Published in final edited form as:

J Behav Med. 2015 April ; 38(2): 237–250. doi:10.1007/s10865-014-9574-5.

Disentangling Multiple Sclerosis & Depression: An Adjusted Depression Screening Score for Patient-Centered Care

Douglas D. Gunzler, PhD¹, Adam Perzynski, PhD¹, Nathan Morris, PhD², Robert Bermel, MD³, Steven Lewis, MS, MBA¹, and Deborah Miller, PhD³

¹Case Western Reserve University, Center for Health Care Research & Policy, MetroHealth Medical Center, Cleveland, Ohio

²Case Western Reserve University, Department of Epidemiology and Biostatistics, Cleveland, Ohio

³Mellen Center for Multiple Sclerosis Treatment and Research, Cleveland Clinic, Cleveland, Ohio

Abstract

Screening for depression can be challenging in Multiple Sclerosis (MS) patients due to the overlap of depressive symptoms with other symptoms, such as fatigue, cognitive impairment and functional impairment, for MS patients. The aim of this study was to understand these overlapping symptoms and subsequently develop an adjusted depression screening tool for better clinical assessment of depressive symptoms in MS patients. We evaluated 3,507 MS patients with a self-reported depression screening (PHQ-9) score using a multiple indicator multiple cause (MIMIC) modeling approach. Our models showed significant differential item functioning effects denoting significant overlap of depressive symptoms with all MS symptoms under study and good model fit. The magnitude of the overlap was especially large for fatigue. Adjusted depression screening scales were formed based on factor scores and loadings that will allow clinicians to understand the depressive symptoms separate from other symptoms for MS patients for improved patient care.

Contact information for the corresponding author: Douglas D. Gunzler, PhD, Case Western Reserve University, Center for Health Care Research & Policy, MetroHealth Medical Center, 2500 MetroHealth Drive, Cleveland, Ohio, 44109-1998, dgunzler@metrohealth.org, Phone: 216-778-2764, Fax: 216-778-3945.

Contact information for the other authors:

Adam Perzynski, PhD, Case Western Reserve University, Center for Health Care Research & Policy, MetroHealth Medical Center, 2500 MetroHealth Drive, Cleveland, Ohio, 44109-1998, atp5@case.edu, Phone: 216-778-2850, Fax: 216-778-3945

Nathan Morris, PhD, Case Western Reserve University, Department of Epidemiology and Biostatistics, 10900 Euclid Ave, Cleveland, OH 44106-4945, njm18@case.edu, Phone: 216-368-5634

Robert Bermel, MD, Mellen Center, Cleveland Clinic Main Campus, Mail Code U10, 9500 Euclid Avenue, Cleveland, OH 44195, BERMELR@ccf.org, Phone: 216-444-8614, Fax: 216-445-6259

Steven Lewis, MS, MBA, Case Western Reserve University, Center for Health Care Research & Policy, MetroHealth Medical Center, 2500 MetroHealth Drive, Cleveland, Ohio, 44109-1998, sal25@case.edu, Phone: 216-778-4398, Fax: 216-778-3945

Deborah Miller, PhD, Cleveland Clinic Main Campus, Mail Code U10, 9500 Euclid Avenue, Cleveland, OH 44195, MILLERD@ccf.org, Phone: 216-444-8600, Fax: 216-445-6259

Conflicts of Interest

Douglas Gunzler, Adam Perzynski, Nathan Morris, Steven Lewis and Deborah Miller declare that they have no conflict of interest. Robert Bermel has received research grants from Novartis.

Keywords

multiple sclerosis; fatigue; depression; structural equation modeling; factor analysis; multiple indicator multiple cause model

Introduction

Multiple sclerosis (MS) is the most common progressive neurological disease of young adults and affects approximately 400,000 persons in the United States (Multiple Sclerosis Association of America, 2014). Depression is the most frequent psychiatric diagnosis in MS patients, with lifetime risk estimated at ~50% (Goldman Consensus Panel, 2005). Patients with MS seem to show increased severity of depressive symptoms compared to patients with other chronic neurological conditions (Wallin et al., 2006).

While MS patients commonly experience depressive symptoms, clinicians cannot reliably distinguish the roles of MS and depression in fatigue, cognitive impairment and functional impairment (Ferrando et al., 2007; Wallin et al., 2006). This can lead to inappropriate clinical decisions (medication selection, escalation, etc.) and to false-positive or false-negative inferences regarding treatment effectiveness. Some patients who “should” respond to certain anti-depressants may not because their presenting symptoms are MS disease symptoms rather than mental health symptoms. In seeking deeper knowledge of symptom origins, we propose methods which should improve our ability to distinguish depression changes in MS patients.

Structural equation modeling (SEM) enables researchers to examine diagnostic overlap for MS and depression and the complex relationships between patients’ physical and mental health states (Bollen, 1989; Kline, 2010). SEM uses a conceptual model, path diagram and system of linked regression-style equations to capture complex and dynamic relationships among a web of variables. In particular, the multiple indicator multiple cause (MIMIC) model, a measurement model with covariates (Asparouhov and Muthén, 2009; Brown, 2006; Kline, 2010; Woods et al., 2009), permits detection and adjustment for differential item functioning (DIF). DIF occurs when people from different groups (e.g., levels of MS-related fatigue) with the same latent trait (level of depression) have a different probability of giving a certain response on a questionnaire or test (e.g., items for sleep problems and fatigue of the PHQ-9).

Patient reported outcomes have an important place in psychiatry and neurology to measure the extent of disease or condition at the individual level, because they reflect the self-reported health state of the patient directly (Alemayehu et al., 2012). However, their effective integration in personalized medicine requires addressing certain conceptual and methodological challenges, since each individual patient may have a different view of how to fill out a test questionnaire, leading to both random and systematic error. Using the MIMIC model, we are able to estimate and adjust for this special type of measurement error, which is also known as DIF.

Our approach will generate depression screening measures which accurately reflect symptom overlap. The resulting tools will allow clinicians to appropriately tailor care management, while accounting for the complex roles of depression and MS symptoms.

Prior studies have assessed whether signs and symptoms of MS may overlap with those of depression with mixed results. Mohr et al. (1997) recommended to omit Beck Depression Inventory items for assessing work ability, fatigue and health concerns, while Aikens et al. (1999) encouraged full application of the scale for assessing depressive symptoms in patients with MS. Mohr et al. (2003) also found using the Beck Depression Inventory that treatment for depression is associated with reductions in the severity of fatigue symptoms, and that this relationship is due primarily to treatment related changes in mood. Crawford (2009) found that the inclusion of somatic items in the Beck Depression Inventory-II did not significantly inflate total scores and thus recommended that there is no reason to exclude those items when using the scale with individuals who have MS. Benedict et al. (2003) validated the short form Beck Depression Inventory-Fast Screen for MS patients, determining that the test is not confounded with neurological symptoms. Chang et al. (2003) performed a psychometric evaluation of the Chicago Multiscale Depression Inventory, supporting meaningful subtypes of depression in MS (mood, depressive cognition and vegetative symptoms), while determining that some items function differently in MS patients compared with depressed patients and controls.

However, in this study, we focus on an improved use of the PHQ-9, a self-reported depression screening tool to be used in connection with expert clinical judgment and/or further rating tools and not as an actual depression diagnosis (Blacker, 2009), in MS patients. We also describe a generalizable approach within a MS population for determining the factor structure of the measure and developing a series of models to evaluate the overlap of multiple symptoms of MS and depression simultaneously. Thus our approach will then allow us to form more practical adjusted depression screening scales based on factor scores and loadings which clinicians can use to better assess depressive symptoms free of the overlap. While our approach focuses on the PHQ-9 and MS symptoms, it is straightforward to generalize to other scales and conditions.

Methods

Study design and KP data base

Cleveland Clinic's Knowledge Program (2008–2013) links patient-reported PHQ-9 data to its EPIC electronic health record, yielding powerful opportunities to study and improve patient care and clinical research. The Mellen Center (2013) for Multiple Sclerosis manages more than 20000 visits and 1000 new patients every year for MS treatment. The Knowledge Program tracks illness severity and treatment efficacy over time across the Mellen Center population.

We use a retrospective cohort study design. Inclusion criteria for our sample includes patients making at least one visit to the Mellen Center with measurements of PHQ-9 score and a timed 25-foot walk available. Data are available for 3507 MS patients from 2008–2011 that meet our inclusion criteria.

Knowledge Program measures assessed

The PHQ-9 is used in both screening and monitoring of depression (Blacker, 2009). Patients specify frequency in the past 2 weeks (0 = not at all to 3 = every day) of nine symptoms, yielding a total score (range: 0–27). Scores of 5, 10, 15, and 20 are validated thresholds for mild, moderate, moderately severe and severe depression. Scores on this self-reported instrument are often used to guide treatment decisions (Kroenke et al., 2001). In particular, a PHQ-9 ≥ 10 has been previously established as a screening cutoff for depressive disorder (Ferrando et al., 2007; Kroenke et al., 2001). The PHQ-9 has been validated using multiple modes for administration, clinical populations, and diverse race/ethnicity groups (Pinto-Meza et al., 2005).

A prior study examined if PHQ-9 scores are biased in MS patients due to the MS symptoms of fatigue and poor concentration (Sjonnese et al., 2012). In the analyses, the authors compared the PHQ-9 items for fatigue and poor concentration in 173 MS patients to 3304 other subjects from the general population in the sample under study. They concluded that there was no evidence to exclude these items from a modified PHQ-9 score. Our study extends this prior work by (1) using a much larger sample of MS patients and only MS patients, eliminating the potential selection bias from a non-comparable control group and (2) conducting psychometric analysis of the measurement properties of depression self-rating and MS disability measures. The Knowledge Program collects eight MS performance scales (Marrie and Goldman, 2007; Schwartz et al., 1999) which are patient-reported disability measures. These include MS-related fatigue, cognitive, mobility and hand function domains with 6 ordinal responses. Reliability, criterion and construct validity have been established for these domain in previous studies of MS patients (Marrie and Goldman, 2007; Schwartz et al., 1999).

In addition to patient-reported measures, objective functional impairment measures of lower (timed 25-foot walk) and upper (9-hole peg test) extremity function are included in the Knowledge Program (Polman and Rudick, 2010). The timed 25-foot walk is a test of quantitative mobility and leg function performance, while the 9-hole peg test is a brief, standardized, quantitative test of arm and hand function. These two measures are combined to form a composite measure of functional impairment in our study (Fischer et al., 1999; Rudick et al., 1996; Whitaker et al., 1995).

Potential overlap of a depression screening scale and MS-related disability measures

PHQ-9 items for sleep problems, fatigue, poor concentration and psychomotor problems have the potential to overlap with symptoms described by MS disability scales (figure 1 panel D) using a unidimensional depression screening scale (we assess this assumption in section 3.2).

PHQ-9 items for sleep problems and fatigue (items 3 and 4) address fatigue. Attempting to separate fatigue as a depressive symptom from MS-related disability in the PHQ-9 is likely problematic for a patient due to the multiple dimensions, such as physical, mental and emotional, that may be involved in describing fatigue symptoms (Krupp, 2004). The PHQ-9 addresses concentration (item 7) and speed of movement and speech (item 8). Concentration

problems that are perceived as memory problems stemming from cognitive impairment in MS can potentially be impaired concentration due to comorbid depression (Siegert and Abernethy, 2005). Further, functional impairment could potentially be associated with concentration problems. Cognitive impairment, functional impairment and depression in MS can potentially lead to difficulties in movement and speech. Thus, potentially items for poor concentration and psychomotor symptoms on the PHQ-9 overlap with MS-related cognitive impairment and functional impairment.

Statistical Analyses

Descriptive Statistics and Computation—We compile descriptive statistics to summarize demographic information and scores on our measures of depression and MS symptoms. Due to sample size and since each ordinal scale (PHQ-9 items, performance scales measures) has at least 4 ordinal categories, the measures can be viewed as approximating an interval scale (Hansson et al., 2009; Huang et al., 2006; Johnson and Creech, 1983; Zumbo and Zimmerman, 1993). We use a maximum likelihood estimator with robust standard errors (MLR option in MPlus) in our analyses (Múthen and Múthen, 2013). To examine the potential for bias based on variable distributions we verified results using a mean and variance adjusted weighted least squares estimator (WLSMV option in MPlus) treating our outcomes (PHQ-9 items) as ordinal categorical measures.

We define $\alpha = 0.05$ for our level of significance in all statistical tests. All statistical tests are two-tailed. SAS Version 9.2 (2008) was used for data cleaning and for compiling descriptive statistics along with analyses with the PHQ-2 in the Appendix. Factor analyses and SEM analyses were carried out using Mplus Version 7 (Múthen and Múthen, 2013). R program (2008) was used to calculate descriptive statistics about the adjusted depression scales in the Appendix and performing the application in the Supplementary Table 1 using the imported final output MPlus dataset.

Exploratory Factor Analysis—Factor analytic techniques are useful for reducing the number of measures under study to a smaller number of factors and detecting a structure in the relationships between measures. Exploratory factor analysis (EFA) of the nine items of the PHQ-9 is used to determine how many latent factors represent aspects of *depression* (Hansson et al., 2009; Huang et al., 2006). For example, PHQ-9 has previously been hypothesized to have a two factor structure representing somatic and affective domains of depression (Kalpakjian et al., 2009). We use an oblique (geomin) rotation. The item factor loadings can be interpreted as the shared variance between an item and *depression* (Raykov and Marcoulides, 2011; Woods et al., 2009). In order to determine the number of latent factors in the PHQ-9, we examine the eigenvalues, which represent the variance accounted for by each underlying factor, in a scree plot. The number of eigenvalues ≥ 1 represent unique latent factors according to Kaiser's rule (Costello and Osborne, 2005). We also use parallel analysis (Horn, 1965), which computes eigenvalues from a random dataset via Monte Carlo simulation with the same numbers of observations and variables as used in the factor analytic model. When the eigenvalues from the random data are larger than the reported eigenvalues for the factor analysis, then the factors are mostly random noise. In MPlus, parallel analysis can only be performed on continuous items. In addition, since

MPlus performs EFA within the SEM framework (Múthen and Múthen, 2012), in order to help determine the number of factors we examined chi-square, comparative fit index (CFI), Tucker Lewis fit index (TLI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) statistical test values (Bentler, 1990; Browne and Cudeck, 1993; Hu and Bentler, 1998; Tucker and Lewis, 1973). A nonsignificant chi-square value, CFI and TLI = 0.95, RMSEA value of = 0.05 and SRMR value = 0.08 represent a good fitting model. The RMSEA statistic is especially useful in that it provides a confidence interval for our point estimate.

MIMIC Modeling Approach—We next assess (and eventually correct for) the overlap in MS and depression symptoms. In this analysis we use as many latent factors as determined by EFA in a series of models that will examine the overlapping symptoms of the two conditions. Namely, confirmatory factor analysis (CFA) is used to form a measurement model representing the latent structure of depression. Then, covariates, in the form of our MS disability measures, are added to the measurement model forming a MIMIC model. We do not make any assumption of causal ordering between each MS symptom we study and depressive symptoms, and therefore specify the relationship as a correlation in our models (see Figure 1 panel D). DIF paths, which are specified in accordance to the hypothesized overlap between these covariates and the items of the PHQ-9 (as described in section 2.3 and shown in Figure 1 panel D), are then evaluated in a series of models to analyze the overlapping symptoms.

We aim to establish a best fitting model in our MIMIC analyses in order to confirm or revise our hypotheses in section 2.3 for the overlap of depression symptoms with other symptoms for MS patients. First, we compare the full MIMIC model to a full constrained model (see Figure 1 legend and panels D and E) to establish if there are overlapping symptoms. Constraining a causal path during analyses involves setting the given path to zero (i.e. removing the path). Then, we examine how individual paths A through F in Figure 1 panel D influences model fit in isolation, in order to constrain to zero any particular paths that are not improving our model fit, and thus are not symptoms of overlap. We make comparisons between models using the chi-square test of difference and by examining the modification indices, where a modification index > 5 represents a clinically significant causal path (Kline, 2010). Accounting for these overlapping symptoms alters the factor loadings and scores in *depression* in a similar manner to how multiple regression can alter the estimated effect of a treatment when controlling for confounders. We build our adjusted depression screening scales using the factor loadings and scores in *depression*, a latent construct of depressive symptoms adjusted for the overlapping symptoms of both conditions under study.

Subgroup Analyses—We test for potential subgroup differences one at a time, by age, sex (male or female), race, MS type (relapsing or progressive) and baseline time since diagnosis, using a Bonferroni correction to control the family-wise error rate, by extending our best fitting model from 2.4.3 in the multigroup confirmatory factor analytic framework (Brown, 2006). In our subgroup analyses we form dichotomous variables for age (threshold is mean $\approx$ median = 46), race (white or other) and baseline time since diagnosis (less or

greater than 10). Figure 2 shows how the model with $E=0$ from 2.4.3 can be extended for subgroup analyses based on sex.

We compare the model fit statistics and indices, for each subgrouping between two models (i.e. MALE model and FEMALE model in Figure 2): (1) freely estimating all parameters across groups (i.e. In Figure 2 parameter estimates from the MALE model will not necessarily equal estimates from the FEMALE model) and (2) constraining the estimates of all factor loadings and item intercepts in the two measurement models to be equal across groups. Therefore, if there is a subgroup difference, the model fit would be significantly worse under the constraint of group equality.

Results

Demographic characteristics of the MS population in the KP data base

The sample mirrors the United States' MS population in that MS is typically diagnosed in patients in their early 30's, Caucasians are of highest risk and females are twice as likely as males to develop MS (Goldman Consensus Panel, 2005; Health Encyclopedia: Multiple Sclerosis Risk Factors, 2005–2013). In our sample, 73% were female, 83% were white, and the average age was 46 ($SD = 12$). These patients had their first MS diagnosis an average of 10 ($SD = 9$) years ago with 81% relapsing and 16% progressive with the remaining patients falling into other categories, or under evaluation for a potential MS diagnosis. We leave these patients who are not relapsing or progressive ($N = 70$) out of our subgrouping analyses based on MS type, due to our uncertainty about their symptoms.

Nearly 30% ($n=1005$) of patients had PHQ-9 ≥ 10 at their entry to the KP. The distribution of PHQ-9 scores represents a wide range of depression severity levels. We summarize the characteristics of the Mellen Center MS population by the PHQ-9 binary threshold of 10 (screening cutoff for depressive disorder) in Table 1.

How many latent factors represent depression in the MS population?

Examining eigenvalues and scree plot showed one latent construct (highest eigenvalue = 4.82, no other eigenvalues ≥ 1) for the PHQ-9 within this MS population with high standardized factor loadings on all nine items (each loading ≥ 0.5). Further, performing parallel analyses by generating 50 random datasets also supports one latent construct, as shown in Figure 1 panel A.

However, the model fit statistics and indices improved for the two factor model over the one factor model (Table 2) providing a mixed message of whether the one or two factor model was a better fit.

In examining the two potential factors in a CFA model, items for anhedonia, feel depressed, feelings of failure and self-harm load on one factor while items for sleep problems, fatigue, appetite change, poor concentration and psychomotor symptoms load on a second factor (Figure 1 panel B). The correlation between the two factors is very large ≈ 0.86 , which implies poor discriminant validity and suggests that a more parsimonious single factor solution is preferred (Brown, 2006; Marsh et al. 2004).

We further examine the expected parameter change values for each modification index for the covariance between error terms for each pair of PHQ-9 items. Expected parameter change represents how much the given parameter would be expected to change if it were freely estimated (Brown, 2006; Muthén and Muthén, 2012). In this case, we flag any expected parameter change in absolute value ≥ 0.10 (Cohen, 1992). Expected parameter change between items for feel depressed and feelings of failure (0.101), sleep problems and fatigue (0.220) and poor concentration and psychomotor symptoms (0.115) are all ≥ 0.10 .

The strong pairwise relationships between related fatigue items (sleep problems and fatigue) and cognitive and functional impairment items (poor concentration and psychomotor symptoms) are in line with our theoretical model for overlapping symptoms. An association between feeling depressed and feelings of failure, both affective symptoms of depression, also has a theoretical basis. The model fit statistics and indices are very similar for the one factor CFA model specifying these three pairwise error term covariances as in Figure 1 panel C as the two factor CFA model in Figure 1 panel B (see Table 2). Introducing correlated residuals is more parsimonious than our two factor solution (i.e. fewer parameters than the two factor model). Although equivalent multi-factor models can be derived for this solution, using correlated residuals maintains the unidimensional interpretation of one depression factor, as suggested by the eigenvalues and PA.

MIMIC model to examine and correct for the overlap of MS and depression symptoms

We applied confirmatory factor analysis to isolate one latent variable *depression* as shown in Figure 1 panel C. Then, we add covariates as well as the differential item functioning paths as shown in Figure 1 panel D. We begin by testing for any overlap of depression and MS symptoms by comparing this model freely estimating paths A through F to a model constraining all paths A through F to zero as shown in Figure 1 panel E (comparing the full MIMIC and full constrained models as described in the Figure 1 legend).

In Table 3, the fit criteria showed that the MIMIC model is a better fitting model than the constrained model, and overall a relatively good fit for the data structure using the criteria as described in 2.4.2. These results were verified by the χ^2 Difference Test comparing the MIMIC to the constrained model ($p < 0.001$).

Estimates for the standardized factor loadings in *depression* are lower in the MIMIC model compared to the constrained model for the items involving overlapping symptoms, sleep problems, fatigue, poor concentration and psychomotor symptoms (Table 4). The decrease in these estimates signifies that after correcting for the overlapping symptoms, these four items contribute less in determining level of *depression*.

Out of the individual paths of overlap from the MS disability covariates to the four items, only path E is not statistically significant. The magnitude of the estimates is large for the overlap of the item for fatigue with MS-related fatigue ($B=0.51$), medium for the item for sleep problems with MS-related fatigue and the items for poor concentration and psychomotor symptoms with MS-related cognitive impairment ($A=0.21$, $C=0.38$ and $D=0.23$) and small for the item for psychomotor symptoms with the 9-hole peg test aspect of MS-related functional impairment (F for peg test= 0.11).

As a result of the statistical evidence in this section, we find overlap of *depression* with all MS symptoms that may be inflating or deflating a MS patient's PHQ-9. Simultaneously, the MIMIC model is estimating the overlap and then correcting the latent factor of *depression*.

Revising the full MIMIC model to determine a best fitting model

We now examine each individual overlapping symptom in isolation, models A, B, C, D, E and F=0, for improvement over our full MIMIC model. From Table 4, evaluating model fit indices and statistics, as well as the chi-square tests of difference and modification indices, we observe improvement over the full MIMIC model only for E=0 (Figure 1 panel F). Bollen and Long (1993), as applied by Sudhahar et al. (2006), state that the appropriate procedure is to constrain this path E to 0 (i.e. remove it) since it is statistically not significant and changes in model fit from the full MIMIC model are not observed.

We now test for improvement for E=0 for a best fitting model through accounting for the subgroup differences as in Figure 2. First, we apply an omnibus test of invariance using a chi-square test of difference to compare, for each subgrouping such as in Figure 2, the model under the constraint of group equality to the model under no such constraint (Brown, 2006). In this case, we do not reject the null hypothesis of group equality for any of the subgroupings (i.e. no statistically significant p-values). Further, we observe similar model test statistics and indices between the two models for each subgrouping.

We also replace the objective performance functional impairment measures with the patient reported counterparts (performance scales mobility and hand function domains). The magnitude of the effect sizes are similar, and we come to similar overall conclusions of significant overlap for MS-related functional impairment with the PHQ-9 item for psychomotor symptoms, but not with poor concentration (E=0.02 for hand function, -0.07 for mobility and F= 0.08 for hand function and 0.12 for mobility).

An adjusted depression screening score can promote better clinical assessment of depressive symptoms in MS patients

Using the results of the best fitting model E=0 we briefly propose and compare three algorithms to calculate an adjusted depression score for MS, accounting for the overlap of depression symptoms with MS-related fatigue, functional impairment and cognitive impairment in patients (see Appendix for more details). Therefore, these scales are self-reported depression screening tools for MS patients that better represent the underlying dimension for depressive symptoms in the PHQ-9, for a more pure estimate of an MS patient's mood. In our application and the discussion section, we discuss the practical use of these for clinical practice (with more technical details provided in the Appendix). It is straightforward to extend our adjusted depression screening score algorithms described below to adjust for additional overlapping symptoms if theoretically and statistically warranted.

Factor scores—A factor score is an individualized score on the latent construct *depression*. In other words, factor scores represent continuous adjusted depression screening scores for MS patients, accounting for the overlap of depression symptoms with MS-related

fatigue and functional and cognitive impairment. Factor scores can be output through MPlus into another data file (such as in a text file) using a few lines of code in the OUTPUT section of a MPlus program. For example, saving factor scores into a text file called FACTORSCORES can be done using the coding:

```
OUTPUT:
SAVEDATA: FILE IS C:\FACTORSCORES.txt;
SAVE IS fscores;
```

Item weights—Factor loadings from the latent construct *depression* can be transformed into item weights to build an adjusted depression screening scale. The factor loadings present a regression-type of beta weight for each item through the shared variance between an item and the latent construct, which is a naïve approach to calculating factor scores.

Downweighting overlapping symptoms—A final scale maintains the same interpretation as the PHQ-9, but simply down-weights the impact of the items of overlap by dividing the factor loadings of *depression* for E=0 over the factor loadings for the CFA model from Table 4. The difference between the PHQ-9 and this adjusted scale is an estimate of how much of the PHQ-9 can be accounted for by the overlapping symptoms.

Application of the revised scoring approaches on two real patients—The potential of these three scoring approaches can be demonstrated through an application with two patients purposively selected from the Knowledge Program Database (Supplemental Table 1). Patient A had a PHQ-9 of 10, and thus screened positive for depressive disorder, but highly endorsed items for sleep problems, fatigue, poor concentration and psychomotor symptoms. Patient B had a PHQ-9 of 6, mostly the result of endorsing other items. The algorithms from 3.5.1–3.5.3 can be calculated from information that is already in the Knowledge Program Database. Therefore, in using these algorithms, a clinician or patient does not have to fill out any additional information during a typical visit. Given the generalizability of our methods, and ease of replacing any of the measures with other measures, in a similar straightforward manner the algorithms can be programmed into other data bases. Also, the algorithms from 3.5.1 and 3.5.2 (factor scores and item weights) can be transformed and matched with current PHQ-9 scores (details provided in Appendix) for immediate use with the same empirical distribution as the current PHQ-9. As a result of MS symptoms potentially inflating the PHQ-9 score, using algorithm 3.5.1 transformed to the standard scoring scale, patient A has a score now significantly below the threshold of 10 and actually scored lower than patient B. The transformed factor loadings approach (3.5.2), a naïve approach to a factor score, leads to a more trivial change. Further, the downweighting overlapping symptoms (3.5.3) algorithm could be used to show that the amount of overlap is of a large magnitude for patient A ($10 - 7.59 = 2.41$), while of a small magnitude for patient B ($6 - 5.30 = 0.70$).

Discussion

The results of our SEM-based analyses show that there is a significant overlap of depression symptoms with other symptoms for MS patients, namely MS-related fatigue and cognitive and functional impairment. The effect size of the overlap was most prominent for fatigue, in line with our a priori hypothesis in 2.3. Based on these analyses, adjusted depression screening scales were formed that can help clinicians prevent over or under prescribing antidepressants and fatigue and MS medication, and providing better tailored care.

By simply adjusting the PHQ-9 through approaches that correct for the overlap of depression symptoms with MS-related fatigue and cognitive and functional impairment, we present improved depression screening scores. No previous study directly evaluated and corrected for the overlap of the PHQ-9 with MS disability scales in a population of MS patients.

In a clinical setting, we suggest adding two more entries in the Knowledge Program data base, one for the PHQ-2 and one for the adjusted scoring algorithm based on factor scores (as described in 3.5.1). The simplest and most efficient approach for clinicians is almost certainly to first use the PHQ-2 for pre-screening for depression in MS patients as shown in 3.5.5 (see Appendix for details), before using the modified algorithm as a screening tool. The two questions of the PHQ-2 showed no evidence of differential item functioning in our analyses, thus a positive PHQ-2 screen combined with a positive adjusted algorithm screen for the transformed factor scores will better identify patients in need of depression treatment as opposed to patients with MS-related symptoms not indicative of depression. This approach will give clinicians information for improved use of the PHQ-9, regarding better diagnosis, treatment decisions and inferences about effectiveness of treatment free of the overlap between depressive symptoms with other symptoms for MS patients. For an example, using these methods, patient A in our application has a PHQ-2 of zero and was shown to have a lower adjusted scale score based on factor scores than PHQ-9, while patient B screened positive for depression via the PHQ-2 (=3) and then had a significantly higher adjusted scale score based on factor scores than original PHQ-9 scoring. Further, clinicians can make use of the scale from algorithm 3.5.3 to provide an estimate of how much of the PHQ-9 score in a MS patient may be accounted for by overlap of depressive symptoms with other symptoms for MS patients. Even for a clinician unwilling to make use of these new scales before further validation, at the very least, we have established that the use of the PHQ-9 without the PHQ-2 for MS patients is ill-advised.

Since the only scale or diagnoses related to depression within this sample is the PHQ-9, there is no way to validate the proposed adjustments within our retrospective sample. However, even if we did have available other measures related to depression that are already validated for use in MS, validation of measures in general would be complex and is the subject of future research. To the best of our knowledge, no prior study has identified any scale or diagnoses related to depression that corrects for overlap with other symptoms for MS patients. The PHQ-9 had previously been validated for use in MS patients (Sjonnese et al., 2012). Along this line, any measure previously considered a valid measure for depression screening in MS should be evaluated in the future for differential item

functioning before considering it valid in MS patients. After finding such a scale or diagnoses, then we could use it to validate our measures. Alternatively, a new study involving clinical diagnostic interview of MS patients by psychiatric experts familiar with MS symptom presentation would provide another benchmark for diagnostic comparison of the validity of self-report scoring approaches.

The assumptions of transforming the algorithms of 3.5.1 and 3.5.2 and matching with the empirical distribution of PHQ-9 scores is that within this MS population, we have correctly identified the empirical distribution (mean, standard deviation, range and cutoffs) of the PHQ-9, but have misclassified the particular scores of individual patients. This further assumes that depression is different in MS patients with some symptoms worse (such as anhedonia and feel depressed) while some symptoms overlap with other symptoms for MS patients and thus inflate PHQ-9 scores. Previous studies have discussed that the etiology of depression is different in MS patients compared to the general population (Chang, 2003). However, if looking to use these scales without these assumptions, in developing a unique empirical distribution with different cutoffs of the adjusted algorithms from the standard scoring of the PHQ-9 our results suggest that in patients with co-morbid neurologic illness like MS, rigorous psychometric evaluation and interpretation is necessary for validating depression diagnosis cutoffs for clinical use. All the proposed score adjustment algorithms could also be useful in further analyses in studying depression within the sample without any further evaluation of diagnosis cutoffs, accounting for the overlap in depression symptoms with other symptoms for MS patients (i.e. as an outcome in a regression model). A limitation to using the MIMIC model is that they test for uniform, but not nonuniform DIF (Woods, et al., 2009). Uniform DIF is constant across *depression*, while non-uniform DIF varies across *depression*. We addressed this limitation through multi-group analyses for subgroup measurement differences.

The study has limited external validity outside the Mellen Center population. In future work, our models will require further validation using other populations and perhaps data from alternate measures and scales (for instance, the Hamilton Rating Scale for Depression (Hamilton, 1960)) and with varied symptom patterns. We did not have in our database the third component of the objective performance functional composite measure (the other two are the timed 25-foot walk and the 9-hole peg test), the Paced Auditory Serial Addition Test, representing cognitive function [19,20,21,22]. Using such a measure could help internally validate cognitive impairment findings. We were, however, able to internally validate our functional impairment findings using the patient-reported counterparts (mobility and hand function domains of the performance scales) for our objective performance measures (timed 25-foot walk, 9-hole peg test).

Other measures of overlap are important to examine in MS patients, such as pain, depending on the depression screening measure or data base, requiring extension of our models for more overlapping symptoms. We tested items of the PHQ-9 that theoretically had an overt possibility of overlap with MS-related disability within our data base. In addition, although we describe symptoms in this paper, the effects of certain treatments for both MS and depression symptoms in MS patients potentially overlap with the PHQ-9 item for appetite.

Analogous SEM-based methods should transfer cleanly to other subgroups and measures. Applying statistical models to the complete KP database could address important clinical questions about multiple neurological conditions, such as spine disease, headaches, sleep disorders, epilepsy and stroke. Analogous SEM approaches to those outlined here can be used to address other neurological conditions and other mental health scales. For example, a similar spine disease adjusted PHQ-9 can be constructed to address overlapping pain.

In future work, we plan to describe and model the causal path for how type of MS and baseline time since diagnosis lead to depression symptoms, accounting for the roles of fatigue and functional and cognitive disability, with an eye toward targeting future interventions. We will also explore the predictive value of the adjusted depression screening scales by linking response to particular anti-depressants, and subgroups of MS patients identified via growth mixture modeling (GMM) (Múthen and Múthen, 2000). Further, we are preparing a tutorial on SEM, factor analysis and MIMIC modeling for analysis of overlapping symptoms in co-occurring conditions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support for this study was provided by a grant from NIH/NCRR CTSA KL2TR000440 and by a grant from Novartis. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. We appreciate the contributions from Drs. Randall Cebul, Thomas Love, Irene Katzan, Neal Dawson, Center for Health Care Research and Policy, Drs. Richard Rudick, and Francois Bethoux, Mellen Center, and Dr. Martha Sajatovic, Departments of Psychiatry and Neurology at Case Western Reserve University School of Medicine.

References

- Aikens JE, Reinecke MA, Pliskin NH, Fischer JS, Wiebe JS, McCracken LM, Taylor JL. Assessing depressive symptoms in multiple sclerosis: is it necessary to omit items from the original Beck Depression Inventory? *J Behav Med.* 1999; 22(2):127–142. [PubMed: 10374139]
- Alemayehu D, Cappelleri JC, Murphy MF. Conceptual and Analytical Considerations toward the Use of Patient-Reported Outcomes in Personalized Medicine. *Am Health Drug Benefits.* 2012; 5(5): 310–317. [PubMed: 24991329]
- Asparouhov T, Muthén B. Exploratory structural equation modeling. *Structural Equation Modeling.* 2009; 16:397–438.
- Benedict RH, Fishman I, McClellan MM, Bakshi R, Weinstock-Guttman B. Validity of the Beck Depression Inventory-Fast Screen in multiple sclerosis. *Mult Scler.* 2003; 9(4):393–396. [PubMed: 12926845]
- Bentler PM. Comparative fit indexes in structural models. *Psychol Bull.* 1990; 107:238–246. [PubMed: 2320703]
- Blacker, D. Psychiatric Rating Scales. In: Sadock, BJ.; Sadock, VA.; Ruiz, P., editors. *Kaplan & Sadock's Comprehensive Textbook of Psychiatry.* 9. Philadelphia, PA: Lippincott Williams & Wilkins; 2009.
- Bollen, KA. *Structural Equations with Latent Variables.* Hoboken, NJ: Wiley; 1989.
- Bollen, KA.; Long, JS. *Testing Structural Equation Models.* Newbury Park, CA: Sage Publications; 1993.
- Brown, T. *Confirmatory Factor Analysis for Applied Research (Methodology in the Social Science).* New York, NY: The Guilford Press; 2006.

- Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. *Testing Structural Equation Models*. Newbury Park, CA: Sage Publications; 1993.
- Casella, G.; Berger, RL. *Statistical Inference*. 2. Pacific Grove, CA: Duxbury Press; 2002.
- Chang CH, Nyenhuis DL, Cella D, Luchetta T, Dineen K, Reder AT. Psychometric evaluation of the Chicago Multiscale Depression Inventory in multiple sclerosis patients. *Mult Scler*. 2003; 9(2): 160–170. [PubMed: 12708812]
- Cohen J. A power primer. *Psychological Bulletin*. 1992; 112(1):155–159. [PubMed: 19565683]
- Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*. 2005; 10:1–9.
- Crawford PW. Assessment of Depression in Multiple Sclerosis Validity of Including Somatic Items on the Beck Depression Inventory–II. *Int J MS Care*. 2009; 11:167–173.
- Ferrando SJ, Samton J, Mor N, Nicora S, Findler M, Apatoff B. Patient Health Questionnaire-9 to Screen for Depression in Outpatients With Multiple Sclerosis. *Int J MS Care*. 2007; 9:99–103.
- Fischer JS, Rudick RA, Cutter GR, Reingold SC. The multiple sclerosis functional composite measure (MSFC): An integrated approach to MS clinical outcome assessment. *Multiple Sclerosis*. 1999; 5:244–250. [PubMed: 10467383]
- Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med*. 2007; 11:596–602.
- Goldman Consensus Panel. The Goldman Consensus statement on depression in multiple sclerosis. *Mult Scler*. 2005; 11:328–337. [PubMed: 15957516]
- Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*. 1960; 23:56–62.
- Hansson M, Chotai J, Nordstöm A, Bodlund O. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *Br J Gen Pract*. 2009; 59(566):e283–e288. [PubMed: 19761655]
- Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
- Hu L, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*. 1998; 3:424–453.
- Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to Measure Depression among Racially and Ethnically Diverse Primary Care Patients. *J Gen Intern Med*. 2006; 21(6):547–552. [PubMed: 16808734]
- Johnson DR, Creech JC. Ordinal measures in multiple indicator models: a simulation study of categorization error. *Am Sociol Rev*. 1983; 48:398–407.
- Kalpakjian CZ, Toussaint LL, Albright KJ, Bombardier CH, Krause JK, Tate DG. Patient Health Questionnaire-9 in Spinal Cord Injury: An Examination of Factor Structure as Related to Gender. *J Spinal Cord Med*. 2009; 32(2):147–156. [PubMed: 19569462]
- Kline, RB. *Principles and Practice of Structural Equation Modeling*. 3. New York, NY: Guilford; 2010.
- Knowledge Program developed at Cleveland Clinic's Neurological Institute. 2008–2013. Retrieved from http://my.clevelandclinic.org/neurological_institute/about/default.aspx
- Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical Care*. 2003; 41:1284–1292. [PubMed: 14583691]
- Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001; 16:606–613. [PubMed: 11556941]
- Krupp, LB. *Fatigue in Multiple Sclerosis*. New York, NY: Demos Medical Publishing; 2004.
- Marrie RA, Goldman M. Validity of performance scales for disability assessment in multiple sclerosis. *Multiple Sclerosis*. 2007; 13:1176–1182. [PubMed: 17623733]
- Marsh HW, Hau KT, Wen Z. In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling A Multidisciplinary Journal*. 2004; 11(3):320–341.
- Mellen Center for Multiple Sclerosis Treatment and Research, Cleveland Clinic, Neurological Institute. 2013. Retrieved from http://my.clevelandclinic.org/neurological_institute/mellen-center-multiple-sclerosis/default.aspx

- Mohr DC, Goodkin DE, Likosky W, Beutler L, Gatto N, Langan MK. Identification of Beck Depression Inventory items related to multiple sclerosis. *J Behav Med.* 1997; 20(4):407–414. [PubMed: 9298438]
- Mohr DC, Hart SL, Goldberg A. Effects of treatment for depression on fatigue in multiple sclerosis. *Psychosom Med.* 2003; 65(4):542–547. [PubMed: 12883103]
- Multiple Sclerosis Association of America. 2014. Retrieved from <http://www.mysaa.org/about-ms/faq/>
- Muthén, LK.; Muthén, BO. *Mplus User's Guide*. 7. Los Angeles, CA: Muthén & Muthén; 2012.
- Muthén B, Muthén L. Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research.* 2000; 24:882–891.
- Pinto-Meza A, Serrano-Blanco A, et al. Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone? *J Gen Intern Med.* 2005; 20:738–742. [PubMed: 16050884]
- Polman CH, Rudick RA. The Multiple Sclerosis Functional Composite: A clinically meaningful measure of disability. *Neurology.* 2010; 74:S8–S15. [PubMed: 20421572]
- R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
- Raykov, T.; Marcoulides, G. *Introduction to Psychometric Theory*. New York, NY: Taylor and Francis Group, LLC; 2011.
- Rudick RA, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J, Lublin F, Miller A, Petkau J, Rao S, et al. Clinical outcomes assessment in multiple sclerosis. *Annals of Neurology.* 1996; 40:469–479. [PubMed: 8797541]
- SAS Institute Inc. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc; 2008.
- Schwartz CE, Vollmer T, et al. Reliability and validity of two self-report measures of impairment and disability for MS. *Neurology.* 1999; 52:63–70. [PubMed: 9921850]
- Siebert RJ, Abernethy DA. Depression in multiple sclerosis: a review. *J Neurol Neurosurg Psychiatry.* 2005; 76:469–475. [PubMed: 15774430]
- Sjonnese K, Berzins S, Fiest KM, Bulloch AG, Metz LM, Thombs BD, Patten SB. Evaluation of the 9-Item Patient Health Questionnaire (PHQ-9) as an Assessment Instrument for Symptoms of Depression in Patients with Multiple Sclerosis. *Postgrad Med.* 2012; 124(5):69–77. [PubMed: 23095427]
- Sudhahar JC, Israel D, Selvam M. Banking Service Loyalty Determination Through SEM Technique. *Journal of Applied Sciences.* 2006; 6:1472–1480.
- Tucker LR, Lewis C. The reliability coefficient for maximum likelihood factor analysis. *Psychometrika.* 1973; 38:1–10.
- Wallin MT, Wilken JA, Turner AP, Williams RM, Kane R. Depression and multiple sclerosis: Review of a lethal combination. *JRRD.* 2006; 43:45–62.
- Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *J Psychopathol Behav Assess.* 2009; 31(4):320–330. [PubMed: 20442793]
- Whitaker JN, McFarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple Sclerosis.* 1995; 1:37–47. [PubMed: 9345468]
- Zumbo BD, Zimmerman DW. Is the selection of statistical methods governed by level of measurement? *Can Psychol.* 1993; 34:390–400.

Appendix: More Details about Adjusted Depression Screening Scale Algorithms

Factor scores

Factor scores include adjustment for the effect of both CFA as in Figure 1 Panel C, since the nine PHQ-9 items load unequally on *depression*, and the adjustment of overlapping

symptoms via covariates and differential item functioning paths from our E=0 model in Figure 1 Panel F. Using MPlus, we calculate factor scores within our sample using the maximum a posteriori method (Múthen and Múthen, 2012), which for continuous outcomes is the widely used regression method: mean=0, variance = 0.56, median = -0.21, range=(-0.84, 2.47), 1st quartile = -0.62, and 3rd quartile = 0.40. These factor scores can then be linearly transformed to fit any potentially useful range (i.e. 0–27, 0–100, etc.)

In our application, and for immediate clinical use, a probability integral transformation (Casella and Berger, 2002) is used to transform these factors scores into the PHQ-9 scores within this population to maintain the interpretation and thresholds of the PHQ-9. This is a one-to-one transformation, so the distribution of PHQ-9 scores in this population will remain the same, just that certain individuals will be matched to different PHQ-9 scores. This transformation is built on continuous cumulative distribution functions and thus there are no special steps required for tied adjusted scores between two or more individuals (unlikely for factor scores). To perform the transformation:

1. Use a kernel density estimator to estimate the density of the PHQ-9 scores in this population.
2. Obtain the cumulative distribution function by numerical integration.
3. Numerically invert the cumulative distribution function for the PHQ-9.
4. Repeat steps 1) and 2) for the factor scores.
5. Transform the factor scores of each individual into a PHQ-9 score for this population by transforming the cumulative distribution function for the factor scores using the inverse of the cumulative distribution function for the PHQ-9.

The resulting factor scores will be distributed the same as the PHQ-9 within this population. For ease of use in a clinical setting, the recommendation is to round these transformed scores off to the nearest whole number. Note, within EHR-based database, every time a new or series of new PHQ-9 scores is entered, we can compute the factor scores. Then, either all subjects can be updated or we can keep prior records the same, and just use the functions to transform the new scores only. Using this approach, only 44% of subjects maintain the same transformed adjusted score as the original PHQ-9 score, with 16% of subjects scoring a transformed adjusted score of two points or more (with three subjects up to 6 points) different than the original PHQ-9 score. Further, 74 subjects had a PHQ-9 score ≥ 10 and now have a transformed adjusted score < 10 , while 65 subjects had a PHQ-9 score < 10 and now have a transformed adjusted score ≥ 10 .

Item weights

Compared to this algorithm in 3.5.2 based on factor loadings, the factor scores algorithm in 3.5.1 results in a more mathematically rigorous individualized score, taking into account item loadings, intercepts, correlation among observed variables, and a predictive equation. However, in the factor scores algorithm, we do not quantify how much each item contributes.

We can directly build a scale using the factor loadings as weights:

$$\text{MS- adjusted PHQ-9 score} = \sum_{i=1}^9 \text{Item } i \text{ Weight} \times \text{Item } i \text{ PHQ-9 Score} \quad (\text{A1})$$

In this case we use the factor loadings in Table 5 for E=0 as item weights and using (A1), in our sample, the mean=3.95 and standard deviation=3.71. Similarly, we can use the probability integral transformation for this algorithm as described for the factor scores algorithm in 3.5.1. In this case, there will be ties (same scores on MS-adjusted PHQ-9), but as explained above this will not be an issue with this transformation. The results here are a lot more conservative in change from the standard scoring, in that 64% of subjects maintain the same score and only 2% have greater than or equal to a two points or more difference from the original PHQ-9. As a result, this score may not be particularly useful in the transformed version, since it is a naïve approach of algorithm 3.5.1, and may be more useful once more rigorous psychometric evaluation of this score is performed.

Downweighting overlapping symptoms

We only downweight the influence of items of the PHQ-9 in which depressive symptoms overlap with other symptoms for MS patients:

$$\text{Item } i \text{ Weight} = \frac{E_i}{CFA_i} \quad (\text{A2})$$

where E_i is the item i standardized factor loading for E=0 and CFA_i is the item i standardized factor loading for the one factor CFA model (see Table 5).

We multiply a patient's score on items for sleep problems by 0.79, fatigue by 0.52, poor concentration by 0.70 and psychomotor symptoms by 0.74. PHQ-9 items for appetite change, feelings of failure and self-harm maintain the same score. Items for anhedonia and feel depressed, have an almost negligible residual DIF effect and we multiply a patient's score on these items by 1.01. We observe for our adjusted PHQ-9, a population mean=5.66, standard deviation = 5.30, and median = 4.07 with a range=(0, 23.32) and 1st quartile = 1.31, and 3rd quartile = 8.53. As noted, an application of 3.5.3 is to subtract this score from the PHQ-9, as an estimate of the amount that items on the PHQ-9 overlap with other symptoms for MS patients.

Using the PHQ-2

The PHQ-2 comprises the first two items of the PHQ-9 and has been used as a depression screening tool or as a pre-screener for the PHQ-9 (Kroenke, 2003). Thus, using this shortened scale will bypass evaluating items for sleep problems, fatigue, poor concentration and psychomotor symptoms. A PHQ-2 score of three has been used as a threshold for depression for screening purposes (Kroenke, 2003). The PHQ-2 is highly correlated with the PHQ-9 in this MS study population (pearson $\rho \approx 0.87$). Further, a receiver operating characteristic (ROC) analysis of PHQ-9 10 vs. PHQ-2 3 in this study population, showed high specificity=95.7 and a large positive predictive value (PPV) = 91.2, and area under the

curve (AUC) = 0.852, though at the expense of the test sensitivity = 63.8. In general, the PHQ-2 has shown wide variability in sensitivity in previous validation studies and more research is needed to see if its diagnostic properties approach those of the PHQ-9 (Gilbody et al., 2007).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Path diagrams for the investigated models

Covariates (fatigue, cognitive impairment and functional impairment) are all MS-related measures. In our study, functional impairment is defined by two correlated objective performance measures (timed 25-foot walk and 9-hole peg test). However, we did not consider models which adjust for only one of the two functional impairment measure to be feasible.

Models	Description
Full Constrained	Estimates association between MS-related fatigue, functional impairment and cognitive impairment and <i>depression</i> while constraining A, B, C, D, E and F = 0.
Full MIMIC	Estimates associations between MS-related fatigue, functional impairment and cognitive impairment and <i>depression</i> while also determining and correcting for the overlap of MS- related fatigue (A and B), cognitive impairment (C and D) and functional impairment (E and F) and PHQ-9 items for sleep problems, fatigue, poor concentration and psychomotor symptoms via DIF paths.
A=0,B=0,C=0,D=0,E=0,F=0	Separate models estimating the association between MS-related fatigue, functional impairment and cognitive impairment and <i>depression</i> while constraining the one specified path (i.e. D=0 constrains D=0) to 0 while freely estimating the five other DIF paths.

Abbreviations: EFA = Exploratory Factor Analysis; CFA = Confirmatory Factor Analysis; MIMIC = Multiple Indicator Multiple Cause

Figure 2.

Multigroup CFA framework for MIMIC model with $E=0$ for Sex

Multigroup modeling allows males and females to have different coefficients for the dashed pathways in the diagram above.

Table 1

Characteristics of the Mellen Center MS population

	PHQ-9 < 10	PHQ-9 10	
	n = 2502	n = 1005	p
PHQ-9	3.64 ± 2.75	15.26 ± 4.40	<0.001
MSPS fatigue	1.62 ± 1.25	3.35 ± 1.12	<0.001
MSPS cognitive	0.86 ± 0.96	2.23 ± 1.30	<0.001
MSPS mobility	1.37 ± 1.58	2.39 ± 1.48	<0.001
MSPS hand function	0.77 ± 0.94	1.79 ± 1.27	<0.001
25-foot timed walk	7.85 ± 10.56	8.83 ± 7.61	0.002
9-hole peg test	23.68 ± 10.66	26.82 ± 12.48	<0.001
age	46.12 ± 11.88	44.47 ± 11.20	<0.001
baseline time since diagnosis	11.80 ± 10.00	10.89 ± 9.37	0.016
gender, n (%)			0.879
female	1836 (74)	740 (74)	
male	666 (27)	265 (26)	
race, n (%)			0.070
caucasian	2112 (85)	821 (82)	
african-american	225 (9)	114 (11)	
other	144 (6)	65 (7)	
MS type, n (%)			0.067
relapsing	2045 (84)	787 (82)	
progressive	383 (16)	177 (18)	

Notes for Table 1: mean ± standard deviation for continuous measures and number of subjects in each category for discrete measures with p-values reported from t-tests and chi-square tests where appropriate.

Table 2

Model fit statistics and indices from the exploratory factor analysis and confirmatory factor analysis models

	1-factor EFA *	2-factor EFA	1-factor CFA with item correlations **	2-factor CFA
X²	754.356 (27)	266.470 (19)	386.57 (24)	401.83 (26)
CFI	0.922	0.973	0.961	0.960
TLI	0.896	0.949	0.941	0.944
RMSEA (95% CI)	0.088 (0.082, 0.093)	0.061 (0.055, 0.068)	0.066 (0.060, 0.071)	0.064 (0.059, 0.070)
SRMR	0.042	0.021	0.032	0.031

p < 0.001 for all X² tests of model fit.

* Same model fit values for the 1-factor CFA model *without* item correlations, with the trivial exception that the **X²** statistic = 754.344, as the 1-factor EFA.

** item correlations are shown in Fig. 1.

Abbreviations: X² = Chi-Square Statistic (degrees of freedom); CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual

Table 3

Model fit statistics and indices from models to estimate the overlap of depressive symptoms with MS-related fatigue, cognitive impairment and functional impairment.

	Constrained	MIMIC	A=0	B=0	C=0	D=0	E=0	F=0
X²	1592.55 (56)	475.45 (48)	573.02 (49)	1207.70 (49)	913.31 (49)	603.12 (49)	479.13 (50)	491.08 (50)
CFI	0.877	0.966	0.958	0.907	0.931	0.956	0.966	0.965
TLI	0.842	0.949	0.938	0.864	0.899	0.935	0.951	0.949
RMSEA (95% CI)	0.088 (0.085, 0.092)	0.050 (0.046, 0.055)	0.055 (0.051, 0.059)	0.082 (0.078, 0.086)	0.071 (0.067, 0.075)	0.057 (0.053, 0.061)	0.049 (0.045, 0.054)	0.050 (0.046, 0.054)
SRMR	0.046	0.028	0.031	0.039	0.036	0.032	0.029	0.030
X² Diff p-value *	<0.001	reference	<0.001	<0.001	<0.001	<0.001	0.159	<0.001
Mod Indices **	--	--	98.27	733.00	430.66	120.79	4.12, 4.26	22.55, 6.07

p < 0.001 for all X² tests of model fit.

In constrained model A, B, C, D, E and F=0 while in the MIMIC model A, B, C, D, E and F are freely estimated.

Better fit is characterized by a higher CFI and TLI, and lower χ^2 , RMSEA and SRMR

* X² Difference Test is for each separate model in comparison to the full MIMIC model.

**

For E=0 and F=0 first modification index is for the 9-hole peg test, second modification index is for the timed 25-foot walk.

Abbreviations: X² = Chi-Square Statistic (degrees of freedom); CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized

Root Mean Square Residual; X² Diff = Chi-Square Difference Statistic; Mod Indices = Modification Indices

Table 4

Standardized factor loadings and estimates (r) of DIF paths for 1-factor CFA with correlations, full constrained, full MIMIC and best fitting (E=0) models.

	Model			
	CFA	Constrained	MIMIC	E=0
<i>depression</i>				
Factor Loadings				
1. anhedonia	0.81	0.79	0.82	0.82
2. feel depressed	0.81	0.77	0.82	0.82
3. sleep problems	0.62	0.65	0.49	0.49
4. fatigue	0.67	0.73	0.34	0.35
5. appetite change	0.66	0.67	0.66	0.66
6. feelings of failure	0.75	0.72	0.75	0.75
7. poor concentration	0.69	0.72	0.48	0.48
8. psychomotor symptoms	0.58	0.61	0.43	0.43
9. self-harm	0.51	0.47	0.51	0.51
Estimates (r)				
DIF paths				
A	-	0	0.21 *	0.21 *
B	-	0	0.52 *	0.52 *
C	-	0	0.38 *	0.38 *
D	-	0	0.23 *	0.23 *
E (peg test)	-	0	-0.02	0
E (timed walk)	-	0	-0.02	0
F (peg test)	-	0	0.11 *	0.11 *
F (timed walk)	-	0	-0.01	0.00

*
p < 0.001