J Shoulder Elbow Surg (2026) 35, 270–279

JOURNAL OF
SHOULDER AND
ELBOW
SURGERY

www.elsevier.com/locate/ymse

What patients with proximal humerus fractures really want and what commonly used outcome scores measure

Sam Razaeian, MD[a,b,*], Laura Hösl, MD[b], Birgitt Wiese, MMath[c], Dafang Zhang, MD[d], Christian Krettek, MD, FRACS, FRCSEd[e], Nael Hawi, MD, MBA[f]

[a]*Department for Trauma, Hand and Reconstructive Surgery, Saarland University, Homburg, Saarland, Germany*
[b]*Department of Trauma Surgery, Hannover Medical School, Hannover, Germany*
[c]*Hannover Medical School, MHH Information Technology (MIT), Hannover, Germany*
[d]*Department of Orthopaedic Surgery, Brigham and Women's Hospital, Boston, MA, USA*
[e]*Hannover Humerus Registry (HHR), Traumastiftung gGmbH, Hannover, Germany*
[f]*Orthopaedic and Surgical Clinic Braunschweig (OCP), Braunschweig, Germany*

Background: This study aims (1) to identify patient-reported questionnaire items, independent of age or gender, that reflect healthy shoulder function and treatment satisfaction in patients with proximal humerus fracture (PHF), and (2) to compare these items and their weighted importance with items measured by the most frequently used outcome measures.
Methods: Patients who sustained a PHF from June 2016 to September 2021 were surveyed with a 29-item questionnaire on their perceptions of item importance as a measure of shoulder function and treatment outcome. Items were generated from the following outcome measures: American Shoulder and Elbow Surgeons score, Constant Score, Neer Score, Oxford shoulder Score (OSS), Quick-DASH, and the University of California, Los Angeles shoulder rating score. A mean difference of at least 10% between gender and age groups (<60 vs. ≥60 years) was defined as clinically significant. Items that were rated as at least 90% important without a clinically and statistically significant mean difference between the groups were defined as essential items.
Results: One hundred forty-six patients with mean age 60.8 years (range 20-92 years) completed the questionnaires. Only 6 out of 29 items were identified as essential items. These include: being pain-free, being able to sleep/having no pain in bed at night, using a knife and a fork at the same time, putting on a coat/to dress, managing toileting, washing under both arms. None of the scoring systems covered all these items with appropriate weighting of scoring points. The OSS most closely covered patient interest with the most appropriate weighting of points.
Conclusion: We identified 6 items from daily life that are of essential importance for patient-reported healthy shoulder function and treatment satisfaction regardless of age and gender. Until a reliable and valid scoring system for PHF is developed that includes these items, we recommend using the OSS, as it most closely reflects patient-reported interests.

This study is authorized by the local ethical committee of Hannover Medical School (no. 3222-2016).

*Reprint requests: Sam Razaeian, MD, Department for Trauma, Hand and Reconstructive Surgery, Saarland University, Kirrberger Str. 100, Homburg, Saarland 66421, Germany.
E-mail address: sam.r@hotmail.de (S. Razaeian).

https://doi.org/10.1016/j.jse.2025.03.029

Level of evidence: Basic Science Study; Validation of Outcome Instruments

Keywords: Proximal humerus fracture; PROM; patient reported outcome; score; value; scoring system

Available evidence over the last 2 decades has not yet enabled a consensus on the optimal treatment of proximal humerus fractures (PHFs).[5-9,16] Factors that make treatment comparison and consensus building difficult include the lack of a reliable classification system, the wide range of functional levels of the injured population, and the increasing implant options that trigger treatment variation. Furthermore, there is a lack of a disease-specific outcome measurement tool.[10,16]

PHFs are most commonly seen in the geriatric population. For the elderly population in particular, it is reasonable to use a shared treatment decision-making strategy based on a variety of factors, including age, comorbidity, functional level, bone quality, and individual patient preferences.[2,3] Such an approach requires patient-centered outcome measurements that in fact measure these preferences with an appropriate age- and gender-independent weighting of points, in order to be broadly applicable to the spectrum of patients with this disorder. Whether currently utilized outcome measurement systems accomplish this is uncertain and has been recently called into question in other parts of the musculoskeletal system.[13]

The fact that the American Shoulder and Elbow Surgeons (ASES) multicenter taskforce studying PHF did not reach any consensus on which outcome measures to include in future studies is telling.[17] Their study aimed to report on the most frequently used outcome measures to make recommendation for future research. Instead, they confirmed a lack of homogeneity in the use of outcome measures across the PHF literature, where 22 different outcome measures were used in over 70 clinical trials.[17] In the absence of evidence for the usage of a single outcome instrument, it has been recommended to use at least 3 common outcome measures and one general health score until the optimal scoring systems are determined.[17]

However, standardization of outcome measures would be important as it improves the surgeon's ability to interpret the evidence and evaluate treatment effects.[17] The lack of standardization complicates cross-study comparisons and confounds the treatment decision-making process.[17] In order to be able to make recommendations for the usage of existing scoring systems or to develop future patient-derived scoring systems, it would be essential to identify age- and gender-independent patient preferences.

This study aims (1) to identify age- and gender-independent patient-reported prerequisites for healthy shoulder function and treatment satisfaction in patients with PHF, and (2) to compare their reported items and their weighted importance with those of the most frequently used patient-reported outcome measurement instruments.

Materials and methods

All adult patients who sustained a PHF from June 2016 to September 2021 and were included in the abovementioned observational registry study of a supraregional Level I Trauma Center, regardless of treatment modality (nonoperative and surgical), were contacted once by letter in October 2022. Figure 1 shows the course of study inclusion in a flow chart. The letter provided background information about the aim of the study and a 29-item questionnaire that was designed to survey patients' perceptions of item importance in daily life for healthy shoulder function and high treatment satisfaction. The questionnaire was developed based on elements from pre-existing questionnaires, selected to represent important elements in function and satisfaction.

The following 3 questions (Q) were asked in writing:

Q1: How much do you need your injured shoulder to do this aspect of your life?

Q2: How important is this aspect of your life for you to be able to say: *"My injured shoulder has a healthy function of 100%."*

Q3: How important is this aspect of your life for you to be able to say: *"I was 100% satisfied with the treatment of my injured shoulder."*

Q1 aimed to identify items for which the affected shoulder joint must actually be used. In this way, broadly defined items such as "to use public transport," which can also involve compensatory movements of other joints, should be unmasked.

Q2 and Q3 were aimed at items that are a prerequisite for a self-perceived 100% functionally healthy shoulder and treatment satisfaction, respectively. In addition, a free text field was offered to give the option of naming items not listed on the questionnaire.

Items of the following 6 commonly used functional outcome scoring systems[17] were assigned to the 3 queries Q1-3: ASES score, standard non-normalized Constant score (CS), Neer Score, Oxford shoulder Score (OSS), QuickDASH (qDASH), and the University of California, Los Angeles (UCLA) shoulder rating Score. Additionally, their items were categorized into the following subdomains: Pain, range of motion, strength, and activities of daily life (ADL). Strength of correlation between age and these subdomains was analyzed. Items of the category "pain" were not assigned to Q1, as this would not have made logical sense in terms of content.

Although the DASH score is one of the most frequently used scores, it was omitted and the qDASH was included instead in order to avoid responder survey fatigue.

Response options were given on a percentage scale from 0 to 100%. In Q1, 0% was defined as "not at all" and 100% as "very much." In Q2 and Q3, 0% were defined as "unimportant" and

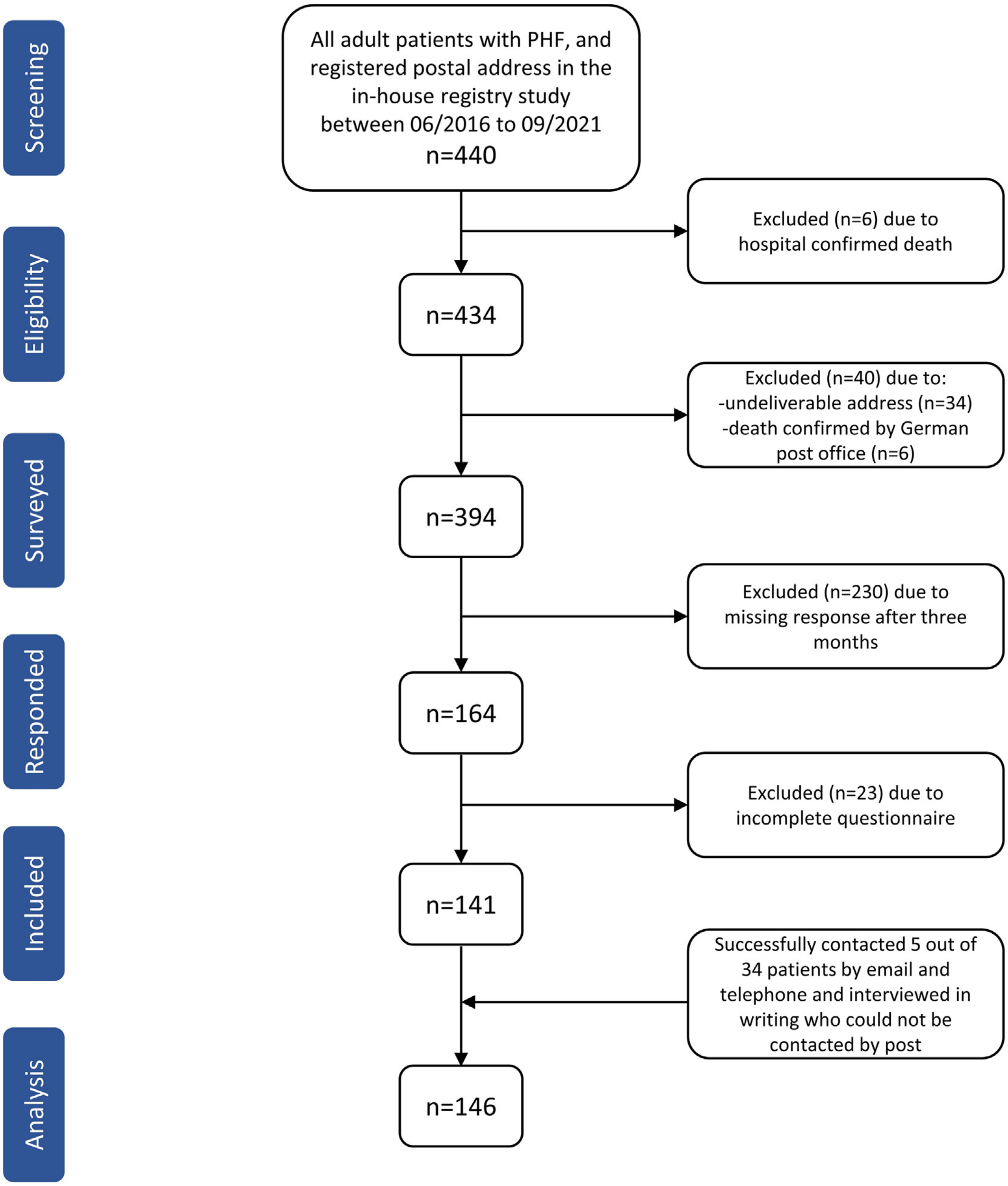

Screening
Eligibility
Surveyed
Responded
Included
Analysis
All adult patients with PHF, and registered postal address in the in-house registry study between 06/2016 to 09/2021
n=440
Excluded (n=6) due to hospital confirmed death
n=434
Excluded (n=40) due to:
-undeliverable address (n=34)
-death confirmed by German post office (n=6)
n=394
Excluded (n=230) due to missing response after three months
n=164
Excluded (n=23) due to incomplete questionnaire
n=141
Successfully contacted 5 out of 34 patients by email and telephone and interviewed in writing who could not be contacted by post
n=146

Figure 1 Flow chart showing course of study inclusion.

100% as "very important." As the survey was based on self-reported patient experiences at the time of sustaining the fracture and the healing process afterwards, the age on the day of the accident and not at the time of the survey was used for further data analysis. We chose the age cut-off of 60 years to define the elderly group, as has been reported in the primary literature and subsequent large meta-analyses.[2,6] The questionnaire in paper form and an attached patient information sheet including background information about the aim of the study were distributed to patients by mail. The inclusion was terminated 3 months after the letters were sent. In the case of undeliverable letters that were returned, attempts were made to contact patients by telephone and email if these were documented. Patients with incomplete surveys were excluded from the analysis.

The questionnaire and patient information sheet were pretested in 10 individuals and revised once before the start of the survey. Comprehension postinterview probing was performed as a cognitive pretesting method to ensure comprehensibility of the questions.

A mean difference of at least 10% between age and gender groups was defined as clinically significant. Items that were rated as at least 90% important among all 3 questions (Q1-3) without a clinically and statistically significant mean difference between the gender and age groups were defined as essential core items of high importance.

Table I Demographic data of the participants

	n (%)	Mean age in yr (range)	SD
Age in yr			
<60	65 (44.5)	47.3 (20-59)	10
≥60	81 (55.5)	71.6 (60-92)	9.3
Gender			
Male	51 (34.9)	56 (31-80)	12.3
Female	95 (65.1)	63.3 (20-92)	16.4
Total	146 (100)	60.8 (20-92)	15.5

SD, standard deviation.

For later comparison of the points weighting, percentage weightings in relation to the possible total score of each scoring system were determined and compared with ranked patient-reported item importance regarding functional priorities (Q2).

Since some scoring systems contain several completely different items in one answer option, which in turn appear separately from each other in separate questions in other scoring systems, a direct quantitative comparison between the systems and the patient data is difficult.

For example, the UCLA shoulder rating score has a response option for a question on the topic of function that summarizes the following 5 completely different items and awards 6 points (17.1%) of the total score for them: "most housework, shopping, and driving possible, able to do hair and to dress and undress, including fastening bra." Some of these important items are queried and scored separately in other scoring systems. In order to allow at least a direct comparison of individual items among the scoring systems, the weighting of points is listed several times in such cases. This results in a total percentage score that is over 100% for these scoring systems.

Statistical analyses

Descriptive statistics, including means, mean differences, standard deviations, and ranges were calculated in order to rank the items by importance. Level of importance was classified as follows: very important: ≥90%; important: ≥80%; moderately important: ≥70%; slightly important: ≥50%; unimportant: <50%.

Mann–Whitney *U* test was used as nonparametric tests to compare mean values. For bivariate analysis of correlation between age and subdomains, Spearman's rho was calculated for nonparametric data. Correlation strength was classified as follows: very high: >0.90; high: 0.70-0.89; moderate: 0.50-0.69; fair: 0.30-0.49; low: 0.10-0.29; or very low: 0.10.

A 95% confidence interval was set. A *P* value < .05 and < .01 was considered statistically significant and highly significant, respectively.

For the analyses, SPSS (version 26; IBM, Armonk, NY, USA) and Microsoft Excel 2021 (Microsoft Corp., Redmond, WA, USA) were used.

Results

Fully completed questionnaires from 146 patients were analyzed (Fig. 1). Table I shows details on the age and gender of included subjects.

Clear age-related differences were observed concerning functional priorities (Q2) with clinically and statistically significant mean differences ranging from 10% to 52.1% (Fig. 2).

The number of items with very high functional importance (Q2) decreased notably from 17 to 6 items in the elderly group with significant mean differences between all 4 subdomains (Fig. 2). This notable difference was not observed between the gender groups (11 vs. 9 items) (Fig 3). There were similar findings with regards to items of self-reported importance for treatment satisfaction (Q3) (Supplementary Figures S5-S7).

Among the 4 subdomains, only the category of "pain" was of the highest importance (Q2 and Q3) regardless of age and gender. The subdomain of "range of motion" did not correlate with age. In contrast, there was a moderate to fair significant negative correlation between age and the subdomains of "ADLs" and "strength" (Figs. 2 and 3 and Table II).

Essential items of high importance

Only 6 out of 29 items were identified as essential items of high importance for a self-reported 100% healthy shoulder function and treatment satisfaction, regardless of age and gender. These include 2 pain items and 4 ADL items: being pain-free, being able to sleep/having no pain in bed at night, using a knife and a fork at the same time, putting on a coat/to dress, managing toileting, and washing under both arms (Figs. 2 and 3, Supplementary Figures S1-S7). The 4 ADLs were also the ones that patients reported they needed their broken shoulder very much to perform (Q1) and were among the items reported as essential to generate a high level of treatment satisfaction (Q3) regardless of age and gender (Supplementary Figures S1-S7).

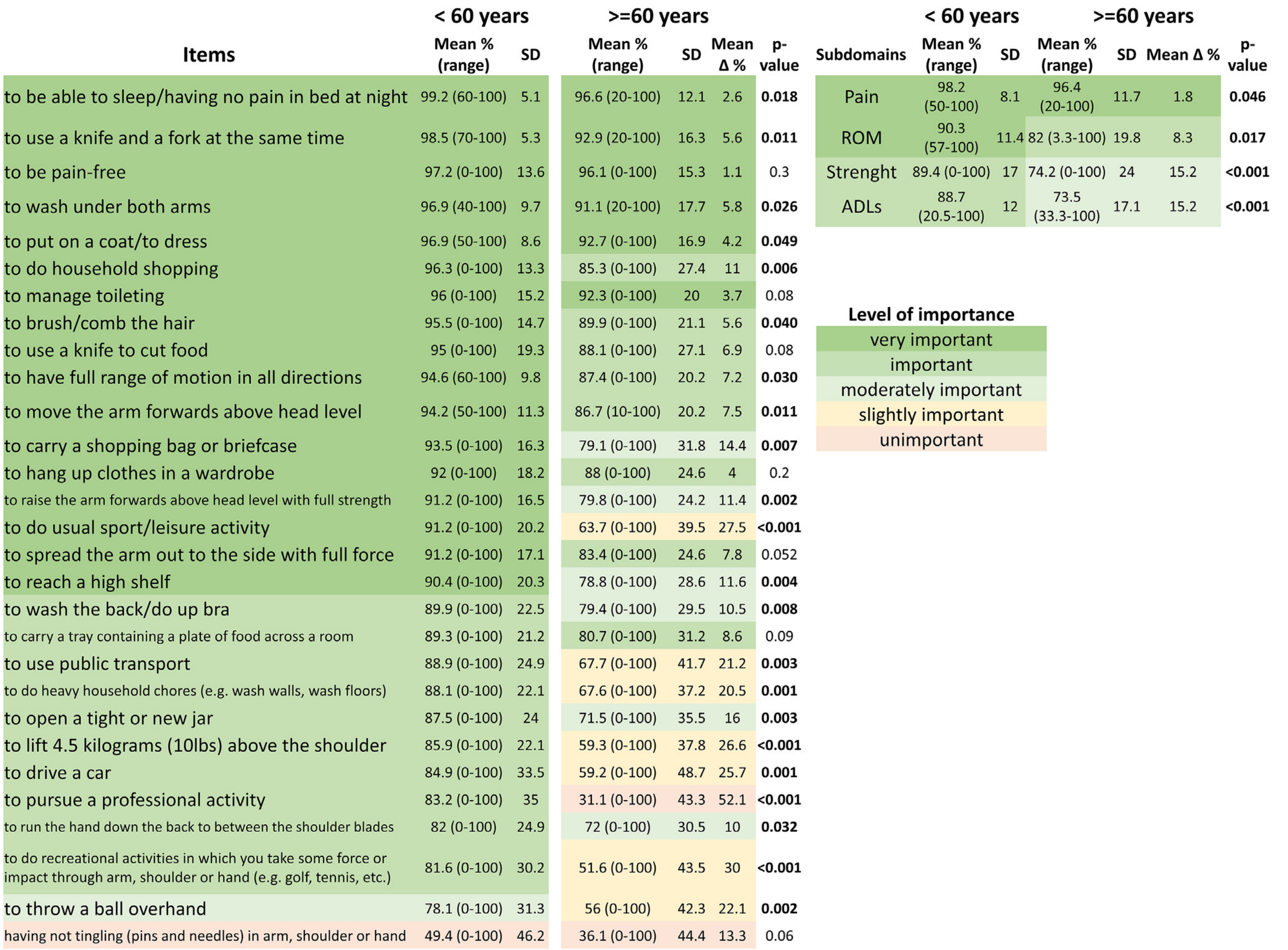
Additional unlisted items

Only 10 patients (6.9%) mentioned a total of 14 additional unlisted different items. Their mean age was 55.3 years (range 26 to 83 years). Seven out of 10 patients were female. With the exception of 2 items from the subdomain of "pain," the remaining items could be assigned to the subdomain "ADL." The following 3 items of self-perceived high importance of 100% in all 3 questions were mentioned twice: "to do cycling," "to write by hand," and "to lift up my child."

Table III lists all details on age, gender, and unlisted items.

Comparison with scoring systems

None of the 6 scoring systems covers all 6 core items (Fig. 4). The OSS covers 5 items, but not the ADL item "managing toileting." The ASES score places the main



Items	< 60 years Mean % (range)	SD	>=60 years Mean % (range)	SD	Mean Δ %	p-value
to be able to sleep/having no pain in bed at night	99.2 (60-100)	5.1	96.6 (20-100)	12.1	2.6	**0.018**
to use a knife and a fork at the same time	98.5 (70-100)	5.3	92.9 (20-100)	16.3	5.6	**0.011**
to be pain-free	97.2 (0-100)	13.6	96.1 (0-100)	15.3	1.1	0.3
to wash under both arms	96.9 (40-100)	9.7	91.1 (20-100)	17.7	5.8	**0.026**
to put on a coat/to dress	96.9 (50-100)	8.6	92.7 (0-100)	16.9	4.2	**0.049**
to do household shopping	96.3 (0-100)	13.3	85.3 (0-100)	27.4	11	**0.006**
to manage toileting	96 (0-100)	15.2	92.3 (0-100)	20	3.7	0.08
to brush/comb the hair	95.5 (0-100)	14.7	89.9 (0-100)	21.1	5.6	**0.040**
to use a knife to cut food	95 (0-100)	19.3	88.1 (0-100)	27.1	6.9	0.08
to have full range of motion in all directions	94.6 (60-100)	9.8	87.4 (0-100)	20.2	7.2	**0.030**
to move the arm forwards above head level	94.2 (50-100)	11.3	86.7 (10-100)	20.2	7.5	**0.011**
to carry a shopping bag or briefcase	93.5 (0-100)	16.3	79.1 (0-100)	31.8	14.4	**0.007**
to hang up clothes in a wardrobe	92 (0-100)	18.2	88 (0-100)	24.6	4	0.2
to raise the arm forwards above head level with full strength	91.2 (0-100)	16.5	79.8 (0-100)	24.2	11.4	**0.002**
to do usual sport/leisure activity	91.2 (0-100)	20.2	63.7 (0-100)	39.5	27.5	**<0.001**
to spread the arm out to the side with full force	91.2 (0-100)	17.1	83.4 (0-100)	24.6	7.8	0.052
to reach a high shelf	90.4 (0-100)	20.3	78.8 (0-100)	28.6	11.6	**0.004**
to wash the back/do up bra	89.9 (0-100)	22.5	79.4 (0-100)	29.5	10.5	**0.008**
to carry a tray containing a plate of food across a room	89.3 (0-100)	21.2	80.7 (0-100)	31.2	8.6	0.09
to use public transport	88.9 (0-100)	24.9	67.7 (0-100)	41.7	21.2	**0.003**
to do heavy household chores (e.g. wash walls, wash floors)	88.1 (0-100)	22.1	67.6 (0-100)	37.2	20.5	**0.001**
to open a tight or new jar	87.5 (0-100)	24	71.5 (0-100)	35.5	16	**0.003**
to lift 4.5 kilograms (10lbs) above the shoulder	85.9 (0-100)	22.1	59.3 (0-100)	37.8	26.6	**<0.001**
to drive a car	84.9 (0-100)	33.5	59.2 (0-100)	48.7	25.7	**0.001**
to pursue a professional activity	83.2 (0-100)	35	31.1 (0-100)	43.3	52.1	**<0.001**
to run the hand down the back to between the shoulder blades	82 (0-100)	24.9	72 (0-100)	30.5	10	**0.032**
to do recreational activities in which you take some force or impact through arm, shoulder or hand (e.g. golf, tennis, etc.)	81.6 (0-100)	30.2	51.6 (0-100)	43.5	30	**<0.001**
to throw a ball overhand	78.1 (0-100)	31.3	56 (0-100)	42.3	22.1	**0.002**
having not tingling (pins and needles) in arm, shoulder or hand	49.4 (0-100)	46.2	36.1 (0-100)	44.4	13.3	0.06

Subdomains	< 60 years Mean % (range)	SD	>=60 years Mean % (range)	SD	Mean Δ %	p-value
Pain	98.2 (50-100)	8.1	96.4 (20-100)	11.7	1.8	**0.046**
ROM	90.3 (57-100)	11.4	82 (3.3-100)	19.8	8.3	**0.017**
Strenght	89.4 (0-100)	17	74.2 (0-100)	24	15.2	**<0.001**
ADLs	88.7 (20.5-100)	12	73.5 (33.3-100)	17.1	15.2	**<0.001**

Level of importance
very important
important
moderately important
slightly important
unimportant

Figure 2 Subgroup analysis of ranked items, and subdomains regarding functional priorities (Q2) for both age groups. *ROM*, range of motion; *Δ*, difference; *SD*, standard deviation; *ADLs*, activities of daily living. Bold values indicate statistically significant mean differences.

emphasis on being pain-free, and the ADLs "washing under both arms" and "being able to use a knife and fork at the same time" are not taken into account. The Neer Score only covers the item "being pain-free." The CS and qDASH do not include any of the 4 essential ADLs. Both the CS and qDASH only include "being pain-free" and "being able to sleep pain-free." The UCLA Shoulder Score only takes into account the ADL "to dress" in addition to "being pain-free". Fig. 5 shows percentage of score coverage of all 29 items.

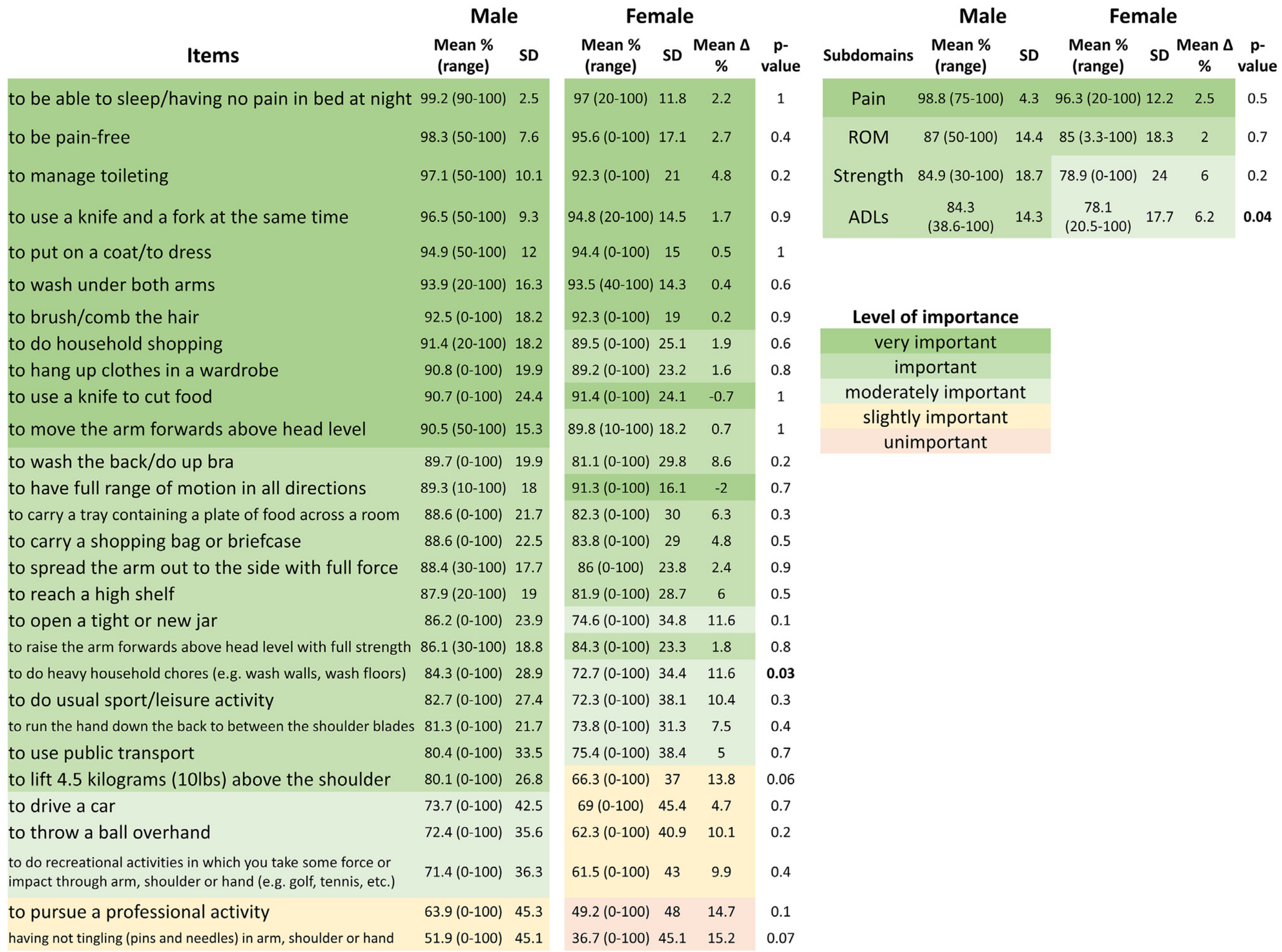
Discussion

Principal findings

To the best of our knowledge, this is the first study that investigated patients' perceptions on item importance of frequently used outcome measurement instruments for PHF. We identified 6 age- and gender-independent items that patients self-report to be highly essential for healthy shoulder function and treatment satisfaction in patients with PHFs.

Our findings could help to overcome a challenge that arises from the epidemiological nature of PHF, which is that the injured population represents a very diverse functional group for whom treatment based on fracture pattern or age alone may not lead to optimal functional outcomes.[3,10]

One of the strengths of this study is that the survey respondents were composed of a cohort of patients who a sustained PHF and who had already gone through the entire healing process. Therefore, we expect that they would be able to provide insights and self-assessments about which questionnaire items are actually relevant in daily life. The fact that only a very small proportion of patients (6.9%) felt it necessary to name additional unlisted items shows that the 29 items of the most commonly used scores examined here do indeed appear to be complete in what is important to these patients, although this might also be a result of



Items	Male Mean % (range)	Male SD	Female Mean % (range)	Female SD	Mean Δ %	p-value
to be able to sleep/having no pain in bed at night	99.2 (90-100)	2.5	97 (20-100)	11.8	2.2	1
to be pain-free	98.3 (50-100)	7.6	95.6 (0-100)	17.1	2.7	0.4
to manage toileting	97.1 (50-100)	10.1	92.3 (0-100)	21	4.8	0.2
to use a knife and a fork at the same time	96.5 (50-100)	9.3	94.8 (20-100)	14.5	1.7	0.9
to put on a coat/to dress	94.9 (50-100)	12	94.4 (0-100)	15	0.5	1
to wash under both arms	93.9 (20-100)	16.3	93.5 (40-100)	14.3	0.4	0.6
to brush/comb the hair	92.5 (0-100)	18.2	92.3 (0-100)	19	0.2	0.9
to do household shopping	91.4 (20-100)	18.2	89.5 (0-100)	25.1	1.9	0.6
to hang up clothes in a wardrobe	90.8 (0-100)	19.9	89.2 (0-100)	23.2	1.6	0.8
to use a knife to cut food	90.7 (0-100)	24.4	91.4 (0-100)	24.1	-0.7	1
to move the arm forwards above head level	90.5 (50-100)	15.3	89.8 (10-100)	18.2	0.7	1
to wash the back/do up bra	89.7 (0-100)	19.9	81.1 (0-100)	29.8	8.6	0.2
to have full range of motion in all directions	89.3 (10-100)	18	91.3 (0-100)	16.1	-2	0.7
to carry a tray containing a plate of food across a room	88.6 (0-100)	21.7	82.3 (0-100)	30	6.3	0.3
to carry a shopping bag or briefcase	88.6 (0-100)	22.5	83.8 (0-100)	29	4.8	0.5
to spread the arm out to the side with full force	88.4 (30-100)	17.7	86 (0-100)	23.8	2.4	0.9
to reach a high shelf	87.9 (20-100)	19	81.9 (0-100)	28.7	6	0.5
to open a tight or new jar	86.2 (0-100)	23.9	74.6 (0-100)	34.8	11.6	0.1
to raise the arm forwards above head level with full strength	86.1 (30-100)	18.8	84.3 (0-100)	23.3	1.8	0.8
to do heavy household chores (e.g. wash walls, wash floors)	84.3 (0-100)	28.9	72.7 (0-100)	34.4	11.6	**0.03**
to do usual sport/leisure activity	82.7 (0-100)	27.4	72.3 (0-100)	38.1	10.4	0.3
to run the hand down the back to between the shoulder blades	81.3 (0-100)	21.7	73.8 (0-100)	31.3	7.5	0.4
to use public transport	80.4 (0-100)	33.5	75.4 (0-100)	38.4	5	0.7
to lift 4.5 kilograms (10lbs) above the shoulder	80.1 (0-100)	26.8	66.3 (0-100)	37	13.8	0.06
to drive a car	73.7 (0-100)	42.5	69 (0-100)	45.4	4.7	0.7
to throw a ball overhand	72.4 (0-100)	35.6	62.3 (0-100)	40.9	10.1	0.2
to do recreational activities in which you take some force or impact through arm, shoulder or hand (e.g. golf, tennis, etc.)	71.4 (0-100)	36.3	61.5 (0-100)	43	9.9	0.4
to pursue a professional activity	63.9 (0-100)	45.3	49.2 (0-100)	48	14.7	0.1
having not tingling (pins and needles) in arm, shoulder or hand	51.9 (0-100)	45.1	36.7 (0-100)	45.1	15.2	0.07

Subdomains	Male Mean % (range)	Male SD	Female Mean % (range)	Female SD	Mean Δ %	p-value
Pain	98.8 (75-100)	4.3	96.3 (20-100)	12.2	2.5	0.5
ROM	87 (50-100)	14.4	85 (3.3-100)	18.3	2	0.7
Strength	84.9 (30-100)	18.7	78.9 (0-100)	24	6	0.2
ADLs	84.3 (38.6-100)	14.3	78.1 (20.5-100)	17.7	6.2	**0.04**

Level of importance

- very important
- important
- moderately important
- slightly important
- unimportant

Figure 3 Subgroup analysis of ranked items, and subdomains regarding functional priorities (Q2) for both gender groups. *ROM*, range of motion; *Δ*, difference; *SD*, standard deviation; *ADLs*, activities of daily living. Bold values indicate statistically significant mean differences.

Table II Spearman's correlation coefficient for correlation analysis between age and subdomains of Q1-3

	Subdomains			
	Pain	Range of motion	Strength	ADLs
Q1	n.m.	−0.24	−0.4	−0.5
P value	n.m.	**.003**	**< .001**	**< .001**
Q2	−0.24	−0.26	−0.42	−0.55
P value	**.003**	**.001**	**< .001**	**< .001**
Q3	−0.23	−0.21	−0.38	−0.5
P value	**.005**	**.011**	**< .001**	**< .001**

Bold values indicate statistical significance with a *P* value < .05. *n.m.*, not measured; *ADLs*, activities of daily living.

survey fatigue However, the finding that none of the scoring systems examined covers all 6 items might underscore the statement that there is no strong evidence for the use of a single outcome measurement.[17]

Our study showed that among the 4 subdomains, only "pain" was of the highest importance regardless of age and gender. The subdomain "range of motion" also did not correlate with age, while there was a moderate to fair significant negative correlation between age and subdomains "ADLs" and "strength." These findings support the ASES, OSS, Neer, and UCLA instruments, in that these instruments have the focus of their total possible scoring points on the subdomain "pain. About one-third of the total points in the Neer, OSS, and UCLA and half of the total points in the ASES scoring system are dedicated to this subdomain. However, the apparent age-dependent decrease in importance in the subdomains of strength and ADLs should be considered in future studies and design of scoring systems. The latter subdomain in particular shows how important it is to identify those ADLs from the large number of existing ADLs that remain important regardless of age.

Of the instruments studied, the OSS came closest to covering patients' priorities. It covers 5 of the 6 age- and gender-independent core items and does so with plausibly equal point weighting. In addition, it covers on average

Table III Details on age, gender, and items of free text field

Age	Gender	Items	Response in %		
			Q1	Q2	Q3
50	male	Pain during weather changes	100	75	100
		Pain after exertion	100	75	100
64	female	to do cycling	100	100	100
		to do gardening work	100	100	100
52	male	to work on the computer	100	100	100
		to write by hand	100	100	100
62	male	to go swimming	100	100	100
83	female	to give my partner a hug	100	100	100
		to lean on both arms	100	100	100
26	female	to carry my child in my arms	70	90	100
		to walk the dog on a lead	80	90	100
55	female	to squeeze a lemon	100	100	90
79	female	to do cycling	100	100	100
32	female	to carry my child in my arms	100	100	100
		to lift my child	100	100	100
		to breastfeed my child	100	100	100
50	female	to write by hand	100	100	100

most of the patient-reported functionally important items (63.6%), although it contains items with identical weighting to the core items (eg, to use public transport and drive a car) that were patient-reported as unimportant, at least by the elderly group. This is not intended to question the completeness or justification for the existence of individual items in this scoring system, but to generate discussion as to whether the individual point weighting should be reconsidered.

Regarding the age- and gender-independent essential items identified in this study, it is noteworthy that all 4 ADLs are activities that usually have to be performed on a daily basis. What can be assumed for going to the toilet, for example, does not necessarily apply to lifting 4.5 kg (10 lbs) or throwing a ball over the shoulder, even though these items are also counted as ADLs and are equally weighted in the ASES scoring system. This problem has recently been highlighted by the widely used Oxford Knee Score, where its questions seem to differ significantly in self-perceived weight for each patient, based on the sociodemographic data, such as age, self-use of a car, and employment, and could possibly contribute to a frequently observed ceiling effect.[13]

Interestingly, the OSS, which best reflects the interests of patients, was used only in about 5 percent of all studies and is not part of the group of recommended outcome measures for future studies by the ASES multicenter taskforce on PHF.[17] Instead, the CS, which along with the Neer Score reflects the interests of patients the worst, was used most frequently by a wide margin in over 60 percent of all studies, and was recommended as an alternative to the DASH score by the taskforce.[17]

In addition to a lack of validity[1,22] and concerns regarding its reliability,[21] criticism of the CS arise from its strength domain that has a high point weighting of up to 25% of the total scoring points.[4,17,22,23] Our study reaffirms this concern as we did not identify "strength" as essential to patients' preferences. In addition, it has to be considered that its importance to patients decreases with increasing age.

The taskforce that confirmed considerable variability in usage of outcome measures across studies on PHF recommended that future studies with higher levels of evidence should use at least 3 outcome scores, preferentially commonly reported ones, and that the ASES score should be considered as one of them to improve cross-study comparison until the optimal scores are determined.[17] The ASES score also lacks reliability and validity for PHF,[15,17,18] although one study by Slobogean found that it correlated the most with physical examination findings of all of the patient reported outcomes evaluated (DASH, Simple Shoulder Test, and OSS) in patients with PHF, assuming a benefit in its use.[17,19] However, the potential for survey fatigue that would result from usage of multiple scoring systems leads us to believe that a different approach may be more advisable.[14] If we want to practice in a patient-centered way and meet the requirements of value-based medicine, outcome scores should be used in the future that actually measure what is really important to patients with an appropriate weighting. As long as there is no clear evidence for the use of a single score, we recommend using the OSS, although usage of multiple patient reported outcomes might reduce ceiling effects Our study findings may be considered in the development of future scoring systems.

Items	Mean % of importance	Point weighting of items in scores in %					
		ASES	Constant	Neer	OSS	qDash	UCLA
to be able to sleep/having no pain in bed at night	97.8	5	2	0	8.33	9.1	0
to be pain-free	96.6	50	15	35	25	9.1	25.7
to use a knife and a fork at the same time	95.4	0	0	0	8.33	0	0
to put on a coat/to dress	94.6	5	0	0	8.33	0	17.1
to manage toileting	93.9	5	0	0	0	0	0
to wash under both arms	93.7	0	0	0	8.33	0	0
to brush/comb the hair	92.4	5	0	0	8.33	0	17.1
to use a knife to cut food	91.2	0	0	0	0	9.1	0
to have full range of motion in all directions	90.6	0	50	35	0	0	0
to do household shopping	90.2	0	0	0	8.33	0	17.1
to be able to move the arm forwards above head level	90.1	0	0	0	0	0	14.3
to hang up clothes in a wardrobe	89.8	0	0	0	8.33	0	0
to be able to spread the arm out to the side with full force	86.8	0	25	0	0	0	0
to carry a shopping bag or briefcase	85.5	0	0	0	0	9.1	0
to be able to raise the arm forwards above head level with full strength	84.9	0	0	0	0	0	14.3
to carry a tray containing a plate of food across a room	84.5	0	0	0	8.33	0	0
to wash the back/do up bra	84.1	5	0	0	0	9.1	17.1
to reach a high shelf	84	5	0	0	0	0	0
to open a tight or new jar	78.6	0	0	0	0	9.1	0
to use public transport	77.2	0	0	0	8.33	0	0
to do heavy household chores (e.g. wash walls, wash floors)	76.8	0	0	0	0	9.1	0
to be able to run the hand down the back to between the shoulder blades	76.4	0	10	7	0	0	28.6
to do usual sport/leisure activity	75.9	5	4	0	0	9.1	28.6
to lift 4.5 kilograms (10lbs) above the shoulder	71.1	5	0	0	0	0	0
to be able to drive a car	70.7	0	0	0	8.33	0	17.1
to throw a ball overhand	65.8	5	0	0	0	0	0
to do recreational activities in which you take some force or impact through arm, shoulder or hand (e.g. golf, tennis, etc.)	65	0	0	0	0	9.1	0
to be able to pursue a professional activity	54.3	5	4	0	0	9.1	0
having not tingling (pins and needles) in arm, shoulder or hand	42	0	0	0	0	9.1	0
	Total in %	**100**	**110**	**77**	**108.3**	**100**	**197**

Figure 4 Comparison of item point weightings of scoring systems and ranked items regarding functional priorities (Q2). Bold items indicate age- and gender-independent essential items of high importance. Color scaling of the point weights is not consistent with the scaling of the item rankings. In order to allow at least a direct comparison of individual items among the scoring systems, the weighting of points is listed several times in cases where an answer option of a scoring system combines several different items that are queried and scored separately in other systems. This results in a total percentage score that is over 100% for these scoring systems. *ASES*, American Shoulder and Elbow Surgeons; *OSS*, Oxford Shoulder Score; *qDash*, QuickDASH *UCLA*, University of California at Los Angeles.

Level of importance	Distribution of items in % (n)	Score coverage of ranked items in % (n)					
		Oxford Shoulder Score	UCLA Shoulder Score	ASES Score	Quick Dash	Constant Score	Neer Score
very important	37.9 (11)	63.6 (7)	45.5 (5)	45.5 (5)	27.3 (3)	27.3 (3)	18.2 (2)
important	24.1 (7)	28.6 (2)	28.6 (2)	28.6 (2)	28.6 (2)	14.3 (1)	0 (0)
moderately important	24.1 (7)	28.6 (2)	42.9 (3)	28.6 (2)	42.9 (3)	28.6 (2)	14.3 (1)
slightly important	10.3 (3)	0 (0)	0 (0)	66.7 (2)	66.7 (2)	33.3 (1)	0 (0)
unimportant	3.4 (1)	0 (0)	0 (0)	0 (0)	100 (1)	0 (0)	0 (0)

Figure 5 Coverage of ranked items (functional priorities (Q2)) in analyzed scoring systems. *UCLA*, University of California at Los Angeles; *ASES*, American Shoulder and Elbow Surgeons.

Limitations

This study has several limitations that need to be considered. First, with a responder rate of 41.6% (n = 164), of which 14% (n = 23) returned an incomplete questionnaire, the study is susceptible to ascertainment bias. It remains unclear to what extent cognitive impairments and/or miscomprehension of the study are responsible for this. Our elderly group (55.5%) had an average age of 71.6 years, but the known epidemiology of PHFs would suggest there exists a much older cohort that seem to be not covered adequately in this study. We believe that the responder rate is also a result of the chosen methodology of written questionnaire and may have been different in the case of systematic interviews. However, such a method would certainly involve substantially higher study personnel burden. Considering age and gender distribution in this cohort with a majority of women and patients over 60 years of age, this may be a representative PHF cohort. But the study design raises the question of whether it is appropriate to subordinate interests of minorities (men and young patients) to those of the majority in order to develop a single, consolidated outcome instrument.

In addition, the threshold of 90% importance and mean difference of at least 10% between the age and gender groups as a definition for essential core items has been arbitrarily chosen by the authors.

Furthermore, this survey was performed on a German cohort. Cultural/regional characteristics might have influenced our results. For example, considering the outcomes in Figure 4, one major difference between the core items on ASES vs. OSS is being able “to use a knife and fork at the same time.” A survey in a country where a knife is only used intermittently might have changed our final conclusion. Another methodological weakness inherent in the questionnaire is that many items are synonymous or representative of other functions. For example, the Neer Score includes “reaching the opposite axilla,” which may seem less important to a patient than the question about the ability to wash under the arms. We also found it challenging to combine items that appeared more than once in some scoring systems. For example, in the Neer Score, the ability to reach the region between the shoulder blades with the hand is rated once in the category “range of motion” as “internal rotation to T6” with 5 points, and once in the category “function” as “reaching brassiere hook” with 2 points, whereby the latter is certainly gender-dependent.

We included commonly used outcome measures in the PHF literature, but did not used the DASH score in order to prevent survey fatigue and the risk of a higher rate of incomplete questionnaires or nonresponders. This must be considered as a limitation as the DASH score is the second-most used outcome measure[17] and has strong reliability and moderately strong validity for assessing patients with PHF with high psychometric properties.[17,20] In retrospect, the DASH score could also have been included instead of the outdated Neer Score. For example, even if only 2 patients added “to write by hand” as an unlisted item, it may have been rated more important if included in the survey. However, 11 out of its 30 items are identical with those of the qDASH, and 2 further items were covered by the other instruments. This would result in a total coverage of 47% (14/30) of its items. Of note, similar to the qDASH, the DASH score does not contain any of the 4 essential ADL items identified in this study.

Finally, Kirkley et al advocated for the development of disease-specific questionnaires through a process that involves (1) disease-specific patient population identification, (2) generation of disease-specific items, (3) item-reduction, (4) pretesting the prototype instrument, (5) determination of reliability, and (6) determination of validity.[11,12] Our questionnaire was developed based on elements from pre-existing instruments, which themselves did not undergo the ideal process for questionnaire development for the PHF population. Therefore, it is possible that the questionnaire used for this study does not best capture patients’ perceptions or expectations.

Conclusion

We have identified 6 items from daily life that are of essential importance for patient-reported healthy shoulder function and treatment satisfaction for PHF regardless of age and gender. Until a reliable and valid scoring system for PHF is developed that includes these items, we recommend using the OSS, as it most closely reflects patient-reported interests with appropriate weighting of points.

Acknowledgments

We would like to thank the Traumastiftung gGmbH for funding this study, and we would like to thank the national project DEAL for enabling Open Access publication.

Disclaimers:

Funding: No funding was disclosed by the authors.
Conflicts of interest: The authors, their immediate families, and any research foundation with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jse.2025.03.029.

References

1. Baker P, Nanda R, Goodchild L, Finn P, Rangan A. A comparison of the Constant and Oxford Shoulder Scores in patients with conservatively treated proximal humeral fractures. J Shoulder Elbow Surg 2008;17:37-41. https://doi.org/10.1016/j.jse.2007.04.019
2. Brorson S, Elliott J, Thillemann T, Aluko P, Handoll H. Interventions for proximal humeral fractures: key messages from a Cochrane review. Acta Orthop 2022;93:610-2. https://doi.org/10.2340/17453674.2022.3495
3. Brorson S, Palm H. Proximal humeral fractures: the choice of treatment. In: Falaschi P, Marsh D, editors. Orthogeriatrics: the management of older patients with fragility fractures. Springer: Cham; 2021. p. 143-53.
4. Constant CR, Gerber C, Emery RJ, Sojbjerg JO, Gohlke F, Boileau P. A review of the Constant score: modifications and guidelines for its use. J Shoulder Elbow Surg 2008;17:355-61. https://doi.org/10.1016/j.jse.2007.06.022
5. Handoll HH, Brorson S. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2015;11:CD000434. https://doi.org/10.1002/14651858.CD000434.pub4
6. Handoll HH, Elliott J, Thillemann TM, Aluko P, Brorson S. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2022;6:CD000434. https://doi.org/10.1002/14651858.CD000434.pub5
7. Handoll HH, Gibson JN, Madhok R. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2003;4:CD000434. https://doi.org/10.1002/14651858.Cd000434
8. Handoll HH, Ollivere BJ. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2010;12:CD000434. https://doi.org/10.1002/14651858.CD000434.pub2
9. Handoll HH, Ollivere BJ, Rollins KE. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2012;12:CD000434. https://doi.org/10.1002/14651858.CD000434.pub3
10. Jawa A, Burnikel D. Treatment of proximal humeral fractures: a critical analysis review. JBJS Rev 2016;4:e2. https://doi.org/10.2106/jbjs.Rvw.O.00003
11. Kirkley A, Alvarez C, Griffin S. The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: the Western Ontario Rotator Cuff Index. Clin J Sport Med 2003;13:84-92. https://doi.org/10.1097/00042752-200303000-00004
12. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med 1998;26:764-72.
13. Luger M, Schopper C, Krottenthaler ES, Mahmoud M, Heyse T, Gotterbarm T, et al. Not all questions are created equal: the weight of the Oxford Knee Scores questions in a multicentric validation study. J Orthop Traumatol 2023;24:44. https://doi.org/10.1186/s10195-023-00722-6
14. Mallon WJ. Outcome instruments and their analysis. J Shoulder Elbow Surg 2024;33:1209-10. https://doi.org/10.1016/j.jse.2024.02.015
15. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons standardized shoulder assessment form, patient self-report section: reliability, validity, and responsiveness. J Shoulder Elbow Surg 2002;11:587-94. https://doi.org/10.1067/mse.2002.127096
16. Razaeian S, Wiese B, Zhang D, Harb A, Krettek C, Hawi N. Non-sensus in the treatment of proximal humerus fractures: uncontrolled, blinded, comparative behavioural analysis between Homo chirurgicus accidentus and Macaca sylvanus. BMJ 2020;371:m4429. https://doi.org/10.1136/bmj.m4429
17. Richard GJ, Denard PJ, Kaar SG, Bohsali KI, Horneff JG, Carpenter S, et al. Outcome measures reported for the management of proximal humeral fractures: a systematic review. J Shoulder Elbow Surg 2020;29:2175-84. https://doi.org/10.1016/j.jse.2020.04.006
18. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: a systematic review of four questionnaires. Arthritis Rheum 2009;61:623-32. https://doi.org/10.1002/art.24396
19. Slobogean GP, Noonan VK, Famuyide A, O'Brien PJ. Does objective shoulder impairment explain patient-reported functional outcome? A study of proximal humerus fractures. J Shoulder Elbow Surg 2011;20:267-72. https://doi.org/10.1016/j.jse.2010.06.005
20. Slobogean GP, Noonan VK, O'Brien PJ. The reliability and validity of the disabilities of arm, shoulder, and hand, EuroQol-5D, health utilities index, and short form-6D outcome instruments in patients with proximal humeral fractures. J Shoulder Elbow Surg 2010;19:342-8. https://doi.org/10.1016/j.jse.2009.10.021
21. van de Water AT, Shields N, Davidson M, Evans M, Taylor NF. Reliability and validity of shoulder function outcome measures in people with a proximal humeral fracture. Disabil Rehabil 2014;36:1072-9. https://doi.org/10.3109/09638288.2013.829529
22. van de Water AT, Shields N, Taylor NF. Outcome measures in the management of proximal humeral fractures: a systematic review of their use and psychometric properties. J Shoulder Elbow Surg 2011;20:333-43. https://doi.org/10.1016/j.jse.2010.10.028
23. Ziegler P, Kuhle L, Stockle U, Wintermeyer E, Stollhof LE, Ihle C, et al. Evaluation of the Constant score: which is the method to assess the objective strength? BMC Musculoskelet Disord 2019;20:403. https://doi.org/10.1186/s12891-019-2795-6