Multiple Sclerosis and Related Disorders 20 (2018) 9–15

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard

The validity and reliability of screening measures for depression and anxiety disorders in multiple sclerosis

Ruth Ann Marrie[a,b,*], Lixia Zhang[b], Lisa M. Lix[b], Lesley A. Graff[c], John R. Walker[c], John D. Fisk[d], Scott B. Patten[e], Carol A. Hitchon[a], James M. Bolton[f], Jitender Sareen[f], Renée El-Gabalawy[c,g], James J. Marriott[a], Charles N. Bernstein[a]

[a] Department of Internal Medicine, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada
[b] Department of Community Health Sciences, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada
[c] Department of Clinical Health Psychology, Max Rady College of Medicine Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada
[d] Nova Scotia Health Authority, Departments of Psychiatry, Psychology & Neuroscience, and Medicine, Dalhousie University, Halifax, Canada
[e] Departments of Community Health Sciences & Psychiatry, Cumming School of Medicine, University of Calgary, Calgary, Canada
[f] Department of Psychiatry, Max Rady College of Medicine Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada
[g] Departments of Anesthesia & Perioperative Medicine, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada

ARTICLE INFO

Keywords:
Multiple sclerosis
Depression
Anxiety
Validity
Reliability
Psychometrics

ABSTRACT

Objective: We aimed to evaluate the validity and reliability of multiple screening measures for depression and anxiety for use in the clinical care of people with multiple sclerosis (MS).

Methods: Participants with MS completed the Patient Health Questionnaire (PHQ-9), Hospital Anxiety and Depression Scale (HADS), Kessler-6 Distress Scale, PROMIS Emotional Distress Depression Short-Form 8a (PROMIS Depression) and Anxiety Short-Form 8a (PROMIS Anxiety), Generalized Anxiety Disorder 7-item Scale (GAD-7), and the Overall Anxiety and Severity Impairment Scale (OASIS). A subgroup repeated the screening measures two weeks later. All participants also completed a Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID). For the screening measures we computed sensitivity, specificity, positive predictive and negative predictive value with SCID diagnoses as the reference standard and conducted receiver operating curve (ROC) analyses; we also assessed internal consistency and test-retest reliability.

Results: Of 253 participants, the SCID classified 10.3% with major depression and 14.6% with generalized anxiety disorder. Among the depression measures, the PHQ-9 had the highest sensitivity (84%). Specificity was generally higher than sensitivity, and was highest for the HADS-D with a cut-point of 11 (95%). In ROC analyses the area under the curve (AUC) did not differ between depression measures. Among the anxiety measures, sensitivity was highest for the HADS-A with a cut-point of 8 (82%). Specificity ranged from 83% to 86% for all measures except the HADS-A with a cut-point of 8 (68%). The AUC did not differ between anxiety measures.

Conclusion: Overall, performance of the depression and anxiety screening measures was very similar, with reasonable psychometric properties for the MS population, suggesting that other factors such as accessibility and ease of use could guide the choice of measure in clinical practice.

1. Introduction

Multiple sclerosis (MS) has a high prevalence of comorbid depression and anxiety disorders throughout the disease course (Marrie et al., 2015; Marrie et al., 2016). Comorbid depressive and anxiety disorders are associated with lower quality of life, and greater pain and health care utilization (Janssens et al., 2003; Fiest et al., 2015; Marrie et al., 2015). Therefore, emphasis has been placed on identifying these disorders promptly, and involving collaborative mental health services if needed.

Multiple potential case identification (aka screening) measures for assessing possible depression and anxiety disorders exist (Williams et al., 2002). However, somatic symptoms of depression such as fatigue, and difficulty sleeping captured in screening measures for depression are also common somatic symptoms of MS. Similar issues arise when screening for anxiety. For example, the Beck Anxiety Inventory (BAI) captures somatic symptoms of anxiety such as dizziness, numbness and tingling (Beck et al., 1988). which are common physical symptoms

* Correspondence to: Health Sciences Center, GF-543, 820 Sherbrook Street, Winnipeg, MB, Canada R3A 1R9.
E-mail address: rmarrie@hsc.mb.ca (R.A. Marrie).

https://doi.org/10.1016/j.msard.2017.12.007
Received 18 November 2017; Accepted 13 December 2017

experienced in MS. This raises the question as to whether these measures adequately identify the depressive and anxiety disorders or whether they are confounded by physical symptoms of the MS, leading to overestimates and misclassification of these disorders. Problems with criterion contamination of depression scales have been reported in MS (Mohr et al., 1997).

A systematic review identified 21 studies which assessed the performance of nine depression screening measures in MS and found that further research was needed to assess the utility of most measures (Hind et al., 2016). A systematic review of screening measures for anxiety found relatively little support for the validity and reliability of three available instruments, the Hospital Anxiety and Depression Scale (HADS), Generalized Anxiety Disorder-7 (GAD-7) and BAI (Litster et al., 2016). Therefore, we aimed to evaluate the validity and reliability of multiple screening measures for depression and anxiety for people with MS.

2. Materials and methods

As detailed elsewhere (Marrie et al., In press), from November 2014 through July 2016 we recruited individuals from the sole provincial MS Clinic with a definite diagnosis of MS (Poser et al., 1983; McDonal et al., 2001; Polman et al., 2005, 2011), who were aged ≥18 years, able to provide informed consent, and with an adequate knowledge of English to complete questionnaires and interviews. Ethics approval was provided by the University of Manitoba Health Research Ethics Board, Victoria General Hospital, the Health Sciences Centre, Seven Oaks General Hospital and St. Boniface Hospital.

After providing informed consent, participants completed questionnaires, and underwent physical assessments as described below. If possible, they participated in the Structured Clinical Interview for DSM-IV-TR Axis I Disorders – Research version (SCID) the same day (First et al., 2002). If not, the SCID was completed within two to four weeks of enrollment. A subgroup of participants completed the screening measures again within two weeks of initial administration.

2.1. Sociodemographic and clinical characteristics

Participants reported their sex, date of birth, ethnicity, and highest level of education attained. Ethnicity was categorized as white or non-white. Education was categorized as less than high school, high school/GED, college, technical/trade, and Bachelor's degree or higher. Participants also reported their age at MS symptom onset. We determined clinical course by medical records review. Participants underwent a neurologic examination for determination of disability status as measured by the Expanded Disability Status Scale (EDSS) (Kurtzke, 1983).

2.2. Screening measures

Each participant completed the Patient Health Questionnaire (PHQ-9) from which we also derived a score for the PHQ-2, the HADS, Kessler-6 Distress Scale, Patient-Reported Outcomes Measurement Information System Emotional Distress Depression Short-Form 8a (PROMIS Depression) and Anxiety Short-Form 8a (PROMIS Anxiety), GAD-7 and Overall Anxiety and Severity Impairment Scale (OASIS) (Zigmond and Snaith, 1983; Spitzer et al., 1999; Norman et al., 2006; Spitzer et al., 2006; Cairney et al., 2007). When selecting these measures we considered properties including face validity, ease of use, availability for self-administration, and copyright restrictions.

The PHQ-9 includes nine items with response options of 0 (not at all) to 3 (nearly every day), and assesses depressive symptoms over the last two weeks (Spitzer et al., 1999). Total scores range from 0 to 27. The PHQ-2 includes the first two items from the PHQ-9 and has been promoted as a briefer screen for depression (Kroenke et al., 2003). Scores range from 0 to 6. The HADS includes 14 items, 7 for depression and 7 for anxiety, which assess symptoms over the past week (Zigmond and Snaith, 1983). Two cut-points are commonly used for the HADS (8, 11) therefore we tested both. Total scores for each of the two subscales range from 0 to 21. The Kessler-6 includes 6 items which measure non-specific distress over the past 30 days; we classified it with depression measures since five of its six items are common depressive symptoms (hopelessness, agitation, depressed mood, low energy, worthlessness). Using the alternative scoring method (https://www.hcp.med.harvard.edu/ncs/k6_scales.php), scores range from 6 to 30. The PROMIS Depression and Anxiety measures include 8 items with response options ranging from 1 (never) to 5 (always) (Pilkonis et al., 2011). These items assess symptoms over the past 7 days. Total scores for the PROMIS Depression measure are transformed into T scores with values ranging from 38.2 to 81.3, while they are transformed into T scores with values ranging from 37.1 to 83.1 for the PROMIS Anxiety measure. A score of 50 is average for the United States general population. The GAD-7 includes 7 items which assess symptoms of anxiety over the last two weeks. Response options range from 0 (not at all) to 3 (nearly every day); total scores range from 0 to 21. The OASIS includes 5 items which assess anxiety and fear over the past week (Norman et al., 2006). Response options range from 0 to 4 and total scores from range 0–20. For all measures, higher scores indicate more severe symptoms.

2.3. Questionnaires assessing related constructs

We assessed fatigue using the Fatigue Impact Scale for Daily Use (D-FIS), a validated instrument which includes 8 items scored on an ordinal scale from 0 (no) to 4 (extreme problem) (Fisk and Doble, 2002). We assessed pain using from MOS-Modified Pain Effects Scale, a valid and reliable instrument with scores ranging from 6 to 30;(Ritvo et al., 1997a, 1997b) higher scores indicate greater pain.

2.4. Interview

The SCID is a semi-structured interview to identify DSM-IV diagnoses including anxiety, and major depression. Trained interviewers, blinded to the results of the screening measures, administered the SCID to determine the current histories of depressive and anxiety disorders. For this study, SCID-based diagnoses of current major depression and generalized anxiety disorder served as the reference standard in analyses of criterion validity. In a complementary analysis, we used a SCID diagnosis of any anxiety disorder in the last month (generalized anxiety disorder, panic disorder, social phobia, specific phobia, anxiety disorder due to general medical condition, anxiety disorder due to substance use, stress disorder) instead of generalized anxiety disorder.

2.5. Analysis

We summarized the characteristics of study participants using frequency (percent [%]) for categorical variables, and mean (standard deviation [SD]) or median (interquartile range [IQR]) for continuous variables. Missing data were not imputed; individuals with missing values for a measure were excluded from analyses of that measure.

Based on the taxonomy proposed by the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) we assessed criterion validity, construct validity (through hypothesis testing), content validity, internal consistency reliability, and test-retest reliability of the selected measures (Mokkink et al., 2010).

Criterion validity indicates how well the scores of the screening tool reflect the reference (criterion) standard. First, we compared depression and anxiety status based on the (i) SCID (criterion standard) and (ii) self-reported screening measures. Based on published cut-points for depression/anxiety for these measures, we computed sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for the screening measures versus the criterion standard. Second, we used receiver operating curve (ROC) analysis to identify the best

cut-point for predicting depression or anxiety (or general psychological distress as appropriate) as this could differ for the MS population from the general population (Stafford et al., 2007; Hedayati et al., 2009). The optimal cut-point was calculated by maximizing Youden's J index (sensitivity + specificity – 1) (Youden, 1950), in which sensitivity and specificity are balanced. We compared the area under the ROC between screening measures using binary logistic regression, separately for depression and anxiety measures.

Construct validity measures the extent to which the measure of interest correlates with measures of other variables in hypothesized ways. We estimated Spearman rank correlations (with 95%CI) between the total scores on the screening measures and pain and fatigue, expecting to find moderate positive correlations. Excessively strong correlations would suggest criterion contamination. We also assessed Spearman rank correlations between the total scores on the screening measures and age, expecting weaker correlations.

Internal consistency reliability refers to the degree to which items on a measure are interrelated and thus reliably measure the construct. We assessed the internal consistency of each screening tool using Cronbach's alpha (α), where an acceptable α was ≥0.70 (Bland and Altman, 1997). Test-retest reliability refers to the reproducibility of the scores over time, for respondents who have not changed. We assessed test-retest reliability by estimating the intraclass correlation coefficient (ICC) with 95%CI between the two administrations of the screening tool, approximately two weeks apart. Based on these 95%CI we classified values of the ICC of < 0.5 as poor, 0.5–0.75 as moderate, 0.75–0.90 as good, and > 0.90 as excellent reliability (Koo and Li, 2016).

2.5.1. Sample size

Assuming a lower bounds sensitivity = 0.75 and specificity ≥0.85 (Spitzer et al., 1999; Stafford et al., 2007), precision = 0.15 and α = 0.05, the required sample size was 247. For assessment of test-retest reliability, if the ICC is ≥0.6 (0.6 being lowest acceptable), precision is 0.1, and α = 0.05, then the required sample size was 158.

Statistical analyses used SAS V9.4 (SAS Institute Inc., Cary, NC).

3. Results

We enrolled 255 persons with MS, of whom 253 completed a SCID and were included in this analysis. Most participants were women with a high school education or better (Table 1). The mean (SD) time between study enrollment and completion of the SCID was 0.10 (0.17) weeks; 80 (32.3%) of the SCIDs were completed the day of enrollment, 168 (66.4%) of the SCIDs were completed within half of a week, 3 (1.2%) more were completed within one week, and the remaining 2 (0.8%) were completed within two weeks. Incomplete responses in the screening measures were uncommon (Table e1).

Table 1
Demographic and clinical characteristics of study participants (n = 253).

Characteristic	Value
Sex, n (%)	
Female	206 (81.4)
Male	47 (18.6)
Age at enrollment, mean (SD)	51.0 (12.9)
Age at symptom onset, mean (SD)	31.3 (11.3)
Race[a], n (%)	
White	216 (85.7)
Other	36 (14.3)
Highest level of education, n (%)	
Less than high school	10 (4.0)
High school/GED	77 (30.4)
College	71 (28.1)
Technical/trade	29 (11.5)
University Bachelor's degree or higher	66 (26.1)
Clinical course, n (%)	
Relapsing remitting	183 (72.3)
Secondary progressive	47 (18.6)
Primary progressive	23 (9.1)
EDSS, median (p25-p75)	4 (3–6)
Time from enrollment to SCID completion (weeks), mean (SD)	0.10 (0.17)

[a] Missing for one participants, EDSS = Expanded Disability Status Scale.

Table 2
Frequency (%) of participants with current depression or anxiety according to each measure used, n (%).

Tool	N (%)
Depression	
SCID major depression	26 (10.3)
PHQ-2	49 (19.8)
PHQ-9	73 (30.0)
HADS-D (cut-point 8)	60 (23.9)
HADS-D (cut-point 11)	20 (8.0)
PROMIS Depression	51 (20.2)
Kessler-6	22 (8.9)
Anxiety	
SCID generalized anxiety disorder	11 (4.3)
SCID any anxiety disorder	43 (17.0)
GAD-7	46 (18.3)
OASIS	54 (21.3)
HADS-A (cut-point 8)	86 (34.1)
HADS-A (cut-point 11)	40 (15.9)
PROMIS Anxiety	49 (19.4)

SCID = Structured Clinical Interview for DSM-IV-TR; SCID anxiety = any DSM-IV anxiety disorder (generalized anxiety disorder, panic disorder, obsessive compulsive disorder, post-traumatic stress disorder, anxiety disorder due to general medical disorder, anxiety disorder due to substance use, stress adjustment disorder, social phobia, specific phobia); SCID generalized anxiety = generalized anxiety disorder; PHQ-2 = Patient Health Questionnaire-2, PHQ-9 = Patient Health Questionnaire-9, GAD-7 = Generalized Anxiety Disorder-7, OASIS = Overall Anxiety and Severity Impairment Scale, HADS-D = Hospital Anxiety and Depression Scale – depression score, HADS-A = Hospital Anxiety and Depression Scale – anxiety score.

Using the SCID, 10.3% of participants were classified as currently meeting the criteria for major depression. The proportion classified as depressed by the depression measures varied, and exceeded the proportion identified as having major depression on the SCID, except for the HADS-D with a cut-point of 11 (Table 2). Using the SCID, 17% of participants were classified as having any anxiety disorder, while 4.3% were classified as having generalized anxiety disorder (Table 3). The proportion classified as anxious by the anxiety measures varied, and exceeded the proportion identified as having a generalized anxiety disorder on the SCID.

3.1. Criterion validity

Performance of the depression measures based on typically recommended cut-points is shown in Table 3. Sensitivity was highest for the PHQ-9 (84%) and lowest for the HADS-D with a cut-point of 11 and the Kessler-6 (both 31%). Specificity was generally higher than sensitivity, and was highest for the HADS-D with a cut-point of 11 (95%), followed by the Kessler-6 (94%) and lowest for the PHQ-9 (76%). The area under the ROC curve (AUC) did not differ between the PHQ-9 (AUC 0.86; 95%CI: 0.80, 0.93) and the HADS-D (AUC 0.84; 95%CI: 0.77, 0.92, p = 0.64), PROMIS Depression (AUC 0.88; 95%CI: 0.82, 0.93, p = 0.63), or Kessler-6 (AUC 0.86; 95%CI: 0.80, 0.92, p = 0.94; Fig. e1).

Based on the ROC analysis, the optimal cut-points for some of the depression screening measures differed from those routinely recommended (Table e2). Specifically, optimal cut-points were 12 for the PHQ-9, 7 for the HADS-D, 57.7 for the PROMIS Depression, and 15 for the Kessler-6.

Performance of the anxiety measures based on the typically recommended cut-points is shown in Table 4. Sensitivity was highest for

Table 3
Test characteristics for previously defined cut-points for depression and anxiety screening measures.

Instrument	Cutpoint ≥	Sens (95%CI)	Spec (95%CI)	PPV (95%CI)	NPV (95%CI)	Accuracy (95%CI)
Depression						
PHQ-2	3	0.68 (0.46, 0.85)	0.86 (0.80, 0.90)	0.35 (0.22, 0.50)	0.96 (0.92, 0.98)	0.84 (0.78, 0.88)
PHQ-9	10	0.84 (0.64, 0.95)	0.76 (0.70, 0.82)	0.29 (0.19, 0.41)	0.97 (0.94, 0.99)	0.77 (0.71, 0.82)
HADS-D	8	0.69 (0.48, 0.86)	0.81 (0.76, 0.86)	0.30 (0.19, 0.43)	0.96 (0.92, 0.98)	0.79 (0.74, 0.85)
	11	0.31 (0.14, 0.52)	0.95 (0.91, 0.97)	0.40 (0.19, 0.64)	0.92 (0.88, 0.95)	0.88 (0.84, 0.92)
PROMIS Depression	T score 60	0.69 (0.48, 0.86)	0.85 (0.80, 0.90)	0.35 (0.22, 0.50)	0.96 (0.92, 0.98)	0.83 (0.79, 0.88)
Kessler-6	19	0.31 (0.14, 0.52)	0.94 (0.90, 0.97)	0.36 (0.17, 0.59)	0.92 (0.88, 0.95)	0.87 (0.82, 0.91)
Anxiety						
GAD-7	10	0.54 (0.23, 0.83)	0.83 (0.78, 0.88)	0.13 (0.05, 0.26)	0.97 (0.94, 0.99)	0.82 (0.76, 0.87)
OASIS	8	0.73 (0.39, 0.94)	0.81 (0.75, 0.86)	0.15 (0.07, 0.27)	0.98 (0.96, 0.99)	0.81 (0.75, 0.85)
HADS-A	8	0.82 (0.48, 0.98)	0.68 (0.62, 0.74)	0.10 (0.05, 0.19)	0.99 (0.95, 1.00)	0.69 (0.63, 0.74)
HADS-A	11	0.64 (0.31, 0.89)	0.86 (0.81, 0.90)	0.18 (0.07, 0.33)	0.98 (0.95, 0.99)	0.85 (0.80, 0.89)
PROMIS Anxiety	T score 60	0.73 (0.39, 0.94)	0.83 (0.78, 0.88)	0.16 (0.07, 0.30)	0.98 (0.96, 0.99)	0.83 (0.77, 0.87)

Sens = sensitivity, spec = specificity, PPV = positive predictive value, NPV = negative predictive value, PHQ-2 = Patient Health Questionnaire-2, PHQ-9 = Patient Health Questionnaire-9, GAD-7 = Generalized Anxiety Disorder-7, OASIS = Overall Anxiety and Severity Impairment Scale, HADS-D = Hospital Anxiety and Depression Scale – depression score, HADS-A = Hospital Anxiety and Depression Scale – anxiety score

Table 4
Construct validity: Correlations of anxiety and depression measures with pain, fatigue, and age (N = 230).

Measure	Pain (95%CI)	Fatigue (95%CI)	Age (95%CI)
Depression			
PHQ-2	0.50 (0.39, 0.59)	0.50 (0.40, 0.59)	−0.28 (−0.39, −0.16)
PHQ-9	0.62 (0.53, 0.69)	0.67 (0.59, 0.73)	−0.27 (−0.39, −0.15)
PROMIS Depression Short Form-8a	0.58 (0.49, 0.66)	0.55 (0.46, 0.64)	−0.20 (−0.33, −0.076)
HADS-D	0.64 (0.55, 0.71)	0.66 (0.57, 0.72)	−0.17 (−0.29, −0.038)
Kessler-6	0.63 (0.54, 0.570)	0.61 (0.53, 0.69)	−0.25 (−0.37, −0.13)
Anxiety			
OASIS	0.48 (0.37, 0.57)	0.48 (0.36, 0.57)	−0.26 (−0.38, −0.14)
GAD-7	0.48 (0.37, 0.57)	0.50 (0.40, 0.59)	−0.34 (−0.45, −0.22)
PROMIS Anxiety Short Form-8a	0.49 (0.38, 0.58)	0.49 (0.39, 0.58)	−0.22 (−0.34, −0.09)
HADS-A	0.53 (0.43, 0.62)	0.54 (0.44, 0.63)	−0.33 (−0.44, −0.02)

PHQ-2 = Patient Health Questionnaire-2, PHQ-9 = Patient Health Questionnaire-9, GAD-7 = Generalized Anxiety Disorder-7, OASIS = Overall Anxiety and Severity Impairment Scale, HADS-D = Hospital Anxiety and Depression Scale – depression score, HADS-A = Hospital Anxiety and Depression Scale – anxiety score.

the HADS-A with a cut-point of 8 (82%) and lowest for the GAD-7 (46%). Specificity ranged from 83% to 86% for all measures except the HADS-A with a cut-point of 8 (68%). The area under the ROC curve (AUC) did not differ between the HADS-A (AUC 0.83; 95%CI: 0.69, 0.98) and the GAD-7 (AUC 0.85; 95%CI: 0.74, 0.95, p = 0.72), the PROMIS Anxiety (AUC 0.85; 95%CI: 0.73, 0.97, p = 0.53), or OASIS (AUC 0.79; 95%CI: 0.65, 0.93, p = 0.19; Fig. e2).

The optimal cut-points for some of the anxiety screening measures differed from those recommended. The optimal cut-points were 7 for the GAD-7, 9 for the HADS-A, and 54.3 for the PROMIS Anxiety measure. The performance of these 'optimal' cut-points is shown in Table e2.

In the complementary analysis using any anxiety disorder as the criterion standard, sensitivities were lower than for analyses using generalized anxiety disorder, as were the AUC (Tables e3, e4).

3.2. Construct validity

Higher scores on all depression and anxiety measures were moderately associated with higher pain and fatigue scores (Table 4). As

Table 5
Reliability of anxiety and depression measures.

Instrument	Internal consistency (Cronbach's alpha) (95% CI)	Test-retest reliability Intraclass Correlation Coefficient (95%CI)[a]
Depression		
PHQ-2	0.80 (0.62, 0.98)	0.86 (0.81, 0.90)
PHQ-9	0.87 (0.80, 0.94)	0.85 (0.80, 0.89)
PROMIS Depression Short Form-8a	0.95 (0.88, 1.0)	0.85 (0.80, 0.89)
HADS-D	0.82 (0.75, 0.90)	0.83 (0.77, 0.87)
Kessler-6	0.87 (0.79, 0.95)	0.87 (0.82, 0.90)
Anxiety		
OASIS	0.90 (0.81, 0.99)	0.73 (0.64, 0.80)
GAD-7	0.92 (0.84, 0.99)	0.76 (0.68, 0.82)
PROMIS Anxiety Short Form-8a	0.95 (0.87, 1.0)	0.79 (0.72, 0.84)
HADS-A	0.86 (0.79, 0.94)	0.83 (0.77, 0.87)

[a] N = 158; PHQ-2 = Patient Health Questionnaire-2, PHQ-9 = Patient Health Questionnaire-9, GAD-7 = Generalized Anxiety Disorder-7, OASIS = Overall Anxiety and Severity Impairment Scale, HADS-D = Hospital Anxiety and Depression Scale – depression score, HADS-A = Hospital Anxiety and Depression Scale – anxiety score.

expected, age was less strongly associated with these measures.

3.3. Reliability

All depression anxiety measures had acceptable internal consistency reliability as measured by Cronbach's alpha (Table 5). Of the depression instruments, the PROMIS depression tool had the highest internal consistency reliability, while the HADS-D had the lowest. Of the anxiety instruments, the PROMIS anxiety tool had the highest internal consistency reliability, while the HADS-A had the lowest.

Test-retest reliability, as measured by an ICC, ranged from 0.83 (HADS-D) to 0.86 (PHQ-2) for the depression instruments (Table 5). On average, test-retest reliability for the anxiety instruments was lower than for the depression instruments, with values ranging from 0.73 (OASIS) to 0.83 (HADS-A).

4. Discussion

We examined the ability of several self-report measures to identify major depressive disorder or generalized anxiety disorder based on the SCID, and examined construct validity, internal consistency reliability and test-retest reliability. At the population level, the HADS-D (11) and HADS-A (11) most closely estimated the prevalence of major depression and generalized anxiety as determined by the SCID. At the individual level, diagnostic performance of the depression measures was good based on ROC analysis, but not excellent (< 0.90), and remarkably similar across measures. Similarly, the diagnostic performance of the anxiety measures was similar across measures, but not as good as that of the depression measures. Internal consistency reliability, construct validity and test-retest reliability were acceptable for all measures.

Generally, the performance of the depression measures as compared to the SCID was not as good as reported in prior studies, which have evaluated the criterion validity of depression measures in smaller North American MS samples ranging in size from 34 to 164 (Honarmand and Feinstein, 2009; Watson et al., 2014; Amtmann et al., 2015; Patten et al., 2015). In the two studies that evaluated the PHQ-2, sensitivities ranged from 70.8% to 80% and specificities ranged from 81% to 93% (Amtmann et al., 2015; Patten et al., 2015), findings comparable to ours. These studies also found that the sensitivities of the PHQ-9 ranged from 93.8% to 95%, higher than we observed, and specificities ranged from 61.2% to 85.9% (Amtmann et al., 2015; Patten et al., 2015). Both studies proposed alternative, higher cut-points for the PHQ-9 to the one recommended for the general population, but they differed with one suggesting 12 (Amtmann et al., 2015), consistent with our findings, while the other study suggested 15 (Patten et al., 2015). Given potential overlap of MS symptoms with somatic symptoms of depression, such higher cut-points may be preferred on theoretical grounds too. In the three studies that evaluated the HADS-D (Pilkonis et al., 2011; Polman et al., 2011; Patten et al., 2015), sensitivities for the cut-point of 11 ranged from 60% to 85%, while specificities ranged from 81% to 96.7%. These sensitivities are considerably higher than we observed at this cut-point. For the PROMIS depression measure, sensitivity at the proposed cut-point has been reported as 79.2% and specificity as 73.3% (Amtmann et al., 2015). The prior study evaluating the PROMIS depression measure also proposed an alternative cut-point (58.8) very similar to our recommended value (57.7). The differences in performance may reflect differences in the administration of the SCID, or our use of a larger, potentially more heterogeneous sample.

Our findings address the identified gap in knowledge regarding the reliability of the HADS, PHQ-9, and PROMIS Depression measures (Hind et al., 2016). Internal consistency reliability of the depression measures was acceptable for all measures (Bland and Altman, 1997). Test-retest reliability was good to excellent for all measures (Koo and Li, 2016).

A recent systematic review of the validity and reliability of anxiety measures found that only the HADS, GAD-7 and BAI had been assessed in MS (Litster et al., 2016); criterion validity of the GAD-7 had not been assessed, and reliability had not been assessed for the HADS or GAD-7. At the proposed cut-points for the general population the sensitivity of the GAD-7 was unacceptably low in our MS population, given that it was designed as a screen specifically for generalized anxiety disorder, while sensitivities for the other measures were modestly better. With the exception of the HADS-A using the cut-point of 8, specificities were similar across instruments and acceptable but lower than desired. In prior studies that evaluated the HADS-A (Honarmand and Feinstein, 2009; Watson et al., 2014), sensitivities for the cut-point of 11 ranged from 43.9% to 90%, while specificities ranged from 92% to 96%, estimates falling within the bounds of our estimates. The anxiety measures were not as effective as the measures of depression in identifying the presence of a current disorder, and performance worsened when we changed the criterion standard to any anxiety disorder, a more heterogeneous group. Although persons meeting the diagnostic criteria for major depression usually have a high level of current distress and interference, some persons with anxiety disorders may have more moderate levels of distress if they are able to avoid the situations that are most difficult for them.

Considering lower and upper bounds estimates, we found that internal consistency reliability was acceptable. One prior study evaluated the internal consistency reliability of the GAD-7, and found it to be 0.75 (Terrill et al., 2015), slightly lower than our lower bounds estimate. Test-retest reliability was moderate to good for the OASIS, GAD-7, and PROMIS Anxiety measures, and good for the HADS-A.

When choosing among these measures, other factors may be considered such as the time required for completion, availability in the public domain, availability of alternative forms, acceptability to patients and use for clinical or research purposes. Missing data were infrequent (0–3.9%) for all measures, and all of them require less than five minutes to complete, supporting their feasibility. The HADS captures depression and anxiety in a single instrument, which may have advantages in a busy clinical practice but costs are incurred if the measure is used for research. The PROMIS tools are increasingly used, and cross-walks are being developed to other measures. Other advantages of the PROMIS scales include development using

contemporary psychometric standards, equivalent forms of varying lengths, normative data from a largcpool which may be used with computerized adaptive testing (Pilkonis et al., 2011).

This study had limitations. All participants were drawn from a clinic population which could introduce bias. However, the MS Clinic delivers all specialized MS care in Manitoba. Some SCIDs were completed in person while others were completed by telephone. Nevertheless, previous studies suggest that findings from in-person and telephone administration of the SCID are comparable (Rohde et al., 1997; Cacciola et al., 1999). We used the SCID for DSM-IV because this was the prevailing version of the DSM criteria at the time the study was funded. Although the DSM-V (American Psychiatric Association, 2013) became available shortly after this study started little research was available on the application of the new criteria in the field. The precision of the sensitivity estimates was modest, reflecting the relatively low prevalence of depression and anxiety in the sample. Replication of our findings will be important. Finally, we did not evaluate responsiveness of the measures to change, which is relevant if these measures are to be used to evaluate symptom severity in interventional studies.

We evaluated the validity and reliability of several measures of depression and anxiety for an MS population. While the diagnostic accuracy of the measures was not as high as desired, all measures shared high negative predictive values; clinicians can be confident in excluding the presence of depression or generalized anxiety disorder in respondents with scores below recommended cut-points. Individuals with elevated scores may or may not meet formal diagnostic criteria for a disorder, but nonetheless may suffer from concerns such as subsyndromal disorders which may also warrant clinical attention (Pietrzak et al., 2012). Overall, performance of the depression measures and anxiety measures was remarkably similar, suggesting that factors such as ease of use should guide the choice of specific measure in clinical practice.

Acknowledgements

This study was funded by the Canadian Institutes of Health Research (THC-135234), Crohn's and Colitis Canada, a Research Manitoba Chair, and the Waugh Family Chair in Multiple Sclerosis (to RAM). Dr. Bernstein is supported in part by the Bingham Chair in Gastroenterology. Dr. Sareen is supported by CIHR #333252. Dr. Katz is supported by Research Manitoba and the Health and Stroke Foundation through the Manitoba Chair in Primary Prevention Research. Dr. Lix is supported by a Research Manitoba Chair. Dr. Zarychanski is supported by a CIHR New Investigator Salary Award. The sponsors had no role in the design and conduct of the study, collection and interpretation of the data, nor in the decision to submit the manuscript for publication.

Members of the CIHR Team in Defining the Burden and Managing the Effects of Psychiatric Comorbidity in Chronic Immunoinflammatory Disease are: Ruth Ann Marrie, James M Bolton, Jitender Sareen, John R Walker, Scott B Patten, Alexander Singer, Lisa M. Lix, Carol A Hitchon, Renée El-Gabalawy, Alan Katz, John D Fisk, Charles N Bernstein, Lesley Graff, Lindsay Berrigan, Ryan Zarychanski, Christine Peschken, James Marriott

Conflicts of interest

Ruth Ann Marrie has conducted clinical trials for Sanofi Aventis.

Carol Hitchon has research funds for unrelated studies from UCB Canada

Jitender Sareen holds stocks in Johnson and Johnson.

Charles Bernstein has consulted to Abbvie Canada, Ferring Canada, Janssen Canada, Pfizer Canada, Shire Canada, Takeda Canada, and Napo Pharmaceuticals and has consulted to Mylan Pharmaceuticals. He has received unrestricted educational grants from Abbvie Canada, Janssen Canada, Shire Canada, and Takeda Canada. He has been on speaker's bureau of Abbvie Canada, Ferring Canada and Shire Canada.

James Marriott has conducted trials for Biogen Idec and Roche.

All other authors have no conflicts of interest to declare.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.msard.2017.12.007.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, 5th ed., text rev. Washington, DC.

Amtmann, D., Bamer, A.M., Johnson, K.L., Ehde, D.M., Beier, M.L., Elzea, J.L., et al., 2015. A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis. J. Psychosom. Res. 79 (6), 550–557.

Beck, A.T., Epstein, N., Brown, G., Steer, R.A., 1988. An inventory for measuring clinical anxiety: psychometric properties. J. Consult Clin. Psychol. 56 (6), 893–897.

Bland, J.M., Altman, D.G., 1997. Statistics notes: cronbach's alpha. Br. Med. J. 314 (7080), 572.

Cacciola, J.S., Alterman, A.I., Rutherford, M.J., McKay, J.R., May, D.J., 1999. Comparability of telephone and In-person structured clinical interview for DSM-III-R (SCID) diagnoses. Assessment 6 (3), 235–242.

Cairney, J., Veldhuizen, S., Wade, T.J., Kurdyak, P., Streiner, D.L., 2007. Evaluation of 2 measures of psychological distress as screeners for depression in the general population. Can. J. Psychiatry 52 (2), 111–120.

Fiest, K.M., Fisk, J.D., Patten, S.B., Tremlett, H., Wolfson, C., McKay, K.A., et al., 2015. Comorbidity is associated with pain-related activity limitations in multiple sclerosis. Mult. Scler. Relat. Disord. 4 (5), 470–476.

First, M., Gibbon, M., Spitzer, R., Williams, J., 2002. User's Guide for the Structured Clinical Interview for DSM-IV-TR Axis I Disorders – Research Version – (SCID-I for DSM-IV-TR, November 2002 Revision). New York Biometrics Research Department, New York State Psychiatric Institute, New York.

Fisk, J.D., Doble, S.E., 2002. Construction and validation of a fatigue impact scale for daily administration (D-FIS). Qual. Life Res. 11 (3), 263.

Hedayati, S.S., Minhajuddin, A.T., Toto, R.D., Morris, D.W., Rush, A.J., 2009. Validation of depression screening scales in patients with CKD. Am. J. Kidney Dis. 54 (3), 433–439.

Hind, D., Kaklamanou, D., Beever, D., Webster, R., Lee, E., Barkham, M., et al., 2016. The assessment of depression in people with multiple sclerosis: a systematic review of psychometric validation studies. BMC Psychiatry 16 (1), 1–18.

Honarmand, K., Feinstein, A., 2009. Validation of the hospital anxiety and depression scale for use with multiple sclerosis patients. Mult. Scler. 15 (12), 1518–1524.

Janssens, A.C.J.W., van Doorn, P.A., de Boer, J.B., Kalkers, N.F., van der Meche, F.G.A., Passchier, J., et al., 2003. Anxiety and depression influence the relation between disability status and quality of life in multiple sclerosis. Mult. Scler. 9, 397–403.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2003. The patient health questionnaire-2: validity of a two-item depression screener. Med. Care 41 (11), 1284–1292.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33, 1444–1452.

Litster, B., Fiest, K.M., Patten, S.B., Fisk, J.D., Walker, J.R., Graff, L.A., et al., 2016. Screening tools for anxiety in persons with multiple sclerosis: a systematic review. Int. J. MS Care 18 (6), 273–281.

Marrie, R., Reider, N., Cohen, J., Stuve, O., Trojano, M., Sorensen, P.S., et al., 2015. The incidence and prevalence of psychiatric disorders in multiple sclerosis: a systematic review. Mult. Scler. J. 21 (3), 305–317.

Marrie, R.A., Elliott, L., Marriott, J., Cossoy, M., Tennakoon, A., Yu, N., 2015. Comorbidity increases the risk of hospitalizations in multiple sclerosis. Neurology 84 (4), 350–358.

Marrie, R.A., Graff, L.A., Walker, J.R., Fisk, J.D., Patten, S.B., Hitchon, C.A., et al., 2017. A prospective study of the effects of psychiatric comorbidity in immune-mediated inflammatory disease: rationale, protocol and participation (In press). JMIR Research Protocols.

Marrie, R.A., Patten, S.B., Tremlett, H., Wolfson, C., Warren, S., Svenson, L.W., et al., 2016. Sex differences in comorbidity at diagnosis of multiple sclerosis: a population-based study. Neurology 86 (14), 1279–1286.

McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.-P., Lublin, F.D., et al., 2001. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. Ann. Neurol. 50 (1), 121–127.

Mohr, D.C., Goodkin, D.E., Likosky, W., Beutler, L., Gatto, N., Langan, M.K., 1997. Identification of beck depression inventory items related to multiple sclerosis. J. Behav. Med. 20 (4), 407–414.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., et al., 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63 (7), 737–745.

Norman, S.B., Cissell, S.H., Means-Christensen, A.J., Stein, M.B., 2006. Development and validation of an overall anxiety severity and impairment scale (OASIS). Depress.

Anxiety 23 (4), 245–249.
Patten, S.B., Burton, J.M., Fiest, K.M., Wiebe, S., Bulloch, A.G., Koch, M., et al., 2015. Validity of four screening scales for major depression in MS. Mult. Scler. J. 21 (8), 1064–1071.
Pietrzak, R.H., Kinley, J., Afifi, T.O., Enns, M.W., Fawcett, J., Sareen, J., 2012. Subsyndromal depression in the United States: prevalence, course, and risk for incident psychiatric outcomes. Psychol. Med. 1–14.
Pilkonis, P.A., Choi, S.W., Reise, S.P., Stover, A.M., Riley, W.T., Cella, D., 2011. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS(R)): depression, anxiety, and anger. Assessment 18 (3), 263–283.
Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., et al., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 69 (2), 292–302.
Polman, C.H., Reingold, S.C., Edan, G., Filippi, M., Hartung, H.P., Kappos, L., et al., 2005. Diagnostic criteria for multiple sclerosis: 2005 revisions to the McDonald criteria. Ann. Neurol. 58 (6), 840–846.
Poser, C.M., Paty, D.W., Scheinberg, L., McDonald, W.I., Davis, F.A., Ebers, G.C., et al., 1983. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann. Neurol. 13 (3), 227–231.
Ritvo, P.G., Fischer, J.S., Miller, D.M., Andrews, H., Paty, D.W., LaRocca, N.G., 1997a. Multiple Sclerosis Quality of Life Inventory: A User's Manual. National Multiple Sclerosis Society, New York.
Ritvo, P.G., Fischer, J.S., Miller, D.M., Andrews, H., Paty, D.W., LaRocca, N.G., 1997b. Multiple Sclerosis Quality of Life Inventory: Technical Supplement. National Multiple Sclerosis Society, New York.
Rohde, P., Lewinsohn, P.M., Seeley, J.R., 1997. Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. Am. J. Psychiatry 154 (11), 1593–1598.
Spitzer, R., Kroenke, K., Williams, J.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166 (10), 1092–1097.
Spitzer, R., Kroenke, K., Williams, J.W., the Patient Health Questionnaire Primary Care Study Group, 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. JAMA 282 (18), 1737–1744.
Stafford, L., Berk, M., Jackson, H.J., 2007. Validity of the hospital anxiety and depression scale and patient health questionnaire-9 to screen for depression in patients with coronary artery disease. Gen. Hosp. Psychiatry 29 (5), 417–424.
Terrill, A.L., Hartoonian, N., Beier, M., Salem, R., Alschuler, K., 2015. The 7-item generalized anxiety disorder scale as a tool for measuring generalized anxiety in multiple sclerosis. Int. J. MS Care 17 (2), 49–56.
Watson, T.M., Ford, E., Worthington, E., Lincoln, N.B., 2014. Validation of mood measures for people with multiple sclerosis. Int. J. MS Care 16 (2), 105–109.
Williams Jr., J.W., Pignone, M., Ramirez, G., Perez Stellato, C., 2002. Identifying depression in primary care: a literature synthesis of case-finding instruments. General. Hosp. Psychiatry 24 (4), 225–237.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 32–35.
Zigmond, A.S., Snaith, R.P., 1983. The hospital anxiety and depression scale. Acta Psychiatr. Scand. 67 (6), 361–370.