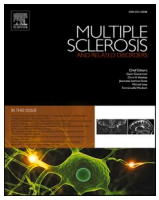
Multiple Sclerosis and Related Disorders 62 (2022) 103753

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

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



Original article

The validity, responsiveness, and score interpretation of the PROMISnq Physical Function – Multiple Sclerosis 15a short form in multiple sclerosis

Paul Kamudoni [a,*], Dagmar Amtmann [b], Jeffrey Johns [c], Karon F. Cook [d], Rana Salem [b], Sam Salek [e], Jana Raab [a], Rod Middleton [f], Pavle Repovic [g], Kevin N. Alschuler [b,h], Gloria von Geldern [h], Annette Wundes [b,h], Amy Barrett [i], Oyebimpe Olayinka-Amao [i], Christian Henke [a]

[a] Global Evidence and Value Development – R&D, Merck Healthcare KGaA, Darmstadt, Germany
[b] Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA
[c] Institute of Medicines Development, Cardiff, United Kingdom
[d] Feral Scholars, Broaddus, Texas, USA
[e] School of Life and Medical Sciences, University of Hertfordshire, Hatfield, United Kingdom
[f] UK MS Register, Swansea University Medical School, Swansea, United Kingdom
[g] Swedish Medical Center, Seattle, WA, USA
[h] Department of Neurology, University of Washington, Seattle, WA, USA
[i] Patient-Centered Outcomes Assessment, RTI Health Solutions, Research Triangle Park, NC, USA

ARTICLE INFO

Keywords:
Multiple sclerosis disability
Physical function
PROMIS

ABSTRACT

Background: A valid, sensitive patient-reported outcome (PRO) measure of physical function (PF) for people with multiple sclerosis (MS) would have substantial value in routine care and clinical research. We now describe development of the PROMISnq Short Form v2.0 PF – Multiple Sclerosis 15a [PROMISnq PF(MS)15a] for assessing PF in relapsing and progressive MS. Also, the validity, reliability, and responsiveness of the PROMISnq PF(MS) 15a is evaluated, minimal important difference (MID) thresholds for score change estimated and a score interpretation guide developed.

Methods: A mixed-methods sequential design was employed. Relevant PF concepts were elicited through semi-structured interviews with people with relapsing MS, and then mapped to the PROMIS PF item bank. Measurement experts integrated results from interviews with people with MS and input from a panel of neurologists to generate a draft short form. Relevance and comprehensiveness of the draft short form were assessed in cognitive debriefing interviews with people with relapsing or progressive MS. Subsequently, item reduction and evaluation of psychometric properties were performed in two observational studies: a cross-sectional study in the US ($n = 296$), and a 96-week longitudinal study in the UK MS Register cohort ($n = 558$). The main outcomes and measures are estimates of: known-groups validity, convergent validity, reliability, responsiveness; MID for worsening.

Results: Factor analyses supported the unidimensionality of the newly derived 15-item short form. Cronbach's alpha (≥ 0.97) and intraclass correlation coefficient (≥ 0.97) of test-retest scores (5–27 days) indicated strong reliability. Convergent validity was demonstrated by moderate-to-strong correlations with scores on related PRO measures. Scores discriminated among patient groups classified by levels of physical health and other criteria. Score changes of 2.3–2.7 points are proposed as MID criteria for minimal worsening in PF.

Conclusion: PROMISnq PF(MS)15a demonstrated reliability, validity and sensitivity to change. Input from patients and clinicians ensured the content is comprehensive and relevant for people with MS.

Abbreviations: AC-API, Assessment Center Application Programming Interface; ANCOVA, analysis of covariance; ANOVA, analysis of variance; CEI, concept elicitation interviews; CDI, cognitive debriefing interviews; EDSS, Expanded Disability Status Scale; ES, effect size; FAMS, Functional Assessment of MS; GHS, Global Health Scale; GPH, Global Physical Health; MID, minimal important difference; MS, multiple sclerosis; MSIS, MS Impact Scale; MSWS, MS Walking Scale; PF, physical function; PGRC, Patient Global Rating of Change; PPMS, primary progressive MS; PRO, patient-reported outcome; PwMS, people living with MS; RRMS, relapsing remitting MS; SD, standard deviation; SPMS, secondary progressive MS; PR-WebEDSS, Patient-Reported Web-based EDSS.

* Corresponding author at: Merck Healthcare KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany.
E-mail address: paul.kamudoni@merckgroup.com (P. Kamudoni).

https://doi.org/10.1016/j.msard.2022.103753
Received 15 November 2021; Received in revised form 16 February 2022; Accepted 18 March 2022
Available online 21 March 2022

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system that results in inflammation, demyelination and neurodegeneration (Filippi et al., 2018). The incidence of MS has been reported at 2.1 cases per 100,000 persons per year, with females accounting for most cases (75%), while the prevalence globally was 35.9 per 100,000 persons in 2020 (Walton et al., 2020). For 85–90% of people living with MS (PwMS), the condition follows a relapsing-remitting course; that is, relapsing-remitting MS (RRMS). People with RRMS may eventually transition to a progressive form of the disease called secondary progressive MS (SPMS). The other 10–15% experience primary progressive MS (PPMS) a progressive form of disease, without remission (Confavreux and Vukusic, 2006; Miller and Leary, 2007; Thompson et al., 1997; Weinshenker et al., 1989).

People with all phenotypes of MS experience a range of debilitating symptoms including cognitive dysfunction, problems with motor control and balance, fatigue, and optic dysfunction, resulting in substantial impairments in health-related quality of life (Rezapour et al., 2017). The assessment of physical function (PF), which relates to the ability to carry out activities that require increasing levels of mobility, strength or endurance (Fries et al., 2006; Haley et al., 1994; Schalet et al., 2016; Stewart and Kamberg, 1992; Wilson and Cleary, 1995), would provide a means for capturing physical limitations associated with MS from a patient's perspective. In PwMS, functional limitations are frequently reported, including impairment to mobility and other aspects of daily life, such as household chores, self-care, and physical activities (Rezapour et al., 2017). Traditionally, the assessment of these limitations in clinical trials are performed using the Expanded Disability Status Scale (EDSS) and the MS Functional Composite, but they do not fully reflect the perspectives of PwMS (European Medicines Agency, 2020). To comprehensively capture MS disability and its impact in clinical practice, the patient's experience and self-reported disease symptoms are useful (European Medicines Agency, 2020; Evans et al., 2018; The Lancet Neurology, 2019).

In our review of existing patient-reported outcome (PRO) measures used in randomized controlled trials and observational research of MS treatments we found no measure specifically developed for assessing the concept of PF in MS. However, we identified at least five multidimensional health-related quality of life measures which included a domain subscale for PF or a related concept (MS quality of life-54 [MSQoL-54], MS impact scale-29 [MSIS-29], MS international quality of life [MUSIQoL], functional assessment of MS [FAMS], and patient reported outcome indices for MS [PRIMUS]), and at least two measures assessing a component/aspect of PF, such as walking ability (i.e., MS walking scale-12 [MSWS-12]) (Hobart et al., 2003) or upper extremity function (arm functions in MS questionnaire [AMSQ]) (Mokkink et al., 2015). Besides the conceptual issues noted above, other concerns for the multidimensional measures identified included: the lack of patient input (e.g., for the MSQoL-54), suboptimal targeting and scaling for PwMS with mild disability (MSIS-29, MSQoL-54) (Cleanthous et al., 2017), and a lack of evidence of responsiveness (MSQoL-54, MUSIQoL, PRIMUS) (Khurana et al., 2017; Moore et al., 2015; Sharafaddinzadeh et al., 2010). The gaps noted highlighted the need for a modern measure assessing PF in PwMS, based on strong patient input.

The emergence of item banks such as the NIH PROMIS or quality of life in neurological disorders (NeuroQoL) has opened up new opportunities for improving the measurement of health domains important across chronic conditions including MS (Cella et al., 2010; Fries et al., 2006). These item banks consist of large pools of well-characterized rigorously developed high-quality questions, calibrated using mathematical models (item response theory). The flexibility and versatility in how item banks are applied and deployed, i.e., dynamically as a computer adaptive test or as a fixed-length short form, scored on a common metric, is a major practical and methodologic advantage. For instance, a fixed-length fatigue short form for use in PwMS, derived from the PROMIS fatigue item bank based on input from PwMS and clinicians, has been previously developed (Cook et al., 2012). In the context of the Food and Drug Administration's clinical outcome assessment drug development tool qualification, the agency agreed to the use of PROMIS item banks as a source of items/questions to derive short forms for assessing PF and fatigue in PwMS (Rose et al., 2014).

Accordingly, based on a mixed-methods research design, we derived a new PF measure that is based on the NIH PROMIS PF item bank and is capable of capturing subtle changes in PF in PwMS (Kamudoni et al., 2020; Kamudoni et al., 2018). The PROMIS PF item bank measures self-reported capability to perform physical activities, and covers functioning related to lower extremities (walking or mobility), upper extremities (dexterity), central regions (neck, back) and instrumental activities of daily living (Rose et al., 2014).

The objectives of this paper include:

- To describe the development of the PROMISnq Short Form v2.0 PF – Multiple Sclerosis 15a [PROMISnq PF(MS)15a], a PROMIS short form for assessing PF in PwMS.
- To evaluate the validity, reliability, and responsiveness of the PROMISnq PF(MS)15a scores.
- To establish minimal important difference (MID) thresholds and develop a score interpretation tool for the PROMISnq PF(MS)15a scores.

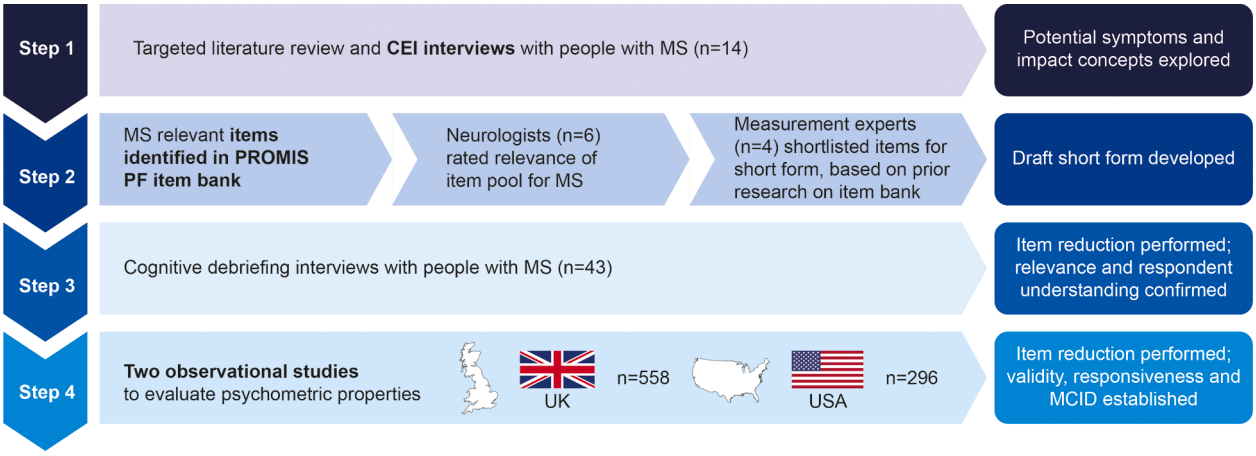


Fig. 1. The development of the PROMISnq PF(MS)15a followed a multi-step mixed methods design. Abbreviations: CEI, Concept Elicitation Interview; MCID, minimal clinically important differences; MS, multiple sclerosis.

2. Methods

2.1. Study design

A mixed-methods design with multiple steps was employed, which included two qualitative studies and two observational studies (Fig. 1). Ethics approvals were obtained for all studies (RTI IRB #14206, Southwest Central Bristol National Research Ethics Service 16/SW/0194, Bristol, UK; Western IRB #20182214, Seattle, WA). All study participants gave informed consent prior to their participation in the studies.

2.2. Development of the new short form

Following a targeted literature review to explore important domains, CEIs were performed with people with relapsing MS, to understand functional limitations related to MS. Subsequently, results from the interviews phase were mapped to the PROMIS PF item bank, and further input obtained from a panel of neurologists and measurement experts including PROMIS investigators, to identify candidate items for the short form that were relevant for PwMS and would optimize item coverage of the PF continuum. Further revision of the candidate item pool and an assessment of relevance, conceptual coverage and respondent understanding were performed in CDIs with people with relapsing and progressive MS.

The PROMISnq PF(MS)15a is scored on a T-score metric, which is standardized for the US general population, to give a mean of 50 and a standard deviation (SD) of 10. Higher scores indicate better PF. Thus, a person who has a T-score of 60 is one SD above the US general population. Scores are calculated based on item response theory (i.e., graded response model), using individuals' item responses. PROMIS T-scores for each participant can be calculated using the assessment center application programming interface, based on response pattern scoring (https://www.healthmeasures.net/score-and-interpret/calculate-scores/scoring-instructions). Alternatively, it can be calculated using a raw-sum-score to T-score crosswalk table. However, this approach may not be as accurate as response pattern scoring using the API.

The new PROMISnq PF(MS)15a is intended for use in research or routine clinical practice in ambulatory PwMS with relapsing or progressive forms of the disease (note: the relevance of different aspects of PF and how these are rated may substantially change with loss of ambulation, e.g., at EDSS score $\geq$7).

2.3. Evaluation of psychometric properties

Psychometric evaluation was performed in two observational studies i.e., a cross-sectional study at two MS centers in the US (US-UW study) and a longitudinal study among members of the UK MS Register in the UK (UK-MSR study). Reliability and construct validity were evaluated in both studies, while responsiveness was evaluated in the UK-MSR study only. MID score thresholds and score interpretation guides were established in the UK-MSR study.

2.4. Participants and procedures

Inclusion criteria for all study participants were:

- Clinician-confirmed MS diagnosis.
- Age 18–65 years.
- Ability to use a computer or tablet.
- Ability to read and write in English, to provide informed study consent and complete study questionnaires.

Exclusion criteria were:

- Self-reported cognitive or physical impairment (e.g., visual impairment) that could interfere with questionnaire or interview completion.
- A patient-reported web-based expanded disability status scale (PR-WebEDSS) $>$ 6.5 (only applied in the observational studies psychometric evalution samples).

Detailed description of the two studies and study procedures are published elsewhere (Kamudoni et al., 2021). In the UK-MSR-study, participants completed assessments at baseline, Weeks 1, 24, 52, 72, and 96 via the register's online portal, from September 2018 to October 2020.

In the US-UW study, prospective data collection took place at the centers from July 2019 to January 2020, and assessments were completed on a tablet computer (Surface PRO device, A4 size or iPad tablet).

2.5. Outcome measures

A draft PROMIS PF(MS) version with 23 items was administered at baseline in the UK-MSR study. Study participants in both studies completed the final PROMISnq PF(MS)15a version (only at Week 52 and later, for the UK-MSR study) (Kamudoni et al., 2020; Kamudoni et al., 2018), PR-WebEDSS [self-reported MS disability] (Leddy et al., 2013), PROMIS v1.2 – Global Health Scale (GHS) [global physical health summary] (Hays et al., 2009), MSIS-29 [physical impact domain] (Hobart et al., 2001). The following assessments were included in one of the two studies only: FAMS [mobility domain] (Cella et al., 1996), MSWS-12 [walking ability] (Hobart et al., 2003), Patient Global Rating of Change (PGRC) – Physical health [change in physical health] (Jaeschke et al., 1989), EuroQoL-5D-3L [self-care, usual activities] (Gusi et al., 2010), Fatigue Severity Scale (FSS) [fatigue severity], (Mills et al., 2009) Modified Fatigue Impact Scale (MFIS) [physical symptom] (Learmonth et al., 2013) (eMETHODS; online supplement). Clinical characteristics were retrospectively extracted from patient records including EDSS, MS phenotype and treatment history.

2.6. Statistical analysis

To assess the psychometric properties of PROMISnq PF(MS)15a, unidimensionality, reliability, validity and responsiveness were examined. Further, MID values were estimated, and a score interpretation tool was developed (i.e., T-score map). 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 UK and the US studies. Unidimensionality was assessed using factor analyses, including a one-factor model confirmatory factor analysis and a bifactor model (eMETHODS; online supplement).

To evaluate ceiling/floor effects, the proportions of the sample with highest/lowest responses across all items were calculated; a proportion of $>$ 0.15 was judged to be problematic (Terwee et al., 2007). Cronbach's coefficient was calculated to assess internal consistency. Intra-class correlation coefficient (mixed effects model for absolute agreement/Model 3 Type 1) (Coons et al., 2009) was calculated between baseline and follow-up scores (5–27 days), to assess test-retest reliability. Reliability of $\geq$ 0.7 is considered adequate for aggregated/group analyses (Reeve et al., 2013).

To assess convergent validity, Spearman's rho was estimated between the PROMISnq PF(MS)15a T-score and scores from related PRO measures; a rho of $>$ 0.4 supports convergent validity (Fayers and Machin, 2013; Prinsen et al., 2018). We anticipated seeing stronger correlations between the PROMISnq PF(MS)15a scores and scores of PROs of closely related concepts such as physical impact (i.e., MSIS-29 physical impact scale), than for relatively more distant concepts such as fatigue (i.e., FSS).

To assess known-groups validity, differences in PROMISnq PF(MS)

Table 1
Hypothesis tested to assess known-groups validity of the PROMISnq PF(MS)15a.

Hypothesis	Patient group
Participants with better health status or physical health report higher PROMISnq PF(MS)15a scores	• GHS global01 excellent/very good/good *versus* fair/poor • GHS global03 physical health excellent/very good/good *versus* fair/poor • GHS GPH summary score of < 50 *versus* ≥ 50
Participants with better mobility (or lower extremity function) report higher PROMISnq PF(MS)15a scores	• MSWS-12 of < 25 *versus* 25–< 50 *versus* ≥ 50 • FAMS mobility of ≤ 15 *versus* 16–22 *versus* > 22 • EQ-5D-3L mobility no problems *versus* some problems
Participants with higher ADL limitations report lower PROMISnq PF(MS)15a scores	• GHS everyday physical activities global question (global06) not at all/a little *versus* moderately/mostly/completely • EQ-5D-3L selfcare no problems *versus* some problems • EQ-5D-3L usual activities no problems *versus* some problems
Participants with lower MS disability report higher PROMISnq PF(MS)15a scores	• EDSS of ≤ 4 *versus* 4.5–6.5 • PR-WebEDSS of ≤ 4 *versus* 4.5–6.5
Participants with progressive disease will report worse PROMISnq PF(MS) 15a scores than those with relapsing disease	• RRMS *versus* SPMS *versus* PPMS

Abbreviations: EDSS, Expanded Disability Status Scale; PR-WebEDSS, Patient-Reported Web-based EDSS; GHS, Global Health Scale; GPH, Global Physical Health; MS, multiple sclerosis; PPMS, primary progressive MS; PwMS, people with MS; RRMS, relapsing remitting MS; SPMS, secondary progressive MS.

15a scores across distinct patient groups were examined based on analysis of variance (ANOVA). Patient groups were defined as shown in Table 1.

Responsiveness was assessed in the UK-MSR study only, based on score changes from Week 52–96. PwMS were classified as improving, unchanged or worsening in their PF using multiple anchors, based on change from Week 52–96, as shown in Table 2. We expected an increase in PROMISnq PF(MS)15a score in the improving group, a score decrease in the worsening group, and a statistically non-significant score change in the unchanged group.

Further, a global question assessing retrospective change in PF at Week 96 was also used as an anchor (a little better/moderately better/very much better [improving]; a little worse/moderately worse/very much worse [worsening]).

Of the 12 variables tested, the ones included in Table 2 fulfilled established criteria for the use of an anchor (Coon and Cook, 2018; Yost et al., 2011). Correlations between anchors and PROMISnq PF(MS)15a score change are reported in eTable 2.

Within-group score change from Week 52–96 in the improving and worsening groups was examined using paired T-tests. The magnitude of score change in each group was calculated based on standard response mean and Cohen's d effect size (ES). ES is interpreted as: small, ES = 0.2; moderate, ES = 0.5; and large ES = 0.8 (Cohen, 1987). Between-group comparisons in score change (i.e., worsening versus unchanged, and unchanged versus improving) were performed using analysis of covariance (ANCOVA) (controlling for baseline score).

Anchor-based approaches were applied to establish MID estimates for the score changes from Week 52–96; these were supported by distribution-based metrics (see online supplement). Patient groups experiencing minimal improvement or minimal worsening were defined based on score change from Week 52–96 based on scores of the measures shown in Table 2.

Table 2
Assessment of responsiveness and minimal important difference based on the score changes from Week 52–96.

Anchor	Responsiveness	Minimal important difference
PGRC – physical health	• A little better/moderately better/better (improving) • A little worse/moderately worse/worse (worsening)	• A little better/moderately better (minimally improving) • A little worse/moderately worse (minimally worsening)
GHS everyday physical activities global question (global06)	• $\geq$ 1-point decrease (worsening) • $\geq$ 1-point increase (improving)	• 1-point decrease (minimally worsening) • 1-point increase (minimally improving)
GHS GPH Summary Score[a]	• $\geq$ 5-point decrease (worsening) • $\geq$ 5-point increase (improving)	• 4.4–9.4 points decrease (minimally worsening) • 4.4–9.4 points increase (minimally improving)
PR-WebEDSS score[b]	• $\geq$ 1-point decrease (improving) • $\geq$ 1-point increase (worsening)	• 1–1.5-point decrease (minimally improving) • 1–1.5-point increase (minimally worsening)
MSIS-29 physical impact score[c]	• $\geq$ 7.5 points decrease (improving) • $\geq$ 7.5 points increase (worsening)	• 7–10 points decrease (minimally improving) • 7–10 points increase (minimally worsening)
MSWS-12 score[d]	• $\geq$ 7.5 points decrease (improving) • $\geq$ 7.5 points increase (worsening)	• 7–10 points decrease (minimally improving) • 7–10 points increase (minimally worsening)

Abbreviations: GHS, Global Health Scale; GPH, Global Physical Health; MS, multiple sclerosis; MSIS, MS Impact Scale; PGRC, Patient Global Rating of Change; PR-WebEDSS, Patient-Reported Web-based Expanded Disability Status Scale.

[a] Current cut-offs are estimated based on Amtmann et al. (2018) calculations are in the online supplement.

[b] Current cut-offs are estimated based on definitions of disability progression from European Medicines Agency (2015) and Rae-Grant et al. (2018).

[c] Current cut-offs are estimated based on Phillips et al. (2014).

[d] Current cut-offs are estimated based on Mehta et al. (2015).

3. Results

3.1. Development of the new short form

Targeted CEIs were undertaken with 14 PwMS and 11 different concepts related to PF issues were reported. These were mapped to 48 items in the PROMIS PF item bank. An expert panel consisting of six neurologists designated 38 of the 48 candidate items as the most relevant. Subsequently, measurement experts, including PROMIS investigators and a patient partner, weighed qualitative evidence from previous stages, item characteristics from previous research on the PROMIS PF item bank, and resolved content overlap, to derive a candidate set of items consisting of 26 items.

In the CDIs, participants ($n = 43$) considered the draft short form (i.e., 26 items) as comprehensive and covering issues important in relation to their PF. Items were deleted (or replaced with items covering similar content from the PROMIS item bank) to address content redundancy, cultural adaption, and comprehension issues; two items from the NeuroQoL mobility bank and a newly drafted item were added to the measure. The relevance, comprehensiveness, and comprehensibility of the 23 items retained in the short form was confirmed in the final round of CDIs.

The instructions were considered clear and easy to understand. Participants considered the item response options as optimal and were able to differentiate among the categories.

Following the CDIs, further revision (i.e., item reduction) of the new short form was carried out based on item-level analyses of the UK-MSR study baseline data. A panel of measurement experts including PROMIS

Table 3
Characteristics of study participants.

Characteristic	UK sample[a]			US sample[a]
	Baseline (n = 558)	Week 52 [b] (n = 440)	Week 96 [b] (n = 390)	Baseline (n = 296)
Age, years				
Mean (SD)	49.89 (9.74)	50.84 (9.27)	50.93 (9.37)	44.50 (11.2)
Median	51.0	52.0	52.0	43.5
Range	19–65	22–65	22–65	21.1–65.6
Gender, n (%)				
Male	133 (23.84)	106 (24.09)	94 (24.10)	75 (25.3)
Female	425 (76.16)	334 (75.91)	296 (75.90)	219 (74.0)
Non-binary	0	0	0	2 (0.7)
Time since				
MS diagnosis, years				
Mean (SD)	10.62 (8.29)	10.85 (8.54)	11.30 (8.67)	9.65 (7.51)
Median	9.0	9.0	9.0	8.22
Range	0–44	0–44	0–44	0.12–37.7
PR- WebEDSS				
Mean (SD)	4.59 (1.89)	4.67 (1.88)	4.58 (1.91)	3.41 (1.7)
Median	5.0	5.0	5.0	3.5
Min-Max	0–6.5	0–6.5	0–6.5	0–6.5
Mild (0–4.0), n (%)	247 (42.73)	185 (42.05)	174 (44.62)	202 (68.2)
Moderate (> 4–6.5), n (%)	331 (52.27)	255 (57.95)	216 (55.38)	94 (31.8)
MS phenotype				
RRMS	374 (67.03)	279 (63.41)	256 (65.42)	280 (94.6)
SPMS	130 (23.30)	116 (26.36)	99 (25.38)	9 (3.0)
PPMS	54 (9.68)	45 (10.23)	35 (8.97)	7 (2.4)

[a] Analysis sample includes respondents with EDSS ≤ 6.5, age ≤ 65 years, and with PPMS, RRMS, or SPMS phenotypes.
[b] This column reports the baseline characteristics of patients who had a follow-up assessment at respective follow-up period. Abbreviations: EDSS, Expanded Disability Status Scale; MS, multiple sclerosis; PPMS, primary progressive MS; RRMS, relapsing remitting MS; SD, standard deviation; SPMS, secondary progressive MS; PR-WebEDSS, Patient-reported Web-based EDSS.

investigators weighed all the evidence (from prior steps of the project) to ensure that the items retained contributed to the measurement of the full continuum of PF, without substantial item overlap, while maintaining content validity. A total of eight items were deleted, reducing the number of items from 23 to 15. A separate manuscript detailing the qualitative development as well as item reduction of the measure is under development.

3.2. Evaluation of psychometric properties

The psychometric properties of the PROMISnq PF(MS)15a were analyzed in a total of 558 PwMS in the UK study and 296 in the US study (eFig. 1), the majority with RRMS (67.0%, UK; 94.6%, US). The mean age of these participants was 49.9 (standard deviation [SD] = 9.8) in the UK and 44.5 (SD = 11.2) in the US (Table 3).

3.3. Unidimensionality

A one-factor confirmatory factor analysis model had good fit, with all goodness of fit statistics within recommended ranges in both samples; only results from the UK MS are reported in this paper (Table 4). Standardized loadings of 0.82–0.96 were obtained. We tested a bifactor model hypothesizing an overarching PF factor and two sub factors: firstly lower-extremity, and secondly, PF unrelated to mobility (combining activities of daily living and upper extremity activities). The bifactor model results showed a strong dominant general factor (OMEGA H = 0.99, ECV 0.95) (eTable 3). These findings support essential unidimensionality of the PROMISnq PF(MS)15a responses, suggesting that use of a single overall score to characterize and measure PF in MS using this measure is appropriate.

3.4. Score distribution and reliability

The mean PROMISnq PF(MS)15a T-score was 39 (SD = 10.6) among UK participants and 45.9 (SD = 10.1) in the US sample. The measure

Table 4
One-factor confirmatory factor analysis of the PROMISnq PF(MS)15a: standardized factor loadings and goodness of fit statistics

Item	Est. Std.	SE
Are you able to stand without losing your balance for several minutes?	0.847	0.014
How much difficulty do you currently have standing up from a low, soft couch?	0.877	0.011
Are you able to hold a plate full of food?	0.818	0.017
Are you able to dress yourself, including tying shoelaces and buttoning?	0.826	0.016
Are you able to walk with a heavy backpack (about 10 lbs/5 kgs) for 20 min?	0.958	0.006
Are you able to carry a laundry basket up a flight of stairs?	0.947	0.006
Are you able to run errands and shop?	0.901	0.010
Are you able to get up from the floor from lying on your back without help?	0.914	0.008
Are you able to push open a heavy door?	0.840	0.014
Are you able to exercise hard for half an hour?	0.876	0.013
Does your health now limit you in hiking a couple of miles (3 km) on uneven surfaces, including hills?	0.983	0.005
Does your health now limit you in doing moderate work around the house like vacuuming, sweeping floors, or carrying in groceries?	0.92	0.008
Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?	0.958	0.007
Does your health now limit you in climbing several flights of stairs?	0.946	0.006
How much difficulty do you have walking on uneven surfaces (e.g., grass, dirt road or sidewalk)?	0.919	0.008

Goodness of Fit Statistics χ^2 statistic: 404.45, df: 90; $\chi 2$ p-value: < 0.001. RMSEA, Estimate (90% CI): 0.081 (0.073, 0.089); CFI: 0.996; TLI: 0.996; SRMR: 0.023.
Abbreviations: χ^2, Chi square; CI, confidence interval; CFI, comparative fit index; RMSEA, root mean square error of approximation; SE, standard error; SRMR, standardized root mean square; TLI, Tucker-Lewis Index.

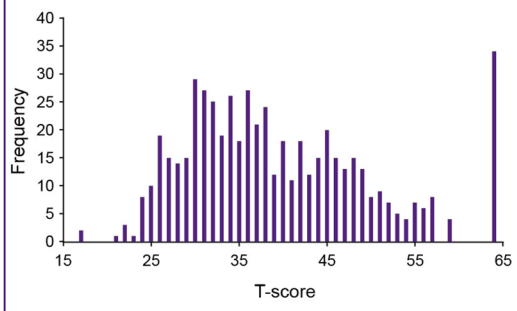
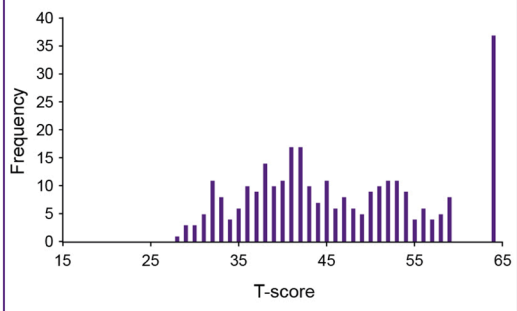
PROMIS PF T-scores	UK sample (n=558)	US sample (n=296)
Mean (SD)	39.0 (10.6)	45.9 (10.1)
Median (min,max)	36.8 (16.0, 63.6)	44.4 (27.9, 63.6)
Ceiling*, %	6.09	12.5
Floor*, %	0.2	0.3
T-scores distribution at baseline		
Cronbach's alpha	0.97	0.96

Fig. 2. Distribution of PROMISnq PF(MS)15a scores and reliability in the two study samples. The histograms depict a broadly normal distribution of the PROMISnq PF(MS)15a scores in the two samples. The spike at the far right in both figures represents participants reporting no limitations on any item/question.
* The proportions of the sample with the highest/lowest response across all items. *Abbreviations:* PF, physical function; SD, standard deviation.

shows optimal scaling, with ceiling and floor effects (i.e., the proportions of the sample with highest/lowest responses across all items) below the critical threshold of 15% (Fig. 2).

The test-retest reliability of the PROMISnq PF(MS)15a T-scores at 5–27 days follow-up, among patients with unchanged GHS physical health (global 03) score, was 0.97 (95% CI, 0.96, 0.98).

3.5. Construct validity

Results for known-groups validity analyses are presented in Table 5. Statistically significant differences were observed in PROMISnq PF(MS) 15a T-scores as hypothesized (ANCOVA Test, $p < 0.01$ for all tests).

The PROMISnq PF(MS)15a showed moderate to strong correlations with scores of related PRO measures (Table 6), supporting the convergent validity of the short form scores in both the US and UK samples.

3.6. Responsiveness

Responsiveness was evaluated in the UK sample only, based on analysis of score changes from Week 52 to Week 96 (Table 7). Broadly, our results supported responsiveness although results varied across anchors and metrics. One anchor, i.e., the PGRC – physical health, showed a significant within-group change in the worsening group only, while four others i.e., the GHS everyday physical activities score, the GHS GPH summary score, the MSIS-29 Physical Impact score, and the MSWS-12 score, showed significant changes for both worsening and improving groups. Among the four anchors with significant changes, ES was ≥ 0.2 (mild effect) in the worsening group only.

Expected differences were observed in comparisons of PROMISnq PF (MS) 15a score change between the worsening *versus* the unchanged groups, as well as the unchanged *versus* the improving groups, for four anchors (i.e., GHS everyday physical activities, GHS GPH summary score, the MSIS-29 Physical Impact, and the MSWS-12) (ANCOVA Test, $p < 0.01$). Results on PR-WebEDSS are presented for descriptive purposes only, but not for evaluating responsiveness due to the small number of observations.

3.7. Minimal important difference thresholds and score interpretation tool

MID estimates for the PROMISnq PF(MS)15a score were calculated based on anchor-based analyses of score changes from Week 52 to Week 96, in the UK-MRS sample only (Table 8). Estimates based on two anchors – GHS everyday physical activities (global06) and the GHS GPH summary score – met all criteria for MID estimation (including ES of 0.2–0.8). The mean change in PROMISnq PF(MS)15a T-score was 2.29 (ES = 0.23) for a 1-point decrease on GHS everyday physical activities (global06), and 2.66 for 4.4–9.4 points decrease on the GHS GPH Summary Score.

In addition, we estimated distribution-based meaningful change metrics in the UK-MSR sample at Week 52 and Week 96, including half-standard deviation and standard error of measurement (Table 8) (Norman et al., 2004). This provided further insights on the MID estimates, in particular for understanding the smallest changes the PROMISnq PF (MS)15a is capable of detecting.

Triangulating the various estimates, a MID estimate of 2.3–2.7 is proposed for worsening PF for the PROMISnq PF(MS)15a.

An abbreviated T-score interpretation guide for the PROMISnq PF (MS)15a is presented in Fig. 3. The full version is available in the online supplement (eFig. 2). The T-score map is displayed as a heatmap that associates PROMISnq PF(MS)15a T-scores with model-predicted responses on the individual items of the short form, and shows the most likely response (i.e., level of limitation) for each T-score (Rothrock et al., 2020). The map was generated using outputs from item response theory analysis (i.e. item characteristic curves), the underlying mathematical modeling approach underpinning the PROMISnq PF(MS)15a's scoring. For any T-score reported, the level of impact on the individual questions can be deduced.

4. Discussion

This paper describes the development of a new PRO measure for assessing PF in relapsing and progressive forms of MS, the PROMIS PF (MS) short form, derived from PROMIS PF and the NeuroQoL item banks. Evidence from mixed-methods research presented in this paper supports the reliability, validity and responsiveness of the new short form. Confidence in the generalizability of the findings is supported by

Table 5
Known-groups validity of the PROMISnq PF(MS)15a: score differences across clinically relevant subgroups.

	UK-MSR Sample, Week 52			US sample (baseline)		
	N	Mean (SE)	F-statistic (*p*-value)	N	Mean (SE)	F-statistic (*p*-value)
(a) Scores for both groups, or US group only						
EDSS score						57.34 ($p < 0.001$)
Mild (0–4.0)	–	–		221	48.04 (0.66)	
Severe (4.5–6.5)	–	–		75	39.58 (0.63)	
PR-WebEDSS score			378.52 ($p < 0.001$)			175.43 ($p < 0.001$)
Mild (0–4.0)	136	46.63 (0.69)		202	50.08 (0.61)	
Severe (4.5–6.5)	165	31.57 (0.40)		94	36.89 (0.64)	
GHS health question (global01)			134.57 ($p < 0.001$)			50.83 ($p < 0.001$)
Fair/poor (1,2)	188	31.86 (0.49)		59	38.16 (0.85)	
Excellent/very good/good (3,4,5)	244	42.21 (0.69)		237	47.82 (0.64)	
GHS physical health question (global03)			213.53 ($p < 0.001$)			89.32 ($p < 0.001$)
Fair/poor (1,2)	226	31.93 (0.44)		96	39.04 (0.71)	
Excellent/very good/good (3,4,5)	206	44.05 (0.72)		200	49.19 (0.68)	
GHS everyday physical activities (global06)			432.31 ($p < 0.001$)			66.1 ($p < 0.001$)
A little/not at all (1,2)	177	28.78 (0.33)		37	34.50 (0.68)	
Moderately/mostly/completely (3,4,5)	255	43.91 (0.56)		259	47.52 (0.60)	
GHS fatigue question (global08)			101.62 ($p < 0.001$)			67.34 ($p < 0.001$)
Severe/very severe (1,2)	112	29.94 (0.56)		74	38.38 (0.73)	
None/mild/moderate (3,4,5)	328	40.43 (0.58)		222	48.40 (0.66)	
GHS GPH summary score			384.11 ($p < 0.001$)			344.65 ($p < 0.001$)
< 50	360	34.57 (0.42)		195	40.75 (0.48)	
≥ 50	72	53.43 (0.83)		101	55.84 (0.77)	
FAMS total score (mobility)						209.05 ($p < 0.001$)
≤ 15	–	–		89	35.89 (0.50)	
16 through 22	–	–		77	43.32 (0.71)	
> 22	–	–		130	54.28 (0.62)	
MS phenotype			80.86 (p <0.001)			3.19 ($p = 0.0426$)
RRMS (1)	279	41.86 (0.62)		280	46.22 (0.59)	
PPMS (2)	45	31.47 (0.93)		7	42.76 (5.18)	
SPMS (3)	116	30.21 (0.55)		9	38.16 (3.11)	
(b) Scores for UK-MSR sample only						
MSWS-12 score			343.57 ($p < 0.001$)			
< 25	90	51.79 (0.72)				
25 to < 50	74	39.90 (0.53)				
≥ 50	101	31.32 (0.45)				
EQ5D selfcare			162.20 ($p < 0.001$)			
none (0)	208	44.11 (0.64)				
slight (1)	167	29.77 (0.39)				
moderate (2)	1	27.57				
EQ5D usual activities			225.98 ($p < 0.001$)			
none (0)	82	51.68 (0.83)				
slight (1)	265	34.67 (0.44)				
moderate (2)	29	25.80 (0.98)				
FSS score			110.31 ($p < 0.001$)			
< 36	76	47.43 (1.23)				
≥ 36	286	34.92 (0.52)				

The number of participants completing each PRO measure varied in the UK MS Register, reflecting the register's design, which allows participants to choose which assessments to complete at each time point.
Abbreviations: EDSS, Expanded Disability Status Scale; PR-WebEDSS, Patient-Reported Web-based EDSS, EQ-5D-3L, Euro Quality of Life; FAMS, Functional Assessment of Multiple Sclerosis; FSS, Fatigue Severity Scale; GHS, Global Health Scale; GPH, Global Physical Health; MS, multiple sclerosis; MSWS, MS Walking Scale; PPMS, primary progressive MS; RRMS, relapsing remitting MS; SPMS, secondary progressive MS.

the breadth of the methods and in the consistency of the results obtained from samples collected in two countries and in different contexts – e.g., people seen in an MS tertiary clinic, living in the US, as well as those enrolled in a registry, living in the UK.

Extensive input was obtained from people with relapsing MS regarding their functional limitations, including the impact on daily life activities, to inform selection of questions/items most relevant to PwMS from the item banks. In a subsequent step, people with relapsing or progressive MS judged the clarity, relevance and appropriateness of items/content considered for inclusion (or included) in the new short form. Insights from item response theory analyses on individual items and from the growing body of published empirical evidence regarding the function of PROMIS PF items in different clinical populations further informed selection of the items.

The scoring algorithm proposed for the PROMISnq PF (MS) 15a uses a single overall score (i.e., a T-score) calculated based on responses from all 15 items. This is underpinned by evidence of essential unidimensionality from the factor and bi-factor analyses, which indicated that despite the existence of different functional aspects (e.g., lower extremity, activities of daily living, and upper extremity) characterizing PF, variance in the PROMIS PF (MS)15a scores is driven by a common underlying concept. Our data suggested limited utility of subdomain scores.

Our findings regarding the relationships between the new short form and MS disability (i.e., based on the PR-Web EDSS and the EDSS), and score differences observed across disability groups were as expected. This provides foundational evidence supporting the validity of the new short form and hints at its potential role as a measure of a patient's perceived disability associated with MS.

PROMISnq PF(MS)15a scores were sensitive to worsening and improvements in PF, over a 44-week follow-up duration, as defined on five different PROs as anchors. Evidence was stronger in relation to PF worsening than for improvement, e.g., ES were smaller for improvement (i.e., Cohen's $d \leq 0.2$, and standard response mean ≤ 0.3). While we

Table 6
Convergent validity of the PROMISnq PF(MS)15a.

	UK-MSR Sample, Week 52		US-UW sample	
	n	Spearman's coefficient	n	Spearman's coefficient
EDSS	–	–	258	–0.63
PR-WebEDSS	318	–0.87	296	–0.75
EQ-5D-3L mobility domain	376	–0.72	–	–
EQ-5D-3L selfcare domain	376	–0.72	–	–
EQ-5D-3L usual activity domain	376	–0.70	–	–
GHS physical health question (global03)	432	0.64	294	0.63
GHS health question (global01)	432	0.57	296	0.61
GHS fatigue question (global08)	432	0.60	296	0.64
GHS everyday physical activities (global06)	432	0.90	296	0.81
GHS GPH summary score	432	0.82	294	0.84
MSIS-29 physical impact score	362	–0.88	296	–0.88
MSWS-12 score	265	–0.93	–	–
FAMS mobility (total score)	–	–	295	0.86
FSS score	362	–0.58	–	–
MFIS (physical score)	–	–	296	–0.83

Abbreviations: EDSS, Expanded Disability Status Scale; EQ-5D-3L, Euro Quality of Life; FAMS, Functional Assessment of MS; FSS, Fatigue Severity Scale; GHS, Global Health Scale; GPH, Global Physical Health; MFIS, Modified Fatigue Impact Scale; MS, multiple sclerosis; MSIS, MS Impact Scale; MSWS, MS Walking Scale.

think more data should be generated to confirm the current findings, particularly in relation to improvements in PF, we also believe the presented results support the majority of clinical and research applications of the measure, which reflect the degenerative nature of MS and are more focused on slowing progression (or worsening).

Clinicians and researchers using the PROMISnq PF(MS)15a to evaluate PF changes over time or to gain a snapshot at a single time point for an individual patient can use the MID estimates provided and the T-score maps to meaningfully interpret the scores of the new short form in a clinically meaningful way.

1 We propose a score change of 2.3–2.7 points, based on our MID analyses, as the threshold for worsening PF on the PROMISnq PF(MS) 15a that is clinically meaningful. This estimate is based on score changes over a 44-week duration, and meets all requisite criteria for MID estimation, including triangulation of findings from multiple anchors.
2 Further, we have developed a T-score map, to provide qualitative descriptors illustrating the typical limitations a patient with any given T-score might be experiencing, based on the 15 items of the PROMIS PF(MS)15a and their related response options.
3 We believe these tools will facilitate the application and integration of the PROMIS PF(MS)15a scores in clinical decision-making.

Consensus has emerged on the need for, and the important role of, a core outcomes set that includes self-reported symptoms and functional status in MS (National Quality Forum, 2021). Such a core set would support a more comprehensive assessment of MS disability in clinical research, routine clinical practice, and/or other settings (European Medicines Agency, 2020; National Institute for Health and Care Excellence, 2019; Nowinski et al., 2017). PF, which encompasses mobility, activities of daily living, and upper extremity function, addresses functional limitations that are proximal to MS and amenable to therapeutic interventions, and which are rated among the most important concerns for PwMS (Larocca, 2011; Martin et al., 2017). Assessing PF as a PRO, as opposed to performance measures or wearable sensors, may have several advantages, including the incorporation of the patient's perception of limitations that are relevant and meaningful to the everyday life context (National Institute for Health and Care Excellence, 2019).

Development of the PROMISnq PF(MS)15a is an important step towards the standardization of outcomes measurement, and emergence of a core outcomes set in MS. Three characteristics of the measure stand out. Firstly, the PROMISnq PF(MS)15a items provide broad conceptual coverage of the key aspects of PF; secondly, the measure targets the full continuum from mild, moderate through to severe levels of impairment of PF; and thirdly, using the PROMIS PF metric ("ruler") for scoring, the new short form's scores can be directly compared with scores based on the full item bank, computer adaptive test scores and scores on at least six other PROMIS PF short forms. This last characteristic will be useful for making comparisons and generalizations across studies employing different measures, as long as they have a mapping to the PROMIS metric (Nowinski et al., 2017).

There are numerous challenges impeding the integration of routine PRO assessments in the neurology clinic and MS care. One such challenge has been the lack of appropriate PRO measures for this context. For PF, the PROMISnq PF(MS)15a is posed to overcome this challenge. It was developed with substantial input from PwMS, it targets the impact of MS on their daily life and function, and, as evidenced in this study, it has robust psychometric properties. Another challenge for the integration of PRO assessment into MS care is a lack of understanding of what PRO scores tell us about the patient's overall disease today and in the future. This is true for the assessment of PF and the many other outcomes that are important for patients and are informative in the management of MS. Though the science of PRO assessment has advanced substantially, the field still lacks understanding of the ways in which PRO assessments complement, supplement, or replace traditional and other novel assessments in MS. These questions may only be fully addressed with more experience from early adopters of PRO measures and research that ascertains the predictive ability of PRO scores and their association with important clinical markers.

5. Limitations and future directions

The current study had some limitations. Our initial qualitative study which informed the selection of content for the short form, included people with relapsing MS only - this may raise concerns about the short form's relevance and appropriateness across all MS phenotypes. This concern was addressed in the subsequent qualitative study which employed cognitive debriefing interviews to assess the relevance and appropriateness of the new short form in people with relapsing and progressive MS.

PwMS older than 65 years and those with an EDSS of > 6.5 were excluded, which limits the generalizability of findings. With respect to age, there are reasons to expect that the current results would be generalizable to those above 65 years. The original development and

Table 7
Responsiveness of PROMISnq PF(MS)15a in UK-MSR Sample: analysis of score changes from Week 52–96 by changes in functional status.

Anchor	PROMISnq PF(MS)15a T-score change (Weeks 52–96)		
	Worsening	Unchanged	Improving
PGRC – physical health			
n	210	99	42
Mean change (SD)	1.30 (3.38)	–0.48 (3.13)	–0.68 (4.48)
T-test statistic; *p*-value	5.58; < 0.001	–1.53; 0.130	–0.98; 0.334
ES (est, 95% CI)	0.15 (–0.04; 0.34)	–0.05 (–0.33; 0.22)	–0.09 (–0.52; 0.33)
SRM	0.39	–0.15	–0.15
ANCOVA, F-statistic; *p*-value unchanged versus worsening/improving	14.03; < 0.001		0.09; 0.767
PR-WebEDSS score[a]			
n	21	64	7
Mean change (SD)	1.30 (2.86)	0.39 (3.09)	–0.86 (3.17)
ES (est, 95% CI)	0.13 (–0.48; 0.73)	0.02 (–0.32; 0.36)	–0.12 (–1.14; 0.90)
SRM	0.46	0.13	–0.27
GHS everyday physical activities (global06)			
n	67	217	56
Mean change (SD)	2.55 (4.05)	0.60 (3.02)	–1.66 (3.70)
T-test statistic; *p*-value	5.16; < 0.001	2.93; 0.004	–3.37; 0.001
ES (est, 95% CI)	0.27 (–0.07; 0.61)	0.05 (–0.14; 0.24)	–0.20 (–0.57; 0.17)
SRM	0.63	0.20	–0.45
ANCOVA, F-statistic; *p*-value unchanged versus worsening/improving	10.66; < 0.001		11.60; < 0.001
GHS GPH Summary score			
n	49	246	45
Mean change (SD)	3.35 (4.82)	0.53 (2.73)	–1.90 (4.14)
T-test statistic; p-value	4.86; < 0.001	3.02; 0.003	–3.07; 0.004
ES (est, 95% CI)	0.29 (–0.11, 0.69)	0.05 (–0.13, 0.23)	–0.17 (–0.59, 0.24)
SRM	0.70	0.19	–0.46
ANCOVA, F-statistic; p-value unchanged versus worsening/improving	16.89; < 0.001		12.57; < 0.001
MSIS-29 physical impact score			
n	61	177	58
Mean change (SD)	1.92 (3.13)	0.31 (3.10)	–1.05 (3.77)
T-test statistic; p-value	4.80; < 0.001	1.34; 0.183	–2.12; 0.038
ES (est, 95% CI)	0.17 (–0.18, 0.52)	0.03 (–0.18, 0.23)	–0.05 (–0.41, 0.30)
SRM	0.61	0.10	–0.28
ANCOVA, F-statistic; p-value unchanged versus worsening/improving	7.70; < 0.001		3.81; 0.024
MSWS-12 score			
n	46	104	42
Mean change (SD)	1.86 (3.55)	0.52 (3.45)	–1.20 (2.82)
T-test statistic; p-value	3.55; < 0.001	1.54; 0.127	–2.75; 0.009
ES (est, 95% CI)	0.28 (–0.13, 0.68)	0.03 (–0.24, 0.30)	–0.15 (–0.60, 0.39)
SRM	0.52	0.15	–0.43
ANCOVA, F-statistic; p-value unchanged versus worsening/improving	4.89; 0.009		4.21; 0.017

Higher PROMISnq PF(MS)15a T-scores indicate better PF. Change scores were calculated as Week 52 – Week 96, i.e., a negative number is consistent with an improvement in PF.

Change scores are W52-W96; ANCOVA adjusted for W52 PROMISnq PF(MS)15a T-score.

[a] For PR-WebEDSS change, within- and between-group analysis of score change was not performed due to the small number of observations in the worsening and the improving groups.

For each anchor measure, three groups were defined (worsening, unchanged, and improving) using the change criteria shown in Table 2.

Abbreviations: ANCOVA, analysis of covariance; CI, confidence interval; ES, effect size; EQ-5D-3 L, Euro Quality of Life 5 Dimension; GHS, Global Health Scale; GPH, Global Physical Health; PF, physical function; PGRC, Patient Global Rating of Change; MSIS, MS Impact Scale; MSWS, MS Walking Scale; SD, standard deviation; SRM, standard response mean; PR-WebEDSS, Patient-Reported Web-based Expanded Disability Status Scale.Minimal important difference thresholds and score interpretation tool.

calibration of the PROMIS PF item bank was conducted in the general population and included adults of all age groups (Rose et al., 2014). Subsequent validation across varied clinical populations suggests the measure's strong psychometric performance might extend to individuals older than 65 years (Schalet et al., 2016). However, future research should directly evaluate this.

Because the content of the PROMISnq PF(MS)15a was derived specifically for use in ambulatory PwMS, the findings do not generalize to PwMS who require use of a scooter or wheelchair for mobility.

Except for the PROMISnq PF(MS) and a few other outcomes, all other assessments in the UK MS Register sample are part of the routine data collected in the register. As such, these assessments were not completed by all study participants at all time points, resulting in different numbers of observations available for various anchors.

The challenges associated with investigating responsiveness and MID in non-interventional cohort study designs in MS have been documented (Hobart et al., 2005; Miller et al., 2016). One of the most significant challenges is the small magnitude of change expected over time in PwMS; e.g., in the current context, changes in the improving groups were negligible, although statistically significant. With this in mind, the use of multiple anchors to define change groups served to increase confidence in the analysis of responsiveness. Future studies of the PROMISnq PF(MS)15a should include evaluations of responsiveness based on a treatment of known efficacy. The PGRC physical health anchor used the full study follow-up duration as a frame of reference i.e., 96 weeks, in contrast to the Week 52–96 score changes on the PROMISnq PF(MS)15a score we have focused on in our responsiveness and MID analyses. Excluding results pertaining to this anchor does not change our broad conclusions.

Table 8
Interpretation of individual-level PROMISnq PF(MS)15a score change; minimal important difference from Week 52–96.

Anchor	PROMISnq PF(MS)15a T-score change (Weeks 52–96)		
	Minimally worsening	Unchanged	Minimally improving
PGRC - physical health			
n	192	99	38
Mean change (SD)	1.03 (3.21)	–0.48 (3.13)	–0.47 (4.58)
Median change	0.96	–0.22	–0.34
95% CI score change	0.46; 1.48	–1.11; 0.14	–1.97; 1.03
ES (est, 95% CI)	0.12 (–0.08, 0.32)	–0.05 (–0.33, 0.22)	–0.07 (–0.52, 0.38)
PR-WebEDSS score			
n	8	83	1
Mean change (SD)	2.56 (3.04)	0.31 (3.02)	–3.0
Median change	3.27	0.19	–3.0
95% CI score change	0.02, 5.01	–0.35, 0.97	–
ES (est, 95% CI)	0.30 (–0.69, 1.28)	0.02 (–0.28, 0.32)	–
GHS everyday physical activities (global06) score			
n	58	217	47
Mean change (SD)	2.29 (3.42)	0.60 (3.02)	–1.27 (3.66)
Median change	2.33	0.56	–0.92
95% CI score change	1.04; 3.19	0.20; 1.00	–2.35; –0.20
ES (est, 95% CI)	0.24 (–0.13, 0.60)	0.05 (–0.14, 0.24)	–0.15 (–0.55, 0.26)
GHS GPH summary score			
n	49	229	40
Mean change (SD)	2.66 (3.19)	0.46 (2.69)	–1.07 (3.40)
Median change	2.42	0.49	–1.08
95% CI score change	1.74; 3.57	0.11; 0.82	–2.15; 0.02
ES (est, 95% CI)	0.22 (–0.18, 0.61)	0.05 (–0.14, 0.23)	0.11 (–0.54, 0.33)
MSIS-29 physical impact score			
n	33	156	31
Mean change (SD)	0.65 (2.15)	0.46 (3.13)	–0.52 (3.54)
Median change	0.43	0.27	–1.30
95% CI score change	–0.11; 1.41	–0.04; 0.95	–1.82; 0.78
ES (est, 95% CI)	0.05 (–0.43, 0.53)	0.03 (–0.18, 0.25)	–0.01 (–0.50, 0.47)
MSWS-12 score			
n	15	99	15
Mean change (SD)	1.35 (2.22)	0.345 (3.48)	–0.06 (2.76)
Median change	1.44	0.01	–0.03
95% CI score change	0.12; 2.58	–0.25; 1.14	–1.58; 0.71
ES (est, 95% CI)	0.23 (–0.48, 0.94)	0.04 (–0.24, 0.31)	–0.17 (–0.86, 0.53)
PROMISnq PF(MS)15a T-score, summary metric	**Week 52**	**Week 96**	
1/3 SD	3.53	3.64	
1/2 SD	5.3	5.5	
IRT SEM	1.73	1.77	
"Traditional" SEM	1.84	1.89	

For each anchor measure, three groups were defined (worsening, unchanged, and improving) using the change criteria shown in Table 2.
Abbreviations: CI, confidence interval; ES, effect size; GHS, Global Health Scale; GPH, Global Physical Health; IRT, item response theory; MS, multiple sclerosis; MSIS, MS Impact Scale; MSWS, MS Walking Scale; PGRC, Patient Global Rating of Change; SD, standard deviation; SEM, standard error of the measurement; WebEDSS, web-based Expanded Disability Status Scale.

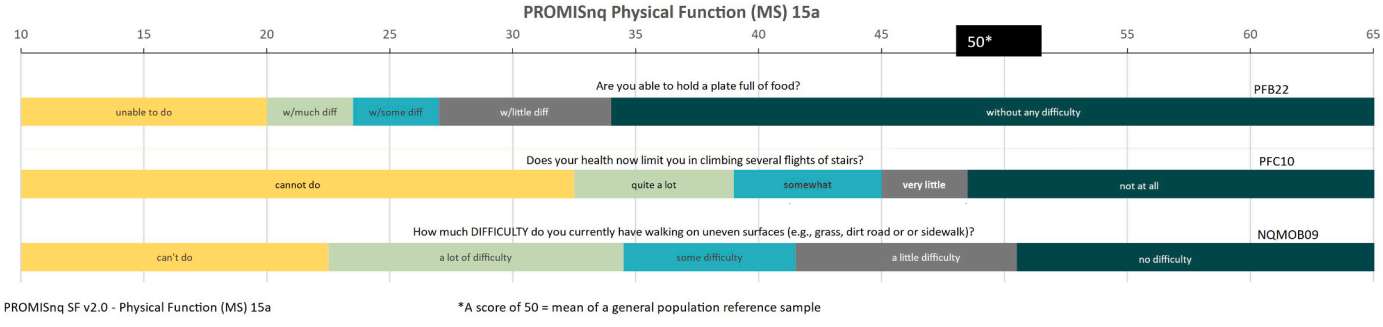


Fig. 3. Score interpretation guide for the PROMISnq PF(MS)15a: score categories based on various functional limitations (abbreviated T-Score Map; full version is available in online supplement). *A score of 50 = mean of a general population reference sample.
The T-map categorizes the PROMISnq PF(MS)15a T-scores in terms of expected responses on each item (i.e., reflecting the five levels of difficulty that constitute the responses for each item).

6. Conclusions

Improvement in the measurement of patient-reported function, particularly the ability to perform various physical activities and activities of daily life, is an important step in advancing our understanding and management of disability related to MS. In this mixed-methods research, we have developed a new PROMIS short form for assessing PF in PwMS, with direct input from PwMS and clinical neurology experts, thus ensuring the comprehensiveness and relevance of the measure in MS populations. Psychometric findings across two observational studies in US and UK populations supported the measure's reliability, validity, and responsiveness. Further, a score interpretation guide was developed, to facilitate the interpretation of PROMISnq PF(MS)15a scores. This tool may also be useful in informing communications between PwMS and healthcare professionals. To identify meaningful individual changes in PF over time, MID estimates for the PROMISnq PF (MS)15a scores are provided.

Author contributions

Conceptualization and methodology: Paul Kamudoni; Christian Henke; Karon F. Cook; Dagmar Amtmann; Amy Barrett, Oyebimpe Olayinka-Amao ; Data acquisition/collection: Amy Barrett, Oyebimpe Olayinka-Amao ; Rod Middleton; Dagmar Amtmann; Rana Salem; Pavle Repovic, Kevin N. Alschuler, Gloria von Geldern, Annette Wundes; Data curation and formal analysis: All authors; Writing - original draft: Paul Kamudoni; Christian Henke; Karon F. Cook; Dagmar Amtmann; Review & editing: All authors; Statistical analysis: Sam Salek; Jeffrey Johns; Karon F. Cook; Rana Salem; Dagmar Amtmann; Qualitative data-analysis: Amy Barrett, Oyebimpe Olayinka-Amao ; Project administration: Jana Raab; Rana Salem; Rod Middleton. All authors had access to the data.

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.

Declaration of Competing Interest

Paul Kamudoni, Christian Henke and Jana Raab are employees of 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. Pavle Repovic has acted as a consultant or speaker for Alexion, Biogen, Celgene, EMD Serono Research & Development Institute, Inc., an affiliate of Merck KGaA, Darmstadt, Germany, Medison, Novartis, Roche, Sanofi Genzyme, and Viela Bio. Annette Wundes has received research funding from Alkermes, Biogen, AbbVie and provided consultancy for AbbVie. 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. Amy Barrett and Oyebimpe Olayinka-Amao are employees of RTI Health Solutions, which received research funding from Merck Healthcare KGaA, Darmstadt, Germany. Jeffrey Johns, Kevin N. Alschuler, Gloria von Geldern and Rod Middleton have nothing to disclose.

Acknowledgments

Robert McBurnley and the members of the Accelerated Cure Project, Merck Healthcare Multiple Sclerosis Patient-Focused Drug Development Patient Council, provided critical input during early development of the short form. Members of the EVOLUTION RMS Phase III study Steering Committee provided valuable clinical insights during the early stages of the research. Ari Gnanasakthy, an employee of RTI Health Solutions, provided important strategic insights throughout the project. Medical editing support was provided by Bioscript Stirling Ltd, 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.2022.103753.

References

Amtmann, D, Bamer, AM, Kim, J, Chung, H, Salem, R, 2018. People with multiple sclerosis report significantly worse symptoms and health related quality of life than the US general population as measured by PROMIS and NeuroQoL outcome measures. Disabil Health J 11(1), 99-107.

European Medicines Agency, 2015. Guideline on clinical investigation of medicinal products for the treatment of Multiple Sclerosis.

Cleanthous, S, Cano, S, Kinter, E, et al., 2017. Measuring the impact of multiple sclerosis: Enhancing the measurement performance of the Multiple Sclerosis Impact Scale (MSIS-29) using Rasch Measurement Theory (RMT). Mult Scler J Exp Transl Clin 3 (3).

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.

Confavreux, C, Vukusic, S, 2006. Natural history of multiple sclerosis: a unifying concept. Brain 129(Pt 3), 606-616.

Cohen, J, 1987. Statistical Power Analysis for the Behavioral Sciences, Edition 1, revised ed. L. Erlbaum Associates.

Cook, KF, Bamer, AM, Roddey, TS, Kraft, GH, 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, CD, Cook, KF, 2018. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 27(1), 33-40.

Coons, SJ, Gwaltney, CJ, Hays, RD, Lundy, JJ, Sloan, JA, Revicki, DA, Lenderking, WR, Cella, D, Basch, E, 2009. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value in Health 12 (4), 419–429.

European Medicines Agency, Qualification opinion on multiple sclerosis clinical outcome assessment (MSCOA). https://www.ema.europa.eu/en/documents/other/qualification-opinion-multiple-sclerosis-clinical-outcome-assessment-mscoa_en.pdf. (Accessed December 2020).

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.

Filippi, M., Bar-Or, A., Piehl, F., et al., 2018. Multiple sclerosis. Nat. Rev. Dis. Primers 4 (1), 43.

Fries, J.F., Bruce, B., Bjorner, J., Rose, M., 2006. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann. Rheum. Dis. 65 (3). Suppliii16-21.

Gusi, N., Olivares, P.R., Rajendram, R., 2010. The EQ-5D health-related quality of life questionnaire. In: Preedy, V.R., Watson, R.R. (Eds.), Handbook of Disease Burdens and Quality of Life Measures, Eds. Springer New York, New York, NY, pp. 87–99.

Haley, S.M., Coster, W.J., Binda-Sundberg, K., 1994. Measuring physical disablement: the contextual challenge. Phys. Ther. 74 (5), 443–451.

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.

Hobart, J., Lamping, D., Fitzpatrick, R., Riazi, A., Thompson, A., 2001. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. Brain 124 (Pt 5), 962–973.

Hobart, J.C., Riazi, A., Lamping, D.L., Fitzpatrick, R., Thompson, A.J., 2003. Measuring the impact of MS on walking ability: the 12-Item MS Walking Scale (MSWS-12). Neurology 60 (1), 31–36.
Hobart, J.C., Riazi, A., Lamping, D.L., Fitzpatrick, R., Thompson, A.J., 2005. How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with some other self report scales. J. Neurol. Neurosurg. Psychiatry 76 (11), 1539–1543.
Jaeschke, R., Singer, J., Guyatt, G.H., 1989. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin. Trials 10 (4), 407–415.
Kamudoni, P., Henke, C., Barrett, A., 2018. A new PROMIS physical function short form for use in relapse and progressive multiple sclerosis types. Qual. Life Res. S73.
Kamudoni, P., Amtmann, D., Johns, J., 2020. The Validity and Applicability of the PROMIS SF v2.1 - Physical Function (MS) 15a: A new PROMIS® Short Form For Assessing Physical Function in Relapsing and Progressive Multiple Sclerosis Types. ACTRIMS-ECTRIMS MS Virtual. Poster1062.
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.
Khurana, V., Sharma, H., Afroz, N., Callan, A., Medin, J., 2017. Patient-reported outcomes in multiple sclerosis: a systematic comparison of available measures. Eur. J. Neurol. 24 (9), 1099–1107.
Larocca, N.G., 2011. Impact of walking impairment in multiple sclerosis: perspectives of patients and care partners. Patient 4 (3), 189–201.
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.
Martin, S.A., Olatinka-Amao, O.C., Henke, C., Gnanasakthy, A., 2017. Conceptual model and instrument review in multiple sclerosis. Value Health 20, A726–A727.
Mehta, L., McNeill, M., Hobart, J., et al., 2015. Identifying an important change estimate for the Multiple Sclerosis Walking Scale-12 (MSWS-12v1) for interpreting clinical trial results. Mult. Scler. J. Exp. Transl. Clin. 1.
Miller, D.H., Leary, S.M., 2007. Primary-progressive multiple sclerosis. Lancet Neurol. 6 (10), 903–912.
Miller, D.M., Bethoux, F., Victorson, D., et al., 2016. Validating Neuro-QoL short forms and targeted scales with people who have multiple sclerosis. Mult. Scler. 22 (6), 830–841.
Mills, R., Young, C., Nicholas, R., Pallant, J., Tennant, A., 2009. Rasch analysis of the Fatigue Severity Scale in multiple sclerosis. Mult. Scler. 15 (1), 81–87.
Mokkink, L.B., Knol, D.L., van der Linden, F.H., Sonder, J.M., D'hooghe, M., Uitdehaag, B.M.J., 2015. The Arm Function in Multiple Sclerosis Questionnaire (AMSQ): development and validation of a new tool using IRT methods. Disabil. Rehabil. 37 (26), 2445–2451.
Moore, F., Vickrey, B., Fortin, K., Lee, L., 2015. Two multiple sclerosis quality-of-life measures: comparison in a national sample. Can. J. Neurol. Sci. 42 (1), 55–63.
Muthén, L.K., Muthén, B.O., 2017. Mplus User's Guide. Eighth Edition. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf.
National Institute for Health and Care Excellence, 2019. Final appraisal document: ocrelizumab for treating primary progressive multiple sclerosis. https://www.nice.org.uk/guidance/ta585/documents/final-appraisal-determination-document-2. (Accessed July 21 2020).
National Quality Forum, NQF Measure Incubator. Multiple sclerosis Quality of Care Performance Measure. https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=89298 (Accessed 30 June 2021).
Norman, G.R., Sloan, J.A., Wyrwich, K.W., 2004. The truly remarkable universality of half a standard deviation: confirmation through another look. Expert Rev. Pharmacoecon. Outcomes Res. 4 (5), 581–585.
Nowinski, C.J., Miller, D.M., Cella, D., 2017. Evolution of patient-reported outcomes and their role in multiple sclerosis clinical trials. Neurotherapeutics 14 (4), 934–944.
Phillips, G.A., Wyrwich, K.W., Guo, S., et al., 2014. Responder definition of the Multiple Sclerosis Impact Scale physical impact subscale for patients with physical worsening. Mult. Scler. J. 20 (13), 1753–1760.
Prinsen, C.A.C., Mokkink, L.B., Bouter, L.M., et al., 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual. Life Res. 27 (5), 1147–1157.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
Rae-Grant, A., Day, G.S., Marrie, R.A., et al., 2018. Comprehensive systematic review summary: disease-modifying therapies for adults with multiple sclerosis: report of the guideline development, dissemination, and implementation subcommittee of the American academy of neurology. Neurology 90 (17), 789–800.
Reeve, B.B., Wyrwich, K.W., Wu, A.W., et al., 2013. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual. Life Res. 22 (8), 1889–1905.
Rezapour, A., Almasian Kia, A., Goodarzi, S., Hasoumi, M., Nouraei Motlagh, S., Vahedi, S., 2017. The impact of disease characteristics on multiple sclerosis patients' quality of life. Epidemiol. Health 39, e2017008.
Rose, M., Bjorner, J.B., Gandek, B., Bruce, B., Fries, J.F., Ware, J.E., 2014. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J. Clin. Epidemiol. 67 (5), 516–526.
Rothrock, N.E., Amtmann, D., Cook, K.F., 2020. Development and validation of an interpretive guide for PROMIS scores. J. Patient Rep. Outcomes 4 (1), 16.
Schalet, B.D., Hays, R.D., Jensen, S.E., Beaumont, J.L., Fries, J.F., Cella, D., 2016. Validity of PROMIS physical function measured in diverse clinical samples. J. Clin. Epidemiol. 73, 112–118.
Sharafaddinzadeh, N., Moghtaderi, A., Kashipazha, D., Majdinasab, N., Shalbafan, B., 2010. The effect of low-dose naltrexone on quality of life of patients with multiple sclerosis: a randomized placebo-controlled trial. Mult. Scler. 16 (8), 964–969.
StataCorp L.L.C., 2017. Stata statistical software: release 15.
Stewart, A.L., Kamberg, C., 1992. Physical functioning. In: Stewart, A.L., Ware, J.E. (Eds.), Measuring Functioning and well-being: the Medical Outcomes Study Approach, Eds. Duke University Press, Durham, NC.
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.
The Lancet Neurology, 2019. Patient-reported outcomes in the spotlight. Lancet Neurol. 18 (11), 981.
Thompson, A.J., Polman, C.H., Miller, D.H., et al., 1997. Primary progressive multiple sclerosis. Brain 120 (Pt 6), 1085–1096.
Walton, C., King, R., Rechtman, L., et al., 2020. Rising prevalence of multiple sclerosis worldwide: insights from the Atlas of MS. third edition Mult. Scler. 26 (14), 1816–1821.
Weinshenker, B.G., Bass, B., Rice, G.P., et al., 1989. The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability. Brain 112 (Pt 1), 133–146.
Wilson, I.B., Cleary, P.D., 1995. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 273 (1), 59–65.
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.