Measuring fatigue in multiple sclerosis clinical trials: PRO measure choice is not immaterial when measuring change.

Jeremy Hobart[1], Tanya King[1], Sonia Sappl[2], Ida Marais[2], James Close[1], Pamela Vo[3], David Andrich[2]

[1] University of Plymouth, Faculty of Medicine and Dentistry, England, UK.
[2] The University of Western Australia, Perth, Australia. [3] Novartis Ltd.

INTRODUCTION

- In MS, fatigue is ubiquitous, burdensome, and frequently measured in clinical trials.
- MS studies have used at least 18 different fatigue patient-reported outcome (PRO) measures.
- Evidence suggests PRO measure choice for trials is not evidence-based[1] and shows measure development quality varies substantially[2].
- These findings imply clinical trial results could be PRO measure dependent. As such, understanding their similarities and differences should facilitate choice.
- We compared the measurement performance of 6[3-8] better developed fatigue measures to determine whether choice matters.

AIM

- To determine the degree to which 6 fatigue PRO measures generated equivalent fatigue estimates.

METHOD

- Table 1 shows the 6 measures and their relative structural characteristics. They differ notably.
- All 6 PRO measures were sent simultaneously to n=740 people with MS.
- Response data were analysed using Rasch measurement theory (RMT[9]) methods. We examined:
  - if each measure was well-functioning enough according to RMT measurement criteria.
  - the degree to which measures of the same fatigue component (overall, motor, cognitive) generated statistically equivalent estimates at group and individual person-levels.

* MFIS=Modified Fatigue Impact Scale; NQoL=NeuroQol Fatigue Scale; PROMIS Fatigue Scale; NFI-MS=Neurological Fatigue Index for MS; FSIQ-RMS=Fatigue Symptoms & Impact Questionnaire - Relapsing MS; FSMC=Fatigue Scale for Motor and Cognitive Functions.

Table 1: The 6 fatigue PRO measures compared

PRO measure*	MFIS	FSMC	NFI-MS	NQol	PROMIS	FSIQ-RMS
Publication year	1997	2009	2010	2011	2012	2019
Total no. items	21	20	23	19	8	20
Response category						
Number	5	5	4	5	5	5
Type	Frequency	Intensity (Degree item applies)	Intensity (Agreement with item)	Frequency	Frequency (7) / Intensity (1)	Intensity (Difficulty)
Scores Reportable: Fatigue variable (no. items)	Overall (21) Motor (9) Cognitive (10) Psychosocial (2)	Overall (20) Motor (10) Cognitive (10)	Summary (10) Motor (8) Cognitive (4) Rest relief (6) Sleep (5)	Overall (19)	Overall (8)	Symptoms (7) Impacts (13)^ Motor (5) Cognitive (5) Coping (5)
Comparisons made						
Overall fatigue	MFIS-21	FSMC-20	NFI-MS Summary -10	NQol-19	PROMIS-8	FSIQ-RMS impacts 13
Motor fatigue	MFIS motor 9	FSMC motor 10	NFI-MS motor 8			FSIQ-RMS motor 5
Cognitive fatigue	MFIS cognitive 10	FSMC cognitive 10	NFI-MS cognitive 4			FSIQ-RMS cognitive 5

* Modified Fatigue Impact Scale; NeuroQol Fatigue Scale; PROMIS Fatigue Scale; Neurological Fatigue Index MS; Fatigue Symptoms & Impact Questionnaire - Relapsing MS; Fatigue Scale for Motor and Cognitive Functions.

Table 2: Measurement performance of fatigue PRO measures

		Internal validity		
	Reliability	Threshold		Item Fit
PRO and measures	PSI[1]	Range	Disordering	Chi sq
Overall fatigue				
MFIS-21	0.965	7.69	0%	4.8%
FSMC-20	0.954	5.73	15.0%	20.0%
NFI-MS-10	0.910	7.06	0%	10.0%
FSIQ-RMS-13	0.930	6.95	0%	7.7%
NQoL-19	0.968	11.46	0%	0.0%
PROMIS-8	0.939	11.68	0%	5.3%
Motor fatigue				
MFIS-9	0.934	7.90	0%	11.1%
FSMC10	0.917	5.59	10%	10.0%
NFI-MS-8	0.896	7.65	0%	0%
FSIQ-RMS-5	0.870	6.14	0%	0%
Cognitive fatigue				
MFIS-10	0.961	10.3	0	0%
FSMC-10	0.930	6.65	10.0%	40%
NFI-MS-4	0.786	6.56	0%	0%
FSIQ-RMS-5	0.858	7.64	0%	20%

[1] PSI=Person Separation Index and is analogous to Cronbach's alpha.

Table 3: Fatigue PRO measure comparisons: correlations and agreement between estimates

	k[1]	MFIS	FSMC	NFI-MS	FSIQ-RMS	NQoL19	PROMIS8
Overall fatigue PRO measure comparisons							
	Correlations (raw and corrected for attenuation for error) [2]						
MFIS	21	1	0.938	0.924	0.924	0.907	0.922
FSMC	20	0.900	1	0.913	0.877	***0.848***	0.882
NFI-MS	10	0.866	0.850	1	0.938	0.892	0.935
FSIQ-RMS	13	0.876	0.826	0.863	1	0.912	***0.968***
NQoL	19	0.877	***0.815***	0.837	0.865	1	0.930
PROMIS	8	0.878	0.835	0.864	***0.905***	0.887	1
	Percent of persons with statistically different estimates [3]						
MFIS	21	-	12.79	7.46	12.62	16.60	9.42
FSMC	20	23.5	-	5.88	15.10	***24.62***	11.26
NFI-MS	10	18.0	17.8	-	4.89	9.18	***4.00***
FSIQ-RMS	13	21.0	31.6	15.9	-	12.43	5.86
NQoL	19	28.4	***37.6***	28.1	20.7	-	6.33
PROMIS	8	18.3	23.1	***11.2***	12.1	14.4	-
Motor / physical fatigue PRO measure comparisons							
	Correlations (raw and corrected for attenuation for error[2]						
MFIS	9	1	0.950	***0.951***	0.911	-	-
FSMC	10	***0.879***	1	0.941	***0.864***	-	-
NFI-MS	8	0.870	0.853	1	0.918	-	-
FSIQ-RMS	5	0.821	***0.771***	0.811	1	-	-
	Percent of persons with statistically different estimates [3]						
MFIS	9	-	6.49	**4.02**	6.32	-	-
FSMC	10	14.7	-	4.18	**7.86**	-	-
NFI-MS	8	**9.8**	11.8	-	4.52	-	-
FSIQ-RMS	5	15.2	**16.5**	12.6	-	-	-
Cognitive fatigue PRO measure comparisons							
	Correlations (raw and corrected for attenuation for error) [2]						
MFIS	10	1	0.942	0.945	0.909	-	-
FSMC	10	***0.891***	1	***0.980***	***0.884***	-	-
NFI-MS	4	0.823	0.840	1	0.963	-	-
FSIQ-RMS	5	0.824	***0.788***	0.791	1	-	-
(PSI [2])		*(0.961)*	*(0.930)*	*(0.789)*	*(0.855)*	-	-
	Percent of persons with statistically different estimates [3]						
MFIS	10	-	8.02	4.78	8.70	-	-
FSMC	10	18.3	-	**1.52**	**9.63**	-	-
NFI-MS	4	12.8	8.4	-	2.55	-	-
FSIQ-RMS	5	**18.6**	17.9	**6.7**	-	-	-

[1] K=number of items in PRO measure
[2] Raw correlations below diagonal. Corrected for error above diagonal in grey. From separate analyses of each PRO measure.
[3] Percent of persons at 5% confidence of a type-1 error below diagonal., 1% confidence of type-1 error above diagonal in grey. Highest and lowest values bolded.

Table 4. Fatigue PRO measure comparisons: subtest analyses[1]

	Person Separation Index		Proportion of variance explained	
PRO measure comparison	All items[2]	Subtests[3]	r [4]	A [5]
Overall fatigue [6]				
Six measures	0.990	0.963	0.976	0.973
Five measures	0.988	0.957	0.996	0.969
Motor fatigue	0.972	0.941	0.942	0.968
Cognitive fatigue	0.976	0.942	1.061	0.966

[1] Subtest analyses determine the degree to which different measures measure the same variable by estimating the variance that is common to all ***items (All items PSI)***, and unique to difference ***measures (Subtests PSI)***. The greater the similarity between the two PSI estimates the less evidence of difference.
[2] All items = sum of items in each comparison group: Overall fatigue 6 measures=91, Overall fatigue 5 measures=78, Motor fatigue=32, Cognitive fatigue=29.
[3] The PSI when each measure in a comparison is reconfigured to form a jumbo item, and the jumbo items form a measure.
[4] r = average latent (error-corrected) correlation among PRO measures compared = proportion of total variance that is common variance among the subscales (common Var / total Var).
[5] A = proportion of non-error variance common among PRO measures compared (common V / (total Var – error Var).
[6] The six overall fatigue PRO measures sub-tested were MFIS-21, FSMC-20, NFI-MS-10, NQoL19, PROMIS-8, FSIQ-RMS-13. For five PRO measures FSIQ-RMS-13 was excluded

INTERPRETATION

- Reliability and validity results indicate all 14 fatigue PRO measures from the 6 different instruments were well functioning enough.
- Correlations and subtest results indicate fatigue measures with notable structural and content differences ostensibly measure the same variable.
- Estimate equivalence results imply comparisons at group and individual levels may come to different conclusions.
- A PRO measure's item number determines its measurement precision and, therefore, inherent ***potential*** ability to quantify between-person differences at one point in time, or to detect within-person change over time.

IMPLICATIONS FOR FATIGUE MEASUREMENT

- The finding that different development rigour, conceptualisation, and structural characteristics of fatigue measures have little impact on the ***nature*** of the variables measured questions whether fatigue measures are specific enough for evaluating therapeutic interventions. This could explain, in part, why few treatments have proven effective. Therefore, how we measure fatigue warrants re-thinking.

IMPLICATIONS FOR PRO MEASUREMENT IN CLINICAL TRIALS

- Clinical trial results are likely to be PRO measure dependent as different measures of the same variables have different potential abilities to detect change. We advise detailed PRO measure comparisons, including comparisons of change, to facilitate choice of which is preferable.
- PRO measures with fewer items may underestimate "true" change in trials. We advise careful consideration.
- Individual-person analyses complement group-level analyses. We advise they are routinely undertaken.
- Rarely used new psychometric methods offer opportunities to advance PRO measure development, evaluation, comparison, selection and understanding. We advise clinical trialists and PRO measure developers to use them routinely and appropriately.

Funding:
This study was funded by Novartis who had no influence over the study design, process, analysis, and interpretation.

Disclosures:
JH: consulting fees, honoraria, support to attend meetings, research support or clinical service support from: Acorda, Bayer Schering, Biogen Idec, BMS, F. Hoffmann-La Roche, Janssen, Merck Serono, Novartis, Oxford PharmaGenesis, Sanofi-Genzyme, Teva. PV: employee at Novartis during this work. TK, JC, SS, IM, DA none.

RESULTS

- The response rate was 73% (538/740). Item-level data completeness was high.
- The sample was: 80% female, aged 25-88 (mean 60yrs), 53% relapsing MS, self-fatigue grading (none=5%; mild=28%; moderate= 47%; severe= 20%).
- Table 2 shows reliability and internal validity estimates for all 14 measures used in the comparisons. All 14 had high reliabilities (PSI $\geq$ 0.86) and good internal validity - in that all measures mapped out continua (threshold range 5.6 to 11.7 logits), had response categories working well (disordering rare), and good item cohesiveness (few out-of-range "fit" values).
- Tables 3 and 4 show that PRO measures of the same fatigue component:
  - were very highly correlated (error-corrected r=0.85-0.98; Table 3).
  - had >96% common variance on subtest analyses (Table 4).
  - generated statistically ***equivalent*** group mean scores (p>0.01).
  - generated statistically ***different*** estimates for individuals 7-38% (mean =18%) of the time, at the 5% level of confidence for type-1 error (Table 3).
- Simulation studies demonstrated that the greater the number of items in a PRO measure, the greater the difference between individual and group-level interpretations from two measures (results not shown).

References:
1. Hobart et al. Do clinical trials prepare to fail by failing to prepare? An examination of MS trials and recommendations for patient-reported outcome measure selection. MSARD 2023 (https://doi.org/10.1016/j.msard.2023.104788
2. Close et al. Measuring fatigue in MS: there may be trouble ahead. Neurology and Therapy 2023 Jun2023 (https://doi.org/10.1016/j.msard.2023.104788).
3. Fisk et al. Measuring the functional impact of fatigue: initial validation of the fatigue impact scale. Clin Infect Dis. 1994;18(Suppl 1):S79-83.
4. Penner et al. The Fatigue Scale for Motor and Cognitive Functions (FSMC): validation of a new instrument to assess multiple sclerosis-related fatigue. MSJ 2009;15(12):1509-17.
5. Mills et al. Development of a patient reported outcome scale for fatigue in multiple sclerosis: the Neurological Fatigue Index (NFI-MS). HQLO. 2010;8:22.
6. Cella et al. Neuro-QOL: brief measures of health-related quality of life for clinical research in neurology. Neurology. 2012;78(23):1860-7.
7. Cook et al. A PROMIS fatigue short form for use by individuals who have multiple sclerosis. Qual Life Res. 2012;21(6):1021-30.
8. Hudgens et al. 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. 2019;22(4): 453–66.
9. Hobart et al. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007;6(12):1094–105