Quality of Life Research (2019) 28:2821–2829
https://doi.org/10.1007/s11136-019-02221-w

Measurement properties of PROMIS short forms for pain and function in orthopedic foot and ankle surgery patients

Anika Stephan[1] · Jens Mainzer[2,3] · Danica Kümmel[1] · Franco M. Impellizzeri[1,4]

Accepted: 29 May 2019 / Published online: 8 June 2019

Abstract

Purpose To examine the measurement properties of the German PROMIS short forms for pain intensity (PAIN), pain interference (PI) and physical function (PF) in orthopedic foot and ankle surgery patients.

Methods Patient-rated outcomes were collected from consecutive patients of our foot and ankle registry before and 6 months after surgery. Measurement properties were tested according to the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN). The German Foot Function Index (FFI-D) served as a legacy measure.

Results 748 patients were included in our cross-sectional sample. Longitudinal and test–retest data were available for 202 and 65 patients, respectively. Construct validity of all short forms was good. All Cronbach's α and intraclass correlation coefficients were > 0.7. The smallest detectable change (SDC) was highest for PF (8.9) and lowest for PI (6.5). Minimal important change was 4 to 5 points and thus smaller than SDC for all instruments. We observed a baseline ceiling effect for PF. PI showed insufficiently correlated change scores with FFI-D disability change scores, and therefore failed the responsiveness testing.

Conclusion Our study showed some adequate psychometric properties, but also certain aspects regarding interpretability and responsiveness that researchers must be aware of when using PROMIS short forms of pain and function in foot and ankle surgery patients.

Keywords PROMIS · Short forms · Psychometric validation · Pain · Function

Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS$^{®}$) aims to provide a common metric of health for many medical conditions [1]. PROMIS item banks are primarily designed for computer adaptive testing (CAT) which has proven to be time-efficient and precise because appropriate questions are selected from an item bank according to the patient's response and estimated score. Despite this methodological and technological progress, PROMIS static short forms of varying lengths remain in use, since they are easy to administer by means of a paper–pencil format, the preferred administration mode for many patients, or serve as static domain items in PROMIS profiles.

The validation of PROMIS pain and function measures in foot and ankle patients has only targeted CAT as the preferred method of administration without any consideration of short forms [2, 3]. Yet valid short forms would allow the use of PROMIS metrics for those groups of patients, practitioners or clinics who still prefer the form of analogue correspondence.

To strengthen the evidence for PROMIS measures in foot and ankle patients, samples with a wide range of conditions and from different populations are required [4]. So far, the validation studies mentioned above used the same data source: the National Orthopaedic Foot & Ankle Outcomes Research database with ten contributing clinics. Therefore, current evidence for PROMIS instruments in foot and ankle patients is solely based on data from approximately 300

✉ Anika Stephan
anika.stephan@kws.ch

[1] Department of Teaching, Research and Development– Lower Extremities, Schulthess Clinic, Lengghalde 2, 8008 Zurich, Switzerland

[2] Foot and Ankle Surgery, Schulthess Clinic, Lengghalde 2, 8008 Zurich, Switzerland

[3] IN MOTION, Richtiarkade 23, 8304 Wallisellen, Switzerland

[4] Faculty of Health, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

US patients undergoing one of six common elective foot and ankle surgeries (i.e. hallux rigidus, hallux valgus, hammertoe, flatfoot deformity, ankle instability, ankle arthritis). There is the need to broaden the evidence of PROMIS measures beyond this principal population.

Our aim is to evaluate if German PROMIS short forms can be regarded as valid tools for assessing pain intensity (PAIN), pain interference (PI) and physical function (PF) in foot and ankle surgery patients. Therefore, we examined the measurement properties of these instruments in a wide range of orthopedic foot and ankle surgery patients according to the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) [5, 6].

Materials and methods

Study design and questionnaire administration

This prospective study was approved by the Cantonal Ethics Committee of Zurich (KEK-ZH no. 2015-0258) and included consecutively enrolled patients of our foot and ankle registry between November 2016 and January 2018. This registry currently documents 70% of all foot and ankle patients surgically treated at our clinic up to 2 years post surgery. Patients are enrolled if they are at least 16 years of age and provide informed consent to use their data for research purposes. Exclusion criteria are living abroad; insufficient knowledge of the German language; cognitive impairment; or ongoing follow-up of former surgeries. Patient-reported outcomes (PROs) were collected from questionnaires administered 1 to 4 weeks before (baseline) and 6 months after surgery. This postoperative follow-up is sufficient for detecting a substantial health change in foot and ankle patients [2]. Starting in May 2017, a subsample of consecutive patients interviewed either at baseline or the 6-month follow-up completed questionnaires with a retest occurring within 14 days (median: 9 days; minimum: 2 days) for reliability testing until a sample size of 30 for each time point was reached (baseline: $n = 35$; 6-month follow-up: $n = 30$). Overall, we assessed PROMIS short form measurement properties with our cross-sectional, longitudinal and test–retest patient populations.

PRO questionnaires

We investigated the German PROMIS short forms for PAIN, PI and PF as provided by the PROMIS Germany research group. Answers are given on five-point verbal rating scales. For PAIN, we used form 3a (v1.0) that assesses pain over a 7-day recall period and current pain [7]. Form 4a (v1.0) defined PI based on the consequences of pain on relevant aspects of one's life over a 7-day recall period [8, 9]. For PF, we used form 4a (v2.0) [10, 11]. Patients reported their current ability to perform various physical activities. Overall scores for PAIN, PI and PF were presented as *T* scores; higher scores indicate more PAIN, higher PI and better PF. A PI and PF score of 50 (10) represents the US general population mean (SD). In contrast, the PAIN score is not centred on a general population mean [1]. Scoring was done using automated response pattern scoring [12] and missing items were not replaced. We decided to use the shortest available short forms because the shorter the instrument, the lower the administrative and respondent burden as well as the easier they are to implement in registry settings such as that found and used within our clinic.

As reference instruments for construct validity, we used a condition-specific instrument that assesses constructs encompassing the PROMIS domains and two single-item questions aimed at assessing the success of surgery as follows.

Specifically, we used the Foot Function Index (FFI), a region-specific instrument assessing pain and disability, which belongs to the most frequently used PRO tools in foot and ankle literature [13]. We used the modified German version FFI-D with 18 items covering subscales for pain (8 items) and disability (10 items) and a reference period of 1 week as a legacy measure [14]. The FFI-D pain subscale assesses the amount of pain at different times of the day and the degree of pain while walking or standing under different conditions. The FFI-D disability subscale measures difficulties in walking under various conditions and during some lower limb motoric tasks as well as restrictions experienced while undertaking leisure activities and depending on the choice of various shoe types. Answers are provided on a ten-point numeric rating scale. Subscales were recorded as the achieved score relative to the maximum achievable score of all answered scale items. The higher the score, the higher the pain or impairment. Two missing items were permitted for each subscale. The German version was shown to be feasible, showed excellent internal consistency and test–retest reliability for both subscales (FFI-D pain: Cronbach's $\alpha = 0.9$, intraclass correlation coefficient (ICC) = 0.97; FFI-D disability: Cronbach's $\alpha = 0.95$, ICC = 0.99) and correlated with comparable instruments in a sample of 53 foot surgery patients aged 18 to 77 years with mainly fore- or hindfoot problems [14]. Test–retest reliability was further quantified by Bland–Altman 95% limits of agreement and showed a non-significant bias for FFI-D pain and disability of 0.3 ± 2.5 and -0.6 ± 2.9 random error, respectively. Recently published reference values for FFI-D are available [15].

At 6 months, patients rated their global treatment outcome (GTO): "*How much did the operation help with your foot problem?*" on a five-point Likert scale ranging from "helped a lot" to "made things worse" [16]. They also defined their state of symptom-specific well-being (SSWB): "*If you had to spend the rest of your life with the symptoms you have now at your foot, how would you feel about it?*" on a five-point Likert scale ranging from "very satisfied" to "very dissatisfied" [17].

PROMIS measurement properties

Structure was assessed in the cross-sectional sample by confirmatory factor analysis with Satorra–Bentler adjustments (CFA). Due to the small number of items per scale, CFA was done for the three short forms together with each scale being considered as one factor, and factors being allowed to covary. Structural validity was demonstrated if data fitted the predefined factor structure and at least 3 fit indices were considered at least "good" (Comparative Fit Index [CFI] > 0.9, Tucker–Lewis Index [TLI] > 0.9, root-mean-square error of approximation [RMSEA] < 0.05, standardized root-mean-square residual [SRMR] < 0.08) [18].

Construct validity was assessed using scale-specific hypotheses testing and considered good if at least 75% of the hypotheses were confirmed. We tested the correlations between PAIN and FFI-D pain, PI and PF with FFI-D disability, and PAIN, PI and PF with SSWB. Correlations were expected to be strong (≥ 0.6) and positive for PAIN and PI, and negative for PF.

Internal consistency was calculated using Cronbach's α with values between 0.7 and 0.95 indicating appropriate internal consistency [19]. Test–retest reliability was assessed with ICC from a single measurement, absolute agreement, two-way mixed-effects model; an ICC ≥ 0.7 was considered appropriate [19]. Agreement was assessed using the Standard Error of Measurement (SEM agreement = $\sqrt{}$(variance due to systematic differences between measurements + residual variance)). The smallest detectable change (SDC) for individuals that can be considered above the measurement error with a 90% confidence level was calculated as SDC90 = 1.65 * $\sqrt{2}$ * SEM agreement [20].

Responsiveness refers to the ability of a questionnaire to detect clinically important changes over time. Longitudinal validity can be considered a measure of responsiveness and is examined by inspecting the correlation of the change score with the change score of the reference instrument [20]. We assessed this aspect with predefined hypotheses: We expected moderate positive correlations ($r \geq 0.5$) between change scores of PAIN and FFI-D pain, of PI and FFI-D disability and moderate negative correlations ($r \leq -0.5$) between change scores of PF and FFI-D disability. We had the same expectations for correlations between change scores of PAIN, PI, PF and the GTO. Hypotheses on effect sizes were included for PI and PF because these PROMIS score outcomes have been previously published for foot and ankle surgery over a comparable follow-up period [21, 22]. Based on these results, we expected Cohen's d of at least ≥ 0.5 with decreasing PI and $d \geq 0.7$ with increasing PF. Responsiveness was considered sufficient if at least 75% of the hypotheses were confirmed.

Floor and ceiling effects were considered absent if percentages were below 15% [19]. To determine the individual-level minimal important change (MIC), we analysed the area under the receiver operating characteristics curve (AUC) with the GTO set as the anchor. Patients who stated that the operation "helped" or "helped a lot" were considered as having a good outcome; all other responses indicated a poor outcome for an invasive intervention such as orthopedic foot or ankle surgery. An AUC ≥ 0.7 was considered adequate [19]. Change scores of the longitudinal sample and its GTO subgroups were calculated for interpretation purposes.

Analyses were performed using StataCorp. 2015 Stata Statistical Software: Release 14 (StataCorp LP, TX, USA). AUC analysis was done with the Stata ROCMIC module [23].

Results

Patient characteristics

Figure 1 outlines the patient selection scheme and Table 1 shows baseline demographics, pain, and functional status. Forty-two percent took painkillers due to foot pain. The percentage of smokers was 19.5%. Twenty-nine percent of the patients were retired and the proportion of part-time workers was 34%. Fifty-five percent of the surgeries targeted the forefoot (mainly scarf-osteotomy as well as Morton's neuroma excision, arthrodesis, cheilectomy), 8% the midfoot (mainly arthrodesis) and 37% the hindfoot (i.e. arthrodesis, calcaneal osteotomies, debridement, exostosis removal).

Structure, construct validity and internal consistency

CFA confirmed the three-dimensional structure of the three PROMIS scales. The CFI was 0.95, SRMR was 0.04 and TLI was close to the cut-off criteria with 0.93, which demonstrates a good fit. RMSEA was 0.106. The covariance between PAIN and PI was 0.65, PAIN and PF − 0.51, and PI and PF − 0.69. Scale-specific hypothesis testing for validity resulted in 100% confirmed hypotheses for PAIN and PF and 89% for PI (Table 2). Cronbach's α was greater than 0.7 for all scales (Table 3).

Test–retest reliability and agreement

ICC was greater than 0.7 for all scales (Table 3). The PF scale had the highest SEM agreement value of 3.8 (SDC90: 8.9) in contrast to PI (SEM agreement: 2.8; SDC90: 6.5).

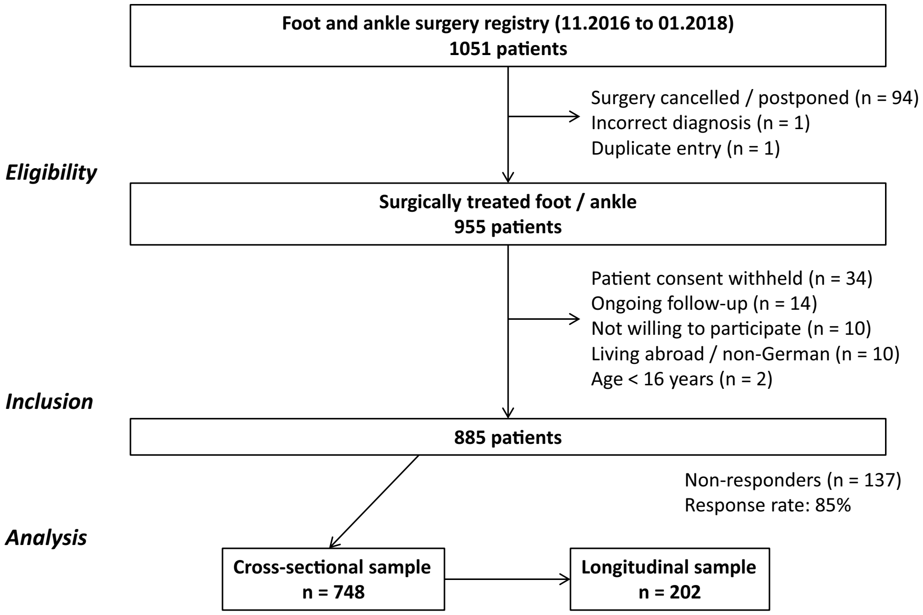
Foot and ankle surgery registry (11.2016 to 01.2018)
1051 patients
Surgery cancelled / postponed (n = 94)
Incorrect diagnosis (n = 1)
Duplicate entry (n = 1)
Eligibility
Surgically treated foot / ankle
955 patients
Patient consent withheld (n = 34)
Ongoing follow-up (n = 14)
Not willing to participate (n = 10)
Living abroad / non-German (n = 10)
Age < 16 years (n = 2)
Inclusion
885 patients
Non-responders (n = 137)
Response rate: 85%
Analysis
Cross-sectional sample
n = 748
Longitudinal sample
n = 202

Fig. 1 Flow chart showing patient eligibility and available sample sizes for assessing the psychometric properties of the German PROMIS short forms

Table 1 Patient characteristics for the three analysed samples at baseline

Characteristics[a]	Cross-sectional (N=748)	Longitudinal (N=202)	Test–retest (N=65)
Age (years)	56.1 (14.9)	55.4 (14.2)	57.7 (17.2)
Gender (female) (n, %)	512 (68)	139 (69)	49 (75)
Height (cm)	164.9 (7.3)	169.7 (9.8)	167.6 (9.9)
Weight (kg)	68.7 (14.5)	73.4 (15.9)	72.8 (16.5)
Body Mass Index (kg/m^2)	25.2 (4.9)	25.4 (4.7)	25.8 (5.3)
PROMIS PAIN (T score)	51.3 (8.1)[b]	51.6 (8.0)	48.0 (10.2)
PROMIS PI (T score)	59.3 (7.6)[b]	59.1 (7.5)	56.0 (9.0)
PROMIS PF (T score)	42.2 (7.6)[b]	42.7 (7.7)	44.4 (9.0)
FFI-D pain	49.8 (20.5)[c]	49.0 (20.8)	53.3 (22.3)
FFI-D disability	52.4 (24.0)[c]	50.2 (23.8)	49.1 (25.6)

PROMIS Patient-Reported Outcomes Measurement Information System, *PAIN* pain intensity, *PI* pain interference, *PF* physical function, *T score* overall PROMIS score calculated per domain, *FFI-D* Foot Function Index German version, *SD* standard deviation

[a] Expressed as mean (SD) unless otherwise stated

[b] All T scores could be calculated for all cases, single PROMIS items were missing in 2 to 4 cases per scale

[c] FFI-D subscale scores could not be calculated for 58 patients (7.8%) due to > 2 missing items

Responsiveness

Hypothesis testing for responsiveness resulted in 83% confirmed hypotheses for PAIN, 33% for PI and 56% for PF. Patients' individual change scores on the PROMIS scales and respective FFI-D scales are plotted in Fig. 2. Table 4 presents the 6-month follow-up scores and mean changes in these scales; PROMIS T scores decreased significantly. Cohen's d exceeded the preset criteria for PI and PF (Table 3). After surgery, we observed floor effects with PAIN and PI, and a ceiling effect with PF (Table 5). Due to a baseline ceiling effect (15%) for PF—which could be presumed to negatively affect the strength of correlation with the FFI-D disability change scores and GTO categories—we conducted a sub-analysis involving the exclusion of cases (n = 29) with a maximum PF baseline score and "good" outcome. Based on this restriction, correlations with FFI-D disability change scores were not affected. However, correlations with the GTO categories increased and hypotheses could not only be confirmed for males as in the original analysis, but as well for the whole sample and for females. Thus, testing for responsiveness with this sub-analysis resulted in 78% confirmed hypotheses in PF.

Table 2 Correlations between PROMIS scales, FFI-D, SSWB and GTO

	Correlation with FFI-D[a]	Correlation with SSWB[b]	Correlation with GTO[b]
PROMIS PAIN			
Baseline	0.68[c]		
6 months	0.64[c]	0.69[d]	
Change	0.56[e,f]		0.52[g]
PROMIS PI			
Baseline	0.70[h]		
6 months	0.67[h]	0.60[d]	
Change	0.42[f,i]		0.47[g]
PROMIS PF			
Baseline	– 0.76[h]		
6 months	– 0.67[h]	– 0.63[d]	
Change	– 0.48[f,i]		– 0.45[g]

PROMIS Patient-Reported Outcomes Measurement Information System, *FFI-D* Foot Function Index German version, *SSWB* symptom-specific well-being, *GTO* global treatment outcome, *PAIN* pain intensity, *PI* pain interference, *PF* physical function

[a] Pearson's correlation coefficient (r)

[b] Spearman's rank correlation coefficient (r_s)

[c] Correlation with FFI-D pain, gender-specific correlations r were ≥ 0.6

[d] Gender-specific correlations $|r_s|$ were ≥ 0.6, except for male PI ($r_s = 0.56$)

[e] Correlation with FFI-D pain change score

[f] Gender-specific correlations $|r|$ were > 0.5 for PAIN and PF (male) and < 0.5 for PI and PF (female)

[g] Gender-specific correlations $|r_s|$ were > 0.5 for PAIN (male) and PF (male), and < 0.5 for PAIN (female), PI and PF (female)

[h] Correlation with FFI-D disability, gender-specific correlations $|r|$ were ≥ 0.6

[i] Correlation with FFI-D disability change score

Minimal important change

The dichotomized GTO was "good" for 82.2% of our patients (Table 4). Change scores of 4 to 5 points were identified as the best cut-off points for the MIC (Table 3).

Discussion

We assessed the measurement properties of the shortest PROMIS forms of pain and function in a diverse sample of orthopedic foot and ankle surgery patients. Our results suggest that construct validity of the PROMIS short forms is good. They have good internal consistency and test–retest reliability, which is in line with results reported for other patient groups [24–26].

Compared to other foot and ankle surgery patients [21, 22] where a PF baseline score of 34 was reported, our patients had higher PF and comparable PI. We believe this result could be attributed to a different distribution of fore- and hindfoot disorders in our study; we have a particularly high proportion of hallux valgus patients who did not necessarily show any signs of intense functional impairment. As expected, FFI-D disability was higher than the population average of around 16 in 50–59-year-old males and 19 in females of the same age range [15]. FFI-D pain was considerably higher than the population average, which is around 14 and 19 in 50–59-year-old males and females, respectively [15]. At baseline, 42% of our sample reported taking painkillers specifically due to foot pain. After 6 months, our sample showed a substantial decline in PAIN and FFI-D pain, the latter still being half a SD above the average of 50–59 year olds [15]. The interpretability of the PAIN score is limited, since PAIN is not centred on a general population mean [1]. Regarding PI and PF, our patient sample achieved almost average values, whereas the foot-specific FFI-D disability score (like the FFI-D pain score) was half a SD above the average of 50–59 year olds.

Internal consistency and test–retest reliability were good for all three PROMIS short forms. We detected measurement errors of three points (PAIN, PI) and four points (PF) on the *T* scale. With the SCD90 being larger than the MIC, clinically relevant changes cannot be distinguished from measurement error for an individual patient. This problem is not uncommon for patient-reported outcome measures [20]. In technical terms, the resolution of the instrument is not good enough to detect the differences we are interested in. Even if we take into account that measurement error is not constant through the score range in IRT-based instruments [27–29], the MIC cannot be detected in any range of PAIN and PF. However, the MIC can be detected in the PI scale for patients who change within the *T* score range of 52 to 72. An illustration of the magnitude of SDCs based on different SEM assumptions and the location of the MIC within these results is highlighted in Fig. 3.

Responsiveness was supported for PAIN and PF, but not for PI. While correlations with the anchor question and effect sizes were acceptable, PI showed insufficiently correlated change scores with those of the FFI-D disability subscale. It should be noted that responsiveness of the FFI-D subscales is unknown. Although the FFI and FFI-D are frequently used for foot and ankle patients, we could not find any studies demonstrating responsiveness, especially within the tight definition outlined by COSMIN (i.e. correlation of the change score with the change score of the reference instrument). Furthermore, we need to critically discuss the thresholds that were set for hypothesis testing. PI failed the chosen criteria for hypothesis testing only slightly by correlating between 0.45 and 0.49 with the anchor question instead of ≥ 0.5. Our threshold is based on the recommendations of Guyatt et al. [30], but more currently recommended

Table 3 Reliability, smallest detectable change, minimal important change and effect size

	Cronbach's α[a]	ICC[a]	$\text{SEM}_{\text{agreement}}$	SDC90	MIC[b]	AUC[a]	Cohen's d
PROMIS PAIN	0.87 (0.85; 0.89)	0.90 (0.84; 0.94)	3.24	7.56	– 4.0[c]	0.79 (0.71; 0.86)	– 1.2[f]
PROMIS PI	0.93 (0.92; 0.94)	0.91 (0.86; 0.94)	2.77	6.45	– 5.0[d]	0.79 (0.71; 0.87)	– 1.2[g]
PROMIS PF	0.91 (0.9; 0.92)	0.81 (0.71; 0.88)	3.82	8.92	4.6[e]	0.80 (0.72; 0.87)	0.8[h]

Sensitivity—the probability that the change score will be above the threshold when improvement is present (true-positive rate) and specificity—the probability that the change score will be below the threshold when improvement is not present (true-negative rate)

ICC intraclass correlation coefficient, $SEM_{agreement}$ agreement standard error of measurement, *SDC90* smallest detectable change calculated with the 90% confidence interval of the $\text{SEM}_{\text{agreement}}$, *MID* minimal important difference, *MIC* minimal important change, *AUC* area under the receiver operating characteristics curve, *PROMIS* Patient-Reported Outcomes Measurement Information System, *PAIN* pain intensity, *PI* pain interference, *PF* physical function

[a] 95% confidence interval in parentheses

[b] MIC was estimated using the smallest sum of squares of 1-sensitivity and 1-specificity. Patients who stated that the operation "helped" or "helped a lot" were considered as having a good outcome (external criterion variable = "1"); all other responses indicated a poor outcome (external criterion variable = "0"). The change score was calculated as baseline minus follow-up. MIC thresholds were finally multiplied with "– 1" to be consistent with the direction of the scales

[c] Sensitivity: 81%; Specificity: 64%

[d] Sensitivity: 74%; Specificity: 69%

[e] Sensitivity: 62%; Specificity: 81%

[f] Males: – 1.0, females: – 1.3

[g] Males: – 0.9, females: – 1.3

[h] Males: 0.7, females: 0.8

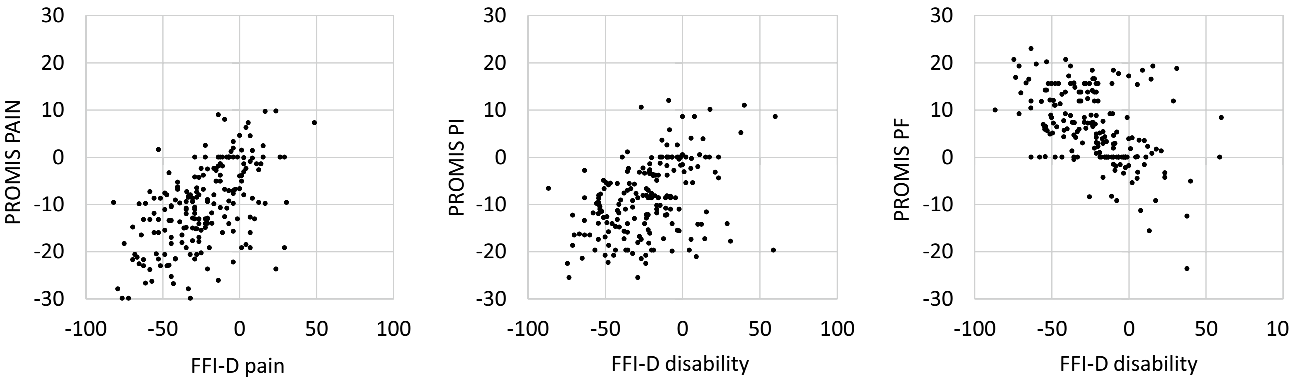

Fig. 2 Scatterplots of individual change scores (*PROMIS* Patient-Reported Outcomes Measurement Information System, *PAIN* pain intensity, *PI* pain interference, *PF* physical function, *FFI-D* Foot Function Index German version)

lower thresholds of 0.3 exist [31]. Even if we had chosen this lower threshold, the confirmed hypotheses would still lie below the requirement of 75% because of the low correlation between the PI and FFI-D disability change scores. Pearson's r ranging from 0.39 to 0.43 failed our predefined criteria of $r \geq 0.5$.

PF revealed a ceiling effect as reported elsewhere [26, 32]. This ceiling effect negatively affected the responsiveness. The PF correlation with the GTO categories did not reach the predefined cut-off because patients with a good outcome, who were already at the end of the PF scale at baseline, could not improve further and therefore had a change score of "0". Indeed, when the ceiling was excluded, this led to an improvement in responsiveness. Overall, this ceiling effect should be considered when interpreting longitudinal PF results.

We do not rate the floor effects of PI and PAIN after surgery as critical, since these scales represent unipolar constructs, where the upper end of the scale indicates severity and the lower end indicates its absence [33]. Nevertheless, it is important to recognize the higher measurement error at the lower end of the scale when interpreting change scores for patients initially reporting low PI and PAIN.

We used the original COSMIN checklist [5] as a guiding framework for conducting our analyses, defining thresholds and reporting key elements. We acknowledge the

Table 4 Follow-up scores at 6 months post surgery in the longitudinal sample and respective change scores

Characteristics[a,b]	Follow-up	Change					
			GTO subgroups				
			“Good outcome“(*n* = 166)		“Bad outcome” (*n* = 36)		
			Helped a lot (*n* = 104)	Helped (*n* = 62)	Helped only little (*n* = 26)	Did not help (*n* = 9)	Made things worse (*n* = 1)
PROMIS PAIN	41.5 (8.6)	– 10.2 (8.8)	– 14.4 (7.6)	– 7.2 (7.7)	– 3.7 (6.4)	– 0.97 (8.1)	0
PROMIS PI	50.4 (7.7)	– 8.7 (8.0)	– 12.0 (7.1)	– 7.0 (7.1)	– 1.0 (6.9)	5.0 (7.9)	Missing
PROMIS PF	48.8 (7.8)	6.1 (7.8)	9.0 (7.2)	5.1 (6.6)	– 0.5 (5.5)	– 1 (5.5)	– 23.6
FFI-D pain	25.0 (23.0)	– 23.8 (25.1)	– 32.3 (24.2)	– 20.2 (21.3)	– 3.7 (19.1)	– 13.5 (31.1)	29.1
FFI-D disability	27.9 (25.0)	– 22.3 (25.9)	– 29.8 (25.3)	– 19.8 (22.7)	– 2.3 (22.4)	– 15.8 (24.5)	37.8

GTO Global Treatment Outcome, *PROMIS* Patient-Reported Outcomes Measurement Information System, *PAIN* pain intensity, *PI* pain interference, *PF* physical function, *FFI-D* Foot Function Index German version

[a] Expressed as mean (SD)

[b] Score changes for PROMIS instruments refer to the *T* scores and score changes for FFI-D refer to the achieved score relative to the maximum achievable score of all answered scale items

Table 5 Frequencies of lowest and highest possible scores indicating floor and ceiling effects for PROMIS scales and FFI-D

	Lowest possible score (%)	Highest possible score (%)
PROMIS PAIN		
Baseline	3.4	1.3
6 months	27.2	0.0
PROMIS PI		
Baseline	7.1	2.8
6 months	35.8	0.0
PROMIS PF		
Baseline	0.3	14.7[a]
6 months	0.0	40.6
FFI-D pain		
Baseline	0	0.0
6 months	9.4	0.0
FFI-D disability		
Baseline	0.0	0.0
6 months	5.1	0.0

PROMIS Patient-Reported Outcomes Measurement Information System, *PAIN* pain intensity, *PI* pain interference, *PF* physical function, *FFI-D* Foot Function Index German version, *Baseline* scores from the cross-sectional sample, *6 months* 6-month follow-up scores from the longitudinal sample

[a] The respective value for the baseline longitudinal sample is 16.3%

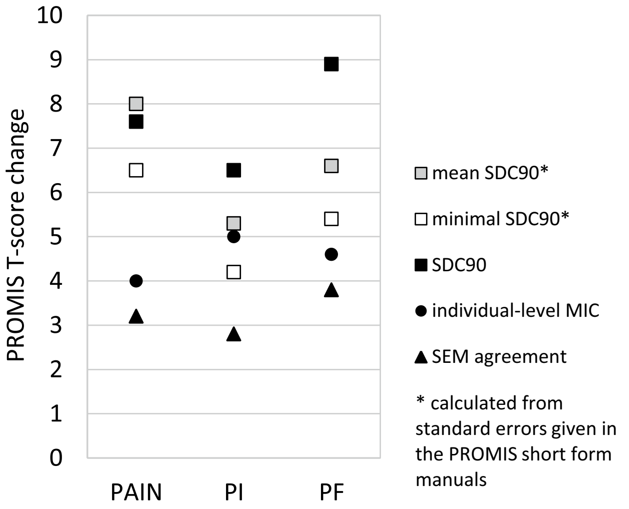

Fig. 3 Location of the MIC, SEM agreement and SDC90, calculated from study data (black symbols). *T* score-specific standard errors given in the respective PROMIS short form manuals were used to calculate alternative SDC90 thresholds based on the mean (grey) and minimal (white) standard error values. *PROMIS* Patient-Reported Outcomes Measurement Information System, *T score* overall PROMIS score calculated per domain, *PAIN* pain intensity, *PI* pain interference, *PF* physical function, *SDC90* smallest detectable change, calculated with a 90% confidence interval, *MIC* minimal important change, *SEM agreement* standard error of measurement

ongoing development of different versions of the checklist for different purposes such as for the study design, determining the risk of bias [6] and reporting. We also appreciate the methodological discussions, particularly concerning the testing of responsiveness, i.e. the appropriateness and interpretation of effect sizes as contributing criteria, guidelines for generating hypotheses and the ratio of hypotheses needed to be confirmed for the instrument being judged responsive [34]. Our study provides a good example of how different tests of responsiveness (i.e. correlation of change scores, correlations with an anchor or thresholds for effect sizes) might pass or miss predefined thresholds and make it difficult to draw an overall conclusion. Thus, we regard our results as preliminary, which should be confirmed or extended by further studies.

Conclusion

Our results confirmed the structural and construct validity of PROMIS short forms for pain and function. Responsiveness was acceptable for PAIN and PF, even if one should consider the initial physical function of patients before using the PF short form because of apparent ceiling effects and loss of responsiveness. On the contrary, PI responsiveness might be insufficient, based on our hypotheses, because the change scores insufficiently correlated with FFI-D disability change scores. While reliability was good, the scale-specific MICs calculated in the present study cannot be distinguished from measurement error, although this degree of error seems reasonable for patient-reported outcomes. When the SDC90 is calculated from the *T* score-specific standard error provided by the PROMIS short form manuals, MIC can be detected in the PI scale for patients who change within the *T* score range of 52 to 72.

In conclusion, the PROMIS short forms for PAIN, PI and PF showed some adequate psychometric properties, but also certain limitations that should be taken into account when using these tools in foot and ankle surgery patients.

Acknowledgements The authors thank Melissa Wilhelmi (Schulthess Clinic) for manuscript editing and critical suggestions as well as the Lower Extremities Research and Development team members Laura Duc, Selina Nauer, Myrta Rüegger and Alexandra Stettler for their valuable assistance with the data collection.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194.
2. Hung, M., Baumhauer, J. F., Brodsky, J. W., Cheng, C., Ellis, S. J., Franklin, J. D., et al. (2014). Psychometric comparison of the PROMIS Physical Function CAT with the FAAM and FFI for measuring patient-reported outcomes. *Foot and Ankle International, 35*(6), 592–599.
3. Hung, M., Baumhauer, J. F., Latt, L. D., Saltzman, C. L., SooHoo, N. F., Hunt, K. J., et al. (2013). Validation of PROMIS (R) Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clinical Orthopaedics and Related Research, 471*(11), 3466–3474.
4. Slullitel, G. A. (2017). CORR Insights((R)): PROMIS Pain Interference and Physical Function scores correlate with the Foot and Ankle Ability Measure (FAAM) in patients with Hallux Valgus. *Clinical Orthopaedics and Related Research, 475*(11), 2781–2782.
5. Mokkink, L. B., Terwee, C. B., Patrick, C. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research, 19*(4), 539–549.
6. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, C. L., Alonso, J., Bouter, L. M., et al. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179.
7. Pain Intensity—Scale. (2016). http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20Scale%20v1.0%20-%20Pain%20Intensity%203a%204-6-2017.pdf. Accessed November 2016.
8. Pain Interference—Short Form 4a. (2016). http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20SF%20v1.0%20-%20Pain%20Interference%204a%206-2-2016.pdf. Accessed November 2016.
9. Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W. H., Choi, S., Revicki, D., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150*(1), 173–182.
10. Physical Function—Short Form 4a. (2016). http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20SF%20v2.0%20-%20Physical%20Function%204a%2011-29-2016.pdf. Accessed November 2016.
11. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E., Jr. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526.
12. HealthMeasures Scoring Service. www.assessmentcenter.net/ac_scoringservice. Accessed May 2018.
13. Sierevelt, I. N., Zwiers, R., Schats, W., Haverkamp, D., Terwee, C. B., Nolte, P. A., et al. (2018). Measurement properties of the most commonly used foot- and ankle-specific questionnaires: The FFI, FAOS and FAAM. A systematic review. *Knee Surgery, Sports Traumatology, Arthroscopy, 26*(7), 2059–2073.
14. Naal, F. D., Impellizzeri, F. M., Huber, M., & Rippstein, P. F. (2008). Cross-cultural adaptation and validation of the Foot Function Index for use in German-speaking patients with foot complaints. *Foot and Ankle International, 29*(12), 1222–1228.
15. Schneider, W., & Jurenitsch, S. (2016). Age- and sex-related normative data for the Foot Function Index in a German-speaking cohort. *Foot and Ankle International, 37*(11), 1238–1242.
16. Impellizzeri, F. M., Mannion, A. F., Naal, F. D., Hersche, O., & Leunig, M. (2012). The early outcome of surgical treatment for femoroacetabular impingement: Success depends on how you measure it. *Osteoarthritis and Cartilage, 20*(7), 638–645.
17. Mannion, A. F., Elfering, A., Staerkle, R., Junge, A., Grob, D., Semmer, N. K., et al. (2005). Outcome assessment in low back pain: How low can you go? *European Spine Journal, 14*(10), 1014–1026.
18. Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

19. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42.
20. de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine*. Cambridge: Cambridge University Press.
21. Ho, B., Houck, J. R., Flemister, A. S., Ketz, J., Oh, I., DiGiovanni, B. F., et al. (2016). Preoperative PROMIS scores predict postoperative success in foot and ankle patients. *Foot and Ankle International, 37*(9), 911–918.
22. Anderson, M. R., Houck, J. R., Saltzman, C. L., Hung, M., Nickisch, F., Barg, A., et al. (2018). Validation and generalizability of preoperative PROMIS scores to predict postoperative success in foot and ankle patients. *Foot and Ankle International, 39*(7), 763–770.
23. Froud, R. (2009). *ROCMIC: Stata module to estimate minimally important change (MIC) thresholds for continuous clinical outcome measures using ROC curves*. Statistical Software Components S457052, Boston College Department of Economics, revised 24 Oct 2014. https://ideas.repec.org/c/boc/bocode/s457052.html. Accessed December 2018.
24. Bartlett, S. J., Orbai, A. M., Duncan, T., DeLeon, E., Ruffing, V., Clegg-Smith, K., et al. (2015). Reliability and validity of selected PROMIS measures in people with rheumatoid arthritis. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0138543.
25. Crins, M. H., Roorda, L. D., Smits, N., de Vet, H. C., Westhovens, R., Cella, D., et al. (2015). Calibration and validation of the Dutch-Flemish PROMIS Pain Interference item bank in patients with chronic pain. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0134094.
26. Liegl, G., Rose, M., Correia, H., Fischer, H. F., Kanlidere, S., Mierke, A., et al. (2017). An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions. *Clinical Rehabilitation*. https://doi.org/10.1177/0269215517714297.
27. Pain Intensity. (2017). A brief guide to the PROMIS® Pain Intensity instruments. http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20Pain%20Intensity%20Scoring%20Manual.pdf. Accessed February 2018.
28. Pain Interference. (2018). A brief guide to the PROMIS© Pain Interference instruments. http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20Pain%20Interference%20Scoring%20Manual.pdf. Accessed February 2018.
29. Physical Function. (2018). A brief guide to the PROMIS® Physical Function instruments. http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20Physical%20Function%20Scoring%20Manual.pdf. Accessed February 2018.
30. Guyatt, G. H., Norman, G. R., Juniper, E. F., & Griffith, L. E. (2002). A critical look at transition ratings. *Journal of Clinical Epidemiology, 55*(9), 900–908.
31. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109.
32. Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., et al. (2015). Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Quality of Life Research, 24*(10), 2333–2344.
33. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.
34. Angst, F. (2011). The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Medical Research Methodology*. https://doi.org/10.1186/1471-2288-11-152.