eNeurologicalSci 19 (2020) 100237

Contents lists available at ScienceDirect

eNeurologicalSci

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

Review article

Measuring upper limb function in MS: Which existing patient reported outcomes are fit for purpose?

James Close[a], Kathryn Baines[b], Laurie Burke[c], Jeremy Hobart[b,*]

[a] Community and Primary Care Research Group, Faculty of Medicine and Dentistry, University of Plymouth, Plymouth Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK
[b] Institute of Translational & Stratified Medicine, Faculty of Medicine and Dentistry, Plymouth Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK
[c] LORA Group LLC, Royal Oak, MD 21662, USA

ARTICLE INFO

Keywords:
Multiple sclerosis
Upper limb/extremity measurement
Patient reported outcome
Concept of interest
Context of use

ABSTRACT

Background: Multiple Sclerosis (MS) clinical trials increasingly focus on progressive and advanced MS, with upper limb function (ULF) as a key outcome. Within clinical trials, Patient Reported Outcomes (PROs) quantify clinical variables and establish meaningfulness of changes. Scientific standards and regulatory criteria (from Food and Drug Administration [FDA]) require PROs be "fit-for-purpose": well-defined and reliable measures of specific concepts in defined contexts.

Objective: To identify, from literature, existing PROs measuring ULF and determine which satisfy scientific and regulatory clinical trials requirements.

Method: We screened PubMed/Web of Science using multiple relevant terms. Abstracts and full texts were screened using suitability criteria. PRO development papers were evaluated using recently expanded Consensus Standards for Measurement Instruments (COSMIN) criteria for content development.

Results: We identified 3619 articles; 485 used 24 different ULF PROs. No PRO satisfied scientific and regulatory requirements as a well-defined measure of a clearly defined construct in a specific clinical context.

Conclusions: Existing ULF PROs don't meet fit-for-purpose criteria. MS clinical trials require new measures with greater emphasis on patient engagement to derive theoretical frameworks, concepts of interest, and contexts of use followed by systematic literature searches, expert input, and qualitative research to support item generation. Until then, trials will miss aspects of meaningful within-patient change and thereby misrepresent (likely underestimating) treatment effects.

1. Background

Clinical trials for people with progressive and advanced multiple sclerosis (MS) have steadily increased over the past decade [1–3]. This represents evidence that MS pathology might be influenced throughout the disease course and the importance of protecting individuals upper limb function (ULF). However, these MS context present measurement challenges. The widely-used EDSS (Expanded Disability Status Scale) [1] quantifies overall function, but does not delineate ULF [2]. The nine hole peg test (9-HPT) (used stand-alone or in the MS functional composite (MSFC)), has valuable measurement properties. Whilst it is an "objective" timed hand function performance test (albeit under subjective control) with equal interval units (seconds), the 9-HPT doesn't measure peoples' ability to perform routine tasks in daily life. Instead, patient-reported outcomes (PROs) are required.

PRO requirements have developed over the last decade. A key development is measurement clarity: PROs must measure clearly defined concepts in specific clinical contexts. The US Food and Drug Administration (FDA) have been instrumental from a regulatory perspective [5,11] – although these are long held measurement science views from some quarters [3].

The fundamental logical requirement is content validity – the extent to which PRO content adequately reflects the measurement construct [4]. The FDA advise that this is achieved using an iterative process of qualitative and quantitative approaches (See Box 1). Content validity is cited the most important measurement property of a PRO [5], and according to FDA guidance, these basic requirements of content validity should be established prior to any statistical (i.e. psychometric") examination.

Here, we review existing ULF PROs against content validity criteria

* Corresponding author.
E-mail addresses: james.close@plymouth.ac.uk (J. Close), kathryn.baines@plymouth.ac.uk (K. Baines), lburke@loragroup.com (L. Burke), jeremy.hobart@plymouth.ac.uk (J. Hobart).

https://doi.org/10.1016/j.ensci.2020.100237
Received 3 December 2019; Received in revised form 8 March 2020; Accepted 12 March 2020
Available online 16 March 2020

Box 1
Summary of FDA guidance for establishing content validity of PROs for clinical trials (left) and corresponding COSMIN checklist items (right).

FDA guidance	Corresponding COSMIN items (Box 1a of revised guidelines)
Concept of interest. The Concept of Interest (COI) should measure a meaningful treatment benefit of symptoms or function. If no PRO instruments exist to assess the COI, new PRO instruments should be developed (or adapted from existing instruments).	***Item 1:*** Is a clear description provided of the construct to be measured?
Conceptual framework. Content validity is underpinned by a conceptual framework. According to the FDA documentation, *"one fundamental consideration in the review of a PRO instrument is the adequacy of the item generation process to support the final conceptual framework of the instrument."* An iterative process of development should be followed. This includes hypothesizing an initial conceptual framework before drafting new instrument on the basis of literature review, expert review and qualitative research in the targeted patient population. In most instances, the final model will describe the COI as domains and subdomains, along with their hypothesized relationships.	**Item 2:** Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?
Context of use. The context of use (COU) specifically defines the target population for the COI (e.g. clinical group/subgroup, sex, age, ethnicity) and the intended context (e.g. clinical, rehabilitation, trials).	**Item 3:** Is a clear description provided of the target population for which the PROM was developed? **Item 4:** Is a clear description provided of the context of use?
Item identification and development. Items should be identified via both [1] literature reviews and [2] qualitative work, with iterative cycles of patient/expert input to adjust both the theoretical framework and items, thus ensuring content validity. In particular, input from the target population (i.e. the COU) should be documented from focus groups or interviews to evaluate items, wording and coverage. According to the FDA, *"Sponsors should provide documented evidence of patient input during instrument development and of the instrument's performance in the specific application in which it is used (i.e., population, condition)… Without adequate documentation of patient input, a PRO instrument's content validity is likely to be questioned."*	**Item 5:** Was the PROM development study performed in a sample representing the target population for which the PROM was developed? **Item 6:** Was an appropriate qualitative data collection method used to identify relevant items for a new PROM? Finally, we added a further item to the checklist (derived from FDA criteria): **Item 6a:** were appropriate literature searches conducted to identify relevant items for a new PROM?
Item reduction and finalisation of instrument. The processes by which an instrument was finalised should be well documented, including the reasons for deleting or modifying items (e.g. via qualitative or statistical approaches).	Not covered by COSMIN standards.
Appropriate documentation. Before a measure is suitable for deployment in a clinical trials, appropriate documentation should be in place. This should include a published development history; a user manual with guidelines for study interpretation and scoring of measure.	Not covered by COSMIN standards.

to inform PRO selections for MS clinical trials, and identify future advancements to satisfy scientific and regulatory criteria. We used Consensus Standards for Measurement Instruments (COSMIN's) checklist and quality indicators [5], which have been updated to incorporate content validity assessments that align with FDA guidance (Box 1).

2. Methods

2.1. Objective

To identify, from literature, existing ULF PROs and determine which meet regulatory and scientific requirements as fit for purpose in MS clinical trials, using expanded COSMIN standards for content development. As many scales are re-validated or adapted for use in other clinical contexts (i.e. beyond the original context of interest), we did not restrict our literature searches to MS, but instead utilized an inclusive search strategy to identify any PROs for ULF. This approach would enable us to identify any ULF PROs that may be suitable for future deployment or validation in an MS context of interest. Any resulting ULF PROs were assessed using the COSMIN checklist for PRO content development – which maps onto FDA criteria (Box 1) – and in this manner, we would determine if any PROs meet regulatory criteria for MS clinical trials.

2.2. Literature review

We searched electronic databases (PubMed, Web of Science) for all articles published prior to November 30th 2017. Search terms were a combination of keywords relating to the concept of ULF [upper extremity, arm, dexterity, motor control, hand control, hand activity] and PROs [scale, instrument, patient reported outcome, questionnaire]. The search was not exclusive to MS, but restricted to articles on human subjects and English language.

Abstracts of returned articles were screened by two independent (blinded) reviewers to identify papers discussing ULF PROs. These were reviewed at full text level to identify cited PROs. During full-text screening, articles were excluded if the instrument was not a PRO, not measuring ULF, developed for children, the PRO or development paper was not accessible (see Fig. 1 for flow diagram).

2.3. Quality analysis

For all PROs meeting inclusion criteria, we assessed the original development paper for the extent to which the PRO met requirements for instrument content development. Historically, established standards for PRO evaluation, such as COSMIN and EMPRO (Evaluating the Measurement of Patient-Reported Outcomes) [6–8], have focused primarily on psychometric properties, and have only covered limited aspects of conceptual development. However, COSMIN standards were recently extended to include comprehensive assessment of content development [5,9].

According to COSMIN guidelines, PRO reviews (such as this paper) should clearly define their scope, including "the construct, target population, and context of use". This provides the frame of reference for evaluating content validity [9]. Here, we assessed ULF PROs in the context of clinical trials, and more specifically MS treatment trials. Aligning with COSMIN, scientific and regulatory requirements and recommendations detailed in Box 1, we targeted the revised COSMIN checklist items associated with PRO content. Other reviews of PRO content development have employed this strategy [10,11]. We focused on the first section of COSMIN's checklist which maps directly to FDA guidance, comprising "Box1a", items 1–6 [6]; see Box 1.

Aligning our methods to recent publications applying these new standards (11), one reviewer (JC) extracted data from studies. A second reviewer (JH) double-checked accuracy of 20% of extracted information. The extracted information was next assessed on a 4-point scale;

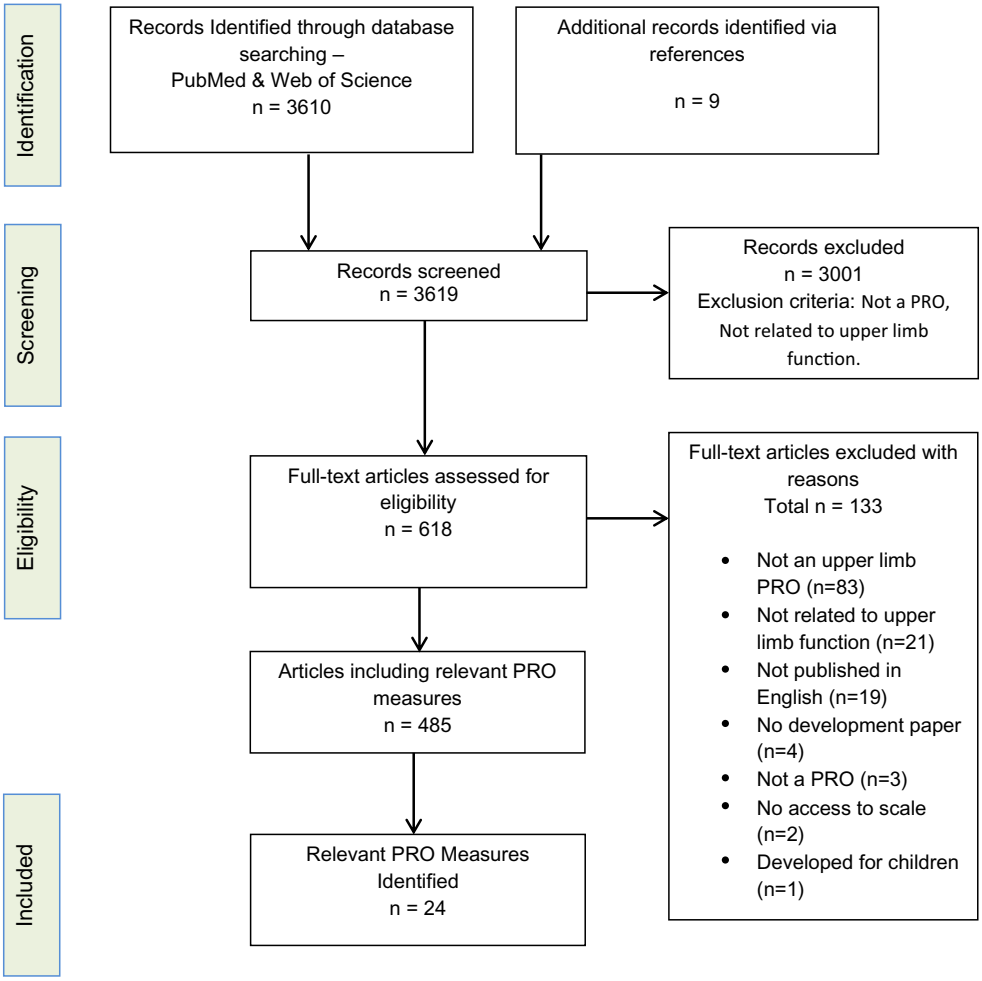

Fig. 1. Flow diagram illustrating literature searches, inclusion/exclusion criteria and identification of 24 UL PRO measures.

"good' (+), "adequate" (+?), "doubtful" (−?), "poor/none" (−), according to the process specified in the COSMIN documentation, '*COSMIN methodology for assessing the content validity of PROMs User manual version 1.0*'.[1] Whilst scoring in COSMIN checklist is item specific, assessment ranges from 'good' (i.e. a widely recognised approach) to 'poor/none' (i.e. approach not documented or missing). Whilst some subjectivity is unavoidable – especially in the middle of the scale – both reviewers independently assessed each item on the checklist, with any inconsistences resolved verbally.

3. Results

Literature searches identified 3610 articles (Fig. 1), with a further 9 articles identified from papers' reference lists. Abstract screening excluded 3001 articles not related to ULF or PROs, and 133 were excluded according to other criteria (see methods and Fig. 1). The remaining 485 articles contained 24 unique ULF PROs.

3.1. PRO descriptions

Table 1 shows descriptive information of the 24 PROs, including original development references, stated intended contexts of use and details about construction.

3.2. Quality analysis

Table 2 summarises PRO content development quality analyses. Supplementary Table 1 provides information informing the quality analysis (extracted from development papers). Table 3 summarises findings across all 24 PROs, highlighting systematic shortcomings in content development quality. Overall findings for each assessment criterion are discussed below.

3.2.1. Concept of interest (COI)

Both COSMIN and FDA require clarity of the measurement concept. Whilst the ULF PROs we reviewed usually documented a concept for measurement, most concepts were loosely described, rather than clearly specified. Therefore, most PROs (17/24) had an 'adequate' COI rating. For instance, MAM-16 and MAM-36 (an expanded scale for general rehabilitation populations) are described as "task oriented and patient-centred tool(s) that measures manual ability for hand use." Neither define what is meant by "manual ability", the specific aspects of "manual ability" measured, nor types of tasks covered by "task oriented". Similar shortcomings were identified across most measures. Two PROs didn't define any COI (UEFI and MAL). One PRO (FLEX-SF) had a broad ambiguous definition ("shoulder function"), hence was rated "poor". Finally, no PROs described a process of patient/expert engagement [12] to define the scope of the COI, any literature searches (if done) were not documented.

[1] https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf. Note that we have reworded the 4-point scale from the original documentation, although the scoring strategy remains the same.

Table 1
Properties of ULF PROs.

Measure	Reference		Disease group	Items	Recall period	Response categories:		Method of:		
	Year	Author				Number	Type	Administration	Development	Scoring
ABILHAND	1998	Penta	None specified	46	NS	3	NOC	Paper/interview	RMT	Summation / RMT
ABILHAND-NMD	2010	Vandervelde	Neuromuscular disorders	22	NS	3	NOC	Paper/interview	RMT	Summation / RMT
ABILHAND-RA	2007	Durez	Rheumatoid arthritis	27	NS	3	NOC	Paper/interview	RMT	Summation / RMT
ABILHAND-SSC	2009	Vanthuyne	Systemic sclerosis	26	NS	3	NOC	Paper/interview	RMT	Summation / RMT
ABILHAND-Stroke	2001	Penta	Chronic stroke	23	NS	3	NOC	Paper/interview	RMT	Summation / RMT
AMHFQ	1999	Alderson	Carpal Tunnel Syndrome	56	Now	7	NOC	Paper	CTT	Simple
AMSQ	2015	Mokkink	Multiple sclerosis	31	Last 2 weeks	6	NOC	Paper	IRT	Summation / IRT
ARM-A	2013	Ashford	Hemiparetic UL	20 (now 21)	Last week	5	NOC	Paper	None	Subscales only; Summation
AUSCAN	2002	Bellamy	Hand osteoarthritis; Hand joint pain	15	Last 2 days	5 (Likert) or 11 (numeric)	Likert/ VAS/ Numerical	Paper/interview	CTT	Not clear
CUE	1998	Marino	Adults with tetraplegia and spinal cord injuries.	32	Average day	7	NOC	Interview	CTT	Not clear
DASH	1996	Hudak	UE musculoskeletal disorders	30 + 8 optional	Last week	5	NOC	Paper	CTT	Summation
Quick DASH	2005	Beaton	UE musculoskeletal problems	11 + 8 optional	Last week	5	NOC	Paper	CTT/RMT	Summation
FLEX-SF	2003	Cook	Shoulder surgery	33 + screening question	NS	5	NOC	Paper	RMT	Summation
MAL	1993	Taub	Stroke	30	NS	6	NOC	Interview	None	Not clear
MAM-16	2005	Chen	Hand impairments	16	NS	5	NOC	Paper	RMT	RMT
MAM-36	2010	Chen	Hand impairments	36	NS	5	NOC	Paper	RMT	RMT
MAP-HAND	2010	Paulsen	Rheumatoid arthritis	18	Last time performed	4	NOC	Paper	RMT	Summation
MAS	2005	Seaton	UE disability	47	Now	5	NOC	Paper	CTT	Summation
MHQ	1998	Chung	Hand disorders	37	Last week	5	NOC	Paper	CTT	Summation
Neuro-QoL	2011	Cella	Neurological disorders	20	NS	5	NOC	Computer	RMT/IRT	Summation
Neuro-QoL SF	2012	Cella	Neurological disorders	8	NS	5	NOC	Paper/computer	RMT/IRT	Summation
UEFI	2001	Stratford	UE dysfunction	20	Now	5	NOC	Paper	CTT	Summation
UEFS	1997	Pranksky	Work-related UE disorders	8	NS	10	NOC	Paper	CTT	Summation
ULFI	2006	Gabel	UL symptoms	25	Last few days	2	Yes/No	Paper	CTT	Summation

Fourth column, '*Disease group*' is the population for which the instrument was originally designed. For sixth column, '*Recall period*', *NS = Not Specified*. For eight column, '*Response categories/type*', *NOC* = Numerically Ordered Categories with descriptors, *VAS* = Visual Analog Scale. For eleventh column, '*Scoring*', *Summation* = item scores summed without weighting or standardisation; *RMT* = Rasch Measurement Theory (i.e. score derived from item responses); *Subscales only* = a simple score for sub-scales, but these are not to be combined; *NC* = Not clear; *UL* = Upper Limb; *UE* = Upper Extremity; *IRT* = Item Response Theory.

Table 2
Quality analysis of content development for 24 ULF-PROs.

PRO instrument	[Q1] COI	[Q2] Conceptual framework	[Q3] Target population	[Q4] COU	[Q5] Representative Sample	Item Generation (Q6 and 6a)		Item reduction method
						[Q6] Qualitative Items	[Q6a] Literature Items	
ABILHAND	+?	- (ICF)	-	−?	−?	-	-	RMT/Statistical
ABILHAND-NMD	+?	- (ICF)	-?	−?	-	-	-	RMT/Statistical
ABILHAND-RA	+?	- (ICF)	+	−?	-	-	-	RMT/Statistical
ABILHAND-SSC	+?	- (ICF)	+	−?	-	-	-	RMT/Statistical
ABILHAND-Stroke	+?	- (ICF)	+	−?	-	-	-	Logical/Statistical
AMHFQ	+?	-	+?	+?	-	-	-	Statistical
AMSQ	+?	−? (ICF)	+?	−?	−?	-	−-?	Expert/Statistical
ARM-A	+	−?	+	+	−?	−?	+	Expert/Delphi
AUSCAN	+?	-	+?	+?	+?	+	−?	Logical/Statistical
CUE	+?	+? (ICF)	+?	+?	-	-	-	Statistical
DASH	+?	−?	+?	−?	-	-	-	Logical/Expert
Quick DASH	+?	−?	+?	-	-	-	-	Logical/Expert
FLEX-SF	-	-	-	+?	-	−?	-	RMT/Statistical
MAL	-	-	-	-	-	-	-	None
MAM-16	+?	−? (ICF)	−?	−?	-	-	-	RMT/Statistical
MAM-36	+?	−? (ICF)	+?	+?	-	-	-	RMT/Statistical
MAP-HAND	+?	−? (ICF)	+?	+?	+?	+?	+?	RMT/Statistical
MAS	+?	+? (ICF)	+?	+?	-	-	-	Expert/Unclear
MHQ	−?	−?	-	+?	-	-	-	Logical/Statistical
Neuro-QoL FM/UEF	−?	-	+?	−?	−?	−?	−?	RMT/IRT/Expert
Neuro-QoL SF	−?	-	+?	+?	−?	−?	−?	IRT/Expert
UEFI	-	- (ICF)	+?	−?	−?	-	-	Statistical/ Redundancy
UEFS	+?	-	+?	−?	−?	-	-	Logical
ULFI	+?	−? (ICF)	-	−?	−	-	+?	Redundancy/ Logical

Quality assignments are as follows: *Good (+), Adequate (+?), Doubtful (−?), Poor/None (−).* See Supplementary Table 1 for details. For the columns:
Q1, '*COI*' (*Concept of interest)*; was a well-defined concept of interest documented?
Q2, '*Conceptual framework*'; was the concept of interest underpinned by a conceptual framework? When appended by '*ICF*', the conceptual framework cited the WHO International Classification of Functioning, Disability and Health.
Q3, '*Target population*'; did the measure have a well-defined target population?
Q4, '*COU*' (*Context of use)*; was the measure developed for use in a well-defined context?
Q5, '*Representative sample*'; was the scale development work conducted in a sample representing the intended context of use?
Q6, *'Qualitative Items';* was qualitative work was used to generate items?
Q6a, *'Literature items';* were literature reviews used to generate items?
The final column, '*Item reduction*', indicates the method used to reduce the number of items (if there was an item reduction stage). RMT = Rasch Measurement Theory; IRT = Item Response Theory. See Supplementary Table 1 for details.

Table 3
Summary of quality analysis of content development for 24 ULF-PROs.

Quality rating	Quality criteria assessed						
	[Q1] COI	[Q2] Conceptual framework	[Q3] Target population	[Q4] COU	[Q5] Representative sample	[Q6] Qualitative items	[Q6a] Literature items
Good (+)	1 (4.2%)	0 (0%)	4 (16.7%)	1 (4.2%)	0 (0%)	1 (4.2%)	1 (4.2%)
Adequate (+?)	17 (70%)	2 (8.3%)	13 (54.2%)	9 (37.5%)	2 (8.3%)	1 (4.2%)	2 (8.3%)
Doubtful (−?)	3 (12.5%)	9 (37.5%)	2 (8.3%)	12 (50%)	7 (29.2%)	4 (16.7%)	4 (16.7%)
Poor/none (−)	3 (12.5%)	13 (54.2%)	5 (20.8%)	2 (8.3%)	15 (62.5%)	18 (75%)	17 (70.8%)

Each column is the number (percent) of sum of PROs scored for each of the seven quality criteria (see Table 1 and methods), according to the ratings. COI = concept of interest; COU = context of use.

3.2.2. Conceptual/theoretical framework

COSMIN guidelines state theoretical models should delimit the concepts boundaries, show relationships between related concepts and subscales, and reflect the current state of science across relevant disciplines. We identified no measures satisfying COSMIN criteria for a robust conceptual framework for ULF. No studies provided complete concepts (e.g. divided into relational domains/subdomains), nor used well-documented systematic literature searches, citations of theoretical papers or expert/patient qualitative input to build a conceptual model.

Around half the PROs (13/24) provided no conceptual framework. Another 9 PROs had doubtful (i.e. ill-defined) conceptual frameworks. Thirteen PROs cited definitions from the WHO International Classification of Functioning, Disability and Health (ICF) [13] for activity limitations, functioning, impairment and participant restrictions (Table 2). However, the ICF is a generic taxonomy of some functions which does not specifically delimit disability in ULF, and the publications we reviewed provided no evidence (or theory) that the ICF was relevant to the specific concept/context being measured. Finally, some measures citing the ICF (e.g. ABILHAND measures) used its definitions to define the broad domain of measurement (e.g. impairment, disability, participation restrictions) rather than to construct their theoretical framework. Whilst Neuro-QOL development involved a detailed process of identifying important diseases, "health-related quality of life" domains, subdomains and items, there was no conceptual framework per se and no elaboration on the "Upper Extremity Function" domain measured.

3.2.3. Target population

Updated COSMIN standards recommend developers provide a clear description of the target population for which the PRO was developed. Seventeen PRO development papers specified target populations achieving "good" (4/24; well-defined clinical population) or "adequate" (13/24; loosely defined clinical population) ratings. Five PROs didn't specify target populations (ABILHAND, FLEX-SF, MAL, MHQ, ULFI).

One PRO, Arm Function in Multiple Sclerosis Questionnaire (AMSQ), was developed specifically for people with MS (Table 1). Another 12 PROs had broad concepts that might encompass MS (e.g. neuromuscular disorders and general functioning: ABILHAND-NMD, DASH, Quick DASH, MAL, MAM-16, MAM-36, MAS, Neuro-QoL, Neuro-QoL-SF, UEFI, UEFS, ULFI). Ten PROs had non-MS contexts (e.g. specific to rheumatoid arthritis, stroke etc.: ABILHAND, ABILHAND-RA, ABLIHAND-SSC, ABILHAND-Stroke, AMHFQ, ARM-A, AUSCAN, CUE, MAP-HAND, MHQ).

3.2.4. Context of use (COU)

Updated COSMIN standards define context of use as the intended application of the PROM. This was only well specified with one measure (ARM-A). Contexts of use for half the PROs were poor/missing (2/24) or doubtful (12/24). Most measures stated that the context of use is to measure the concept of interest for clinical use on patients within the intended population. Thus, whilst clinical targets were frequently cited, the specific clinical purpose, context and samples (e.g. trial; acute care; therapy etc.) were rarely defined (e.g. outcome assessment may require a different instrument to a diagnostic tool). Whether a PRO was developed for discriminative, evaluative or predictive applications was rarely stated. These contextual factors are required (by some) to determine desirable properties of the final instrument. For example, evaluative applications need to include items covering entire range of the scale, whereas with diagnostic applications this may not be necessary [5,9].

3.2.5. Representative sample

COSMIN and FDA recommend items for PROs are generated from samples representative of the context in which the PRO will be used. Two of 24 PROs (AUSCAN; MAP-HAND) generated items from well-described target populations, although the sample representativeness was only deemed COSMIN 'adequate'. Whilst other instruments used representative samples for psychometric evaluations, they did not develop items using representative, diverse samples of the target clinical population. Typically, samples used for item development were extremely limited/absent (15/24 PROs), or doubtful (7/24 PROs) – where either the sample did not adequately represent the target population, or the sample was not well described.

These shortcomings stemmed from broad definitions for 'target population'. For example, ARM-A targeted 'hemiparetic upper limb patients', but items were generated from a small sample ($n = 13$) of questionable representativeness (all were stroke patients, and authors admitted sample was limited). Similarly, ABILHAND targeted people with rheumatoid arthritis (RhA), using a convenience sample of $n = 18$ people with RhA who had hand surgery up to 25 yrs. ago and from other" selected scales" (no further details given). Whilst development papers often described the samples used for psychometric validation, samples used for item generation did not often provide sample characteristics (CUE, FLEX-SF, MAS, MHQ, NeuroQol, NeuroQoL SF, UEFI, UEFS), and therefore the adequacy of the sample could not be assessed.

3.2.6. Qualitative work to generate items

COSMIN and FDA require clarity of the item generation processes. One PRO (AUSCAN) documented well conducted qualitative methods against a representative sample of patients ($n = 50$) across all instrument domains, along with assurances that item saturation was reached (i.e. no further items – or relevant themes – were uncovered in the final set of qualitative interviews). Another PRO (MAP-HAND) used methods deemed 'adequate,' with suitably documented qualitative methods (60 semi-structured interviews; video observations). However, the publication didn't clarify the representativeness of sample, level of saturation reached, and which items were derived from qualitative work. Item generation for many instruments (18/24) used either no qualitative work or undocumented methods.

3.2.7. Literature review to generate items

One PRO (ARM-A) was developed using well-specified literature searches, based on a previous systematic review. Two PROs (MAP-HAND, ULFI) were rated 'adequate', with detailed literature performed to generate items, but exact methods were not specified. Most PRO development publications (17/24) did not conduct document literature searches to generate items. Instead, they cited a small number of pre-existing scales, without documenting how they were identified.

3.2.8. Item reduction / selection for final instrument

Most PROs contain less items than originally examined (sometimes referred to as the item pool). Table 2 (final column) and Supplementary Table 1 show a range of approaches used to reduce or select items from original pool. This included exclusively statistical approaches (Rasch Measurement (e.g. ABILHAND, MAP-HAND, Item Response (e.g. Neuro-QoL) and Classical Test Theory criteria (e.g. AMHFQ) [RMT, IRT, CTT]) with items removed if (for example) they did not "fit" the requirements of a statistic model, were redundant, had malfunctioning response categories, were too easy or uncommon, unable to discriminate or missing values > 10%.

Some PROs (Arm-A, DASH, MAS) used purely qualitative methods. This included patient and caregiver interviews to reduce items by (for example), identifying irrelevant items. Other PROs (e.g. Quick DASH) used a concept-retention approach: selecting the highest ranking item/s from each domain and concept. Some PROs (e.g. DASH) relied purely on expert opinion (i.e. clinicians or researchers) for item reduction. Only one PRO (Neuro-QoL) used a combination of qualitative work with patients (the only instrument using cognitive interviews), expert input and statistical judgement for item reduction. Several PROs (AMHFQ, CUE, MAL, MAS) did not provide clear explanations of their item reduction methods.

3.2.9. Miscellaneous issues

Our review highlighted other issues relevant to quality evaluation. Whilst some might consider these minor omissions, they contribute to an overall lack of measurement clarity and transparency, which is required by investigators from PROs used in high-stakes clinical trials. For example, ARM-A had an undocumented additional item added after psychometric testing. ABILHAND items originally had 4 response categories, reduced to 3 without clarification (then propagated across ABILHAND-derived PROs: ABILHAND-NMD, ABILHAND-RA, ABILHAND-SSC, ABILHAND-Stroke). For these ABILHAND-derived measures, the number of initial items was 56, rather than 57 documented in the original ABILHAND. How CUE and MAL scores are generated is unclear. Two PROs had no (AUSCAN) or unclear (CUE) instructions for handling missing data. Twelve PROs (ABILHAND, ABILHAND-NMD, ABILHAND-RA, ABILHAND-SSC, ABILHAND-Stroke, FLEX-SF, MAL, MAL-16, MAL-36, Neuro-QoL, Neuro-QoL SF, UEFS), defined no recall period, increasing ambiguity and reducing measurement accuracy.

4. Discussion

We aimed to identify all existing ULF PROs and determine their suitability for MS clinical trials. None of 24 PROs satisfied COSMIN or FDA criteria as fit-for-purpose measures in any clinical trials. Whilst many of the PROs may be considered (by some) acceptable for low-stakes exploratory studies, they do not satisfy criteria for registrational or high-stakes clinical trials [14,15]. When suboptimal PROs are used,

it is impossible to determine the extent to which they provide accurate and clinically interpretable quantifications of ULF and of treatment effects. Type II error (failing to detect a true effect or treatment impact underestimation) of an unknown degree is the most likely outcome, unless fit-for-purpose PROs are used. With no adequate ULF PROs, this problem will continue to handicap outcomes of high-stakes clinical trials.

Not surprisingly, COSMIN's recently revised PRO content development checklist hasn't yet been used widely. Nevertheless, two publications using the standards have concluded PRO content validity is "worrisome" [10,11]. Our review mirrors these findings. In particular, instrument development was consistently weak with regard to patient engagement, literature review and expert input to develop a targeted concept of interest for measurement [15]. In the final publications, attention was rarely given to conceptual frameworks (often absent), contexts of use (often loosely defined), use of representative samples to generate items (often no sample), use of qualitative work to generate items (often not performed) and literature searches to generate items (rarely documented).

All these criteria are essential for fit-for-purpose PRO measurement (see, for example, the FDA Roadmap to patient-focused outcome measurement in Clinical Trials – Fig. 2). Regularly used clinical concepts like upper limb function, physical function, disability, wellbeing and quality of life are broad, ambiguous umbrella terms. In trials they are measured indirectly as a score derived from peoples' responses to a set of questions (items), and for any item set to measure a concept accurately, that concept must be clearly defined (to the extent possible), and conceptualized so that the relationship between the score, items and concept is transparent. Moreover, concepts may be context specific. For example, "upper limb function" may be disease dependent (MS v stroke v RhA), or differ across disease subtypes (relapsing v secondary progressive v primary progressive MS). These are empiric questions for examination, determination and clarification, and cannot be assumed. Consequently, if PRO developers wish to satisfy scientific and regulatory criteria, and trialists wish to use fit-for-purpose PROs, developers must define exactly what they seek to measure and the intended context in which the PRO is used. It follows logically that conceptual work and qualitative research should be undertaken in clearly defined samples representing the intended context of use.

Multiple factors may have driven this widespread neglect of concepts and contexts in PRO development. Health measurement evolved from educational measurement and ability testing, where concept and context issues may be less complicated (e.g. maths tests for children). Standard psychometric textbooks predominantly concern statistical methods with minimal sections on "content validity", with typically nothing devoted to defining and conceptualizing variables for measurement [16–19]. Hence, there is limited guidance for PRO developers. Finally, there has been circular thinking, where clinical concepts are "defined" by the items used in the absence of underpinning conceptual work.

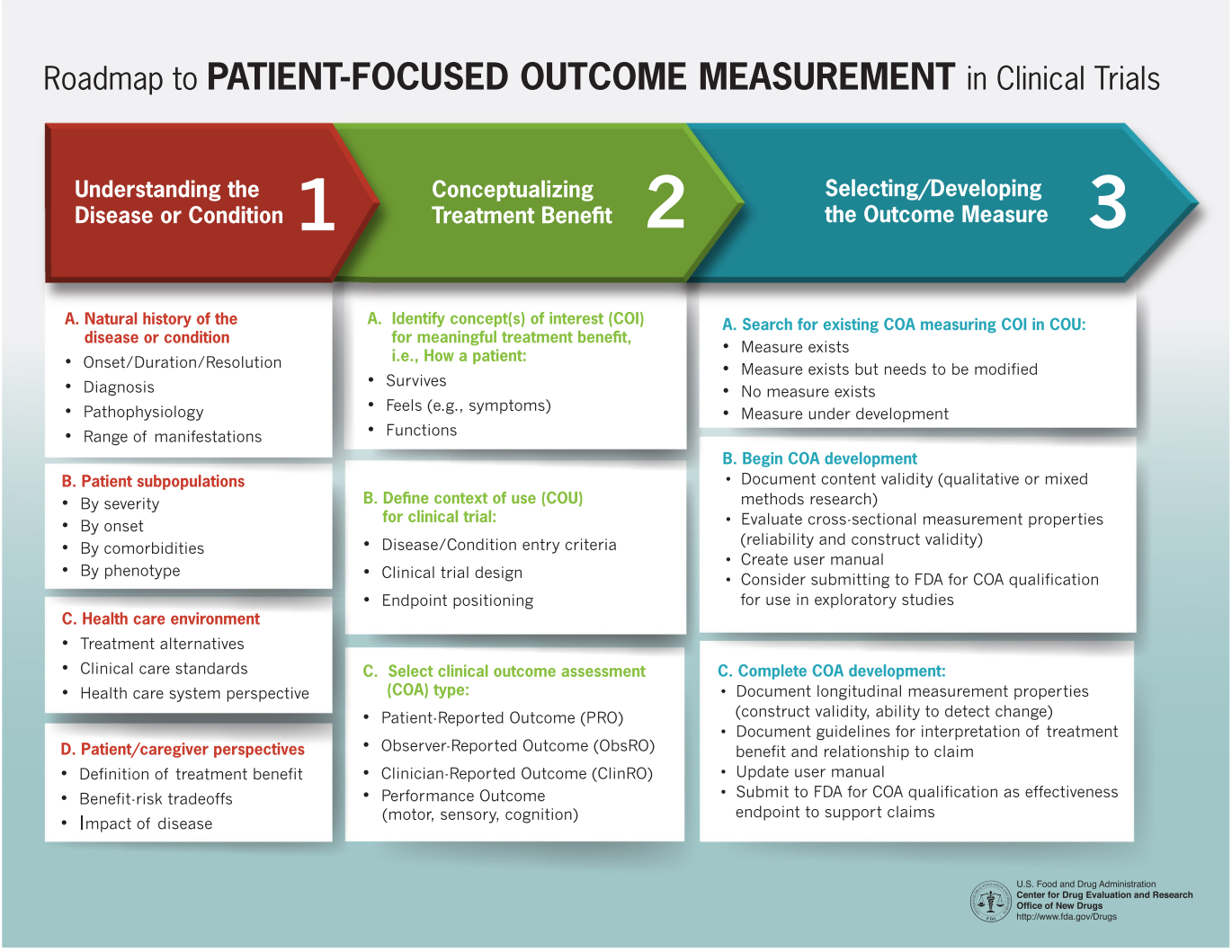

Fig. 2. Roadmap to patient-focused outcome measurement in clinical trials, FDA 2013.

Our review also highlights an over reliance and over interpretation of statistical tests in concluding PRO validity. For example, correlations with other measures and variables, known group differences validity and hypothesis testing do not, despite claims, indicate what is being measured, nor the quality with which a construct is being measured [20–22].

Many scales have an item reduction stage in their development. Typically, an item set is administered to a sample, response data analysed and items failing statistical criteria discarded. The advantage of this approach is that binary applications of statistical criteria is easy. The disadvantage is that results are sample-specific. This can lead to different PROs from the same pool in different diseases. We think it more correct when psychometric statistics are used as hypothesis tests. Specifically, when PROs have strong conceptual bases the PRO is a hypothesis of how a variable might be measured. Testing PRO data against statistical tests identifies discrepancies from statistical requirements and indicate (diagnose) issues in the hypothesis for investigation [23]. In this regard Rasch measurement theory (RMT), as articulated by Andrich and Rasch, is the most advantageous statistical method as it articulates, mathematically, requirements for measurement from PROs [24].

Our study has limitations. Our searches were only performed up until November 2017, and due to difficulties comprehensively identifying all PRO development papers with systematic literature searches, our review may be incomplete. Furthermore, we only evaluated original validation papers. According to FDA guidance,[2] construct validity could, in theory, be established *post-hoc*. However, his has not been done for the 24 PROs identified to our knowledge, and certainly not in MS. We only assessed the content development of original PRO references for the first six items of the revised COSMIN checklist as these are a pre-requisite for subsequent psychometric evaluations. Indeed, COMSIN's guidance is that unless an adequate target population is involved in PRO development the *overall* development is deemed “inadequate”. Similarly, COSMIN's protocol uses a ‘worst score counts’ method, because poor study methodological aspects aren't overcome by sound psychometric profiles [10,11]. This view mirrors FDA's stance that other types of validity or reliability testing will not overcome problems with content validity. Due to the systematic shortcomings across all ULF PROs, examining further psychometric evaluations was deemed redundant, and therefore beyond the scope of this study.

MS clinical trials must use fit for purpose ULF PROs to ensure results accurately reflect treatment effects; failure to do so will have damaging implications for people with MS, health care funders, and sponsors by undermining treatment development and licensing. Specifically, research is required to define and conceptualise ULF in MS, identify important ULF concepts for measurement, articulate those concepts as cohesive item sets, and determine empirically whether and how they differ across MS types. This work is required to determine the suitability of existing PROs, in addition to providing solid conceptual foundations for a next generation of fit-for-purpose PROs. Without this work, MS trials measuring ULF will have type II error of unquantifiable magnitude. The current state of play is indeed “worrisome” [10,11].

Funding

This work was supported by Jeremy Hobart research funds, with JC being further supported by the National Institute for Health Research (NIHR) Applied Research Collaboration (ARC) South West Peninsula. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Declaration of Competing Interest

None.

Acknowledgements

We thank Georgina Rule (GR), Jessica Mills (JM), and Louise Barrett (LB) who contributed to aspects of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ensci.2020.100237.

References

[1] J.F. Kurtzke, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS), Neurology. 33 (11) (1983) 1444–1452.
[2] Y. Zhang, A. Salter, G. Cutter, O. Stupsilonve, Clinical trials in multiple sclerosis: milestones, Ther. Adv. Neurol. Disord. 11 (2018) (1756286418785499).
[3] Thurstone L. Attitudes can be measured. Am. J. Sociol.33(4):529–54.
[4] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, et al., The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes, J. Clin. Epidemiol. 63 (7) (2010) 737–745.
[5] C.B. Terwee, C.A.C. Prinsen, A. Chiarotto, M.J. Westerman, D.L. Patrick, J. Alonso, et al., COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study, Qual. Life Res. 27 (5) (2018) 1159–1170.
[6] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, et al., The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study, Qual. Life Res. 19 (4) (2010) 539–549.
[7] C.A.C. Prinsen, L.B. Mokkink, L.M. Bouter, J. Alonso, D.L. Patrick, H.C.W. de Vet, et al., COSMIN guideline for systematic reviews of patient-reported outcome measures, Qual. Life Res. 27 (5) (2018) 1147–1157.
[8] J.M. Valderas, M. Ferrer, J. Mendivil, O. Garin, L. Rajmil, M. Herdman, et al., Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures, Value Health 11 (4) (2008) 700–708.
[9] C.B. Terwee, C.A. Prinsen, A. Chiarotto, Vet HCD, L.M. Bouter, J. Alonso, et al., COSMIN methodology for assessing the content validity of PROMs, User manual. version 1.0 (2018).
[10] A. Chiarotto, R.W. Ostelo, M. Boers, C.B. Terwee, A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain, J. Clin. Epidemiol. 95 (2018) 73–93.
[11] A. Chiarotto, C.B. Terwee, S.J. Kamper, M. Boers, R.W. Ostelo, Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review, J. Clin. Epidemiol. 102 (2018) 23–37.
[12] E.M. Perfetto, E.M. Oehrlein, M. Boutin, S. Reid, E. Gascho, Value to whom? The patient voice in the value discussion, Value Health 20 (2) (2017) 286–291.
[13] Organization WH, International Classification of Functioning, Disability and Health, (2001).
[14] Administration FaD, Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, (2009).
[15] Administration FaD, Roadmap to Patient-focused Outcome Measurement in Clinical Trials, (2013).
[16] J.C. Nunnally, I.H. Bernstein, Psychometric Theory, 3rd ed., xxiv McGraw-Hill, New York, 1994 (752 p.).
[17] D.L. Streiner, G.R. Norman, J. Cairney, Health measurement scales : a practical guide to their development and use, Fifth edition, xiii Oxford University Press, Oxford, 2015 (399 pages).
[18] A. Anastasi, S. Urbina, Psychological Testing, 7th ed, xiii Prentice Hall, Upper Saddle River, N.J., 1997 (721 p).
[19] R.L. Thorndike, Applied Psychometrics, ix Houghton Mifflin, Boston, 1982 (390 p).
[20] A. Stenner, M. Smith, Testing construct theories, Percept. Mot. Skills 55 (1982) 415–426.
[21] A. Stenner, M. Smith, D. Burdick, Towards a theory of construct definition, J. Educ. Meas. 20 (4) (1983) 305–316.
[22] J.C. Hobart, S.J. Cano, J.P. Zajicek, A.J. Thompson, Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations, Lancet Neurol. 6 (12) (2007) 1094–1105.
[23] J. Hobart, S. Cano, R. Baron, A. Thompson, S. Schwid, J. Zajicek, et al., Achieving valid patient-reported outcomes measurement: a lesson from fatigue in multiple sclerosis, Multiple Sclerosis (Houndmills, Basingstoke, England) 19 (13) (2013) 1773–1783.
[24] J. Hobart, S. Cano, Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods, Health Technol. Assess. 13 (12) (2009) 1–177 iii, ix-x.