A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis

Dagmar Amtmann *, Alyssa M. Bamer, Kurt L. Johnson, Dawn M. Ehde, Meghan L. Beier, Jamie L. Elzea, Charles H. Bombardier

Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA

ARTICLE IN PRESS

Contents lists available at ScienceDirect

Journal of Psychosomatic Research

A R T I C L E  I N F O

Article history:
Received 30 January 2015
Received in revised form 20 August 2015
Accepted 26 August 2015
Available online XXX

Keywords:
Depression
PHQ
CESD
PROMIS
Neuro-QOL
Beck depression
SCID
Diagnostic accuracy

A B S T R A C T

Background: Depression is one of the most prominent and debilitating symptoms in individuals with multiple sclerosis (MS), yet there is currently no consensus on the best instruments for depression screening in MS. More head to head comparisons of available screening instruments are needed to advise MS researchers and clinicians.

Methods: A cross-sectional comparison of the effectiveness of screening for MDD using multiple patient reported outcome (PRO) screeners against a modified SCID telephone interview was completed in 164 individuals with MS. Stratum goals were set for depression levels to ensure participation by people with borderline and higher levels of depression. Criterion standard was a modified SCID MDD module. PRO measures included the PHQ-9, BDI-FS, PROMIS depression, Neuro-QOL depression, M-PHQ-2, PHQ-2, and CESD.

Results: 48 (29%) individuals met the modified SCID criteria for MDD. The sensitivity of the PRO measures ranged from 60% to 100% while specificity ranged from 46% to 86%. The ROC area for the PRO measures ranged from 0.79 to 0.83. Revised (higher) cutoff scores were suggested by the ROC analyses for most self-reported screeners.

Limitations: Enrollment was stopped early because of difficulties with recruitment. Several SCID recording could not be reviewed and diagnosis confirmed.

Conclusions: CESD-10 and PHQ-9 had the best diagnostic performance using optimal cutoffs, but no one PRO measure stood out as significantly better than any other. Even when revised cutoff scores were used, none of the self-reported screeners identified people with MDD with adequate accuracy. More accurate self-reported screeners would facilitate diagnosing of MDD for both research and clinical purposes.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Multiple sclerosis (MS) is a chronic neurologic disease characterized by inflammation, demyelination, and neurodegeneration in the central nervous system. People with MS often report physical, cognitive and psychological symptoms, of which depression is one of the most prominent and debilitating [49]. The life-time prevalence of major depressive disorder (MDD) in individuals with MS has been estimated at over 50% [38]. In addition, the 12-month prevalence of MDD is approximately twice that of the general population at 15.7% [38]. Depression is associated with a host of poor outcomes in people with MS, including poorer overall health, non-adherence to disease modifying medications [54], loss of employment or reduction in work hours [39], an increased risk of suicidal ideation and completed suicide [17, 18, 42], greater cognitive dysfunction [4], and an overall reduction in quality of life [15, 25]. Despite these poor outcomes in MS, MDD remains under-recognized and undertreated [16, 20, 25].

Numerous instruments are available to assess depressive symptoms in people with MS, and they can also be used to identify depression cases in need of treatment. In the MS literature these measures are commonly referred to as depression “screening” instruments or measures [19, 34]. However, only a few published studies have compared the agreement between depression measures and structured diagnostic interviews for MDD in people with MS. In a series of newly diagnosed individuals with MS, the original Beck Depression Inventory (BDI) [6] (cutoff 13) produced 71% sensitivity, 79% specificity when compared to the Diagnostic Interview Schedule for DSM-III disorders [53]. A more recent similar study of the BDI-II in a clinical MS population produced similar results at 85% sensitivity and 76% specificity compared to the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) [56]. A two-item measure adapted from the Primary Care Evaluation of Mental Disorders (PRIME-MD) [51, 58] was compared to MDD diagnoses derived from the Structured Clinical Interview for DSM-IV Disorders (SCID) [23] and reported a 99% sensitivity and 87% specificity [35] in one study and 80% sensitivity and 93% specificity in a second
[40]. The depression subscale of the Hospital Anxiety and Depression Scale (HADS) has also been validated against the SCID or the Schedules for Clinical Assessment in Neuropsychiatry. A cutoff of eight in two studies found 90% sensitivity and 87.3% specificity [28] and 85% sensitivity and 82% specificity [40] while a third study found 77% sensitivity and 81% specificity at a cutoff of 11 [56]. Recently a study of the Center for Epidemiologic Studies Depression scale (CESD) found that it provided 95% sensitivity and 73% specificity compared to the SCID [40]. These findings provide important information regarding the utility of each measure, however many instruments have only been examined once and in small clinical samples and still more depression instruments are available that have not been examined.

There is currently no consensus on best practices for what instruments to use to assess depression in people with MS. As noted by the American Academy of Neurology in their recent review of the evidence for depression screening measures in MS [34], a number of depression measures used in the MS field lack strong evidence for their utility in identifying cases of depression, particularly relative to other commonly used measures. More head to head comparisons of measures are needed to advise researchers and clinicians. Among the many common measures whose utility for case identification in MS have not been well studied are the Beck Depression Inventory-Fast Screen (BDI-FS) [5], Patient Health Questionnaire-9 (PHQ-9) [30], and the CESD [57]. The BDI-FS is made up of a subset of items from the Beck scale proposed for use in MS [7], while the CESD and the PHQ-9 are two of the most commonly used instruments in the literature. Additionally, new instruments were recently developed with modern psychometric methods, and their clinical utility as MDD screeners in MS has yet to be examined. These include the Patient Reported Outcomes Measurement Information Systems (PROMIS®) and Neurological Quality of Life (Neuro-QOL) depression item banks. Both instruments provide population norms and have an added benefit of the availability of administration through computerized adaptive testing (CAT) [24,41]. The Neuro-QOL measure was specifically developed for use in neurological populations, including MS, and the psychometric properties of the PROMIS Depression short form have been examined in people with MS and found to be acceptable [2,46].

A recently published evidence-based guideline on the assessment and management of psychiatric disorders in individuals with MS emphasized the need for research comparing different self-report and diagnostic instruments for identifying psychiatric disorders, including MDD [34]. Therefore, the purpose of this study was to: (1) examine the correspondence between the standard diagnostic interview (SCID) and multiple self-report depression measures which are commonly used as tools for identifying MDD in MS; and (2) examine the published cutoffs for each measure and potentially identify optimal cutoffs for identifying people with MS for MDD.

2. Methods

2.1. Participants

Between September 2011 and March 2012 individuals with MS were recruited through invitation letters, print advertisements, and referrals from active research studies at the University of Washington (UW) in Seattle, WA. Individuals were sent invitation letters if they had participated in past research studies at the UW and indicated interest in future studies or were members of the UW disability registry. Individuals were required to be 18 years or older, self-report a definitive MS diagnosis by a physician, be able to read and understand English, and have access to a telephone from which they could answer sensitive questions. In order to ensure the participant pool represented all levels of depressive symptoms and included more participants with borderline depression, enrollment targets were stratified by PHQ-9 scores at the time of screening. This was done in order to ensure that the performance of the instruments was examined in participants across the whole depression continuum and individuals with no or low depressive symptoms, who are the most likely to volunteer for research studies, were not overrepresented. Initial recruitment targets were set at 200 total with 10% having no or minimal depressive symptoms (PHQ-9 score ≤5), 20% mild (PHQ-9 score 5–9), 30% moderate (PHQ-9 score 10–14), 30% moderate-severe (PHQ-9 score 15–19), and 10% severe (PHQ-9 score ≥20). Initial mailings to past research participants were done through random selection. However, individuals reporting a moderate to high level of depressive symptoms in past survey studies were specifically targeted in later recruitment mailings in order to try to meet recruitment goals for those strata. If a participant did not respond to the invitation recruitment mailing an attempt was made to call them approximately two weeks after the mailing. Once recruitment goals were met for a stratum, individuals scoring within that range upon screening were determined ineligible for study enrollment. All procedures, including written informed consent, were approved by the UW Human Subjects Division. Participants were paid $25 upon study completion.

2.2. Procedures

At screening, potential participants completed the PHQ-9 on the telephone in order to determine study eligibility. Those who were eligible and interested were mailed a packet containing the six self-report depression measures described below. They opened and completed the self-report measures at the scheduled time of their telephone interview. The average time between initial screening into the study and the interview (that included both the SCID and responding to self-report measures) was 10 days (range: three to 29 days). The research interviewer was available for questions but did not otherwise participate or interact with the participants while they responded to the self-reported measures. Upon completion of the self-report measures, the researcher conducted the SCID MDD module with the participant [22]. Thus, all self-report measures were completed within minutes of the SCID as they were done while the interviewer waited on the phone prior to administering the interview. Participants were then instructed to mail the self-report packet back to research staff.

2.3. Measures

2.3.1. Disease and demographic characteristics

Participants were asked to provide their age, gender, race and ethnicity, education level, and employment status. In addition, participants reported the year of their MS diagnosis and completed a self-report version of the mobility section of the Expanded Disability Status Scale (EDSS) [9] in order to estimate MS severity level.

2.3.2. Self-report depression measures

Participants completed six different paper and pencil self-report measures of depressive symptoms. The following measures were selected for inclusion: PROMIS Depression short form (SF), Neuro-QOL depression SF, PHQ-9, modified PHQ-2, BDI-FS, and the CESD. Order of administration was counterbalanced to address potential order effects [14].

2.3.2.1. PROMIS depression (PROMIS-D) SF. The Patient Reported Outcomes Measurement Information System (PROMIS®, www.nihpromis.org) depression SF version 1.0 includes eight items that were selected from the PROMIS-D item bank using CAT simulation results, item information, and content [41]. The PROMIS-D item bank was developed using item response theory (IRT), and scores on the SF are directly comparable to CAT scores or other SF scores. Unlike most depression measures, the PROMIS-D item bank does not measure behavioral and somatic indicators [41]. All eight questions are rated on a five point Likert scale (1 = never; 5 = always), and respondents are asked to recall how they felt over the past seven days. Scores are reported on a T-score metric (mean = 50; standard deviation (SD) = 10) that is

Please cite this article as: D. Amtmann, et al., A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis, J Psychosom Res (2015), http://dx.doi.org/10.1016/j.jpsychores.2015.08.007
centered on a sample that is representative of the United States general population by age, gender, and race/ethnicity [11,32]. The short form was scored following the PROMIS instructions using the raw score/scale score look-up tables [43]. Higher scores indicate higher levels of depression. PROMIS-D SF has been used in MS [2,46]. No studies to date have identified an optimal cutoff on the PROMIS-D in identifying a MDD. However, prior research linking the PROMIS-D to the PHQ-9 and CESD suggests that a score of 59.9 is equivalent to a PHQ-9 score of 10, and a score of 56.2 is equivalent to a CESD score of 16 [12]. These two cutoffs have been recommended for use in identifying those with at least moderate depression [31,44,50].

2.3.2.4. PHQ-9. On the Patient Health Questionnaire-9 (PHQ-9) respondents are asked to rate how often they have been bothered by each depressive symptom over the last two weeks on a four point scale [30]. The nine items parallel the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for MDD [1]. Items are summed, and total scores range from zero to 27, with scores of 10 or greater suggesting a Major Depressive Episode (MDE) [30,50]. A scoring algorithm paralleling the SCID DSM-IV diagnosis of MDD has also been recommended for identifying those with MDD [50]. The PHQ-9 has been used to measure depression in multiple medical populations, including MS [13,21,48].

2.3.2.5. Modified PHQ-2 (M-PHQ-2). This screener is based on the first two questions of the Primary Care Evaluation of Mental Disorders (PRIME-MD) depression section [51], and captures depressed mood and anhedonia, the two essential criteria of DSM-IV MDD. The two questions are identical to the PHQ-2, with changes to the response options only. Rather than the four response options available on the PHQ-2, patients were asked to answer with either a “yes” or “no.” Answering yes to either question has been used to indicate probable MDD in primary care patients, [51] urgent care veteran patients, [58] and people with MS [35]. We included this modified measure specifically to try and replicate the results found by Mohr et al. in their 2007 study in MS patients.

2.3.2.6. BDI-FS. The Beck Depression Inventory-Fast Screen has seven groups of statements, and respondents are asked to choose the statement that most reflects how they were feeling within the last two weeks [5]. Each statement set has scores ranging from zero to three, and scores are summed with a final score range of zero to 21. A cutoff of four or greater is recommended for identifying those with at least mild depression [5]. The BDI-FS was specifically designed to be used in medical settings and, therefore, does not include somatic symptoms. Preliminary evidence supports the use of the BDI-FS in persons with MS [7].

2.3.2.7. CESD-20 and CESD-10. The full version of the Center for Epidemiologic Studies Depression Scale has 20 items, and respondents are asked to rate the frequency of both positive and negative feelings within the “past week” on a four-point scale [44]. Scores are summed and can range from zero to 60. A cutoff of 16 has been recommended for identifying those with clinically significant levels of depressive symptoms [44,57]. The CESD was developed for use in primary care settings but evidence supports its use in individuals with MS [55]. The CESD-10 is a SF that includes half of the original CESD scale items [3]. As with the full version, scores are summed across all items. The score ranges from zero to 30 and a cutoff of 10 or higher on the SF has been recommended to identify those with clinically significant depressive symptoms [3]. In this study, the CESD-10 was not administered separately, rather the subset of ten items from the full CESD were scored separately in order to create a CESD-10 score.

2.3.3. MDD diagnosis

The criterion standard in this study was a diagnosis of MDD during the past month based on telephone administration of SCID [22]. The SCID has been validated for use over the telephone, and the interview questions and decision trees are directly comparable to in person administration [45,47]. Bachelor’s level research assistants underwent training in SCID administration by observing videos and completing in person training with a doctoral level Advanced Registered Nurse Practitioner and a Masters level research therapist. The questions from the MDD module of the SCID were asked using the specific wording and established guidelines for administration and coding of SCID responses [22,23]. All recorded SCIDs were reviewed in early weeks of the study to confirm that each administrator was achieving 90% inter-rater reliability (IRR) with the research therapist. Ongoing feedback was provided to research staff in order to improve interview technique and accuracy. The trained research assistants conducted the SCID interviews and based on the published protocols rendered a preliminary diagnosis of MDD. All telephone interviews were recorded for quality assurance, but 14 recordings were unintelligible. One of two clinicians with extensive experience in SCID administration listened to all interview recordings to confirm or revise the MDD status assigned by research assistants. This final diagnosis of MDD was used in the analyses because the authors consider the diagnosis by clinicians more accurate due to their greater clinical training and experience with diagnosing depression in people with MS. Because the diagnosis of MDD was made in most cases by the clinicians, rather than the person who administered the SCID, we refer to our procedure as the “modified SCID” in the results. Both the research assistants and clinicians were unaware of the participants’ self-reported depression scale scores when conducting or listening to the SCID interviews.

2.4. Analyses

In order to evaluate the performance of each self-report depression measure we compared previously recommended published cutoffs or scoring algorithms with the results of the modified SCID MDD diagnosis.

Please cite this article as: D. Amtmann, et al., A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis, J Psychosom Res (2015), http://dx.doi.org/10.1016/j.jpsychores.2015.08.007
Criterion validity was evaluated by calculating sensitivity, specificity, positive and negative likelihood ratios, and Cohen’s kappa. Sensitivity is defined as probability of screening positive for MDD among those who have MDD (using the modified SCID interview as the gold standard). Specificity is similarly defined but is the probability of screening negative on the PRO measure among those who do not have MDD according to their modified SCID interview. Positive likelihood ratios are defined as sensitivity divided by 1 minus specificity (LR+ = sensitivity/(1-specificity)) and negative likelihood ratios are defined as one minus sensitivity divided by specificity (LR− = 1-sensitivity)/specificity). A positive likelihood ratio (LR+) of greater than 1 indicates the test result is associated with the disease while a negative likelihood ratio (LR−) less than 1 indicates that the result is associated with absence of the disease. A LR+ greater than 10 has been suggested to indicate a conclusive increase in likelihood of disease while a LR− of <0.1 suggests a large and often conclusive decrease in likelihood of disease [26]. In addition, receiver operating characteristic curves (ROC) were generated for each measure and the area under the ROC curve was calculated. Optimal cutoff scores for each measure (with exception of PHQ-2 and the M-PHQ-2) were calculated by maximizing the Youden index (sensitivity + specificity − 1) [60]. For each of the newly calculated optimal cutoffs, performance was again evaluated by examining sensitivity, specificity, positive and negative likelihood ratios, and Cohen’s kappa. A sensitivity analysis evaluating optimal cutoffs only in individuals with clinician confirmed diagnoses was conducted to evaluate the potential impact of using the research assistants’ initial evaluation in the 14 cases without intelligible recordings. Sensitivity analyses using the research assistants’ initial evaluation of MDD on the SCID interview for all participants were also conducted. All analyses were conducted using Stata 12.1 software [52].

3. Results

3.1. Participants

A total of 372 individuals from past research studies were mailed invitation letters, 28 were referred from active research studies, and 11 responded to print advertisements. Of these, 50 declined to participate, 72 were determined ineligible at the time of enrollment (70 due to strata reaching recruitment capacity), three were deceased, one was unable to complete the interview due to severe comorbid conditions, and 119 did not respond or were lost to follow-up. A total of 166 individuals completed the study, but two cases were excluded from the main analyses after clinician reviews determined that these SCID interviews lacked the detail required to make a diagnosis. The mean age of the sample was 53.2 years, and mean MS duration was 15 years. Consistent with the distribution of MS in the population, the majority of participants were female (77%), white (90%), married or living with a partner (64%), and unemployed (79%; see Table 1 for more sample characteristics). Of the 164 participants, 14 had missing or inaudible audio recordings of their SCID interviews. For these 14 participants the research assistant’s initial assessment of MDD status was used as the criterion standard as they could not be re-evaluated by the research clinicians. Of the 150 SCID interviews that were re-evaluated by the research clinicians, 20 had their MDD status changed from the initial research assistant’s assessment (87% agreement). Analyses to identify optimal cutoffs were repeated after dropping these 14 individuals and all identified optimal cutoffs remained consistent with initial analyses inclusive of these individuals. Results of sensitivity analyses using the research assistants’ initial MDD assessment for all 166 completed interviews are provided in Appendix 1. Recruitment was ceased prior to reaching the initial goal of 200 participants due to difficulties in recruitment and resource limitations. The only screening stratum that reached capacity during recruitment was the lowest PHQ-9 stratum (PHQ-9 score <5). The distribution of PHQ-9 intake screening scores for the final 164 participants is provided in Table 1.

### Table 1

<table>
<thead>
<tr>
<th>Characteristics of sample of persons with MS (N = 164).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ± SD number (%)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
</tr>
<tr>
<td><strong>MS duration (years)</strong></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
</tr>
<tr>
<td>0.5 Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td><strong>Married/living with partner</strong></td>
</tr>
<tr>
<td><strong>Currently employed</strong></td>
</tr>
<tr>
<td><strong>Race</strong></td>
</tr>
<tr>
<td>White non-Hispanic</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td><strong>Education</strong></td>
</tr>
<tr>
<td>High School (or less)</td>
</tr>
<tr>
<td>Some College/Associates Degree</td>
</tr>
<tr>
<td>Bachelor’s Degree</td>
</tr>
<tr>
<td>Advanced Degree</td>
</tr>
<tr>
<td><strong>EDSS Group (n = 163)</strong></td>
</tr>
<tr>
<td>Mild (&lt;4.0)</td>
</tr>
<tr>
<td>Moderate (4.5–6.5)</td>
</tr>
<tr>
<td>Severe (≥7.0)</td>
</tr>
<tr>
<td><strong>PHQ-9 Screening Intake Group</strong></td>
</tr>
<tr>
<td>None (&lt;5)</td>
</tr>
<tr>
<td>Mild (5–9)</td>
</tr>
<tr>
<td>Moderate (10–14)</td>
</tr>
<tr>
<td>Moderate–severe (15–19)</td>
</tr>
<tr>
<td>Severe (≥20)</td>
</tr>
</tbody>
</table>

SD: Standard deviation; EDSS: Expanded disability status scale; PHQ: Patient health questionnaire.

3.2. Depression measures comparison

Of the 164 participants included in the primary analyses, 48 (29%) were identified as meeting criteria for a major depressive disorder during the past month using the modified SCID criterion standard. The sensitivity, specificity, likelihood ratios, and kappa for all previously defined or published cutoffs for each of the PRO depression measures are presented in Table 2. The percentage of people identified as having significant depression by the PRO measures ranged from 27.4% (PHQ-9 MDD algorithm) to 67.7% (M-PHQ-2). Sensitivity values (true positives) ranged from 60.4% (PHQ-9 MDD scoring algorithm) to 100% (M-PHQ-2 and CESD) while specificity values (true negatives) ranged from 45.7% (M-PHQ-2) to 86.2% (PHQ-9 MDD scoring algorithm). Only 13% of the sample was incorrectly classified as not being depressed (compared to the modified SCID) by any one of the eight cutoffs presented in Table 2. Alternatively, 50% of individuals were incorrectly identified as having significant depression (compared to the modified SCID) by at least one cutoff in Table 2, and 31% were incorrectly categorized by four or more of the measures in Table 2. Kappa values ranged from 0.33 (PRIME-MD-2 and BDI) to 0.49 (PHQ-2). With the exception of the PHQ-2 and M-PHQ-2 (binary outcomes), the ROC curves for each PRO measure are presented in Fig. 1 and underlying sensitivity and specificity at all cutoffs for each measure is provided in Appendix 2. The area under the ROC curves is also provided for comparison in Table 3.

Optimal cutoffs for each PRO measure in Fig. 1 were identified by choosing the score on each that maximized the Youdin Index (i.e., maximized the sum of the sensitivity and specificity). With the exception of one of the PROMIS-D cutoffs, in all cases the identified optimal cutoff for people with MS was higher than that suggested by the scale developers. In addition, none of the identified cutoffs (optimal or developer suggested) met the Youden Index threshold (sensitivity + specificity − 1 ≥ 0.8) suggested for an adequate diagnostic test [60]. The sensitivity and specificity values for the identified optimal cutoffs ranged from 81% to 94% for sensitivity and 68% to 83% for specificity (Table 3). Kappa values for the optimal cutoffs ranged from 0.48 to 0.56 (Table 3).
Table 2

Test characteristics for previously defined cutoffs or scoring methods for patient reported outcome measures in an MS sample (N = 164).

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Cutoff or scoring method</th>
<th>Percent at or above cutoff (n)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Youden Index</th>
<th>LR+ (95% CI)</th>
<th>LR- (95% CI)</th>
<th>Kappa (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROMIS Depression SF</td>
<td>≥59.9</td>
<td>42.1% (69)</td>
<td>79.2% (65.0–89.5)</td>
<td>73.3% (64.3–81.1)</td>
<td>0.525</td>
<td>2.96 (2.12–4.14)</td>
<td>0.28 (0.16–0.50)</td>
<td>0.46 (0.08)</td>
</tr>
<tr>
<td>PROMIS Depression SF</td>
<td>≥56.2</td>
<td>64.0% (105)</td>
<td>97.9% (88.9–99.9)</td>
<td>50.0% (40.6–59.4)</td>
<td>0.479</td>
<td>1.96 (1.16–2.4)</td>
<td>0.04 (0.01–0.29)</td>
<td>0.36 (0.06)</td>
</tr>
<tr>
<td>PHQ-9 algorithm</td>
<td>≥10</td>
<td>54.8% (90)</td>
<td>86.0% (78.6–91.9)</td>
<td>4.66</td>
<td>4.38 (2.63–7.29)</td>
<td>0.46 (0.32–0.66)</td>
<td>0.47 (0.08)</td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td>≥2 on either item</td>
<td>33.7% (56)</td>
<td>81.2% (72.7–87.7)</td>
<td>0.518</td>
<td>0.36 (0.23–0.56)</td>
<td>0.49 (0.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-PHQ-2</td>
<td>“Yes” to either item</td>
<td>67.7% (111)</td>
<td>100.0% (92.6–100.0)</td>
<td>0.457</td>
<td>1.84 (1.56–2.18)</td>
<td>0.0 (N/A)</td>
<td>0.33 (0.06)</td>
<td></td>
</tr>
<tr>
<td>BDI-FS</td>
<td>≥4</td>
<td>66.3% (108)</td>
<td>97.9% (88.9–99.9)</td>
<td>0.494</td>
<td>1.85 (1.55–2.20)</td>
<td>0.04 (0.01–0.31)</td>
<td>0.33 (0.06)</td>
<td></td>
</tr>
<tr>
<td>CESD-20</td>
<td>≥16</td>
<td>62.8% (103)</td>
<td>100.0% (92.6–100.0)</td>
<td>0.526</td>
<td>2.11 (1.74–2.55)</td>
<td>0.0 (N/A)</td>
<td>0.39 (0.06)</td>
<td></td>
</tr>
<tr>
<td>CESD-10</td>
<td>≥10</td>
<td>67.1% (110)</td>
<td>100.0% (92.6–100.0)</td>
<td>0.466</td>
<td>1.87 (1.58–2.22)</td>
<td>0.0 (N/A)</td>
<td>0.34 (0.06)</td>
<td></td>
</tr>
</tbody>
</table>

BDI-FS: Beck depression inventory fast screen; CESD: Center for epidemiologic studies depression scale; CI: Confidence interval; LR: Likelihood ratio; MDD: Major depressive disorder; PHQ: Patient health questionnaire; PROMIS: Patient reported outcomes measurement information systems; SE: Standard error; SF: Short form.

4. Discussion

The purpose of this study was to examine the correspondence between the diagnosis of MDD based on the SCID and six self-report instruments used to identify people with MDD. Our results suggest that none of the self-report instruments identified people with MDD with adequate accuracy, suggesting a need for development of a more accurate screening instrument or scoring algorithm that results in more accurate diagnosis in persons with MS. Depending on their purpose, the researchers/clinicians may choose one measure over another based on the measure’s other characteristics, such as whether somatic indicators should be included, use of the specific instrument in previous studies for comparability, ability to administer via CAT, availability of norms, or the length of questionnaire. This finding is generally consistent with the recent evidence-based guidelines on the assessment and management of psychiatric disorders in individuals with MS [34], which found insufficient evidence to support or refute any of the depression screening measures (BDI, M-PHQ-2, CES-D) examined in their review. It was also notable that the BDI-FS did not outperform many of the other measures, despite preliminary evidence for it in MS [7] and its inclusion in the widely used MACFIMS protocol [8]. Using the Youden index with previously published cutoffs, the PHQ-9 was slightly more effective in identifying individuals with MDD in the current study than the other instruments. However, if using the optimal cutoffs in Table 3, the CESD-10 and PHQ-9 performed slightly better than the other measures based solely on the Youden Index. All measures except for the PHQ-9 MDD scoring algorithm overestimated the percentage of people with MDD diagnosis, many by over 30%. Of those who were mis-categorized as significantly depressed by at least four of the measures in Table 2, scores on the PHQ-9 ranged from four to 24 with a mean of 12, and scores on the CESD-20 ranged from nine to 50 with a mean of 27. This indicates that some individuals can score very high on the self-report measures, yet still not have significant depression when interviewed by a clinician.

Our results are directly comparable to previously published results on the PHQ-9, M-PHQ-2 and CESD-20. For the rest of the measures we couldn’t directly compare our results to previous studies because either the information was not available in people with MS (PROMIS), has never been studied before (NeuroQOL), or a different version or scoring of the instrument was used (BDI, PHQ-2). For M-PHQ-2 we did not...
replicate the previously published results by Mohr et al. [35]. In our sample sensitivity was similar, but specificity was considerably lower indicating that the instrument overestimated the number of people with MDD by almost 40% [35]. For CESD-20, our results at the published cutoff of 16 for sensitivity were similar to that of Pandya et al. [37] and slightly higher than that of Patten et al. [40]. However, our specificity is substantially lower than that of Patten et al. at that cutoff (85.4% versus 52.6%). Interestingly, in both Patten et al. and this study, optimal cutoffs on the CESD were substantially higher (21 and 22 respectively). In addition, both studies found that raising the cutoff also increased the performance of the PHQ-9. As with the CESD, PHQ-9 sensitivity in this study was similar to Patten et al. at the recommended cutoff but sensitivity was substantially lower (88% vs 61%). In addition, it is important to note that though we could not directly compare our results with other studies as described above, the ranges of specificity and sensitivity found in the current study are very similar to those of other published studies which examined other measures in MS [28,53,56].

Our results also suggest that higher cutoffs for all measures except PROMIS improve diagnostic effectiveness in people with MS. However, even the new cutoffs do not reach the recommended effectiveness suggested for an adequate diagnostic test [60]. The sensitivity and specificity achieved using the different cutoffs can be helpful for researchers and clinicians in selecting the instrument and the cutoff that best accomplishes their goals. In addition, although the measures may not reach the recommended effectiveness for a diagnostic test, this does not undermine their use as tools for monitoring depressive symptoms in epidemiological studies.

Finally, it is interesting to note that the relatively recent addition of IRT-based Neuro-QOL instrument that was developed specifically for people with neurologic conditions (including people with MS) does not seem to provide any advantage over the IRT-based PROMIS-D that was developed to be used across different populations. This suggests that IRT-based instruments are population invariant providing the calibrating samples are representative of the population. It also suggests that disease specific instruments do not necessarily result in greater precision of measuring universal constructs, such as depression, that appear to be the same in people with MS as in otherwise healthy people or people with other health conditions.

4.1. Limitations

This study had several limitations. First, we did not conduct a sample size calculation prior to completing the study, but rather chose our estimated target for enrollment with the aim to examine functioning across the depression continuum and based on available study resources. When compared to similar studies in MS, such as Patten et al. [40], our sample size is sufficient to estimate sensitivity and specificity with adequate precision. We believe that it was not the sample size, but the sampling strategy of recruiting more people with moderate levels of depressive symptoms that affected the confidence intervals around our estimates. We also did not reach our target of 200 participants due to difficulties with recruitment and this reduced sample size may have resulted in wider confidence intervals around our accuracy estimates. Second, bachelor’s level research staff conducted the SCIDs resulting in two cases that had to be dropped because the SCID was stopped prematurely. In addition, in 14 cases where the recording failed, it was not possible to review the SCID by the experienced clinicians, and the diagnosis assigned by the research staff was used in the analyses. To examine the impact of these cases we re-ran all the analyses without the 14 cases and found that even when dropping these cases, the identified optimal cutoffs for all measures were identical, suggesting these 14 cases had minimal (if any) impact on our results. In addition, in the standard administration of the SCID the diagnosis of MDD is made by the person who administered the SCID. In this study we changed some of the initial diagnoses by the research assistants based on review by more experienced clinicians, resulting in non-standard administration. In our experience, many studies use expert oversight of raters and this procedure resulted in more accurate data. However, for full disclosure we also provided the analyses using the preliminary diagnosis in Appendix 1. Third, we used the inclusive method to score the SCID and the self-report measures [59]. That is, we scored potentially transdiagnostic symptoms (e.g., fatigue, poor concentration) as contributing to a diagnosis of depression, if present, and did not attempt to ascertain whether transdiagnostic symptoms were more likely to be attributable to the primary effects of MS than to MDD. Some research suggests transdiagnostic symptoms such as fatigue should be down-weighted to account for this overlap [27]. On the other hand, depression screening measures constructed to exclude transdiagnostic symptoms, such as the BDI-FS, PROMIS and Neuro-QOL, did not perform significantly better than measures that included these symptoms, such as the PHQ-9. Whether the hospital anxiety and depression scale (HADS) [61], another commonly used measure developed to exclude such symptoms, would have had a different outcome is unknown and worthy of future research. Fourth, the PHQ-9 instrument was administered twice. Once by phone at the initial screening to assign participant to intake severity category, and second on paper along with all remaining self-report instruments that were only administered once. It is possible that the repeated administration of the PHQ-9 (average 10 days later) bias the results, although test-retest reliability of the PHQ-9 has been reported to be excellent (ICC = 0.96) [33] and we do not believe this has affected its diagnostic effectiveness. Nevertheless, perhaps selecting a different instrument, one not examined in this study, would have been preferable. Fifth, we did not collect information about prior or current treatment of depression, nor did we exclude individuals based on those criteria. This may have led...
to an overestimation in the accuracy of the screening instruments, as exclusion of these individuals in a sensitivity analyses by Patten et al. [40] resulted in decreased predictive value of the instruments. Finally, because we set enrollment targets related to the level of depression, in this study we could not examine the prevalence of MDD in MS or the positive and negative predictive values of the measures. The stratified enrollment targets may also explain why the self-report measures did not perform as highly in our study as in some other studies and why our confidence intervals around our estimates are rather than similar studies with fewer participants (see [40]). While this strategy was a disadvantage in some respects, it can also be considered a strength because sufficient numbers of people along the whole continuum of depression are unlikely to be achieved in clinic-based studies. In addition, if these self-report measures are to be used as case finding instruments in borderline cases, our stratification approach should provide more accurate estimates of sensitivity and specificity than in studies that enroll clinic based samples that include either mostly people with low depressive symptoms (such as found in the general medical practice) or mostly those with high depressive symptoms (such as those found in mental health clinics).

5. Conclusions
In summary, our results as well as the Minden et al. [34] evidence based review suggest that there is much work to be done to improve accurate identification of MDD using screening measures in this population. It is possible that because of complexity of diagnosing MDD self-reported screening instruments are limited in their ability to accurately identify MDD in people with MS. It may be useful to examine whether a set of questions drawn from all the instruments would result in more accurate, however development of such an instrument would involve potentially difficult to resolve intellectual property issues. While the measures examined in this study are commonly used in MS clinical care and research and have published cutoffs, no instrument clearly outperformed others or could be recommended as a gold standard. Future research may want to examine the use of these measures in specific contexts (e.g., outpatient MS specialty care center, epidemiologic study), including determining acceptable/desired specificity and sensitivity for that specific context a priori. Because the current literature is equivocal with regards to the importance of including or excluding transdiagnostic symptoms (e.g., fatigue, concentration difficulties), it will be important for future investigators to examine the relationship of individual items to MDD diagnosis (i.e., which of the items are the best predictors of MDD diagnosis) to better understand the role of such symptoms in diagnosing depression.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jpsychores.2015.08.007.

Acknowledgments
The contents of this manuscript were developed under a grant from the Department of Education, NIDRR grant number H133B080025. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. This review was also supported in part by a grant from the National Multiple Sclerosis Society, grant number MB 0008. Neither funding agency played any role in the study design or the Federal Government. This review was also supported in part by a Department of Education, NIDRR grant number H133B080025. A.T. Beck, R.A. Steer, G.K. Brown, BDI-Fast Screen for Medical Patients: manual, Pearson, San Antonio, TX, 2000.


D. Grimes, K.F. Schulz, Re...
Please cite this article as: D. Amtmann, et al., A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis, J Psychosom Res (2015), http://dx.doi.org/10.1016/j.jpsychores.2015.08.007