



## A comparison of the measurement properties of the PROMIS Fatigue (MS) 8a against legacy fatigue questionnaires

Paul Kamudoni<sup>a,\*</sup>, Jeffrey Johns<sup>b</sup>, Karon F. Cook<sup>c</sup>, Rana Salem<sup>d</sup>, Sam Salek<sup>b,e</sup>, Jana Raab<sup>a</sup>, Rod Middleton<sup>f</sup>, Christian Henke<sup>a</sup>, Dagmar Amtmann<sup>d</sup>

<sup>a</sup> Global Evidence & Value Development - R&D, Merck Healthcare KGaA, Darmstadt, Germany

<sup>b</sup> Institute of Medicines Development, Cardiff, UK

<sup>c</sup> Feral Scholars, Broadus, TX, USA

<sup>d</sup> Department of Rehabilitation Medicine, University of Washington, Seattle, USA

<sup>e</sup> School of Life and Medical Sciences, University of Hertfordshire, Hatfield, UK

<sup>f</sup> UK MS Register, Swansea University Medical School, Swansea, UK

### ARTICLE INFO

#### Keywords:

Fatigue  
Multiple sclerosis  
Outcome measurement  
Quality of life  
PROMIS

### ABSTRACT

**Background:** Amidst the growing number of patient-reported outcome (PRO) measures of fatigue being used in multiple sclerosis (MS) clinical trials and clinics, evidence-based consensus on the most appropriate and generalizable measures across different settings would be beneficial for clinical research and patient care. The objective of this research was to compare the validity and responsiveness of scores from the PROMIS Fatigue (MS) 8a with those of the Fatigue Severity Scale (FSS) and the Modified Fatigue Impact Scale (MFIS), across US and UK MS populations.

**Methods:** Two observational studies were performed in MS populations as part of a PRO measure development project, including a cross-sectional study in two tertiary US MS centers ( $n = 340$ ) and a 96-week longitudinal study in the UK MS Register cohort ( $n = 352$ ). In post-hoc analyses, we examined relative validity, based on ability to discriminate across patient groups with different fatigue levels or functional status at baseline (i.e., ANOVA-F PRO<sub>x</sub> ÷ ANOVA-F PROMIS (MS) 8a), and relative responsiveness, based on baseline-to-Week-52 score change (effect sizes) across fatigue or functional status response groups.

**Results:** Mean ± standard deviation (SD) age was  $44.6 \pm 11.3/50.0 \pm 9.7$ ; and 72.9%/77.3% were female (US/UK samples). The mean PROMIS Fatigue (MS) 8a T-score ± SD at baseline was  $57.7 \pm 10.5/58.9 \pm 9.3$  (US/UK samples). Compared with the PROMIS Fatigue (MS) 8a, relative validity (anchor: Global Health Score [GHS] fatigue global question) was 85% for MFIS symptom score, 48% for MFIS total score, and 44% for the FSS. Relative to the FSS, PROMIS Fatigue (MS) 8a scores were more sensitive to worsening (effect size = -0.43 versus -0.18) as well as improvement (effect size = 0.5 versus 0.2) in fatigue ( $\geq 1$ -point increase/decrease in GHS fatigue global question) over 52 weeks of follow-up. A similar pattern of score changes was observed based on a second anchor.

**Conclusion:** The PROMIS Fatigue (MS) 8a scores showed higher responsiveness to fatigue changes than those of the FSS. The PROMIS measure also had higher precision in differentiating levels of fatigue compared to the FSS, the MFIS physical, and MFIS total scores. These differences have practical implications for the application of these questionnaires in both clinical practice and research settings (e.g., sample size estimation in clinical trials).

**Abbreviations:** ANCOVA, analysis of covariance; ANOVA, analysis of variance; ES, effect size; FSIQ-RMS, Fatigue Symptoms and Impacts Questionnaire - relapsing multiple sclerosis; FSS, Fatigue Severity Scale; GHS, Global Health Score; GPH, Global Physical Health; MFIS, Modified Fatigue Impact Scale; MID, minimal important difference; MS, multiple sclerosis; NFI-MS, neurological fatigue index – multiple sclerosis; PPMS, primary progressive multiple sclerosis; PRO, patient-reported outcome; PROMIS, Patient-Reported Outcome Measurement Information System; PR-WebEDSS, Patient-Reported Web-based Expanded Disability Status Scale; PwMS, people with multiple sclerosis; REI, relative efficiency index; RRMS, relapsing remitting multiple sclerosis; SPMS, secondary progressive multiple sclerosis; SRM, standard response mean.

\* Corresponding author.

E-mail address: [paul.kamudoni@merckgroup.com](mailto:paul.kamudoni@merckgroup.com) (P. Kamudoni).

<https://doi.org/10.1016/j.msard.2022.104048>

Received 7 January 2022; Received in revised form 11 April 2022; Accepted 8 July 2022

Available online 9 July 2022

2211-0348/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Fatigue is a common and disabling symptom of the chronic neurodegenerative disease multiple sclerosis (MS), and is reported by more than 80% of People with MS (PwMS) within the first year of disease onset and throughout its course (Kister et al., 2013). Moreover, 69% of PwMS regard fatigue as either the most important or one of the most important symptoms of their disease (Fisk et al., 2015), making its measurement and management in clinical practice, as well as its consideration in clinical research, important. Currently, the pathogenesis of fatigue is not fully understood, several pathways with linkages to disease mechanism have been suggested, including demyelination and secondary axonal degeneration, as well as pro-inflammatory cytokines (Comi et al., 2001; Penner and Paul, 2017). In addition, psychological factors such as mood disorders, motivation, and arousal, and peripheral factors including physiological changes such as muscle contractility, excitability, and sleep disorders, also play a role in fatigue in MS (Langskov-Christensen et al., 2017; Rudroff et al., 2016).

A growing number of patient-reported outcome (PRO) measures are available for assessing fatigue in MS, including legacy instruments such as the Modified Fatigue Impact Scale (MFIS) and the Fatigue Severity scale (FSS), as well as a new generation of measures relying on modern test methods such as the Fatigue Symptoms and Impacts Questionnaire - Relapsing Multiple Sclerosis (FSIQ-RMS), Neurological Fatigue Index - Multiple Sclerosis (NFI-MS) and the PROMIS SF v1.0 - Fatigue - Multiple Sclerosis 8a (PROMIS Fatigue (MS) 8a). These differ in terms of conceptual focus (e.g., fatigue severity versus impacts) and dimensional structure (e.g., inclusion of subscales for specific aspects of fatigue). The emergence of item banking to measure PROs provides new options for measuring health domains (including fatigue). Measures that are well tailored/targeted to specific populations, (Cella et al., 2010) brief (Evans et al., 2018) and show good precision (Bingham et al., 2019) can be derived from PRO item banks. The PROMIS Fatigue (MS) 8a is one such item bank-based measure. It was derived from the PROMIS fatigue item bank, which was developed with input from clinicians and PwMS; this measure has demonstrated good content validity and strong psychometric properties in MS populations (Cook et al., 2012).

There is still limited robust evidence-based recommendations regarding appropriate PRO measures for use across different settings in MS, e.g., clinical practice, clinical research or performance measurement (Kamudoni et al., 2021; Wang and Woodruff, 2015). For example, previous analyses comparing the FSS and the MFIS, such as the work of Amtmann et al. (2012) or Learmonth et al. 2013, provide useful insights for choosing between the two measures. The ongoing US Food and Drug Administration (FDA) clinical outcome assessment qualification of the PROMIS Fatigue (MS) 8a is a small albeit helpful step in consolidating the measurement of fatigue in clinical trials (Critical path institute, 2009). Ultimately, the selection of the most appropriate measures is context specific and depends on the goals of PRO assessments. Critically, however, measure selection should take into account validity evidence as well as practical aspects such as respondent burden, interpretability of scores, generalizability and actionability of results.

The objective of this research was to compare the validity and responsiveness of PROMIS Fatigue (MS) 8a scores with those of the MFIS and the FSS, across US and UK MS populations.

## 2. Methods

This was a post-hoc analysis of two observational studies in PwMS, carried out to develop and validate PROMIS short-form measures in MS, including a cross-sectional study in two tertiary MS centers in the US and a 96-week longitudinal study in the UK MS Register cohort. Further details about the observational studies are published elsewhere (Kamudoni et al., 2021).

### 2.1. Participants and procedures

Study participants had a clinician-confirmed MS diagnosis, were 18–65 years old, were able to use a computer or tablet, and were able to read and write in English. Exclusion criteria included the use of a wheelchair or scooter as the main form of mobility and cognitive or other impairments (e.g., visual) that could interfere with questionnaire completion. In addition, PwMS with a patient-reported expanded disability scale (PR-WebEDSS) score > 6.5 and MS phenotypes other than relapsing remitting MS (RRMS), primary progressive MS (PPMS), and secondary progressive MS (SPMS) were excluded from the analysis sample. The post-hoc analysis sample was comprised of the subset of the enrolled study participants who completed both the PROMIS Fatigue (MS) 8a measure and the FSS or the MFIS.

The UK MS Register is among the largest MS registers in Europe (Ford et al., 2012). Data are collected via a web portal, with the possibility of linkage with healthcare records provided by participating National Health Service (NHS) neurology clinics (Ford et al., 2012). Participants are recruited through the registry's portal or through the 48 participating NHS neurology centers across the UK. All members of the registry were sent an emailed invitation to join the study. In addition to the standard routine assessments in the registry (every 6 months), study participants completed various PRO instruments (described below) over 96 weeks; that is at baseline, Weeks 1, 24, 52, 72, and 96. (eMethods Online Supplement; eTable 1). Data were collected between September 2018 and October 2020; the current analyses are based on data collected from baseline through Week 52.

The cross-sectional US-UW study was conducted at the MS Center at the University of Washington Medical Center – Northwest and the Swedish Neuroscience Institute, in Seattle, Washington, USA. PwMS were invited to join the study via post (if they had an active registration at one of the two centers) or during their routine attendance at the clinics. Prospective data collection took place between July 2019 and January 2020 in the clinic using an iPad® tablet.

### 2.2. Outcome measures

#### 2.2.1. PROMIS SF v1.0 - Fatigue (MS) 8a

The PROMIS Fatigue (MS) 8a was developed as a measure of fatigue experience and impacts over 7 days, in PwMS. The short form's eight items were derived from the PROMIS fatigue item bank based on input from PwMS and clinical experts (Cook et al., 2012). The PROMIS Fatigue (MS) 8a is scored on a T-score metric, which has a mean of 50 and a standard deviation (SD) of 10; higher scores indicate higher fatigue. The T-score metric is referenced to the US general population with respect to race/ethnicity, age, education, and sex; for example, a T-score of 40 would be one SD below the US general population.

#### 2.2.2. Fatigue Severity Scale (FSS)

The FSS is a 9-item measure of fatigue experience and related impacts. It is based on a 7-day recall period and was developed for use across chronic conditions including MS (Amtmann et al., 2012; Krupp et al., 1989; Learmonth et al., 2013; Mills et al., 2010). The summary score is calculated as a raw sum (or mean) of the item scores; a higher score indicates higher fatigue (range: 9–63). A score of > 36 is indicative of severe fatigue (Andreassen et al., 2011). A minimal important difference (MID) of 4.5–9.9 (i.e., 0.5/1.1, times 9) has been reported in PwMS (Wang and Woodruff, 2015).

#### 2.2.3. Modified Fatigue Impact Scale (MFIS)

The MFIS was developed as a measure of the impacts of fatigue on quality of life over 4 weeks in PwMS (Amtmann et al., 2012; Elbers et al., 2012; Learmonth et al., 2013; Rietberg et al., 2010). The measure's 21 items cover three subdomains: physical, cognitive, and psychologic functioning. Domain scores and a total score are calculated as raw item sum scores; higher scores represent higher fatigue impact.

2.2.4. Other measures

Participants completed other PRO measures, including the PR-WebEDSS (Leddy et al., 2013) and the PROMIS v1.2 – Global Health Scale (GHS) (Hays et al., 2009) (eMETHODS Online Supplement; Outcome Measures). In addition, data on clinical characteristics were retrospectively extracted from the participants’ records.

2.3. Statistical analysis

We performed various analyses to compare the measurement properties of the PROMIS Fatigue (MS) 8a, the MFIS, and the FSS, including evaluation of floor and ceiling effects, relative validity, and responsiveness.

Software used included STATA v15.1 (StataCorp, 2017), Software MPLUS v8.2 (Muthén and Muthén, 2017), and R v3.33 (R Core Team, 2018). Analyses were performed separately for the US and the UK studies.

The proportions of the sample with highest/lowest responses across all items were calculated to evaluate ceiling/floor effects, for each PRO measure. A proportion of > 0.15 was judged to be an indicator of problematic ceiling or floor effect (Terwee et al., 2007).

Reliability was assessed based on Cronbach’s alpha coefficient (internal consistency reliability) and item-response theory-based scale information (score precision by fatigue level across the continuum). Scale information relates to “traditional” reliability as follows:  $Reliability = 1 - \frac{1}{Information(\theta)}$

We assessed known groups validity of each PRO measure by testing hypothesized score differences across distinct patient groups, using analysis of variance (ANOVA). The groups were defined based on:

- GHS fatigue question (none/mild/moderate; severe/very severe)
- GHS general health question (excellent/very good/good; fair/poor)
- GHS physical health question (excellent/very good/good; fair/poor)
- PR-WebEDSS (0–4; 4.5–6.5)
- EDSS (0–4.0; 4.5–6.5)
- GHS Global Physical Health (GPH) summary score (< 50; ≥ 50).

Subsequently, we calculated a relative validity index, as the ratio of F-statistics from the between-group ANOVA tests performed on each PRO measure (as described above) (i.e.,  $F\text{-statistic}_{FSS \text{ or } MFIS} / F\text{-Statistic}_{PROMIS \text{ Fatigue (MS) 8a}}$ ); using the PROMIS Fatigue (MS) 8a as denominator. A ratio of less than 100% indicated less discriminatory power compared to the PROMIS Fatigue (MS) 8a.

Further, we performed analyses of known-groups validity and relative validity in subgroups defined based on PR-WebEDSS (i.e., 0–4, 4.5–6.5) and MS phenotype (relapsing MS; secondary and progressive MS). These analyses were performed only for the GHS fatigue question.

The longitudinal design of the UK MS Register allowed us to evaluate and compare responsiveness of the PROMIS Fatigue (MS) 8a and the FSS. For each PRO measure, score change from baseline to Week 52 across participant groups experiencing differing levels of change in fatigue or functional status was examined. Participants were classified as improving or worsening based on baseline to Week 52 score changes in:

- GHS fatigue question (≥ 1-point decrease; ≥ 1-point increase)
- GHS GPH summary score (≥ 5-point decrease; ≥ 5-point increase).

The appropriateness of the selected anchors was assessed based on multiple criteria. Spearman’s correlations between longitudinal anchors and change scores from PROMIS Fatigue (MS) 8a and FSS (at Week 52) are shown in the Online supplement (Coon and Cook, 2018; Yost et al., 2011).

For each PRO measure, the within-group score change was assessed using paired T-tests, between baseline and Week 52 scores. Standard Response Mean (SRM) and Cohen’s d effect size (ES) were calculated for

**Table 1**  
Characteristics of study participants.

Characteristic	UK sample <sup>a</sup> (n = 352)	UK sample <sup>b</sup> (n = 246)	US sample <sup>c</sup> (n = 340)
<b>Age</b>			
Mean (SD)	50.0 (9.7)	51.0 (9.5)	44.6 (11.3)
Median	51	53	43.9
Range	22-65	22-65	20.9-65.6
<b>Gender, n (%)</b>			
Male	80 (22.7)	58 (23.6)	90 (26.5)
Female	272 (77.3)	188 (76.4)	248 (72.9)
Non-binary	0	0	2 (0.6)
<b>Time since MS diagnosis, years</b>			
Mean (SD)	10.14 (7.90)	10.57 (8.20)	9.72 (7.77)
Median	8.0	8.5	8.13
Range	0-38	0-38	0.12-37.72
<b>Patient-reported WebEDSS</b>			
Mean (SD)	4.58 (1.20)	4.75 (1.98)	3.54 (1.76)
Median	5.0	6.0	3.5
Min-Max	0-6.5	0-6.5	0-7.5
Mild (0-4.0), n (%)	153 (43.5)	92 (37.4)	223 (65.6)
Moderate (> 4-6.5), n (%)	199 (56.5)	154 (62.6)	109 (32.1)
Severe (> 6.5)	-	-	8 (2.4)
<b>MS phenotype</b>			
RRMS	236 (67.1)	152 (61.79)	283 (83.2)
SPMS	78 (22.2)	65 (26.42)	12 (3.5)
PPMS	38 (10.8)	29 (11.79)	8 (2.4)
Other	-	-	15 (4.4)
Unknown	-	-	22 (6.5)
<b>PROMIS Fatigue (MS) 8a</b>			
Mean (SD)	58.9 (9.3)	58.9 (9.1)	57.7 (10.5)
Median	60.1	60.0	58.6
Range	34.1-80.7	34.1-80.7	34.1-80.7
<b>Modified Fatigue Impact Scale (total)</b>			
Mean (SD)	-	-	36.12 (22.3)
Median	-	-	36
Range	-	-	0-84
<b>Modified Fatigue Impact Scale (physical)</b>			
Mean (SD)	-	-	17 (10.8)
Median	-	-	18
Range	-	-	0-36
<b>Fatigue Severity Scale</b>			
Mean (SD)	44.7 (13.6)	45.3 (13.8)	-
Median	46.5	48.0	-
Range	9-63	9-63	-

**Abbreviations:** EDSS, Expanded Disability Status Scale; FSS, Fatigue Severity Scale; GHS, Global Health Score; MS, multiple sclerosis; PPMS, primary progressive MS; PROMIS, Patient-Reported Outcome Measurement Information System; RRMS, relapsing remitting MS; SD, standard deviation; SPMS, secondary progressive MS; WebEDSS, patient-reported web-based Expanded Disability Status Scale.

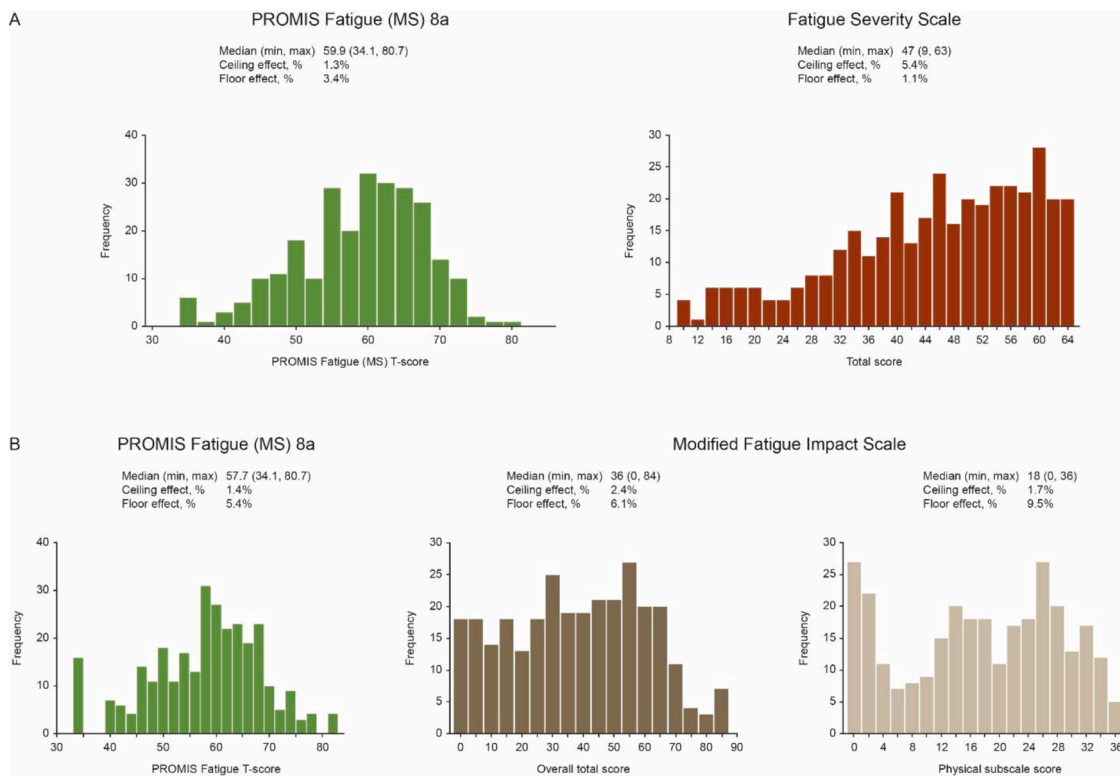
<sup>a</sup> Analysis sample includes respondents with EDSS ≤ 6.5, age ≤ 65 years, and with PPMS, RRMS, or SPMS phenotypes, and FSS and GHS at baseline

<sup>b</sup> Responsiveness analysis sample includes respondents with EDSS ≤ 6.5, age ≤ 65 years, and with PPMS, RRMS, or SPMS phenotypes, and FSS and GHS at baseline and week 52.

<sup>c</sup> All respondents, with web-EDSS assessment.

each group. ES was interpreted as: small, ES = 0.2; moderate, ES = 0.5; and large ES = 0.8 (Cohen, 1987). Between-group comparisons in change scores i.e., worsening versus unchanged, and unchanged versus improving, were performed using analysis of covariance (ANCOVA; controlling for baseline score), for each PRO measure.

Relative responsiveness was assessed by comparing the PRO measures’ precision at detecting across group differences in score change (worsening vs. unchanged vs. improving). The ratio of F-statistics from ANCOVA of score change across the three groups was calculated ( $F\text{-statistic}_{FSS} / F\text{-Statistic}_{PROMIS \text{ Fatigue (MS) 8a}}$ ) (Fayers and Machin, 2013;



**Fig. 1.** Distribution of PRO scores at baseline in (A) UK Sample ( $n = 352$ ) and (B) US sample ( $n = 340$ ) PRO, patient-reported outcome; PROMIs, patient-reported outcome measurement information system; MS, multiple sclerosis.

Ware and Gandek, 1998). A ratio of less than 100% indicated stronger responsiveness for the PROMIS Fatigue (MS) 8a relative to the FSS.

### 3. Results

#### 3.1. Participant characteristics and baseline scores

Of study participants included in the two observational studies (Kamudoni et al., 2021), the sample used in the current analysis included 340 (US-UW) and 352 (UK MS Register) patients. Further details about the study population and the study data are available in the previously published article (Kamudoni et al., 2021). The mean age of study participants in the analysis sample was 50.0 (SD = 9.7) in the UK MS Register analysis sample and 44.6 (SD = 11.3) in the US-UW analysis sample (Table 1). The majority had RRMS (67.1% in the UK MS Register sample and 83.2% in the US-UW sample). In the US-UW sample, most participants had a PR-webEDSS of 0–4 (65.6%), while most participants in the UK MS Register sample had a PR-webEDSS of > 4–6.5 (56.5%).

#### 3.2. Score distribution and reliability

Fig. 1 shows ceiling and floor effects for each measure by sample (A = UK Sample; B = US Sample). For the UK Sample, floor effects are higher for PROMIS Fatigue (MS) 8a compared to the FSS (3.4% and 1.1%, respectively). However ceiling effects were lower for PROMIS Fatigue (MS) 8a compared to the FSS (1.3% and 5.4%, respectively). For all three measures, floor/ceiling effects were well below the critical values of 15%.

Cronbach’s alpha coefficient was 0.95 or greater for all three PRO measures. Information function plots for the PRO measures are presented in Fig. 2. Compared to PROMIS Fatigue (MS) 8a, the MFIS total score had information-calculated reliability of 0.95 or greater range of the fatigue continuum, particularly, at high levels of fatigue. On the other hand, the PROMIS Fatigue (MS) 8a had high reliability (i.e., > 0.95) across a wider range of scores relative to the FSS.

#### 3.3. Comparative validity

We examined known groups based on score differences across participant groups referencing multiple anchors. Scores of all three PRO measures showed statistically significant differences across participant groups, for all anchors, except the MS phenotype (ANOVA Test,  $p < 0.01$ ).

The PROMIS Fatigue (MS) 8a discriminated better among fatigue levels (anchor: GHS fatigue question) relative to MFIS total and physical scores (Table 2). On the other hand, both MFIS total and physical scores discriminated better across non-fatigue anchors, including the self-reported disability levels (PR-webEDSS) and summary physical health (GHS GPH Summary score).

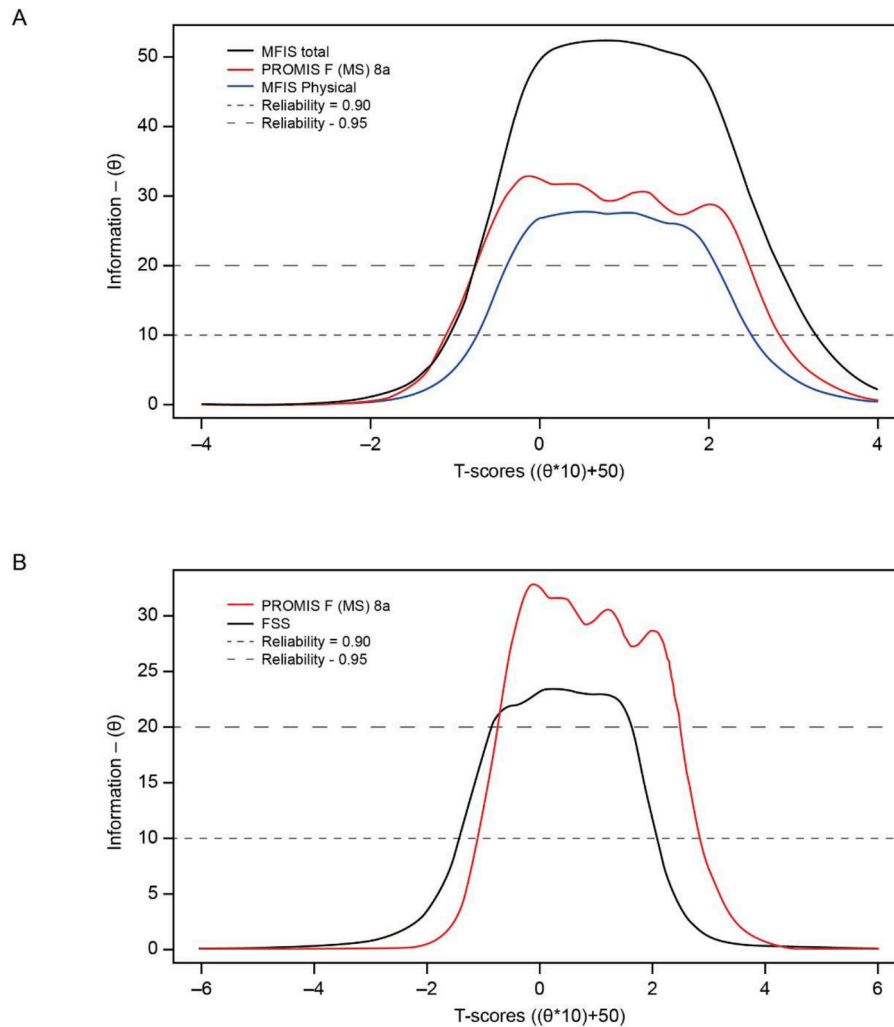
The PROMIS Fatigue (MS) 8a performed better relative to the FSS in discriminating fatigue levels (anchor: GHS fatigue question) and summary physical health (GHS GPH Summary score), but not across self-reported disability levels (PR-webEDSS; Table 3).

In subgroup analyses, performed for the GHS fatigue question anchor only, the PROMIS Fatigue (MS) 8a showed better discrimination of participants across fatigue levels than the MFIS (both physical or total) or the FSS, in all subgroups, i.e., relapsing and progressive types, and mild as well as moderate disability (PR-WebEDSS). This is consistent with findings for the overall study sample (eTables 3 and 4).

#### 3.4. Comparative responsiveness

Analysis of responsiveness was performed for the PROMIS Fatigue (MS) 8a and the FSS only, based on the UK MS Register sample; longitudinal data on the PROMIS Fatigue (MS) 8a versus MFIS were not available, given the cross-sectional design of the US-UW sample.

We examined PROMIS Fatigue (MS) 8a and the FSS score changes from baseline to Week 52 across patient groups experiencing different levels of change in fatigue or functional status, based on the GHS fatigue question and the GHS GPH Summary score (Table 4).



**Fig. 2.** Scale information function plots based on item response theory characteristics. (A) the PROMIS fatigue (MS) 8a and the MFIS; (B) the PROMIS Fatigue (MS) 8a and the FSS. FSS, Fatigue Severity Scale; PROMIS, patient-reported outcome measurement information system; MFIS, modified fatigue impact scale; MS, multiple sclerosis.

Overall, the PROMIS Fatigue (MS) 8a was sensitive to worsening as well as improvement in fatigue levels and summary physical health. The FSS exhibited weak sensitivity, which was limited to improvements in fatigue (Table 4).

We observed statistically significant within-group score changes of a mild-to-moderate magnitude (ES 0.4–0.5), for both the worsening and the improving fatigue groups for the PROMIS Fatigue (MS) 8a. For the FSS, although score changes were statistically significant for both directions, the magnitude of change was small; ES was  $\leq 0.2$  for both worsening as well improving groups.

Scores of both PRO measures showed discrimination between the improving and the unchanged/stable groups, as well as between the unchanged and the worsening groups (ANCOVA,  $p < 0.01$ ). The REI comparing the two PRO measures, based on the ratio of F-statistic from the respective between-group ANCOVAs, indicated that the PROMIS Fatigue (MS) 8a score outperformed the FSS scores in discriminating among groups experiencing different levels of change.

#### 4. Discussion

In this research, PROMIS Fatigue (MS) 8a scores showed better discrimination of fatigue levels compared with the FSS and the MFIS total and MFIS physical subscale scores. PROMIS Fatigue (MS) 8a scores were more sensitive to changes in fatigue compared with FSS scores.

Scale floor effects were lower for the FSS, while ceiling effects were lower for the MFIS, relative to the PROMIS Fatigue (MS) 8a. The fact that these results were observed based on two separate populations, i.e., attendees at tertiary clinics in the US, and members of the UK MS Register, lends confidence in the generalizability of these results to other samples of PwMS. Moreover, our subgroup analyses by disability and MS phenotype supported our conclusions in relapsing and progressive MS populations, as well as in mild and moderate disability (eTables 3 and 4 and ).

Previous research has compared the MFIS and the FSS, while no study has compared these two PRO measures with the PROMIS Fatigue (MS) 8a. In a small study ( $n = 86$ ) based on the North American Research Committee on Multiple Sclerosis patient registry, MFIS scores showed stronger correlation with mobility, according to the MS Walking Scale-12 and the Six-Minute Walk test and stronger correlation with cognition according to the Symbol Digit Modalities Test, relative to the FSS (Learnmonth et al., 2013). On the other hand, the FSS scores showed better precision than the MFIS based on standard error of measurement and 6-month coefficient of variation (Learnmonth et al., 2013). Amtmann et al. (2012) employed modern test methods to examine the psychometric properties of the MFIS and the FSS in a sample of community-living people with MS. In this study by Amtmann et al. (2012), floor effects were low for both the FSS (0.9%) and MFIS (1.1%), whereas ceiling effects were higher for the FSS (6.8%) compared with

**Table 2**  
Comparative validity of PROMIS Fatigue (MS) 8a against the MFIS based on score differences across distinct participant subgroups (US-UW Sample).

Anchor	N	PROMIS Fatigue (MS) 8a T-score			MFIS Physical Sub-score			MFIS Total Score			Relative validity index <sup>a</sup> , %	
		Score Mean (SD)	F-statistic	p-value	Score Mean (SD)	F-statistic	p-value	Score Mean (SD)	F-statistic	p-value	MFIS Total	MFIS Physical
<b>ClinEDSS</b>			18.918	0.000		34.887	0.000		18.873	0.000	82.74	161.26
0–4.0	222	56.0			15.1			33.4				
4.5–6.5	75	61.9			23.3			46.1				
<b>PR-WebEDSS</b>			128.578	0.000		222.108	0.000		161.103	0.000	120.82	153.33
0–4.0	175	52.7			10.9			24.8				
4.5–6.5	165	63.6			24.6			50.2				
<b>GHS General Health question (Global01)</b>			46.967	0.000		59.986	0.000		53.192	0.000	101.85	116.61
Fair/poor (1,2)	76	64.8			25.4			52.5				
Excellent/very good/good (3,4,5)	264	56.0			15.3			32.7				
<b>GHS Physical Health question (Global03)</b>			76.169	0.000		96.510	0.000		73.652	0.000	86.28	119.00
Fair/poor (1,2)	115	64.2			24.7			50.3				
Excellent/very good/good (3,4,5)	223	54.7			13.8			30.3				
<b>GHS Fatigue question (Global08r)</b>			362.441	0.000		352.710	0.000		327.575	0.000	47.99	85.05
Severe/very severe (1,2)	226	63.3			23.1			48.3				
None/mild/moderate (3,4,5)	114	47.4			6.6			15.1				
<b>GHS GPH Summary score</b>			301.531	0.000		420.987	0.000		330.638	0.000	107.03	137.07
≤ 50	232	62.9			23.1			47.9				
> 50	108	47.5			5.7			14.1				

Abbreviations: GHS, Global Health Score; GPH, Global Physical Health; MFIS, Modified Fatigue Impact Scale; MS, multiple sclerosis; PROMIS, Patient-Reported Outcome Measurement Information system; PR-WebEDSS, patient-reported web-based Expanded Disability Status Scale; SD, standard deviation.

<sup>a</sup> For relative validity, less than 100% represents better performance for the PROMIS Fatigue (MS) 8a.

**Table 3**  
Comparative validity of PROMIS Fatigue (MS) 8a against the FSS based on score differences across distinct participant subgroups (UK MS Register sample).

Anchor	PROMIS Fatigue (MS) 8a T-score			Fatigue Severity Scale			Relative validity index <sup>a</sup> (%)
	n	Mean (SE)	F-statistic (df); p-value	n	Mean (SE)	F-statistic (df); p-value	
<b>PR-WebEDSS score</b>	352		107.11 (351); < 0.001	352		11.21 (351); < 0.001	10
0–4.0	153	53.76 (0.75)		153	37.94 (1.11)		
4.5–6.5	199	62.87 (0.51)		199	49.85 (0.77)		
<b>GHS General Health (Global01)</b>	352		48.86 (351); < 0.001	352		40.03 (351); < 0.001	82
Fair/poor (1,2)	142	64.25 (0.51)		142	51.55 (0.81)		
Excellent/very good/good (3,4,5)	210	55.30 (0.65)		210	40.03 (0.96)		
<b>GHS Physical Health question (Global03)</b>	352		69.93 (351); < 0.001	352		45.87 (351); < 0.001	66
Fair/poor (1,2)	189	63.62 (0.48)		189	50.48 (0.76)		
Excellent/very good/good (3,4,5)	163	53.45 (0.71)		163	37.95 (1.07)		
<b>GHS Fatigue question (Global08r)</b>	352		226.54 (351); < 0.001	352		100.58 (351); < 0.001	44
Severe/very severe (1,2)	97	67.56 (0.51)		97	55.12 (0.90)		
None/mild/moderate (3,4,5)	255	55.62 (0.53)		255	40.70 (0.81)		
<b>GHS GPH Summary score</b>	352		9.23 (351); < 0.001	352		3.83 (351); < 0.001	42
≤ 50	294	61.37 (0.43)		294	47.75 (0.66)		
> 50	58	46.41 (1.09)		58	29.10 (1.79)		

Abbreviations: df, degree of freedom; FSS, Fatigue Severity Scale; GHS, Global Health Score; GPH, Global Physical Health; MFIS, Modified Fatigue Impact Scale; MS, multiple sclerosis; PROMIS, Patient-Reported Outcome Measurement Information system; PR-WebEDSS, patient-reported web-based Expanded Disability Status Scale; SE, standard error.

<sup>a</sup> For relative validity, less than 100% represents better performance for the PROMIS Fatigue (MS) 8a.

the MFIS (0.7%). The MFIS scores were more highly correlated with scores on measures of depression and other health concepts in comparison with the FSS scores; scores of both the MFIS and FSS demonstrated strong known groups validity. In addition, based on test information from the item response theory analysis, MFIS appeared to measure with more precision at higher levels of fatigue (Amtmann et al., 2012).

The differences we found among the three PRO measures should be interpreted with contrastin characteristics of the measures. First, the three measures have important conceptual differences. The MFIS was designed to cover the impacts of fatigue on multiple aspects, i.e., cognition, physical, and psychosocial. The FSS and the PROMIS Fatigue

(MS) 8a are unidimensional scales that define fatigue severity in terms of the impacts and manifestations of fatigue. Second, in contrast to the MFIS and the FSS, the PROMIS Fatigue (MS) 8a was developed based on modern test measurement methods and explicitly sought to include items that targeted the full continuum of fatigue, from low to high severity. Further, while the MFIS uses a 4-week recall period, the FSS and the PROMIS Fatigue (MS) 8a use a 1-week recall period. The response burden of the scales also varies. The PROMIS Fatigue (MS) 8a, the FSS, and the MFIS physical subscale have similar numbers of items (8, 9, and 9, respectively). The MFIS total, however, has 21 items.

In determining the appropriateness of PRO measures in a given setting and context of use, it is important to consider the measurement

**Table 4**

Comparative responsiveness of PROMIS Fatigue (MS) 8a versus FSS: score changes over a 52-week duration; UK MS Register sample (N = 246).

Anchor	Participant subgroups (n)	PROMIS Fatigue (MS) 8a			Fatigue Severity Scale			Relative responsiveness (%) <sup>b</sup>
		Mean Change (SD) <sup>a</sup>	Cohen's d ES	ANCOVA F-statistic	Mean Change (SD) <sup>a</sup>	Cohen's d ES	ANCOVA F-statistic	
		BL – Week 52			BL – Week 52			
<b>GHS Fatigue question (Global08r)<sup>c</sup></b>	Improving (42)	4.01 (4.30) <sup>****</sup>	0.50 (0.08, 0.95)	32.38 <sup>****</sup>	2.36 (7.33) <sup>*</sup>	0.20 (-0.23,0.63)	10.96 <sup>****</sup>	33.8
	Stable (146)	0.45 (4.51)	0.05 (-0.18, 0.38)		-0.77 (8.37)	-0.05 (-0.28, 0.17)		
	Worsening (58)	-3.72 (5.62) <sup>****</sup>	-0.43 (-0.79, -0.06)		-2.40 (7.89) <sup>*</sup>	-0.18 (-0.54,0.19)		
<b>GHS GPH Summary score<sup>d</sup></b>	Improving (25)	3.97 (4.52) <sup>***</sup>	0.39 (-0.17, 0.95)	23.02 <sup>****</sup>	3.36 (6.99) <sup>*</sup>	0.22 (-0.33, 0.79)	12.38 <sup>****</sup>	53.8
	Stable (176)	0.37 (4.81)	0.04 (-0.17, 0.25)		-0.67 (8.28)	-0.05 (-0.26, 0.16)		
	Worsening (45)	-3.19 (6.01) <sup>****</sup>	-0.37 (-0.79, 0.05)		-2.64 (7.84) <sup>*</sup>	-0.19 (-0.61,0.22)		

**Abbreviations:** ANCOVA, analysis of covariance; BL, baseline; ES, effect size; FSS, Fatigue Severity Scale; GHS, Global Health Scale; GPH, Global Physical Health; MS, multiple sclerosis; PROMIS, Patient-Reported Outcome Measurement Information System; SD, standard deviation.

<sup>a</sup> Baseline to Week 52 score change in respective subgroups tested using paired T-test.

<sup>b</sup> For relative responsiveness, less than 100% represents better performance for the PROMIS Fatigue (MS) 8a. Statistical significance of paired T-test or ANCOVA test:

<sup>\*</sup>  $p \leq 0.05$

<sup>\*\*</sup>  $p \leq 0.01$

<sup>\*\*\*</sup>  $p \leq 0.001$

<sup>\*\*\*\*</sup>  $p \leq 0.0001$ .

<sup>c</sup> **GHS fatigue question (Global08r)** response groups: **improving** =  $\geq 1$ -point increase, **worsening** =  $\geq 1$ -point decrease.

<sup>d</sup> **GHS GPH Summary score** response groups: **improving** =  $\geq 5$ -point increase, **worsening** =  $\geq 5$ -point decrease.

properties of available scales, the objectives of the research, the intended use of the data, practical aspects such as administrative and patient burden, measurement setting, and the context of use. The current findings are consistent with previous research in supporting the reliability and validity of the MFIS, the FSS and the PROMIS Fatigue (MS) 8a as measures of fatigue severity and/or impact in MS. In the context of clinical practice, clinical research or drug development, when a single score is needed to summarize the severity of fatigue as a symptom, the PROMIS Fatigue (MS) 8a would be recommended over the MFIS or the FSS, as it is brief, and provides better discrimination of fatigue levels. This recommendation would apply to both relapsing and progressive MS populations, as well as in mild-moderate disability (PR-EDSS 0–4 and 4.5–6.5). On the other hand, where a more detailed evaluation of fatigue impact covering a longer duration of time (~4 weeks) is needed, i.e., where separate domain scores for physical, cognitive, psychosocial impacts are required, the MFIS may be a better option. Although, the MFIS total showed marginally better information (reliability) at the upper end of fatigue, and similarly, the FSS in the lower end of the fatigue continuum, which was replicated in the results on floor/ceiling effects, we observed that this did not confer advantages in measurement of fatigue in our samples.

**4.1. Strengths and limitations**

In this research, two of the most widely used legacy instruments for measuring fatigue in MS were compared head-to-head. The inclusion of samples representing different contexts of use and settings (i.e., a clinic and a registry population) and different countries (i.e., US and UK), supports the generalizability of our findings. Although our data did not include PwMS older than 65 years or those with a PR-webEDSS > 6.5, or those needing a scooter or wheelchair for mobility; the full range of fatigue, from low to high levels, were observed in our samples. Given the design of the original studies from which this current work is based, our data did not include all three PRO measures in a single sample or any longitudinal data on the MFIS (MFIS was not assessed in the longitudinal UK MS Register sample). As such, we were unable to directly compare the MFIS with the PROMIS Fatigue (MS) 8a in terms of responsiveness. Previous evidence has supported the responsiveness of MFIS scores (Rietberg et al., 2010). Similarly, we were unable to make direct comparisons between the FSS and the MFIS; this was undertaken in the previous studies, which we have cited above.

Recently, researchers have applied modern measurement methods to develop new PRO measures for assessing fatigue in MS, including the

FSIQ-RMS and the NFI-MS (Hudgens et al., 2019; Mills et al., 2010), increasing the options available to researchers. Future research should compare scores of these new measures along with those included in the current study with respect to psychometric performance, practicality, and interpretability. Such research would be an important step towards standardization of fatigue assessment in PwMS. The development of “cross-walks” that associate scores from different PRO measures to a common PRO metric would be beneficial for comparison of results based on different PRO measures. For example, PROMIS fatigue scores are based on a common metric shared across short forms, the full item bank, and the computer-adaptive administration. Moreover, other fatigue PRO measures such as the MFIS or the Functional Assessment of Chronic Illness Therapy-fatigue have been linked to this metric (Lai et al., 2014; Noonan et al., 2012).

**5. Conclusions**

Development of empirically driven recommendations on the most suitable PRO measures for different settings and context of use is a key step in standardization of measurement of important outcomes in MS such as fatigue. Our findings indicate better psychometric performance for the PROMIS Fatigue (MS) 8a relative to the MFIS and the FSS in both a clinic and a registry population. Scores on the PROMIS Fatigue (MS) 8a provided better discrimination among fatigue levels than did those of the FSS or the MFIS physical and total scores. The PROMIS scores also were more responsive to changes in fatigue levels over a 52-week period compared with those of the FSS. Based on our findings, we recommend the PROMIS Fatigue (MS) 8a in situations where brevity is a key consideration (e.g., routine clinical practice), and where the primary interest is in an overall assessment of fatigue severity.

**Funding**

This study was sponsored by Merck Healthcare KGaA, Darmstadt, Germany (CrossRef Funder ID: 10.13039/100009945). The sponsor was involved in the study design, data collection and analysis.

**CRedit authorship contribution statement**

**Paul Kamudoni:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Jeffrey Johns:** Writing – original draft. **Karon F. Cook:** Conceptualization, Methodology, Writing – original draft. **Rana Salem:** Data curation, Formal analysis, Writing –

original draft, Project administration. **Sam Salek:** Formal analysis, Writing – original draft. **Jana Raab:** Writing – original draft, Project administration. **Rod Middleton:** Data curation, Funding acquisition, Writing – original draft, Project administration. **Christian Henke:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Dagmar Amtmann:** Conceptualization, Methodology, Data curation, Funding acquisition, Writing – original draft, Writing – review & editing.

### Declaration of Competing Interest

Paul Kamudoni, Christian Henke and Jana Raab are employees of the Merck Healthcare KGaA, Darmstadt, Germany. Karon Cook has provided consultancy to Merck Healthcare KGaA, Darmstadt, Germany. Sam Salek has a consultancy contract with Merck Healthcare KGaA, Darmstadt, Germany and unrestricted educational grants from GSK and the European Haematology Association. Dagmar Amtmann has received research funding from EMD Serono Research & Development Institute, Inc., an affiliate of Merck KGaA, Darmstadt, Germany. Rana Salem has received research funding from EMD Serono Research & Development Institute, Inc., an affiliate of Merck KGaA, Darmstadt, Germany. Jeffrey Johns and Rod Middleton have nothing to disclose

### Acknowledgments

Contributions from non-authors: Amy Barrett, Bimpe Olayinka-Amao; Pavle Repovic, Kevin N. Alschuler, Gloria von Geldern, and Annette Wundes helped with the data acquisition for the study. Amy Barrett and Bimpe Olayinka-Amao implemented the qualitative research aspects of the study. The authors would like to thank Ankit Turakhiya, PhD, of Bioscript Group Ltd, Macclesfield, UK for providing editorial support, funded by Merck Healthcare KGaA, Darmstadt, Germany.

### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.msard.2022.104048](https://doi.org/10.1016/j.msard.2022.104048).

### References

- Amtmann, D., Bamer, A.M., Noonan, V., Lang, N., Kim, J., Cook, K.F., 2012. Comparison of the psychometric properties of two fatigue scales in multiple sclerosis. *Rehabil. Psychol.* 57 (2), 159–166.
- Andreasen, A.K., Stenager, E., Dalgas, U., 2011. The effect of exercise therapy on fatigue in multiple sclerosis. *Mult. Scler.* 17 (9), 1041–1054.
- Bingham, C.O., Gutierrez, A.K., Butanis, A., et al., 2019. PROMIS fatigue short forms are reliable and valid in adults with rheumatoid arthritis. *J. Patient Rep. Outcomes* 3 (1), 14.
- Cella, D., Riley, W., Stone, A., et al., 2010. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J. Clin. Epidemiol.* 63 (11), 1179–1194.
- Cohen, J., 1987. *Statistical Power Analysis for the Behavioral Sciences*, 1st ed. L. Erlbaum Associates, revised ed.
- Comi, G., Leocani, L., Rossi, P., Colombo, B., 2001. Physiopathology and treatment of fatigue in multiple sclerosis. *J. Neurol.* 248 (3), 174–179.
- Cook, K.F., Bamer, A.M., Roddey, T.S., Kraft, G.H., Kim, J., Amtmann, D., 2012. A PROMIS fatigue short form for use by individuals who have multiple sclerosis. *Qual. Life Res.* 21 (6), 1021–1030.
- Coon, C.D., Cook, K.F., 2018. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual. Life Res.* 27 (1), 33–40.
- Critical path institute, 2009. The patient-reported outcome consortium. <https://c-path.org/programs/proc/>. (Accessed March 2022).

- Elbers, R.G., Rietberg, M.B., van Wegen, E.E., et al., 2012. Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties. *Qual. Life Res.* 21 (6), 925–944.
- Evans, J.P., Smith, A., Gibbons, C., Alonso, J., Valderas, J.M., 2018. The national institutes of health patient-reported outcomes measurement information system (PROMIS): a view from the UK. *Patient Relat. Outcome Meas.* 9, 345–352.
- Fayers, P.M., Machin, D., 2013. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. Wiley.
- Fisk, J.D., Pontefract, A., Ritvo, P.G., Archibald, C.J., Murray, T.J., 2015. The impact of fatigue on patients with multiple sclerosis. *Can. J. Neurol. Sci.* 21 (1), 9–14. *Journal Canadien des Sciences Neurologiques*.
- Ford, D.V., Jones, K.H., Middleton, R.M., et al., 2012. The feasibility of collecting information from people with Multiple Sclerosis for the UK MS Register via a web portal: characterising a cohort of people with MS. *BMC Med. Inform. Decis. Mak.* 12, 73.
- Hays, R.D., Bjorner, J.B., Revicki, D.A., Spritzer, K.L., Cella, D., 2009. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual. Life Res.* 18 (7), 873–880.
- Hudgens, S., Schuler, R., Stokes, J., Eremenco, S., Hunsche, E., Leist, T.P., 2019. Development and validation of the FSIQ-RMS: a new patient-reported questionnaire to assess symptoms and impacts of fatigue in relapsing multiple sclerosis. *Value Health* 22 (4), 453–466.
- Kamudoni, P., Johns, J., Cook, F.C., et al., 2021. Standardizing fatigue measurement in multiple sclerosis: the validity, responsiveness and score interpretation of the PROMIS SF v1.0 – Fatigue (MS) 8a. *Mult. Scler. Relat. Disord. J.* <https://doi.org/10.1016/j.msard.2021.103117>.
- Kister, I., Bacon, T.E., Chamot, E., et al., 2013. Natural history of multiple sclerosis symptoms. *Int. J. MS Care* 15 (3), 146–158.
- Krupp, L.B., LaRocca, N.G., Muir-Nash, J., Steinberg, A.D., 1989. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch. Neurol.* 46 (10), 1121–1123.
- Lai, J.S., Cella, D., Yanez, B., Stone, A., 2014. Linking fatigue measures on a common reporting metric. *J. Pain Symptom Manag.* 48 (4), 639–648.
- Langeskov-Christensen, M., Bisson, E.J., Finlayson, M.L., Dalgas, U., 2017. Potential pathophysiological pathways that can explain the positive effects of exercise on fatigue in multiple sclerosis: a scoping review. *J. Neurol. Sci.* 15 (373), 307–320.
- Learmonth, Y.C., Dlugonski, D., Pilutti, L.A., Sandroff, B.M., Klaren, R., Motl, R.W., 2013. Psychometric properties of the fatigue severity scale and the modified fatigue impact scale. *J. Neurol. Sci.* 331 (1–2), 102–107.
- Leddy, S., Hadavi, S., McCarren, A., Giovannoni, G., Dobson, R., 2013. Validating a novel web-based method to capture disease progression outcomes in multiple sclerosis. *J. Neurol.* 260 (10), 2505–2510.
- Mills, R.J., Young, C.A., Pallant, J.F., Tennant, A., 2010. Development of a patient reported outcome scale for fatigue in multiple sclerosis: the neurological fatigue index (NFI-MS). *Health Qual. Life Outcomes* 8, 22.
- Mills, R.J., Young, C.A., Pallant, J.F., Tennant, A., 2010. Rasch analysis of the modified fatigue impact scale (MFIS) in multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* 81 (9), 1049–1051.
- Muthén, L.K., Muthén, B.O., 2017. *Mplus User's Guide*, 8th ed. Muthén & Muthén, Los Angeles, CA <https://www.statmodel.com/download/usersguide/MplusUserGuideVer8.pdf>. (Accessed December 2020).
- Noonan, V.K., Cook, K.F., Bamer, A.M., Choi, S.W., Kim, J., Amtmann, D., 2012. Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the modified fatigue impact scale and the PROMIS fatigue short form. *Qual. Life Res.* 21 (7), 1123–1133.
- Penner, I.K., Paul, F., 2017. Fatigue as a symptom or comorbidity of neurological diseases. *Nat. Rev. Neurol.* 13 (11), 662–675.
- R Core Team. R: a language and environment for statistical computing, (2018).
- Rietberg, M.B., Van Wegen, E.E., Kwakkel, G., 2010. Measuring fatigue in patients with multiple sclerosis: reproducibility, responsiveness and concurrent validity of three Dutch self-report questionnaires. *Disabil. Rehabil.* 32 (22), 1870–1876.
- Rudroff, T., Kindred, J.H., Ketelhut, N.B., 2016. Fatigue in multiple sclerosis: misconceptions and future research directions. *Front. Neurol.* 7, 122.
- StataCorp LLC. *Stata statistical software: release 15* (2017).
- Terwee, C.B., Bot, S.D., de Boer, M.R., et al., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* 60 (1), 34–42.
- Wang, X.S., Woodruff, J.F., 2015. Cancer-related and treatment-related fatigue. *Gynecol. Oncol.* 136 (3), 446–452.
- Ware, J.E., Gandek, B., 1998. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA project approach. *International quality of life assessment. J. Clin. Epidemiol.* 51 (11), 945–952.
- Yost, K.J., Eton, D.T., Garcia, S.F., Cella, D., 2011. Minimally important differences were estimated for six patient-reported outcomes measurement information system-cancer scales in advanced-stage cancer patients. *J. Clin. Epidemiol.* 64 (5), 507–516.