

RESEARCH

Open Access



Validity of the Musculoskeletal Tumor Society Score for lower extremity in patients with bone sarcoma or giant cell tumour of bone undergoing bone resection and reconstruction surgery in hip and knee

Nikolai Sherling¹, Müjgan Yilmaz², Christina Enciso Holm², Michael Mørk Petersen^{2,3} and Linda Fernandes^{1*}

Abstract

Background The Musculoskeletal Tumor Society Score (MSTS) is widely used to evaluate *functioning* following surgery for bone and soft-tissue sarcoma. However, concerns have been raised about its content validity due to the lack of patient involvement during item development. Additionally, literature reports inconsistent results regarding data quality and structural validity. This study aimed to evaluate content, structural and construct validity of the Danish version of the MSTS for lower extremity (MSTS-LE).

Methods The study included patients from three complete cohorts ($n = 87$) with bone sarcoma or giant cell tumour of bone who underwent bone resection and reconstruction surgery in hip and knee. Content validity was evaluated by linking MSTS items to frameworks of *functioning*, core outcome sets and semi-structured interviews. Data quality, internal consistency and factor analysis were used to assess the underlying structure of the MSTS. Construct validity was based on predefined hypotheses of correlation between the MSTS and concurrent measurements.

Results Content validity analysis revealed concerns regarding the MSTS. The MSTS did not sufficiently cover patient-important functions, the item Emotional acceptance could not be linked to the framework of *functioning*, the items Pain and Emotional acceptance pertained to domains beyond *functioning* and items' response options did not match items. A two-factor solution emerged, with the items Pain and Emotional acceptance loading highly on a second factor distinct from *functioning*. Internal consistency and construct validity showed values below accepted levels.

Conclusion The Danish MSTS-LE demonstrated inadequate content validity, internal consistency, and construct validity. In addition, our analyses did not support unidimensionality of the MSTS. Consequently, the MSTS-LE is not a simple reflection of the construct of *functioning* and the interpretation of a sum score is problematic. Clinicians and researcher should exercise caution when relying solely on MSTS scores for assessing lower extremity function.

*Correspondence:
Linda Fernandes
life@kp.dk

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Alternative outcome measurements of *functioning* should be considered for the evaluation of postoperative function in this patient group.

Keywords Musculoskeletal tumor society score, Bone sarcoma, Validity, Physical function, Outcome measures

Background

One commonly used outcome measurement in patients treated surgically for soft tissue or bone sarcoma is the Musculoskeletal Tumor Society Score (MSTS) [1]. The MSTS was developed to evaluate postoperative function aiming to permit comparisons of end-results from different surgical treatments [1, 2]. To be useable, an outcome measurement should demonstrate content validity, i.e. include items important to the patient group and items relevant for the construct to be measured [3]. No study has asked patients with bone sarcoma what functions and activities in daily life they consider important and compared that to the MSTS. Additionally, the construct of the MSTS has not been compared to established frameworks. Therefore, it is unknown whether the items of the MSTS measure functions and activities that are important to the patient group or whether they reflect the construct of *functioning*. Despite the lack of evidence for content validity, the MSTS for the lower extremity (MSTS-LE) has shown consistent results for construct validity, with moderate to high correlations to other measurements of *functioning*, for example to the Toronto Extremity Salvage Score (TESS) and the Short Form 36 physical function [4–7]. Conversely, the results for internal structure of the MSTS are inconsistent. Some studies have found ceiling effects for both MSTS single items and sum score [4, 6, 7] while others did not find ceiling effects [5, 8]. Three studies have tested the MSTS for structural validity, with the conclusion of a one-factor solution, i.e., unidimensionality, for the MSTS [5, 6, 9]. However, the studies showed eigenvalues close to cut-off for a two-factor solution and all three studies showed moderate to low factor loadings for the items Pain and Emotional acceptance [5, 6, 9]. Based on results of low factor loadings for Pain and Emotional acceptance, one could question their reflection of *functioning*. If the MSTS is to be used in future research and clinical practice for the evaluation of *functioning*, further evidence of its ability to reflect *functioning* is needed.

The aim of this study was therefore to evaluate the validity of the MSTS-LE, more specifically its content validity, data quality, internal consistency, structural and construct validity.

Methods

Design, inclusion, and patient characteristics

Data for this project was extracted from three cohorts including patients with bone sarcoma or giant cell tumour of bone in the lower extremity going through

bone tumour resection and reconstruction with a tumour prosthesis in the hip or knee (Table 1). Assessments were completed once for each patient, i.e., the study design was cross-sectional. All three cohorts ($n=87$) were used for analyses of internal structure. Cohort one ($n=30$) was in addition tested for content and construct validity.

Cohort one included 30 patients enrolled from a complete cohort of 72 patients [10]. The patients had undergone surgery between 2006 and 2016 at the Musculoskeletal Tumor Section, Department of Orthopedic Surgery, University Hospital Rigshospitalet, Copenhagen. The included patients ($n=30$) were interviewed using the Patient Specific Functional Scale (PSFS) and assessed using the MSTS and concurrent outcome measurements at mean 7 (range, 2–12) years after surgery.

Cohort two included 24 patients enrolled from a complete cohort of 50 patients [11]. The patients had undergone surgery between 1985 and 2005 at the Musculoskeletal Tumor Section, Rigshospitalet, Copenhagen. The included patients ($n=24$) were assessed using the MSTS at mean 15 (range, 4–29) years after surgery.

Cohort three included 33 patients enrolled from a national cohort using the Global Modular Replacement System (GMRS) only as tumour prosthesis for reconstruction of bone [12]. The patients had undergone surgery between 2005 and 2013 at the Musculoskeletal Tumor Sections at Rigshospitalet, Copenhagen, and Aarhus University Hospital, Aarhus. The included patients ($n=33$) were assessed using the MSTS at mean 5 (range, 1–11) years after surgery.

Musculoskeletal Tumour Society Score (MSTS)

The Danish MSTS-LE was used [1, 7]. It comprises six items (Pain, Function, Emotional acceptance, Supports, Walking ability and Gait) and is administered by a clinician. Each item is scored on a 5-point Likert scale, ranging from 0 (worst possible score) to 5 (best possible score) [1]. The items have unique response options for 0 through 5 (Table 2). A sum score for the six items is calculated (maximum 30 points) and normalised to a 0–100 score.

Semi-structured interview used for the evaluation of content validity

Patient Specific Functional Scale (PSFS) is designed to identify patient-important functions and activities. It is valid for use in numerous diseases and conditions and can be administered as a semi-structured interview or as a patient reported outcome (PRO) [13, 14]. We chose

Table 1 Patient characteristics of the three included cohorts (n = 87)

	Cohort 1 n = 30	Cohort 2 n = 24	Cohort 3 n = 33	Difference between cohorts (p-value)
Sex				0.30
Male	16 (53)	17 (71)	17 (52)	
Female	14 (47)	7 (29)	16 (48)	
Age at MSTs scoring	51 (19–83)	50 (12–79)	42 (24–75)	0.16
Years since surgery	7 (2–12)	15 (4–29)	5 (1–11)	< 0.001
MSTS score (0–100)	66 (27–93)	71 (27–90)	73 (30–97)	0.34
Histology				0.01
Osteosarcoma	9 (30)	15 (63)	11 (33)	
Chondrosarcoma	14 (47)	1 (4)	9 (27)	
Ewing sarcoma	1 (3)	3 (13)	1 (3)	
Giant cell tumour	3 (10)	5 (21)	7 (21)	
Other*	3 (10)	-	5 (15)	
Surgery				< 0.01
Proximal femur	12 (40)	2 (8)	5 (15)	
Distal femur	14 (47)	14 (58)	15 (45)	
Proximal tibia	4 (13)	6 (25)	13 (39)	
Total femur	-	2 (8)	-	
Prosthesis				< 0.001
GMRS	20 (67)	1 (4)	33 (100)	
Segmental	8 (27)	-	-	
Mega C	2 (7)	-	-	
HMRS	-	17 (71)	-	
Endorotational System	-	4 (17)	-	
Not identified	-	2 (8)	-	
Major revision**	3 (10)	12 (50)	5 (15)	< 0.001

Values presented as n (%) or mean (min–max). *synovial-, myofibro-, angio-, and leiomyosarcoma with osseous involvement. **change or removal of bone-anchored parts of the implant. MSTs Musculoskeletal Tumor Society Score. GMRS Global Modular Replacement System from Stryker; Segmental from Zimmer Biomet; Mega C from Link; HMRS Howmedica Modular Reconstruction System from Stryker; Endorotational System from Link. One way ANOVA was applied for analyses of differences for continuous variables and Chi-square for categorical variables

the semi-structured interview modality, carried out by a physiotherapist (LF). The patients were asked to identify up to five important functions or activities they were unable to perform or had difficulties with because of the condition. Once identified, the activities were categorised and listed. Activities that included the same type of movement were categorised into a meaningful concept [15]. For example, “walking on uneven surfaces”, “walking fast” or “walking long distances” were categorised into *Walking*. *Sport* represented any sports-related activity, for example playing golf, water polo, or swimming. Running was a separate category, as it could be either sports related or a means of moving quickly from one place to another, e.g., run to catch a bus. After identifying important activities, the patients were asked to score

them for level of difficulty on a 11-point scale (0=unable to perform the activity, 10=able to perform activity at the same level as before surgery) [13, 16]. For each category, the number of patients identifying the activity and the median level of difficulty was presented in a chart. Individual mean PSFS scores were also used for the evaluation of construct validity.

Concurrent outcome measurements

Numeric Rating Scale (NRS) is a valid and widely used tool for the measurements of pain intensity among patients with varying conditions [17, 18]. The patients were asked to score current pain intensity (0=no pain, 10=worst pain imaginable).

Toronto Extremity Salvage Score (TESS) is a patient-specific PRO developed to account for the heterogeneity of *functioning* in patients with bone and soft-tissue sarcoma [19, 20]. It is unidimensional and comprises 30 questions about daily tasks, work/school and leisure time [19]. Difficulties performing the activities are scored on a 5-point Likert scale (1=impossible to do, 5=not at all difficult). The total score is calculated as a percentage of the maximum score. The Danish version has shown acceptable comprehensibility, test-retest reliability and construct validity [21].

The EORTC QLQ-C30 is a multidimensional PRO measuring quality of life (QoL) in patients with cancer [22]. The QLQ-C30 is widely used, shows robust psychometric properties, has population-based reference data and is translated into Danish [22–24]. It consists of 30 questions scored on Likert scales [4]. We used the sum score and sub scores for physical functioning, emotional functioning and pain in the analyses, normalised to 0–100-points. A high sum score represents a high QoL, a high functioning score represents high levels of functioning and a high pain score represents high levels of pain [22, 24].

30-second chair stand test (CST) assesses muscle power and strength of the lower extremity, it can predict deterioration of function and can be used in people with different diseases and ages [25–30]. It has shown good reliability (ICC>0.80) and a measurement error of 1 repetition [26]. The patients were asked to stand up and sit down from a 45 cm chair as many times as possible in 30 s. A standardised protocol from the Association of Danish Physiotherapists was used.

6-minute walk test (6MWT) assesses walking capacity and has been used on patients with numerous diagnoses, including bone sarcoma [10, 26, 31–35]. It has shown ICCs of >0.90 and measurement errors between 14 and 30 m [26, 34]. The patients were asked to walk as fast as possible back and forth on a 20-meter walking track in an enclosed corridor at the hospital. A standardised

Table 2 Items and response options of the Musculoskeletal Tumour Society Score for lower extremity

Score	Pain	Function	Emotional acceptance	Supports	Walking ability	Gait
5	No pain / no medication	No restriction / no disability	Enthusied / would recommend to others	None / no supports	Unlimited / same as preoperative	Normal / no alterations
4	Intermediate	Intermediate	Intermediate	Intermediate / occasional use of brace	Intermediate	Intermediate
3	Modest / non-disabling / non-narcotic analgesic	Recreational restriction / minor disability	Satisfied / would do again	Brace / mostly brace	Limited / significantly less	Minor cosmetic / cosmetic alterations only
2	Intermediate	Intermediate	Intermediate	Intermediate / occasional cane/crutch	Intermediate	Intermediate
1	Moderate / intermittently disabling / intermittent narcotics	Partial occupational restriction / major disability	Accepts / would repeat reluctantly	One cane or crutch / mostly cane/crutch	Inside only / cannot walk outside	Major cosmetic / minor functional deficit
0	Severe / continuously disabling / continuous narcotics	Total occupational restriction / Complete disability	Dislikes / would not repeat	Two canes or crutches / always cane/crutches	Not independently / can walk only with assistance or wheelchair bound	Major handicap / major functional deficit

protocol from the Association of Danish Physiotherapists was used.

Analysis

Demographic data, PRO scorings, and physical tests were presented as number (%), mean (SD), median (range) values as appropriate for different scales. A sample size of at least five observations per item and at least 100 observations has been suggested for determining structural validity [36–38]. We were able to include 87 patients from three complete cohorts between 1985 and 2016. The MSTS scorings had no missing data. For the data collection of concurrent measurements, there was one patient in Cohort 1 that declined physical tests (CST, 6MWT) at the hospital because of logistical reasons, and one patient had internally missing data in QLQ-C30 physical functioning. Different statistical analyses were applied for different psychometric evaluations. Analyses were performed using IBM SPSS v.29.

Content validity is defined as the degree to which the content of a PRO is an adequate reflection of the construct to be measured [36, 39]. Since the MSTS intends to measure *functioning* [1], the six items and their response options should be a reflection of *functioning*. An international consensus for quality rating of PROs has recommend three overarching criteria for the evaluation of content validity: relevance, comprehensiveness, and comprehensibility [3]. Relevance includes an evaluation of items’ relevance for the construct and the population of interest. To evaluate items’ relevance for the construct of *functioning*, MSTS items were listed and, wherever possible, linked to codes of the International Classification of Functioning, Disability and Health (ICF) [40]. To evaluate MSTS items’ relevance for the population of interest, we linked MSTS-items to activities identified in the PSFS. Comprehensiveness includes an evaluation of whether

key concepts are included in an outcome measurement. Key concepts can be found in core outcome sets [41–45]. Since there is no specific core outcome set for patients that undergo bone sarcoma surgery, we chose to link MSTS items to key concepts defined in core outcome sets for cancer and primary total knee and hip joint replacement [43, 44]. Comprehensibility was evaluated by linking response options of the MSTS to the ICF and PSFS. Response options should match items to meet quality standards [3]. The linking processes were done independently by two of the authors (NS, LF) following recommendations for ICF-linking of outcome measures [15].

Data quality. Missing data of individual items, central tendency, distribution of item-scoring and floor- and ceiling effects were described. Floor- and ceiling effects were defined as present if >15% of patients scored the lowest or highest possible score, respectively [46].

Internal consistency has been defined as the degree of interrelatedness amongst the individual items [39]. The analysis requires a unidimensional scale of at least three items [39]. If our analysis of structural validity suggested >1 dimension, internal consistency was tested separately for each dimension [46]. Inter-item correlation, item-total correlation, and Cronbach’s alpha if item deleted were determined [47]. An inter-item correlation between 0.20 and 0.50 is recommended [36]. The item-total correlations assume that patients with a high total score also have high scores on all items [36]. If an item shows an item-total correlation of <0.30 it does not help greatly in distinguishing between patients with high and low scores and can be removed. A Cronbach’s alpha if item deleted shows the value for remaining items that are still in the analysis. A high value indicates that the deleted item is redundant and a low value that there is room for more items under the same construct. A Cronbach’s

alpha value between 0.70 and 0.90 is commonly considered acceptable interrelatedness [48].

Structural validity has been defined as the degree to which the scores of a PRO are an adequate reflection of the dimensionality of the construct to be measured [39]. Initially, the data was tested for suitability for factor analysis. Inter-item correlation coefficients between 0.20 and 0.80, overall correlation of a Kaiser-Meyer-Olkin (KMO) of >0.50 (ideally >0.80) and a significant Bartlett's test of sphericity have been recommended as prerequisites for factor analysis [36, 37]. We applied a principal component analysis (PCA). The number of latent factors extracted was based on the shape of a scree plot (elbow and levelling), Kaiser's criterion (eigenvalue >1) and the cumulative percentage of explained variance after each factor (ideally 70–80%) [37, 49, 50]. Oblique rotation (direct oblimin) method was applied since our factor correlation matrix showed a coefficient above the suggested cut-off 0.32 [37, 49, 50]. There is no consensus on threshold for sufficient loading of an item to a factor, but with a sample size of at least 100 patients, a loading of >0.30 is usually considered significant [50]. Items that load substantially (>0.3) on more than one factor are called complex variables and need to be taken into consideration [50].

Construct validity is defined as the degree to which the score of an outcome measurement is consistent with hypotheses of expected relationships to other PROs [39]. High correlations are expected when measurements of the same construct and with the same mode of administration are compared (convergent). Conversely, lower correlations are expected when different constructs are compared (divergent). Previously published results of correlations between the MSTS and concurrent outcome measures were used as guidance when formulating predefined hypotheses [19]. MSTS sum scores were expected to have high correlations to scorings from TESS, PSFS and QLQ-C30 physical function, as they all measure *functioning* subjectively [7, 51]. MSTS sum score was expected to correlate at a moderate level with QLQ-C30 sum score, since it is a multidimensional measurement [52]. Concurrent measurements of more narrow constructs (e.g., pain, walk capacity, emotional function) were expected to have high correlations to single items of the MSTS but low correlations to MSTS sum score [53]. The research group formulated hypotheses of correlation prior to analyses. Cut-offs for high (≥ 0.60), moderate (>0.30 to <0.60) and low (≥ 0.30) correlation were applied [40]. For a positive rating of hypothesis testing, at least 75% of predefined hypotheses should be confirmed [46]. The Spearman's rank correlation coefficient test was used.

Results

Content validity

Semi-structured interview. The patients ($n=30$), identified a total of 94 important activities which they found impossible or difficult to perform. These single activities were categorized into 12 meaningful concepts (Fig. 1). The three most frequently identified activities were Walking ($n=14$), Sports ($n=19$) and Running ($n=20$), with median (min–max) difficulty levels of 3.5 (0–5) points, 1 (0–7) point, and 0 (0–6) points, respectively.

Items' relevance for the construct of functioning. All MSTS-items, except for Emotional acceptance, could be linked to ICF-codes (Table 3). The item Function was considered a wide concept and could be linked to any ICF-code under the domains (b) and (d).

Items' relevance for the included sample. Two of six MSTS-items could be linked to PSFS (Table 3). The MSTS-item Function could be linked to any activity identified in the PSFS.

Key concepts. The MSTS-items Pain and Functioning were linked to the different domains Pain and Function defined in both core outcome sets [43, 44]. The domain 'patient satisfaction', in the core outcome set for joint replacement, was partly linked to the MSTS-item Emotional acceptance, since one response option included the word 'satisfied'.

Comprehensiveness. The response options for the items Pain, Function and Walking Ability changed content throughout the scale (Table 3). The response options 'disabling' and 'disability' could be linked to several ICF-codes and the response option 'recreational' could be linked to several activities identified in the PSFS (Table 3).

Data quality

Item median values ranged from 3 to 5 and all response options were used (Table 4). None of the items showed floor effects, but all items, except for Function, showed ceiling effects (Table 4). There were no internal missing values.

Internal consistency

Three inter-item correlation coefficients exceeded 0.50 (Supports and Walking ability, $r=0.60$; Supports and Gait, $r=0.55$; Walking ability and Gait, $r=0.55$) (Table 5). As our PCA did not support unidimensionality, but a two-factor solution, the item-total and the Cronbach's alpha was only tested for Factor 1. The item Function showed the lowest item-total correlation ($r=0.45$) but did not fall below the limit of <0.30 (Table 5). The items Supports and Walking ability showed Cronbach's alpha, if item deleted, below accepted values between 0.70 and 0.90 (Table 5).

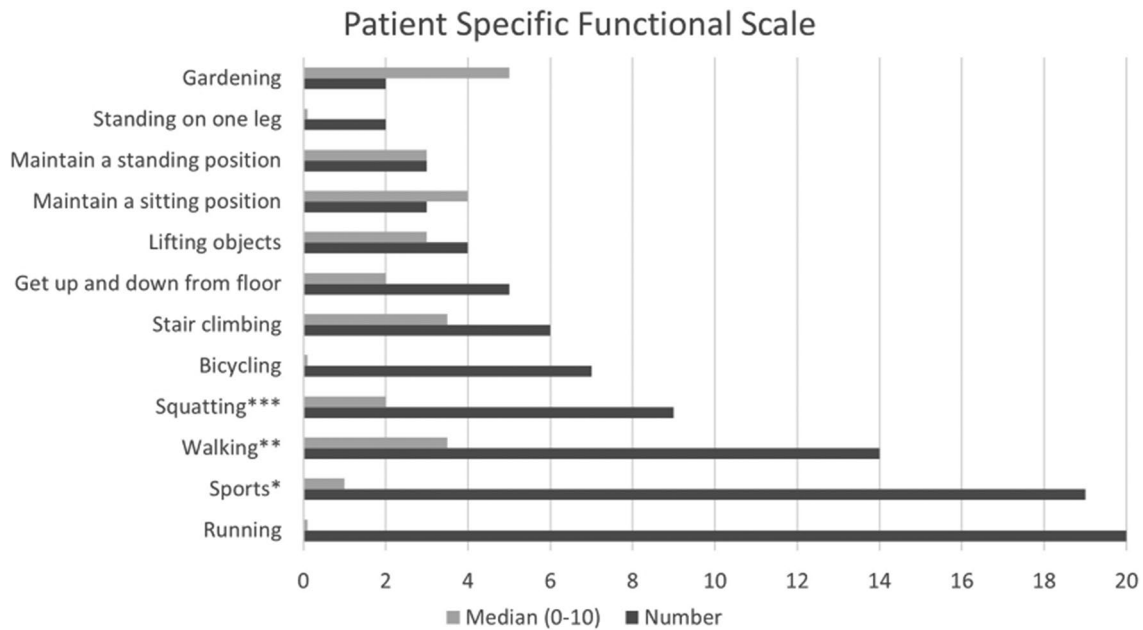


Fig. 1 Number of activities (dark grey bar) the patients found important and were unable to perform or had difficulties with because of the condition. Median score (light grey bar) of the level of difficulty ranging from 0–10 points (0=unable to perform the activity, 10=able to perform activity at same level as before surgery). ***Squatting: This includes the isometric position in a squat and the dynamic squat. **Walking: This is a summary of walking in various speeds and distances in diverse terrain. *Sports: This includes various sports activities such as soccer, swimming, golf, tennis, badminton, dancing, water polo and skiing

Table 3 Content validity. Linkage of items and response options of the MSTs to ICF and PSFS

MSTS items and response options	ICF code	ICF description	Linked to ICF	Linked to PSFS
Pain	b280	Sensation of pain	1	0
Severity	b280	Sensation of pain	1	0
Disabling	d410-d699	Mobility, Self-care, Domestic Life	>1	>1
Analgesics	e1101	Drugs	1	0
Function	b410-d999	Body Functions and Activities & Participation	>1	>1
Disability	d410-d699	Mobility, Self-care, Domestic Life	>1	>1
Recreational	d920	Recreation and leisure	1	>1
Occupational	d845	Acquiring, keeping, and terminating a job	1	0
Emotional acceptance	No code		0	0
Supports	d465	Moving around using equipment	1	0
Walking ability	d450	Walking	1	1
Walking distance	d450	Walking	1	1
Wheelchair / walking with assistance	d465	Moving around using equipment	1	0
Gait	b770	Gait pattern functions	1	0

Linked to ICF and Linked to PSFS describe whether MSTs-items and response options could be linked (1) or not (0) to ICF-codes or to activities identified in the PSFS. If an MSTs-item or response option could be linked to several ICF-codes or activities in the PSFS, the link number was >1. MSTs Musculoskeletal Tumor Society Score. ICF International Classification of Functioning, Disability and Health. PSFS Patient Specific Functional Scale

Structural validity

The inter-item correlation between Pain and Gait showed a low correlation ($r=0.19$). Since this study was not a data reduction exercise and the two items had acceptable correlations to remaining items, they were retained. The KMO was 0.79 and the Bartlett’s test was significant ($p<0.001$) suggesting adequate data for the performance of a factor analysis.

The scree plot illustrated a steep slope for Factor 1 (eigenvalue 2.904), intermediate slope for Factor 2

(eigenvalue 1.017) and almost flat slope for Factor 3 (eigenvalue 0.685) (Fig. 2). The cumulative percentage of total variance explained was 48.4% for Factor 1 and 65.4% for Factor 1 and 2. Based on eigenvalues, cumulative percent and the scree plot, a two-factor solution was suggested for the analysis of factor-loading pattern.

The factor loading pattern for a two-factor solution showed high loadings for Supports, Gait, Walking ability and Function to Factor 1, but not for Pain and Emotional acceptance (Table 6). The items Walking ability

Table 4 Data quality of the Musculoskeletal Tumor Society Score

MSTS	n	Mean	SD	Median	Min – max	Floor Effect	Ceiling Effect
Pain	87	3.67	1.15	4	0–5	1.1%	21.8%
Function	87	2.71	1.36	3	0–5	4.6%	8.0%
Emotional acceptance	87	3.59	1.19	4	0–5	1.1%	25.3%
Supports	87	3.85	1.71	5	0–5	5.7%	62.1%
Walking ability	87	3.77	0.96	4	0–5	1.1%	21.8%
Gait	87	3.39	1.34	3	0–5	3.4%	26.4%
Sum score	87	20.98	5.32	22	6–30	1.1%	2.3%

Floor- and ceiling effect represent the percentage of patients scoring the lowest and highest response options

Table 5 Inter-item correlation of all six MSTS items (n = 87)

Inter-item correlation							Item-total	Cronbach's alpha
	1	2	3	4	5	6		
1	1.00	0.28	0.39	0.25	0.45	0.19	-	-
2		1.00	0.32	0.39	0.46	0.31	0.45	0.77
3			1.00	0.21	0.40	0.26	-	-
4				1.00	0.60	0.55	0.64	0.67
5					1.00	0.55	0.69	0.68
6						1.00	0.58	0.70

Item-total correlation and Cronbach's alpha if item deleted for Factor 2 including the four items Function, Supports, Walking ability and Gait. MSTS-items: 1. Pain. 2. Function. 3. Emotional acceptance. 4. Supports. 5. Walking ability. 6. Gait

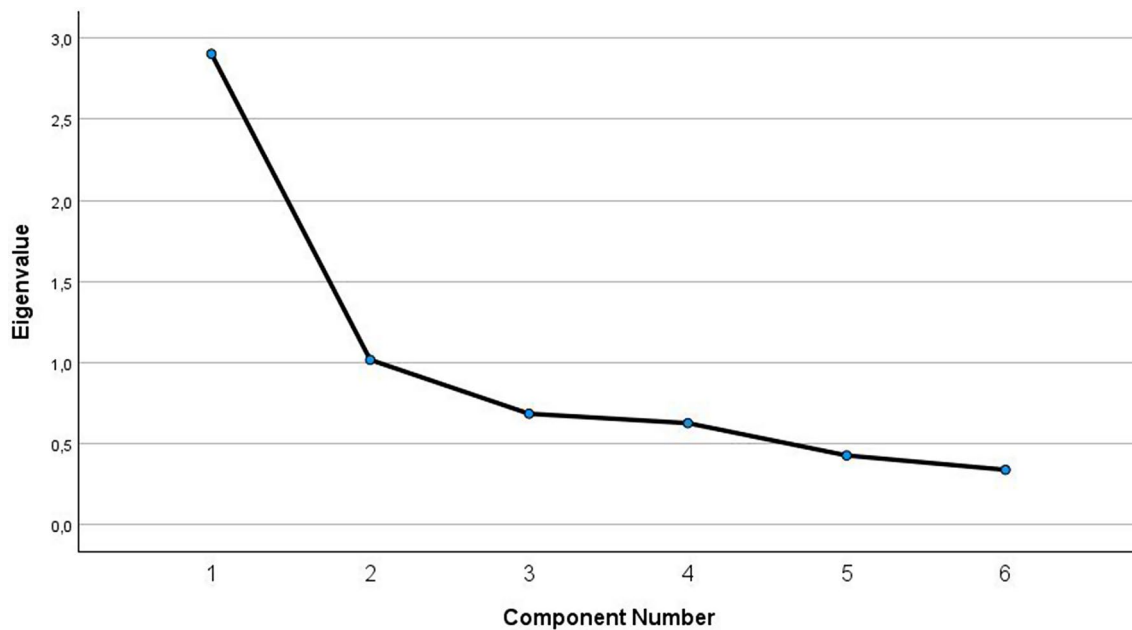


Fig. 2 Scree plot of the principal component analysis

and Function loaded >0.30 on two factors, thus identified as complex variables.

Construct validity

Six out of 13 (46%) predefined hypotheses were ascertained (Table 7). The TESS, QLQ-C30 sum score, QLQ-C30 physical functioning sub score and pain ratings showed high correlations to the MSTS (Table 7). The QLQ-C30 sum score was not expected to have a high

correlation to the MSTS, since it measures QoL and not *functioning* only. The MSTS showed a low correlation to the PSFS, which was unexpected since both should reflect the construct of *functioning*. Also, the MSTS item Walking ability had an unexpectedly low correlation to walking capacity (6MWT).

Table 6 MSTS factor loading pattern by principal components analysis with loadings sorted by size (n = 87)

Items of the MSTS		Factor	
		1	2
4	Supports	0.89	-0.08
6	Gait	0.87	-0.11
5	Walking ability	0.69	0.32
2	Function	0.43	0.36
1	Pain	-0.03	0.83
3	Emotional acceptance	-0.01	0.81

Extraction method: Principal components analysis using oblique rotation method (direct oblimin) with Kaiser normalization. Factor loadings ≥ 0.3 are in bold print. Rotation converged in six iterations MSTS Musculoskeletal Tumor Society Score

Table 7 Hypotheses testing (n = 30)

	MSTS sum score	Expected level and direction of correlation	Hypothesis-confirmed
TESS	0.73	High (+)	Yes
PSFS	-0.16	High (+)	No
QLQ-C30 physical functioning	0.80	High (+)	Yes
QLQ-C30 sum score	0.71	Moderate (+)	No
NRS current pain	-0.68	Low (-)	No
6MWT	0.15	Low (+)	Yes
CST	0.34	Low (+)	No
MSTS Pain			
NRS current pain	-0.66	High (-)	Yes
QLQ-C30 pain	-0.68	High (-)	Yes
MSTS Function			
TESS	0.51	High (+)	No
QLQ-C30 physical functioning	0.62	High (+)	Yes
MSTS Emotional acceptance			
QLQ-C30 emotional functioning	0.55	High (+)	No
MSTS Walking ability			
6MWT	0.14	Moderate (+)	No

Spearman's correlation coefficients (rho) between MSTS sum score and concurrent measurements and the MSTS-items Pain, Emotional accept, Function and Walking ability and concurrent measurements. Expected positive (+) or expected negative (-) direction of correlation. MSTS Musculoskeletal Tumor Society Score. TESS Toronto Extremity Salvage Score. PSFS Patient Specific Functional Scale. QLQ-C30 EORTC Quality of Life Questionnaire Cancer-30. NRS Numeric Rating Scale. 6MWT Six-minute walk test. CST 30-second chair stand test

Discussion

The MSTS-LE showed insufficient content validity. The internal consistency and hypothesis testing were below acceptable levels. We found ceiling effects in five of six items and, in contrast to other studies, our analyses supported a two-factor solution.

The evaluation of content validity showed that there were concerns with the three quality criteria; relevance, comprehensiveness, and comprehensibility. The item

Emotional acceptance was not relevant to the construct of *functioning*, the item Function was relevant to the construct but had a too broad and unspecific content. Pain and Function should pertain to separate constructs. Three items did not have matching response options and many patient-important activities identified in the interview were not represented in the MSTS. The MSTS has been criticised for not involving patients' perception of function in the development of items and response options [19]. We used a semi-structured interview to evaluate the MSTS-items' relevance to the population of interest. Our results showed that the patients reported many more functions and activities that were important for them, than those in the MSTS. For example, recreational activities such as gardening, bicycling, hiking, and different sports activities were considered important, but not specifically named in the MSTS. For the measurement of *functioning*, an alternative to the MSTS could be the TESS [54]. The items of the TESS were development based on input from patients with bone and soft-tissue sarcoma [19]. Comparing the TESS to our interview, the TESS includes kneeling, walking, gardening, and recreational activities also found in our interviews, suggesting that TESS has a more relevant content than the MSTS for this patient group. The evaluation of the items' relevance to the construct of *functioning* showed that the item Emotional acceptance could not be linked to the ICF. This suggests that Emotional acceptance does not reflect *functioning* and should not be part of PROs with *functioning* as the construct of interest. Further, the linking process of the item Function was of concern, as it could be linked to many ICF-codes, reflecting several functions, resulting in a very broad and unspecific content. This was supported by a relatively low item-total correlation for the item Function, suggesting that the content is unspecific and there is scope for more items under the same construct [48]. An unspecific content will make interpretation difficult. The items Pain and Function were linked to important key concepts, but were defined as two separate domains in the core outcome sets suggesting that they reflect different constructs [43, 44]. When separate constructs are measured, they should either pertain to different PROs, or they should be treated separately in multidimensional scales [36]. Based on the unspecific content of the item Function and the potential mix of different constructs within the MSTS-LE, a sum score should be interpreted with caution. Moreover, the evaluation of comprehensibility of response options showed similar results as Lee et al., with concerns about the formulations for Pain, Function and Walking ability [4]. The item Pain relates to the intake of analgesics rather than the perception of pain and the items Function and Walking ability change content throughout the scale. One main requirement in formulating items and

their response options is that they should be simple, easy to understand and the response options should match their items [3]. Since the response options for three items of the MSTS-LE change in content, the response options are difficult to interpret and do not match the items.

The factor analysis in our study supported a two-factor solution. In contrast to our results, two earlier studies considered the MSTS to be unidimensional, i.e. consisting of one factor only [5, 6]. The scree plots in the two earlier studies showed elbow shapes located at the second factor, similar to our study, but their eigenvalues for the second factors were just below 1, whereas ours was just above 1. Determining the number of factors, and thereby the dimensionality of a measurement, can be difficult when scree plots do not take a characteristic sharp elbow shape and eigenvalues are close to the cut-off value 1. One of the earlier studies discussed the possibility of a two-factor solution but decided to let the eigenvalue < 1 for a second factor determine the unidimensionality of the MSTS [5]. Values close to cut-offs can lead to different conclusions in different studies, which in this case indicates that the MSTS is not sufficiently robust between samples. Further, looking at the factor-loading patterns, it is doubtful whether the MSTS can be supported as a unidimensional measurement of *functioning*. Our study clearly showed that the items Pain and Emotional acceptance had low loadings to Factor 1 and high loadings to Factor 2, indicating that Pain and Emotional acceptance are explained by another underlying construct than *functioning*. This is supported by the three earlier studies showing lower loadings for Pain and Emotional acceptance compared to the other items of the MSTS although, they never tested a two-factor solution and investigated whether Pain and Emotional acceptance had a better fit to a second factor [5, 6, 9]. Since Pain and Emotional acceptance can be vaguely explained by the underlying construct of *functioning*, they should be treated as a separate factor. The MSTS-LE should therefore not be considered a unidimensional measurement of *functioning*, but rather a multidimensional measurement where the dimensions should be treated separately with separate subscores rather than a sum score, as is current practice.

One limitation in our study was sample size. It is recommended that at least 100 patients are included when performing factor analyses [36–38]. With the data available ($n=87$) one could consider increasing the limit for an item to contribute sufficiently to a factor from >0.30 to >0.50 [36]. By doing so, the items Function would not load sufficiently to Factor 1. This leaves the item Function a complex variable only, not pertaining to any of the two factors, which complicates the interpretation of the MSTS even further. Another limitation is time from surgery to assessment point. In all three cohorts

time from surgery varied widely and for most included patients many years had elapsed. Time from surgery can affect which patients could be included from the complete cohorts. Because around 60–80% of the patients in the three cohorts were alive at inclusion [11, 12, 55], the cohorts could comprise patients with a better outcome of physical function than the background population. Including a subgroup with a better function from the total population has presumably biased the results to better scorings of the MSTS and can possibly explain our high ceiling effects.

Conclusions

The MSTS showed insufficient content validity and when asking patients, other functions than those included in the MSTS were of importance. Our findings do not support the MSTS as a unidimensional measurement of *functioning*, but a two-factor solution. Thus, MSTS sum scores should be interpreted with caution. We suggest that alternative outcomes, such as the TESS and objective measurements, are considered for the evaluation of *functioning* in clinical practice and future research.

Abbreviations

MSTS	Musculoskeletal Tumor Society Score
MSTS-LE	Musculoskeletal Tumor Society Score – Lower Extremity
TESS	Toronto Extremity Salvage Score
GMRS	Global Modular Replacement System
PSFS	Patient Specific Functional Scale
PRO	Patient Reported Outcome
NRS	Numeric Rating Scale
QLQ-C30	The EORTC Quality of Life Questionnaire Core 30, v. 3
QoL	Quality of Life
CST	30-second chair stand test
6MWT	6-minute walk test
ICF	International Classification of Functioning, Disability and Health
KMO	Kaiser-Meyer-Olkin
PCA	Principal Component Analysis

Acknowledgements

The authors thank the patients participating in the three cohorts included.

Author contributions

N.S., M.M.P. and L.F. contributed to concept and design. M.Y., C.E.H. and L.F. collected the data. N.S. and L.F. contributed to data analysis, statistical analysis, and manuscript preparation. N.S. and L.F. contributed to literature search, manuscript editing, and manuscript review. M.Y., C.E.H. and M.M.P. revised the manuscript. All authors reviewed and approved the manuscript.

Funding

LF received funding from Vissing Fonden, Aalborg, Denmark (grant number 85969).

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was carried out according to the Helsinki Declaration. Approvals from the Danish Data Protection Agency (VD-2018-20-6594, 2013-41-2591, 2013-41-2591), the Capital Regional Committee on Health Research Ethics (H-18032141) and the Danish Health and Medicines Authority (no. 3-3013-

894/1, 3-3013-1045/1/) were obtained prior to inclusion. Patients in Cohort 1 and 2 received oral and written information and signed informed consent prior to inclusion. In case of minors under the age of 16, informed consent to participate was obtained from the parents. Cohort 3 was a retrospective cohort, using data from registers. In Denmark, informed consent in registered based studies is deemed unnecessary according to national legislation [56].

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Midwifery, Physiotherapy, Occupational Therapy, and Psychomotor Therapy, Faculty of Health, University College Copenhagen, Copenhagen, Denmark

²Musculoskeletal Tumor Section, Department of Orthopedic Surgery, University Hospital Rigshospitalet, Copenhagen, Denmark

³Institute of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Received: 29 August 2023 / Accepted: 24 July 2024

Published online: 16 August 2024

References

- Enneking WF, Dunham W, Gebhardt MC, Malawar M, Pritchard DJ. A system for the functional evaluation of reconstructive procedures after surgical treatment of tumors of the musculoskeletal system. *Clin Orthop Relat Res* 1993;241–6.
- Amino K, Kawaguchi N, Matsumoto S, Manabe J, Furuya K, Isobe Y. Functional Evaluation of Limb Salvage Operation for Malignant Bone and Soft Tissue Tumors Using the Evaluation System of the Musculoskeletal Tumor Society. *New Developments for Limb Salvage in Musculoskeletal Tumors*. 1989:27–30.
- Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research* 2018;March.
- Lee SH, Kim DJ, Oh JH, Yoo KH, Kim HS. Validation of a functional evaluation system in patients with musculoskeletal tumors. *Clin Orthop Relat Res* 2003;217–26.
- Rebolledo DCS, Vissoci JRN, Pietrobon RP, de Camargo OP, Baptista AM. Validation of the Brazilian version of the musculoskeletal tumor society rating scale for lower extremity bone sarcoma. *Clin Orthop Relat Res*. 2013;471:4020–6.
- Iwata S, Uehara K, Ogura K, Akiyama T, Shinoda Y, Yonemoto T et al. Reliability and Validity of a Japanese-language and Culturally Adapted Version of the Musculoskeletal Tumor Society Scoring System for the Lower Extremity. *Clin Orthop Relat Res*. 2016;474:2044–52. 2016.
- Saebye CKP, Keller J, Baad-Hansen T. Validation of the Danish version of the musculoskeletal tumour society score questionnaire. *World J Orthop*. 2019;10:23–32.
- Janssen SJ, Pereira NRP, Raskin KA, Ferrone ML, Hornicek FJ, van Dijk CN, et al. A comparison of questionnaires for assessing physical function in patients with lower extremity bone metastases. *J Surg Oncol*. 2016;114:691–6.
- Mallet J, El Kinani M, Crenn V, Ageneau P, Berchoud J, Varenne Y et al. French translation and validation of the cross-cultural adaptation of the MSTs functional assessment questionnaire completed after tumor surgery. *Orthopaedics & Traumatology: Surgery & Research*. 2023;109(3):103574.
- Fernandes L, Holm CE, Villadsen A, Sørensen MS, Zebis MK, Petersen MM. Clinically important reductions in physical function and quality of life in adults with Tumor prostheses in the hip and knee: a cross-sectional study. *Clin Orthop Relat Res*. 2021;479:2306–19.
- Holm CE, Bardram C, Riecke AF, Horstmann P, Petersen MM. Implant and limb survival after resection of primary bone tumors of the lower extremities and reconstruction with mega-prostheses fifty patients followed for a mean of fourteen years. *Int Orthop*. 2018;42:1175–81.
- Yilmaz M, Sørensen MS, Saebye CKP, Baad-Hansen T, Petersen MM. Long-term results of the global modular replacement system tumor prosthesis for reconstruction after limb-sparing bone resections in orthopedic oncologic conditions: results from a national cohort. *J Surg Oncol*. 2019;120:183–92.
- Stratford P, Gill C, Westaway M, Binkley J. Assessing disability and change on individual patients: a report of a patient specific measure. *Physiotherapy Can*. 1995;47:258–63.
- Barten JA, Pisters MF, Huisman P, Takken T, Veenhof C. Measurement properties of patient-specific instruments measuring physical function. *J Clin Epidemiol*. 2012;65:590–601.
- Cieza A, Geyh S, Chatterji S, Kostanjsek N, Üstün B, Stucki G. ICF linking rules: an update based on lessons learned. *J Rehabil Med*. 2005;37:212–8.
- Berghmans DDP, Lenssen AF, van Rhijn LW, de Bie RA. The patient-specific functional scale: its reliability and responsiveness in patients undergoing a total knee arthroplasty. *J Orthop Sports Phys Ther*. 2015;45(7):550–6.
- Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), short-form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), short Form-36 Bodily Pain Scale (SF-Arthritis Care Res (Hoboken). 2011;63(Suppl 1):S240–52.
- Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, et al. Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: a systematic literature review. *J Pain Symptom Manage*. 2011;41:1073–93.
- Davis AM, Wright JG, Williams JI, Bombardier C, Griffin A, Bell RS. Development of a measure of physical function for patients with bone and soft tissue sarcoma. *Qual Life Res*. 1996;5:508–16.
- Willeumier JJ, van der Wal CWPG, van der Wal RJP, Dijkstra PDS, Vliet Vlieland TPM, van de Sande MAJ. Cross-cultural adaptation, translation, and Validation of the Toronto Extremity Salvage Score for Extremity Bone and soft tissue tumor patients in Netherlands. *Sarcoma*. 2017;2017:6197525.
- Saebye CKP, Safwat A, Kaa AK, Pedersen NA, Keller J. Validation of a Danish version of the Toronto Extremity Salvage Score questionnaire for patients with sarcoma in the extremities. *Dan Med J*. 2014;61:A4734.
- Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. The EORTC QLQ-C30 Scoring Manual. Vol. (3rd Edition). 2001.
- Juul T, Petersen MA, Holzner B, Laurberg S, Christensen P, Grønvald M. Danish population-based reference data for the EORTC QLQ-C30: associations with gender, age and morbidity. *Qual Life Res*. 2014;23:2183–93.
- Koller M, Aaronson NK, Blazeby J, Bottomley A, Dewolf L, Fayers PM, et al. Translation procedures for standardised quality of life questionnaires: the European Organisation for Research and Treatment of Cancer (EORTC) approach. *Eur J Cancer*. 2007;43(12):1810–20.
- Orange ST, Marshall P, Madden LA, Vince RV. Can sit-to-stand muscle power explain the ability to perform functional tasks in adults with severe obesity? *J Sports Sci*. 2019;37:1227–34.
- Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. *Osteoarthritis Cartilage*. 2012;20:1548–62.
- Crockett K, Ardell K, Hermanson M, Penner A, Lanovaz J, Farthing J, et al. The relationship of knee-extensor strength and rate of torque development to sit-to-stand performance in older adults. *Physiotherapy Can*. 2013;65:229–35.
- Slaughter SE, Wagg AS, Jones CA, Schopflocher D, Ickert C, Bampton E, et al. Mobility of vulnerable elders study: Effect of the sit-to-stand activity on mobility, function, and quality of life. *J Am Med Dir Assoc*. 2015;16:138–43.
- Tveter AT, Dagfinrud H, Moseng T, Holm I. Health-related physical fitness measures: reference values and reference equations for Use in Clinical Practice. *Arch Phys Med Rehabil*. 2014;95:1366–73.
- Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport*. 1999;70:113–9.
- Burr JF, Bredin SSD, Faktor MD, Warburton DER. The 6-Minute Walk Test as a predictor of objectively measured aerobic fitness in healthy working-aged adults. *Phys Sportsmed*. 2011;39:133–9.
- Dam JC, van Bekkering E, Bramer WP, Beishuizen JAM, Fiocco A, Dijkstra M. Functional outcome after surgery in patients with bone sarcoma around the knee; results from a long-term prospective study. *J Surg Oncol*. 2017;115:1028–32.
- Galiano-Castillo N, Arroyo-Morales M, Ariza-García A, Sánchez-Salado C, Fernández-Lao C, Cantarero-Villanueva I, et al. The six-minute walk test as a measure of health in breast cancer patients. *J Aging Phys Act*. 2016;24:508–15.
- Bohannon RW, Crouch R. Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: a systematic review. *J Eval Clin Pract*. 2017;23:377–81.

35. Schmidt K, Vogt L, Thiel C, Jäger E, Banzer W. Validity of the six-minute walk test in cancer patients. *Int J Sports Med.* 2013;34(7):631–6.
36. de Vet HCW, Terwee CB, Lidwine B, Mokkink LB, Knol DL. *Measurement in Medicine.* Vol. 1st ed. New York: 2011.
37. Park JH, Kim JI. Practical Consideration of Factor Analysis for the Assessment of Construct Validity. *J Korean Acad Nurs.* 2021;51(6):643–7.
38. Suhr DD. Principal Component Analysis vs. exploratory factor analysis. *Stat Data Anal.* 2005;30:203–30.
39. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford P, Knol DL, 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.* 2010;63:737–45.
40. Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil.* 2000;81:515–20.
41. Chiarotto A, Ostelo RW, Turk DC, Buchbinder R, Boers M. Core outcome sets for research and clinical practice. *Braz J Phys Ther.* 2017;21:77–84.
42. Ramsey I, Eckert M, Hutchinson AD, Marker J, Corsini N. Core outcome sets in cancer and their approaches to identifying and selecting patient-reported outcome measures: a systematic review. *J Patient Rep Outcomes* 2020;4.
43. Ramsey I, Corsini N, Hutchinson AD, Marker J, Eckert M. A core set of patient-reported outcomes for population-based cancer survivorship research: a consensus study. *J Cancer Surviv.* 2021;15:201–12.
44. Singh JA, Dowsey MM, Dohm M, Goodman SM, Leong AL, Voshhaar MMJHS, et al. Achieving consensus on total joint replacement trial outcome reporting using the OMERACT filter: endorsement of the final core domain set for total hip and total knee replacement trials for endstage arthritis. *J Rheumatol.* 2017;44:1723–6.
45. Singh JA, Dowsey MM, Choong PF. Patient endorsement of the Outcome measures in Rheumatology (OMERACT) total joint replacement (TJR) clinical trial draft core domain set. *BMC Musculoskelet Disord* 2017;18.
46. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34–42.
47. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury.* 2011;42(3):236–40.
48. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ.* 2011;2:53–5.
49. Brown JD. Choosing the Right Number of Components or factors in PCA and EFA. *JALT Testing&Evaluation SIG Newsl.* 2009;2009(13):3–19.
50. Brown JD. Choosing the right type of Rotation in PCA and EFA. *JALT Testing&Evaluation SIG Newsl.* 2009;13:3.
51. Kim HS, Yun J, Kang S, Han I. Cross-cultural adaptation and validation of the Korean Toronto Extremity Salvage Score for extremity sarcoma. *J Surg Oncol.* 2015;112(1):93–7.
52. Saebye CKP, Fugloe M, Nymark H, Safwat T, Petersen A, Baad-Hansen MM. Factors associated with reduced functional outcome and quality of life in patients having limb-sparing surgery for soft tissue sarcomas - a national multicenter study of 128 patients. *Acta Oncol.* 2017;56(2):239–44.
53. Marchese VG, Rai SN, Carlson CA, Pamela SH, Spearing EM. Assessing functional mobility in survivors of lower-extremity sarcoma: reliability and validity of a new assessment tool. 2007;49(2):183–9.
54. Kask G, Barner-Rasmussen I, Repo JP, Kjaldman M, Kilk K, Blomqvist C, et al. Functional outcome measurement in patients with lower-extremity soft tissue sarcoma: a systematic literature review. *Ann Surg Oncol.* 2019;26(13):4707–22.
55. Holm CE, Soerensen MS, Yilmaz M, Petersen MM. Evaluation of tumor-prostheses over time: Complications, functional outcome, and comparative statistical analysis after resection and reconstruction in orthopedic oncologic conditions in the lower extremities. *SAGE Open Med [Internet].* 2022 Apr 21 [cited 2023 Aug 8];10:20503121221094190. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9047786/>
56. Bekendtgørelse af lov om videnskabetisk behandling af sundhedsvidenskabelige forskningsprojekter og sundhedsdatavidenskabelige forskningsprojekter [Internet]. LBK 1338 af 01/09/2020. <http://www.retsinformation.dk/eli/ta/2020/1338>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.