gregory L. Carter declares that he has no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or National Research Committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Figure 1. ROC curves of all patient-reported measures vs a SCID diagnosis of MDE

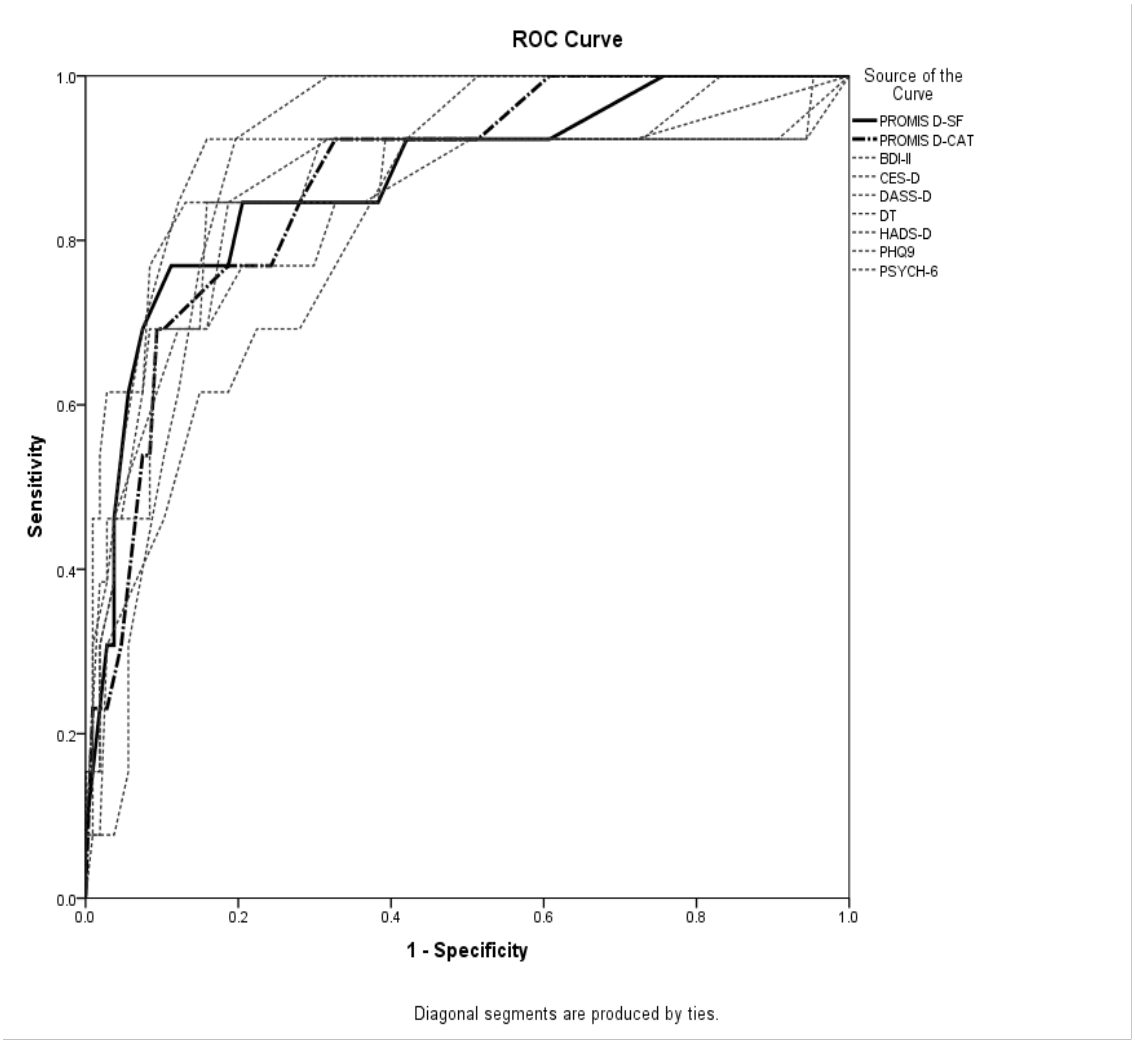

Table 1. Number of items* answered on the PROMIS Depression CAT

	Total Sample			SCID				PROMIS-D-CAT Categories					
			Cumulative	MDE		No MDE		Normal		Mild		Moderate	
N items	n=132	%	%	n=18	n%	n=114	n%	n=92	n%	n=33	n%	n=7	n%
2	10	7.6	7.6	1	5.6	9	7.9	3	3.3	7	21.2	0	0
3	40	30.3	37.9	9	50.0	31	27.2	16	17.4	19	57.6	5	71.4
4	25	18.9	56.8	5	27.8	20	17.5	19	20.7	4	12.1	2	28.6
5	5	3.8	60.6	2	11.1	3	2.6	5	5.4	0	0	0	0
6	13	9.8	70.5	1	5.6	12	10.5	12	13.0	1	3.0	0	0
7	8	6.1	76.5	0	0	8	7.0	6	6.5	2	6.1	0	0
8	2	1.5	78.0	0	0	2	1.8	2	2.2	0	0	0	0
9	1	.8	78.8	0	0	1	0.9	1	1.1	0	0	0	0
10	4	3.0	81.8	0	0	4	3.5	4	4.3	0	0	0	0
11	1	.8	82.6	0	0	1	0.9	1	1.1	0	0	0	0
12	23	17.5	100.0	0	0	23	20.2	23	25.0	0	0	0	0

* Number of items in other depression measures: BDI-II 21, CES-D 20, DASS-D 21 (7 for depression subscale), DT (1), HADS-D 14 (7 for depression subscale), PHQ-9 9, PSYCH-6 6, PROMIS-D-SF 8

Table 2. Correlations between PROMIS depression measures and established measures

	PROMIS-D-CAT T Score		PROMIS-D-SF score	
	n	Rho*	n	Rho*
BDI-II	129	.789	124	.748
CES-D	129	.810	125	.774
DASS-D	130	.802	125	.764
DT	132	.621	127	.578
HADS-D	132	.626	127	.589
PHQ-9	128	.670	124	.618
PSYCH-6	129	.658	125	.609

*All significant at p<.001.

Table 3. Area Under the Curve for each measure vs the SCID diagnoses.

Measure	**n**	**AUC (95% CI)**
PROMIS-D-CAT T score	132	.84 (.75-.93)
PROMIS-D-SF	127	.83 (.71-.95)
BDI-II	129	.81 (.68-.94)
HADS-D	132	.82 (.69-.94)
PHQ-9	128	.82 (.69-.95)
CES-D	129	.83 (.69-.97)
DT	132	.83 (.73-.93)
DASS-D	130	.86 (.73-.98)
PSYCH-6	129	.92 (.87-.97)

Table 4. Diagnostic accuracy of PROMIS measures compared with SCID diagnosis of MDE

Score	Sensitivity	95%CI	Specificity	95%CI	Positive Predictive Value	95%CI	Negative Predictive Value	95%CI
PROMIS-D-CAT								
49	.89	.72-1.00	.51	.41-.61	.22	.18-.27	.97	.92-1.00
50	.89	.72-1.00	.52	.42-.61	.23	.18-.27	.97	.92-1.00
51	.89	.72-1.00	.63	.54-.72	.28	.22-.34	.97	.94-1.00
52	.89	.74-1.00	.66	.57-.75	.29	.17-.41	.97	.94-1.00
53	.83	.67-1.00	.70	.62-.78	.31	.24-.39	.96	.93-1.00
54	.72	.50- .89	.74	.65-.82	.30	.22-.40	.94	.90-.98
55 (Mild)	.72	.50- .89	.76	.68 –.76	.32	.23 - .43	.95	.91–.98
56	.67	.44- .89	.77	.69-.84	.31	.22-.43	.94	.90-.98
57	.67	.44-.89	.78	.70-.85	.32	.22-.44	.94	.90-.98
58	.67	.44-.89	.81	.73-.88	.35	.25-.48	.94	.90-.98
59	.61	.39-.83	.89	.83-.95	.48	.33-.67	.94	.90-.97
60	.61	.39-.83	.90	.85-.96	.50	.34-.83	.94	.90-.97
61	.50	.28-.72	.92	.87-.96	.50	.32-.72	.92	.89-.95
62	.50	.28-.72	.93	.89-.96	.53	.35-.75	.92	.89-.96
63	.33	.11- .56	.96	.91-.99	.56	.29-.83	.90	.87 - .93
64	.33	.11-.56	.96	.91-.99	.56	.29-.83	.90	.87-.93
65 (Moderate)	.22	.06-.44	.97	.94-1.00	.60	.20-1.00	.89	.87-.92
66	.22	.06-.44	.97	.94-1.00	.60	.20-1.00	.89	.87-.92
67	.22	.06-.44	.98	.95-1.00	.67	.25-1.00	.89	.87-.92
PROMIS-D-SF								
T-score (raw score)								

48.2 (11)	.87	.67-1.00	.44	.35-.53	.17	.14-.21	.96	.91-1.00
49.8 (12)	.87	.67-1.00	.51	.42-.60	.19	.15-.23	.97	.92-1.00
51.2 (13)	.87	.67-1.00	.57	.47-.66	.21	.17-.26	.97	.93-1.00
52.3 (14)	.80	.60-1.00	.62	.51-.71	.22	.16-.28	.96	.92-1.00
53.4 (15)	.73	.53-.93	.67	.58-.76	.23	.16-.30	.95	.91-.99
54.3 (16)	.73	.53-.93	.71	.63-.71	.26	.18-34	.95	.91-.99
55.3(17) Mild	.73	.53-.93	.79	.71-.86	.31	.22-.43	.96	.92-.99
56.2 (18)	.67	.40-.87	.80	.72-.87	.31	.21-.43	.95	.91-.98
57.1 (19)	.67	.40-.87	.84	.77-.90	.36	.24-.50	.95	.91-.98
58.8 (21)	.67	.40-.87	.88	.82-.94	.44	.29-.61	.95	.92-.98
59.72 (22)	.67	.40-.87	.89	.83-.95	.45	.30-.65	.95	.92-.98
60.7 (23)	.60	.33-.80	.93	.88-.97	.53	.35-.75	.95	.91-.97
61.6 (24)	.53	.27-.73	.95	.90-.98	.57	.36-.82	.94	.91-.96
62.5 (25)	.40	.20-.60	.96	.93-.99	.60	.33-.89	.92	.90-.95
63.5 (26)	.40	.20-.60	.96	.93-.99	.60	.33-.89	.92	.90-.95
64.4 (27)	.27	.07-.47	.96	.93-.99	.50	.17-.86	.91	.88-.93
65.4 (28)	.27	.13-.53	.97	.94-1.00	.60	.29-1.00	.91	.89-.94
66.4 (29)	.27	.13-.53	.97	.94-1.00	.60	.29-1.00	.91	.89-.94
67.4 (30) Moderate	.20	.07-.47	.98	.96-1.00	.67	.29-1.00	.90	.89-.93
68.3 (31)	.13	.07-.33	.99	.97-1.00	.75	.33-1.00	.90	.89-.92
69.3 (32)	.07	.07-.27	1.00	1.00-1.00	1.00	1.00-1.00	.89	.89-.91

References

1. Institute of Medicine Report 2008, (IOM) (2008). *Cancer care for the whole patient: Meeting psychosocial health needs*. The National Academies Press: Washington, DC:

2. Alex J Mitchell, Nadia Hussain, Lorraine Grainger, Paul Symonds (2011). Identification of patient-reported distress by clinical nurse specialists in routine oncology practice: a multicentre UK study. Psychooncology 20: 10. 1076-1083 Oct.

3. Holland, J.C. and Bultz, B.D., 2007. The NCCN guideline for distress management: a case for making distress the sixth vital sign. *Journal of the National Comprehensive Cancer Network*, *5*(1), pp.3-7.

4. American College of Surgeons Commission on Cancer. Cancer Program Standards 2012 Version 1.1: Ensuring patient-centered care. Available at: http://www.facs.org/cancer/coc/programstandards2012.html. Accessed March 2017

5. Linda E Carlson, Amy Waller, Alex J Mitchell (2012). Screening for distress and unmet needs in patients with cancer: review and recommendations. J Clin Oncol 30: 11. 1160-1177 Apr.

6. Lambert SD, Clover K, Pallant JF, Britton B, King MT, Mitchell AJ, Carter G. Making Sense of Variations in Prevalence Estimates of Depression in Cancer: A Co-Calibration of Commonly Used Depression Scales Using Rasch Analysis. Journal of the National Comprehensive Cancer Network. 2015 Oct 1;13(10):1203-11.

7. Mitchell, A.J., Meader, N., Davies, E., Clover, K., Carter, G.L., Loscalzo, M.J., Linden, W., Grassi, L., Johansen, C., Carlson, L.E. and Zabora, J., 2012. Meta-analysis of screening and case finding tools for depression in cancer: evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *Journal of affective disorders*, *140*(2), pp.149-160.

8. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S. and Cook, K., 2010. Initial adult health item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS™) network: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), p.1179.

9. Choi, S.W., Reise, S.P., Pilkonis, P.A., Hays, R.D. and Cella, D., 2010. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), pp.125-136.

10. Baum, G., Basen-Engquist, K., Swartz, M.C., Parker, P.A. and Carmack, C.L., 2014. Comparing PROMIS computer-adaptive tests to the Brief Symptom Inventory in patients with prostate cancer. *Quality of Life Research*, *23*(7), pp.2031-2035.

11. Cella, D., Choi, S., Garcia, S., Cook, K.F., Rosenbloom, S., Lai, J.S., Tatum, D.S. and Gershon, R., 2014. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research*, *23*(10), pp.2651-2661.

12. Lazenby M, Ercolano E, Grant M, Holland JC, Jacobsen PB, McCorkle, R. Supporting Commission on Cancer–Mandated Psychosocial Distress Screening with Implementation Strategies. *Journal of Oncology Practice*, 2015:11(3), e413-e420. doi: 10.1200/JOP.2014.002816.

13. Maxim, L. D., Niebo, R., & Utell, M. J. (2014). Screening tests: a review with examples. *Inhalation Toxicology*, *26*(13), 811–828. http://doi.org/10.3109/08958378.2014.955932

14. Mitchell, A.J., 2007. Pooled results from 38 analyses of the accuracy of distress thermometer and other ultra-short methods of detecting cancer-related mood disorders. *Journal of Clinical Oncology*, *25*(29), pp.4670-4681.

15. Clover, K. Carter, G, Adams, C. Hickie I, Davenport T. Concurrent validity of the PSYCH-6, a very short scale for detecting anxiety and depression, among oncology outpatients. Australian New Zealand Journal of Psychiatry 2009a,43:7,682-688.

16. Wagner, L.I., Schink, J., Bass, M., Patel, S., Diaz, M.V., Rothrock, N., Pearman, T., Gershon, R., Penedo, F.J., Rosen, S. and Cella, D., 2015. Bringing PROMIS to practice: brief and precise symptom screening in ambulatory cancer care. *Cancer*, *121*(6), pp.927-934.
17. Stone, A.A., Broderick, J.E., Junghaenel, D.U., Schneider, S. and Schwartz, J.E., 2016. PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. *Journal of clinical epidemiology*, *74*, pp.194-206.

18. Clover KA, Rogers K, Britton B, Oldmeadow C, Attia J, Carter GL. Reduced prevalence of pain and distress during four years of screening with QUICATOUCH in Australian oncology patients (European Journal of Cancer Care, 2017)

19. Beck AT, Steer RA, Brown GK. Manual for The Beck Depression Inventory Second Edition (BDI-II). San Antonio: Psychological Corporation; 1996.

20. Alex J Mitchell, Nick Meader, Paul Symonds (2010) Diagnostic validity of the Hospital Anxiety and Depression Scale (HADS) in cancer and palliative settings: a meta-analysis. J Affect Disord 126: 3. 335-348 Nov
21. Vodermaier, A; Millman, RD (2011). Accuracy of the Hospital Anxiety and Depression Scale as a screening tool in cancer patients: a systematic review and meta-analysis. Supportive Care in Cancer, Volume 19, Issue 12, pp 1899-1908

22. Zigmond, AS; Snaith, RP (1983). "The hospital anxiety and depression scale". Acta Psychiatrica Scandinavica 67 (6): 361–370. doi:10.1111/j.1600-0447.1983.tb09716.x. PMID 688082

23. National Comprehensive Cancer Network (2003) Distress management. Clinical practice guidelines. J Natl Compr Canc Netw. 2003 Jul;1(3):344-74.

24. Hickie, I.B., Koschera, A., Davenport, T.A., Naismith, S.L., Scott, E.M. Comorbidity of common mental disorders and alcohol or other substance misuse in Australian general practice. Med J Aust, 175 (2001), pp. S31–S36

25. First MB, Gibbon M, Spitzer RL, Williams JB. Structured Clinical Interview for DSM-IV-TR Axis I Disorders: Research Version. Biometrics Research Department, New York State Psychiatric Institute, 2002. First et al, 2002

26. Dancy CP, Reidy J. Statistics without maths for psychology. Harlow: Pearson Education Limited. 2004.

27. Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), p.1285.

28. Pilkonis, P.A., Choi, S.W., Reise, S.P., Stover, A.M., Riley, W.T. and Cella, D., 2011. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*, *18*(3), pp.263-283.

29. Fischer, H.F., Klug, C., Roeper, K., Blozik, E., Edelmann, F., Eisele, M., Störk, S., Wachter, R., Scherer, M., Rose, M. and Herrmann-Lingen, C., 2014. Screening for mental disorders in heart failure patients using computer-adaptive tests. *Quality of Life Research*, *23*(5), pp.1609-1618.

30. Centre for Epidemiologic Studies. Centre for Epidemiologic Studies Depression Scale (CES-D). Rockville: National Institute of Mental Health; 1970.

31. Clover, K. Carter, G, Mackinnon, A. Adams, C. Is my patient suffering clinically significant emotional distress? Demonstration of a probabilities approach to evaluating algorithms for screening for distress. Supportive Care in Cancer 2009b;17:1455-1462 .

32. Health Measures (2016). Assessment Centre Pricing Information. http://www.healthmeasures.net/images/LearnMore/Assessment_Center_Pricing_Information_07252016.pdf. Accessed 29 May 2017.

33. Krebber AM, Buffart LM, Kleijn G, Riepma IC, Bree R, Leemans CR, Becker A, Brug J, Straten A, Cuijpers P, Verdonck-de Leeuw IM. Prevalence of depression in cancer

patients: a meta-analysis of diagnostic interviews and self-report instruments. Psycho-Oncology. 2014 Feb 1;23(2):121-30.