Clin Orthop Relat Res (2023) 00:1-13
DOI 10.1097/CORR.0000000000002798

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Clinical Research

OPEN

How Are Age, Gender, and Country Differences Associated With PROMIS Physical Function, Upper Extremity, and Pain Interference Scores?

Constantin Yves Plessen MSc[1,2], Gregor Liegl PhD[1], Claudia Hartmann MBA[1], Marilyn Heng PhD[3], Alexander Joeris PhD[4], Aaron J. Kaat PhD[5], Benjamin D. Schalet PhD[5], Felix Fischer PhD[1], Matthias Rose PhD[1]

Received: 31 October 2022 / Accepted: 5 July 2023 / Published online: 30 August 2023

Abstract

Background The interpretation of patient-reported outcomes requires appropriate comparison data. Currently, no patient-specific reference data exist for the Patient-Reported Outcome Measurement Information System (PROMIS) Physical Function (PF), Upper Extremity (UE), and Pain Interference (PI) scales for individuals 50 years and older.

Questions/purposes (1) Can all PROMIS PF, UE, and PI items be used for valid cross-country comparisons in these domains among the United States, the United Kingdom, and Germany? (2) How are age, gender, and country related to PROMIS PF, PROMIS UE, and PROMIS PI scores? (3) What is the relationship of age, gender, and country across

For the Advancing Outcomes and Building Expertise in Research for Trauma Consortium.
The institution of one or more of the authors (Charité – Universitätsmedizin Berlin) has received, during the study period, funding from AO Trauma, AO Foundation, Switzerland. One author (AJ) is an employee of AO Trauma, AO Foundation, Switzerland.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
Ethical approval for this study was obtained from Charité – Universitätsmedizin Berlin, Berlin, Germany (EA4/212/20).
This work was performed at the Department of Psychosomatic Medicine, Charité – Universitätsmedizin Berlin, Germany for the AO Trauma, AO Foundation, Switzerland; in collaboration with AO Innovation Translation Center, Clinical Science, AO Foundation, Switzerland; the Department of Orthopaedic Surgery, Orthopaedic Trauma Service, Massachusetts General Hospital, Boston, MA, USA; and the Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

[1]Department of Psychosomatic Medicine, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Department of Clinical, Neuro-, and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[3]Department of Orthopaedic Surgery, Orthopaedic Trauma Service, Massachusetts General Hospital, Boston, MA, USA

[4]AO Innovation Translation Center, Clinical Science, AO Foundation, Duebendorf, Switzerland

[5]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

C. Y. Plessen ✉, Department of Psychosomatic Medicine, Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany, Email: constantin-yves.plessen@charite.de

individuals with PROMIS PF, PROMIS UE, and PROMIS PI scores ranging from very low to very high?

Methods We conducted telephone interviews to collect custom PROMIS PF (22 items), UE (eight items), and PI (eight items) short forms, as well as sociodemographic data (age, gender, work status, and education level), with participants randomly selected from the general population older than 50 years in the United States (n = 900), United Kingdom (n = 905), and Germany (n = 921). We focused on these individuals because of their higher prevalence of surgeries and lower physical functioning. Although response rates varied across countries (14% for the United Kingdom, 22% for Germany, and 12% for the United States), we used existing normative data to ensure demographic alignment with the overall populations of these countries. This helped mitigate potential nonresponder bias and enhance the representativeness and validity of our findings. We investigated differential item functioning to determine whether all items can be used for valid cross-cultural comparisons. To answer our second research question, we compared age groups, gender, and countries using median regressions. Using imputation of plausible values and quantile regression, we modeled age-, gender-, and country-specific distributions of PROMIS scores to obtain patient-specific reference values and answer our third research question.

Results All items from the PROMIS PF, UE, and PI measures were valid for across-country comparisons. We found clinically meaningful associations of age, gender, and country with PROMIS PF, UE, and PI scores. With age, PROMIS PF scores decreased (age ß$_{\text{Median}}$ = -0.35 [95% CI -0.40 to -0.31]), and PROMIS UE scores followed a similar trend (age ß$_{\text{Median}}$ = -0.38 [95% CI -0.45 to -0.32]). This means that a 10-year increase in age corresponded to a decline in approximately 3.5 points for the PROMIS PF score—a value that is approximately the minimum clinically important difference (MCID). Concurrently, we observed a modest increase in PROMIS PI scores with age, reaching half the MCID after 20 years. Women in all countries scored higher than men on the PROMIS PI and 1 MCID lower on the PROMIS PF and UE. Additionally, there were higher T-scores for the United States than for the United Kingdom across all domains. The difference in scores ranged from 1.21 points for the PROMIS PF to a more pronounced 3.83 points for the PROMIS UE. Participants from the United States exhibited up to half an MCID lower T-scores than their German counterparts for the PROMIS PF and PROMIS PI. In individuals with high levels of physical function, with each 10-year increase in age, there could be a decrease of up to 4 points in PROMIS PF scores. Across all levels of upper extremity function, women reported lower PROMIS UE scores than men by an average of 5 points.

Conclusion Our study provides age-, gender-, and country-specific reference values for PROMIS PF, UE, and PI scores, which can be used by clinicians, researchers, and healthcare policymakers to better interpret patient-reported outcomes and provide more personalized care. These findings are particularly relevant for those collecting patient-reported outcomes in their clinical routine and researchers conducting multinational studies. We provide an internet application (www.common-metrics.org/PROMIS_PF_and_PI_Reference_scores.php) for user-friendly accessibility in order to perform age, gender, and country conversions of PROMIS scores. Population reference values can also serve as comparators to data collected with other PROMIS short forms or computerized adaptive tests.

Level of Evidence Level II, diagnostic study.

Introduction

Patient-reported outcome measures (PROMs) play a crucial role in evaluating health outcomes, and the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative provides standardized instruments for assessing various health domains. Adequate reference values in any population where a PRO score will be used are essential for accurate interpretation and comparability. However, existing research on PROMIS scores suggests that age, gender, and country might impact PROMIS Physical Function (PF), Upper Extremity (UE), and Pain Interference (PI) scores [8, 13, 36]. Women generally tend to report lower PF and higher PI scores than men [24, 34], and older patients exhibit a decline in PF and an increase in PI as they age [24]. Notably, cultural, social, and healthcare-related factors contribute to variations in PROMIS scores across countries, emphasizing the importance of country-specific norms for accurate interpretation and comparability.

To fill the knowledge gaps regarding the influence of age, gender, and country on PROMIS PF, UE, and PI scores, we aimed to provide patient-specific reference values for clinicians and researchers working with older populations. We focused on individuals aged 50 and older to address the lack of precise PROMIS reference values for people in this age group, who experience the highest prevalence of surgeries and lowest level of PF [3, 22, 28]. By providing these patient-specific reference values, we can enhance the clinical utility, interpretability, and comparability of PROMIS scores across different patient populations and studies.

We therefore asked: (1) Can all PROMIS PF, UE, and PI items be used for valid cross-country comparisons in these domains among the United States, the United Kingdom, and Germany? (2) How are age, gender, and country related to PROMIS PF, PROMIS UE, and PROMIS PI scores? (3) What is the relationship of age, gender, and country across individuals with PROMIS PF, PROMIS UE, and PROMIS PI scores ranging from very low to very high?

Patients and Methods

Study Design and Setting

We conducted a population-based telephone survey of the general population aged 50 and older in the United States, United Kingdom, and Germany. The aim of the survey was to collect reference data for the PROMIS PF, UE, and PI item banks.

Data Collection

We collected data using computer-assisted telephone interviews in which interviewers followed a structured questionnaire displayed on a computer screen while conducting the interview. This method of administration was necessary because collecting data from older participants using online platforms was not considered viable. We did not expect any mode-of-administration effects based on previous research [17, 21, 27, 32]. Computer-assisted telephone interviewing is a method of data collection that combines telephone interviewing with computer technology. In this process, interviewers follow a structured questionnaire displayed on a computer screen while conducting the interview. The interviewer records the responses directly into the computer system during the conversation.

Survey Methodology

The survey, conducted by USUMA GmbH Social Research Institute, used a stratified random sampling approach using the Arbeitskreis Deutscher Markt und Sozialforschungsinstitute (Working Group of German Market and Social Research Institutes) Telephone Sample Selection framework. This framework facilitates adequate sampling of private households and individuals living in them if they are reachable by telephone. This framework ensured the representativeness and precision of our data by allowing us to collect responses from at least 75 individuals per cell, characterized by country, gender, and age group, across the United States, Germany, and the United Kingdom. This targeted approach allowed us to obtain sufficiently precise reference value estimates, enhancing the validity and robustness of our study outcomes.

To further improve the representativeness of our sample, we used the Kish Selection Grid to randomly select target persons in households containing multiple individuals. This approach ensures equal selection probability for all target persons, specifically individuals 50 years and older. Using these methods, we could acquire a random sample of our target population.

For the survey, we approached participants in the United States and United Kingdom via an existing database of private household telephone numbers. However, in Germany, where a centralized telephone number database is absent, we used an ADM-maintained algorithm. This method generates telephone numbers by correlating area codes with population densities, thereby ensuring a representative sample.

Participants

Participants were eligible if they were at least 50 years old; there were no other exclusion criteria. We focused on participants older than 50 years because older individuals tend to have a higher prevalence of surgeries and lower PF [22]. Furthermore, existing reference values for PROMIS PF, UE, and PI scores are predominantly based on general adult populations [13], and there is a lack of age-specific reference data for older individuals. By focusing on this age group, our study addresses a gap, providing much-needed reference values for clinicians and researchers working with older populations [28]. These age-specific reference values enable better interpretation of patient-reported outcomes in an aging population and are particularly relevant for orthopaedic subspecialties, such as arthroplasty and geriatric trauma, where most patients are older adults [18, 26].

Overall, we collected data from 2726 individuals from the United States, United Kingdom, and Germany aged 50 to 98 years. Their sociodemographic variables were broadly comparable (Table 1). The response rates varied across the three countries, from 12% in the United States to 22% in Germany, which might pose a risk for response bias. However, the characteristics of our sample closely mirror the normative data for each country, suggesting the representativeness and validity of our findings. PROMIS PF, UE, and PI T-scores differed among countries regarding their distribution (Table 2). In addition, we found considerable floor and ceiling effects in all countries for PROMIS UE and PI scores. Almost 50% of United States individuals had the highest T-score estimate of 57.4, while 39% of United Kingdom individuals had this maximum score. At the same time, the PROMIS PI had considerable floor effects, with up to 43% of individuals reaching the minimum T-score of 40.7.

Background of the PROMIS

The PROMIS initiative developed self-reported instruments for more than 100 relevant health outcomes based on item response theory [5]. Compared with traditional test theory approaches, the use of item response theory in PROMIS provides the advantage of separating the construct (PRO)

Table 1. Summary statistics of sociodemographic information

Sociodemographic factors	United States (n = 900)	Germany (n = 921)	United Kingdom (n = 905)
Age in years, median (range)	70 (50 to 98)	69 (50 to 98)	69 (50 to 98)
Sex, female, % (n)[a]	5 (450)	50 (456)	50 (456)
Work status, % (n)			
Full-time	19 (173)	17 (159)	15 (132)
Part-time	6 (53)	6 (59)	11 (101)
Self-employed	9 (79)	5 (47)	5 (42)
Student	0 (1)	0 (0)	0 (0)
Retired, in early retirement	58 (522)	68 (624)	63 (568)
Job seeker or not employed	3 (30)	2 (20)	4 (33)
Other	4 (40)	1 (10)	3 (27)
Education, % (n)[b]			
Less than high school degree	7 (64)	20 (185)	31 (280)
High school graduate	21 (188)	25 (230)	31 (281)
Some college	23 (203)	14 (125)	27 (241)
Bachelors	24 (213)	5 (44)	9 (80)
Masters	16 (144)	26 (239)	2 (16)
Doctorate	5 (45)	3 (31)	0 (2)
Unknown	5 (43)	4 (35)	1 (5)
Marital status, % (n)[c]			
Married or living with partner	55 (486)	57 (525)	57 (503)
Divorced	14 (120)	12 (108)	12 (106)
Widowed	18 (162)	19 (176)	18 (161)
Single	13 (117)	12 (109)	13 (119)
Unknown	0 (1)	0 (0)	0 (0)
PROMIS global physical health T-score			
Median (range)	50 (23.4 to 63.3)	50 (23.4 to 63.3)	50 (23.4 to 63.3)
PROMIS global mental health T-score			
Median (range)	52.8 (25.8 to 64.6)	48.6 (25.8 to 64.6)	48.6 (25.8 to 64.6)

We assessed PROMIS Global Mental and Physical Health with two-item short forms: PROMIS Global Health PROMIS Scale v1.2, Global Health Physical and Mental 2a.

[a] Data are missing for 2 individuals from each country.

[b] Data are missing for 32 individuals from Germany.

[c] Data are missing for 14 individuals from the United State, 3 individuals from Germany, and 16 individuals from the United Kingdom.

from the respective measures (PROM) [4]. All items of an item bank are calibrated along a latent-trait continuum of the target construct using information on the item's location and slope. Because this method defines latent scales, PRO estimates based on any item subset of an item bank can be directly compared using the same scale. Thus, any combination of items, including short forms or combinations of items selected as part of computerized adaptive testing, can be used in health assessments. In the item response theory framework, it is also possible to link legacy PROMs to the corresponding PROMIS metric that represents the same construct, facilitating instrument comparisons. This is already possible for several measures; for example, the Short Form-36 Health Survey and Health Assessment Questionnaire Disability Index is linked to the PROMIS PF, and the Oswestry Disability Index is linked to PROMIS PI [31, 35]. Hence, PROMIS establishes a theoretical and empirically tested framework for PROs and provides comprehensive measures to assess these PROs.

PROMIS scores are reported as T-scores, with a mean of 50 ± 10 representing the distribution of scores in the United States general population based on the 2020 Census demographic distributions. This convention provides the advantages of easy interpretability of the resulting scores and comparability across PROMIS measures and other measures linked to a specific PROMIS metric on the same scale.

Table 2. Summary statistics of PROMIS measures

PROMIS summary statistics	United States (n = 900)	Germany (n = 921)	United Kingdom (n = 905)
PROMIS PF T-score			
Floor, % (n)	1 (0)	1 (0)	2 (0)
Ceiling, % (n)	14 (123)	13 (123)	12 (107)
Median (range)	47.7 (14.4 to 63.8)	48.8 (11.6 to 63.8)	46.9 (9.2 to 63.8)
PROMIS UE T-score			
Floor, % (n)	0 (2)	0 (2)	0 (3)
Ceiling, % (n)	49 (437)	42 (390)	39 (357)
Median (range)	55.9 (14.8 to 57.4)	49.5 (14.8 to 57.4)	48.2 (14.8 to 57.4)
PROMIS PI T-score			
Floor, % (n)	40 (356)	31 (287)	43 (391)
Ceiling, % (n)	2 (14)	0 (2)	3 (28)
Median (range)	48.6 (40.7 to 77.0)	51.4 (40.7 to 77.0)	48.6 (40.7 to 77.0)

Floor % (n) = percentage (number) of individuals with the minimum T-score; ceiling % (n) = percentage (number) of individuals with the maximum T-score.

However, this comparison with reference data originally collected in the United States might be less useful in other cultural contexts [5, 10]. In addition, PROMIS instruments are usually used in clinical settings, where comparison against the general population might have limited relevance. A comparison with more meaningful reference groups (for example, same age, gender, and country) might improve our understanding of PROMIS scores for patients, clinicians, and researchers. Such patient-specific reference data seem essential for making scores more interpretable for clinical decision-making. For example, previous research indicates that constructs associated with the level of PF might be influenced by age [19, 20], suggesting the need for age-specific reference values. In particular, older age groups are typically affected by low PF, with a declining trend beginning at approximately 50 years old [19, 20]. Moreover, the scores of several PRO domains differ considerably regarding country and gender [20].

Variables

Sociodemographic Information

We collected sociodemographic information on age, gender, education, work status, and marital status. Moreover, four items were administered to assess physical and mental health (PROMIS Global Health PROMIS Scale, version 1.2, Global Health Physical and Mental 2a [12]).

PROMIS Short Forms

We selected specific items from each PROMIS item bank by focusing on their relevance to the study's objectives, ensuring the chosen items adequately captured the constructs of interest. Our selection process involved consultations with experts in the field and a thorough review of the existing research [25, 29] to identify the most appropriate items for our study. Our selection of items therefore provides a comprehensive assessment of the PROMIS PF, UE, and PI scales.

PROMIS PF

The PROMIS PF version 2.0 item bank evaluates a wide variety of activities, from self-care (for example, daily living tasks) to more complex activities requiring a variety of physical abilities, including strenuous activities such as playing sports or jogging. Overall, the PROMIS PF bank contains 165 items using a 5-point ordinal response scale, and includes items about neck and back function, upper and lower extremity function, and ability to do instrumental activities of daily living, including housework and shopping [19, 29]. Higher T-scores indicate higher, meaning better, levels of physical function. We collected 22 items selected by two experts (BDS and AJK), covering different aspects and levels of physical function. Depending on clinical population and procedure used to estimate the minimum clinically important difference (MCID), this difference in T-scores is estimated to be between 3.4 and 4.6 for PROMIS PF [16, 30].

PROMIS UE

We included eight of the PF items as an additional domain measuring activities demanding the use of the upper extremities (such as writing, pressing buttons, and opening containers) from the PROMIS UE version 2.0 item bank (containing 31 items in total) [14].

These items were: PFA14r1 (Are you able to carry a heavy object [over 10 pounds/5 kg]?), PFA29r1 (Are you able to pull heavy objects [10 pounds/5 kg] toward yourself?), PFA34 (Are you able to wash your back?), PFA36 (Are you able to put on and take off a coat or jacket?), PFB13 (Are you able to carry a shopping bag or briefcase?), PFB26 (Are you able to shampoo your hair?), PFB28r1 (Are you able to lift 10 pounds [5 kg] above your shoulder?), and PFB34 (Are you able to change a light bulb overhead?).

PROMIS PI

The PROMIS PI item bank measures the influence of self-reported pain on important parts of one's life; for example, how pain prevents people from engaging in social, cognitive, emotional, physical, and recreational activities. The PROMIS PI item bank version 1.1 consists of 40 items in total [1]. Higher T-scores imply more (and worse) pain interference. Depending on the clinical population, the MCID T-score for PROMIS PI is estimated to be between 3.4 and 5.5 [2, 16]. We selected eight items covering different aspects and levels of pain interference over the previous 7 days, asking PAININ12 (how much did pain interfere with the things you usually do for fun?), PAININ13 (how much did pain interfere with your family life?), PAININ22 (how much did pain interfere with work around the home?), PAININ3 (how much did pain interfere with your enjoyment of life?), PAININ31 (how much did pain interfere with your ability to participate in social activities?), PAININ34 (how much did pain interfere with your household chores?), PAININ36 (how much did pain interfere with your enjoyment of social activities?), and PAININ9 (how much did pain interfere with your day-to-day activities?).

Ethical Approval

Ethical approval for this study was obtained from Charité Universitätsmedizin Berlin, Berlin, Germany (EA4/212/20).

Statistical Analyses

PROMIS Scoring

Following the PROMIS scoring guidelines [9], we estimated T-scores based on observed item responses using the expected a posteriori estimator of the PROMIS Graded Response Models, which is calibrated to the United States general population.

Score Differences

We modeled the 50th percentile and 95% CIs using median regressions, with the United States as the reference group to investigate country differences in PROMIS PF, UE, and PI scales. We included age and gender as predictors. We also calculated standardized mean differences (Cohen d) for these country differences.

Reference Values

We applied plausible value imputation, a method designed for analyzing latent trait scores and accounting for differences in measurement precision in the context of item response theory [11, 37]. Plausible value imputation treats the latent trait level as missing and incorporates the uncertainty associated with the estimation of the latent trait by drawing multiple "plausible values" from the latent trait's posterior distribution. In our study, we used plausible value imputation in the following manner. First, we estimated latent trait scores based on the PROMIS item parameters and available response data. Second, we approximated the posterior distribution using a normal distribution with mean = T-score and SD = standard error. Third. we created 25 datasets by randomly drawing latent trait measurements from the distribution of plausible values for each individual. Fourth, we performed quantile regressions separately on each dataset. Finally, we combined the results from the separate analyses to obtain the final results according to the Rubin rules, considering variability across the imputed datasets [23].

The use of plausible value imputation allowed us to obtain a smooth distribution of outcomes, which provides more accurate and unbiased estimates of population parameters [11]. This approach also enabled us to make more valid inferences about the relationships between latent traits and other variables of interest in our study.

This approach enabled us to consider the different levels of measurement precision across the entire range of the latent construct, therefore providing a more accurate representation of the underlying latent constructs.

We used quantile regressions to model the 1st, 5th, 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, 95th, and 99th percentiles and their respective standard errors for the PROMIS PF, UE, and PI scales. We included age, gender, and country as predictors in the regression model to allow stratification of reference values. We fitted a series of regression models, including linear, quadratic, and cubic effects for continuous predictors as well as potential interaction effects between predictors. We selected the most appropriate model balancing complexity and parsimony by comparing the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The estimated percentiles and their respective standard errors were pooled over the imputed datasets according to the Rubin rules [23]. Those pooled estimates were used to calculate 95% CIs for the reference values, assuming a normal distribution. Additionally, we developed an interactive internet application to provide general-population reference values for the PROMIS PF, UE, and PI, adjusted for age, gender, and country, which is available at: www.common-metrics.org/PROMIS_PF_and_PI_Reference_scores.php [6].

Differential Item Functioning

We explored differential item functioning (DIF) among countries as a prerequisite for valid across-country comparisons [33]. We used the R package Lordif to investigate DIF for the PROMIS PF, UE, and PI items [7]. We flagged items exceeding a change in Nagelkerke pseudo-$r^2 > 0.02$ and visually inspected the DIF impact, namely, the absolute difference between the item characteristic curves between countries weighted by the score distribution.

Open Science Practices

All components necessary for reproducible data analysis (preregistration, open data, and code) have been made accessible via the Open Science Framework (https://osf.io/2cuf7/?view_only=64ad89db4b0c45878a5f358954835b3a).

Sample Size Rationale

We determined the appropriate sample size for this population-based study in terms of the precision of mean scores. Assuming an SD of 10, a sample size of 900 allows for an estimate of the population T-score mean of each country within ± 0.65 T-scores (95% CI) and the T-score mean in each gender and age group (n = 75 of 150) within ± 2.3/1.6 T-scores (95% CI).

Results

Validity of Across-country Comparisons Among the PROMIS PF, UE, and PI Items Among the United States, United Kingdom, and Germany

All items across the PROMIS PF, UE, and PI measures were valid for across-country comparisons. This is because any observed variations in the functioning of all items were negligible. Three PROMIS PF items (PFA11, PFA34, and PFM26) and one PROMIS UE item (PFA34) were flagged for DIF, indicating these items might not adequately measure the same construct in all three countries. However, a visual inspection of these items revealed a negligible impact of DIF. Because the T-scores of each estimate were very similar when accounting or not accounting for DIF, we did not remove any items from further analysis.

Differences in PROMIS Scores by Age, Gender, and Country

There were clinically meaningful associations of age, gender, and country of residence with all included PROMIS measures: PF (Supplemental Table 1; http://links.lww.com/CORR/B203), UE (Supplemental Table 2; http://links.lww.com/CORR/B204), and PI (Supplemental Table 3; http://links.lww.com/CORR/B205).

With increasing age, we found a decline in PROMIS PF (age ß$_{Median}$ = -0.35 [95% CI -0.40 to -0.31]) and PROMIS UE scores (age ß$_{Median}$ = -0.38 [95% CI -0.45 to -0.32]). For PROMIS PF, this translated to a decrease in 3.5 points after 10 years, approximately the MCID for an individual [30]. A decline of 7.0 points over 20 years indicates a substantial shift in the level of PF. We also found a small increase in PROMIS PI scores with increasing age (age ß$_{Median}$ = 0.10 [95% CI 0.04 to 0.16]). The accumulated effect of 40 years would be equal to the MCID on a group level; however, this does not mean that the effect is irrelevant on an individual level [2]. Additionally, women had, on average, worse scores than men for the PROMIS PF (gender ß$_{Median}$ = -3.30 [95% CI -4.20 to -2.30]), PROMIS UE (gender ß$_{Median}$ = -5.32 [95% CI -6.71 to -3.92]), and PROMIS PI (gender ß$_{Median}$ = 1.39 [95% CI 0.46 to 2.32]).

For the comparison between the United States and United Kingdom, we found higher T-scores for the United States in all domains. For the PROMIS PF, this difference was -1.21 points (95% CI -2.40 to -0.01; Cohen d = 0.14 [95% CI 0.05 to 0.23]). For the PROMIS UE, the difference was -3.83 points (95% CI -5.44 to -2.23; Cohen d = 0.25 [95% CI 0.15 to 0.34]), and for PROMIS PI it was -0.48 points (95% CI -2.06 to 1.11; Cohen d = 0.12 [95% CI 0.03 to 0.21]). Participants in the United States had lower T-scores than those in Germany for the PROMIS PF (1.13 points [95% CI 0.03 to 2.23]; Cohen d = 0.14 [95% CI 0.05 to 0.23]) and PROMIS PI (1.93 points 95% CI 1.17 to 2.70]; Cohen d = 0.14 [95% CI 0.05 to 0.23]). These differences among age, gender, and country indicate that stratification of reference data is warranted.

Relationship Among Age, Gender, and Country Across Varying Levels of PROMIS Scores

The PROMIS scores for PF, UE, and PI in the United States, United Kingdom, and Germany were analyzed across the entire distribution, from the lowest to the

Table 3. Quantile regression coefficients for PROMIS PF 2.0 items

	Percentiles with 95% CIs												
Parameter	**1%**	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**99%**
United States	28.8 (24.8 to 32.8)	35.1 (32.3 to 37.9)	40.6 (38.5 to 42.8)	46.2 (44.5 to 48.0)	50.4 (48.8 to 52.0)	53.4 (52.0 to 54.8)	55.8 (54.4 to 57.2)	58.3 (56.9 to 59.8)	61.7 (60.1 to 63.3)	65.3 (63.5 to 67.1)	70.0 (67.6 to 72.4)	73.2 (70.5 to 75.9)	76.7 (71.6 to 81.7)
United Kingdom	27.7 (24.7 to 30.7)	32.3 (29.5 to 35.1)	38.3 (36.0 to 40.6)	45.2 (43.3 to 47.0)	49.6 (48.0 to 51.2)	52.4 (51.0 to 53.8)	54.8 (53.4 to 56.1)	57.1 (55.7 to 58.5)	60.3 (58.7 to 61.8)	63.7 (61.9 to 65.5)	68.2 (65.8 to 70.7)	71.0 (68.3 to 73.6)	74.8 (70.7 to 78.9)
Germany	31.7 (28.1 to 35.3)	37.7 (34.8 to 40.6)	43.6 (41.5 to 45.7)	49.0 (47.3 to 50.7)	52.4 (50.9 to 53.9)	54.8 (53.5 to 56.2)	57.0 (55.7 to 58.3)	59.1 (57.7 to 60.4)	62.0 (60.5 to 63.5)	65.5 (63.8 to 67.2)	70.2 (67.9 to 72.5)	72.9 (70.6 to 75.3)	76.5 (72.3 to 80.8)
Age	-0.2 (-0.3 to -0.1)	-0.2 (-0.3 to -0.1)	-0.2 (-0.3 to -0.2)	-0.3 (-0.4 to -0.2)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.4 (-0.4 to -0.3)	-0.4 (-0.5 to -0.3)	-0.4 (-0.5 to -0.3)	-0.4 (-0.5 to -0.3)	-0.2 (-0.4 to -0.1)
Gender	-4.4 (-7.3 to -1.4)	-2.6 (-4.6 to -0.7)	-3.5 (-5.0 to -2.0)	-3.7 (-4.9 to -2.6)	-3.5 (-4.6 to -2.4)	-3.4 (-4.4 to -2.4)	-3.2 (-4.1 to -2.3)	-3.0 (-4.0 to -2.1)	-3.1 (-4.2 to -2.0)	-3.1 (-4.3 to -1.9)	-2.9 (-4.6 to -1.3)	-2.6 (-4.7 to -0.5)	-1.4 (-4.7 to 1.9)

Reference values for the United States, United Kingdom, and Germany with percentiles ranging from 1% to 99%. Each value in a country row contains the reference values for men aged 50 years. Each column contains the corresponding percentiles. Age: With each additional year beginning at age 50 years, the value from the age row and the same percentile column must be subtracted. Gender: Negative values indicate lower values for women and positive values represent higher values for men. Patient-specific reference values can be calculated by combining the numeric value from a country, age, and gender row for the percentile of interest. For instance, for an 80-year-old woman from the United Kingdom in the 20th percentile, one would have to identify the 20th percentile from the United Kingdom ($T_{UK\ 20\%}$ = 45.2), subtract the corresponding gender value, which can be found in the gender row ($T_{gender\ 20\%}$ = -3.7), and subtract the value for age, which can be calculated by multiplying age above 50 years with the corresponding age value ($T_{age80\ 20\%}$ = 80-50 * -0.3 = -9). This results in T = 32.5 for an 80-year-old woman from the United Kingdom in the 20th percentile.

highest percentiles, revealing differences and patterns across countries, genders, and ages. A trend of decreasing PF and UE T-scores with increasing age was observed, while PI T-scores showed a negligible association with age. In terms of gender, men generally scored higher on PF and UE than women, but had slightly lower PI scores, although the differences were relatively small and not consistent across all percentiles. Differences in scores were also observed across countries, with individuals in the United States tending to have higher PF and UE scores, and those in Germany having higher PI scores.

For PROMIS PF (Table 3), scores across all percentiles were generally higher in Germany and the United States than in the United Kingdom. Age showed a consistent, negative effect on PF scores, indicating that older individuals tended to have lower PF scores. Gender was also associated with PF scores, with men typically scoring higher than women, although the differences became less pronounced at higher percentiles. This gender difference exceeded the magnitude of an MCID for those with the lowest PROMIS PF levels (Table 3) and decreased to one-third of the MCID in the 99th percentile.

For PROMIS UE (Table 4), the United States also showed generally higher scores across all percentiles, followed by Germany and the United Kingdom. The association with age was again negative, indicating decreased UE function with increasing age, and similar to PF, men tended to score higher than women in UE function, with the gender difference decreasing at higher percentiles.

In contrast to PF and UE, PI revealed there was relatively little association of age with scores (Table 5). The gender effect was positive, suggesting that women experienced slightly more pain interference than men, a trend seen with PF and UE. However, the magnitude was smaller because these gender differences increased to only half the MCID for the 95th percentile. When assessing the distribution of PI scores across all percentiles, Germany generally had higher scores than both the United States and United Kingdom up to the 90th percentile. However, past this percentile, the trend inverted, and Germany exhibited lower scores, suggesting less pain interference in the upper echelons of the distribution than in the other countries. This pattern underscores a divergence in pain interference experiences in different sections of the population.

Further patient-specific reference values for PROMIS PF (Table 3), UE (Table 4), and PI (Table 5) are depicted for the United States, United Kingdom, and Germany as percentiles ranging from 1% to 99% with their respective 95% CIs. These values represent the average PROMIS T-scores for the respective percentiles in 50-year-old men. Additionally, the regression coefficients for age and gender are included. For each additional year of age above 50

Table 4. Quantile regression coefficients for PROMIS UE items

	Percentiles with 95% CIs												
Parameter	**1%**	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**99%**
United States	25.8 (21.1 to 30.5)	34.4 (31.7 to 37.1)	40.2 (37.5 to 42.9)	47.0 (45.0 to 49.1)	51.2 (49.3 to 53.1)	54.7 (52.8 to 56.7)	58.0 (56.1 to 60.0)	60.8 (59.0 to 62.7)	63.4 (61.6 to 65.3)	65.4 (63.6 to 67.3)	67.5 (65.7 to 69.3)	69.5 (66.9 to 72.0)	73.7 (69.9 to 77.6)
United Kingdom	25.9 (22.0 to 29.8)	31.2 (28.6 to 33.9)	36.9 (34.2 to 39.7)	44.3 (42.2 to 46.4)	48.4 (46.5 to 50.4)	51.7 (49.9 to 53.5)	54.6 (52.8 to 56.5)	57.5 (55.6 to 59.4)	60.3 (58.6 to 62.1)	63.0 (61.1 to 64.9)	65.9 (63.8 to 68.0)	68.1 (65.3 to 70.9)	72.8 (69.2 to 76.4)
Germany	28.7 (24.1 to 33.3)	36.7 (33.9 to 39.6)	42.5 (40.1 to 45.0)	48.9 (47.0 to 50.9)	52.3 (50.5 to 54.1)	55.1 (53.4 to 56.8)	57.5 (55.7 to 59.3)	59.8 (58.1 to 61.4)	62.0 (60.2 to 63.7)	64.1 (62.4 to 65.9)	66.7 (64.8 to 68.6)	68.8 (66.3 to 71.3)	73.4 (69.6 to 77.2)
Age	-0.2 (-0.3 to -0.0)	-0.2 (-0.3 to -0.1)	-0.2 (-0.3 to -0.2)	-0.3 (-0.4 to -0.2)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.3)	-0.3 (-0.4 to -0.2)	-0.3 (-0.3 to -0.2)	-0.2 (-0.3 to -0.1)	-0.1 (-0.2 to -0.1)	-0.1 (-0.2 to 0.0)
Gender	-3.3 (-6.5 to -0.1)	-4.0 (-6.0 to -2.1)	-4.3 (-6.1 to -2.6)	-4.4 (-5.9 to -3.0)	-4.8 (-6.1 to -3.6)	-4.9 (-6.1 to -3.7)	-4.8 (-6.0 to -3.6)	-4.6 (-5.8 to -3.3)	-4.2 (-5.5 to -2.9)	-3.4 (-4.8 to -2.0)	-2.4 (-3.7 to -1.0)	-1.7 (-3.3 to -0.0)	-1.1 (-4.3 to 2.1)

Reference values for the United States, United Kingdom, and Germany with percentiles ranging from 1% to 99%. Each value in a country row contains the reference values for men aged 50 years. Each column contains the corresponding percentiles. A low percentile indicates low upper extremity functioning, while a high percentile models high upper extremity functioning. Age: With each additional year beginning at age 50 years, the value from the age row and the same percentile column must be subtracted. Gender: Negative values indicate lower values for women and positive values higher values for men. To obtain patient-specific reference values, these values must be calculated. For instance, for an 80-year-old woman from the United Kingdom in the 20th percentile—low upper extremity functioning, one would have to identify the 20th percentile from the United Kingdom ($T_{UK\ 20\%}$ = 44.3), subtract the corresponding gender value, which can be found in the gender row ($T_{gender\ 20\%}$ = -4.4), and subtract the value for age, which can be calculated by multiplying age above 50 years with the corresponding age value ($T_{age80\ 20\%}$ = 80-50 * -0.3 = -9). This results in T = 30.9 for an 80-year-old woman from the United Kingdom in the 20th percentile.

years, the corresponding age coefficient must be added, and for women, the respective gender coefficient.

We provide all patient-specific reference values based on an additive model including age, gender, and country. This additive model was chosen because it had the best fit for our quantile regressions based on AIC and BIC coefficients. The additive model exhibits the best fit across the 0.01 to 0.99 quantiles for PROMIS PF, as shown by the AIC (Supplemental Fig. 1; http://links.lww.com/CORR/B206) and BIC (Supplemental Fig. 2; http://links.lww.com/CORR/B207). This was similarly observed for PROMIS UE, where the AIC (Supplemental Fig. 3; http://links.lww.com/CORR/B208) and BIC (Supplemental Fig. 4; http://links.lww.com/CORR/B209) also favor the additive model within these quantiles. The same holds true for PROMIS PI, where the additive model presents the best fit, as demonstrated by the AIC (Supplemental Fig. 5; http://links.lww.com/CORR/B210) and BIC (Supplemental Fig. 6; http://links.lww.com/CORR/B211).

Internet Application

We created an internet application to provide a better user experience for clinicians wanting to calculate patients' specific reference values (Fig. 1). This application is available at: www.common-metrics.org/PROMIS_PF_and_PI_Reference_scores.php. Clinicians can receive patient-specific plots and tables for PROMIS PF, UE, and PI scores after inputting their patients' country, age, and gender.

Discussion

PROMs are vital in assessing health outcomes, and age, gender, and country are known to impact scores. Despite this, the lack of detailed, patient-specific reference values, especially for those aged 50 years and above, currently limits the accurate interpretation and comparability of these scores. Recognizing the knowledge gaps in the application of PROMIS scores across different demographics and countries, we posed three questions. We sought to determine whether all PROMIS PF, UE, and PI items could be universally applied across three countries; to explore the association of age, gender, and country of residence with these scores; and to understand the relationship of these variables across a range of PROMIS PF, UE, and PI scores. Our findings demonstrate notable associations between age, gender,

Table 5. Quantile regression coefficients for PROMIS PI items

	Percentiles with 95% CIs												
Parameter	**1%**	**5%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**99%**
United States	28.0 (24.2 to 31.8)	32.7 (30.5 to 34.8)	35.1 (32.8 to 37.4)	38.0 (35.8 to 40.1)	40.9 (38.8 to 43.0)	44.5 (42.8 to 46.1)	47.3 (46.0 to 48.5)	49.1 (47.9 to 50.3)	51.1 (49.8 to 52.5)	54.5 (52.7 to 56.4)	60.8 (58.3 to 63.4)	66.6 (63.5 to 69.6)	76.6 (72.2 to 81.1)
United Kingdom	27.9 (24.0 to 31.8)	32.6 (30.3 to 34.8)	34.9 (32.9 to 36.9)	37.8 (35.8 to 39.8)	40.4 (38.2 to 42.6)	44.2 (42.4 to 46.0)	47.6 (46.1 to 49.1)	49.9 (48.6 to 51.3)	52.6 (51.1 to 54.1)	56.5 (54.6 to 58.4)	63.5 (61.0 to 66.0)	69.6 (66.4 to 72.9)	78.4 (74.3 to 82.5)
Germany	28.7 (25.1 to 32.3)	33.7 (31.4 to 36.0)	36.3 (34.3 to 38.4)	40.2 (38.1 to 42.3)	43.4 (41.4 to 45.5)	46.6 (45.1 to 48.1)	48.8 (47.7 to 50.0)	50.4 (49.3 to 51.4)	52.0 (50.8 to 53.2)	54.7 (53.0 to 56.3)	59.4 (57.2 to 61.7)	64.5 (61.6 to 67.5)	71.6 (67.7 to 75.5)
Age	0.0 (-0.1 to 0.2)	0.0 (-0.0 to 0.1)	0.1 (-0.0 to 0.1)	0.1 (0.0 to 0.2)	0.1 (0.0 to 0.2)	0.1 (0.1 to 0.2)	0.1 (0.0 to 0.1)	0.1 (0.0 to 0.1)	0.1 (0.0 to 0.1)	0.1 (0.0 to 0.1)	0.0 (-0.0 to 0.1)	-0.0 (-0.1 to 0.1)	0.0 (-0.1 to 0.1)
Gender	0.5 (-2.4 to 3.4)	0.6 (-1.0 to 2.2)	0.7 (-0.8 to 2.2)	1.2 (-0.2 to 2.6)	1.6 (0.2 to 3.1)	1.5 (0.4 to 2.7)	1.3 (0.4 to 2.2)	1.6 (0.7 to 2.4)	1.9 (0.9 to 2.9)	2.5 (1.3 to 3.8)	2.4 (0.8 to 4.1)	2.1 (-0.0 to 4.2)	0.5 (-2.6 to 3.7)

Reference values for the United States, United Kingdom, and Germany with percentiles ranging from individuals with very low levels for PROMIS PI (lowest 1%) to very high levels (99%). Each value in a country row contains the reference values for men aged 50 years. Each column contains the corresponding percentiles. Age: With each additional year beginning at age 50 years, the value from the age row and the same percentile column must be subtracted. Gender: Negative values indicate lower values for women and positive values indicate higher values for men. To obtain patient-specific reference values, these values must be calculated. For instance, for an 80-year-old woman from the United Kingdom in the 20th percentile, one would have to identify the 20th percentile from the United Kingdom (20% = 37.8), add the corresponding gender value, which can be found in the gender row ($T_{gender\ 20\%}$ = 1.2), and add the value for age, which can be calculated by multiplying age older than 50 years by the corresponding age value ($T_{age80\ 20\%}$ = 80-50 * 0.1 = 3). This results in T = 42 for an 80-year-old woman from the United Kingdom in the 20th percentile.

and country of residence and PROMIS measures, highlighting shifts in PF and PI over time and between genders. In practical terms, these results can inform clinicians' interpretations of patient-reported outcomes, allowing for a more-nuanced understanding of an individual patient's health-related quality of life. When considering patient care, these findings can aid healthcare providers in making more informed decisions and better patient assessments by considering a patient's age, gender, and country of residence.

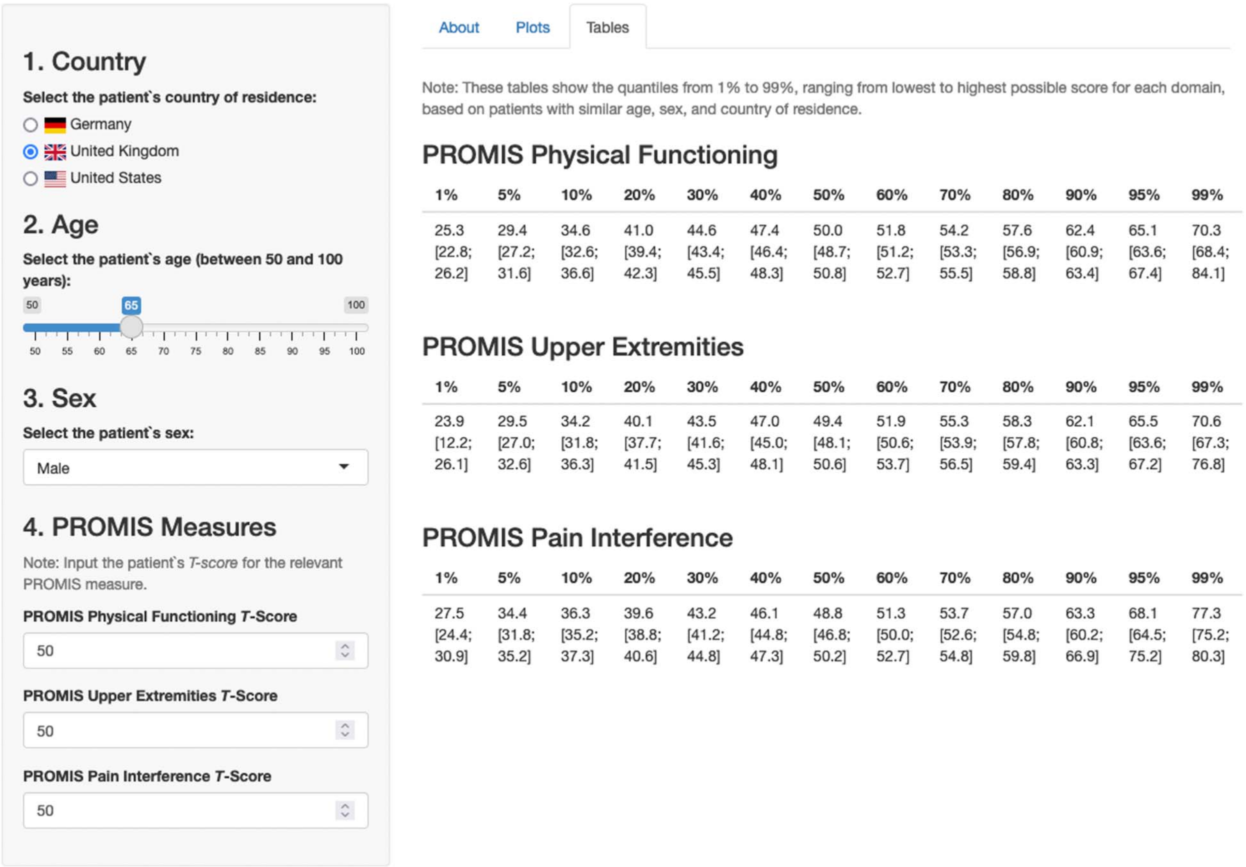
Limitations

First, the response rates varied across the three countries: 14% for the United Kingdom, 12% for Germany, and 22% for the United States. This disparity in response rates might be attributed to differences in technical access options between countries. However, because we used random selection models and robust sampling methods to sample participants in each country, representativeness of the presented data can likely still be assumed.

Second, as is the case in any study, nonrandom nonresponse remains a potential threat to the representativeness of the sample. Especially in older people, reasons for nonresponse might be related to low scores in the outcome measures (for example, PROMIS PF). This indicates that especially in the lowest percentiles of PROMIS PF and UE and the highest percentiles of PI, variance in scores might be somewhat greater than shown in our results. Thus, the quantile regression coefficients of age, gender, and country for the most extreme (worst) quantiles of the physical functioning distributions might be considered conservative, lower-bound estimates. Importantly, however, this does not affect the results of our median regression, and values across quantiles are likely still representative for "typical" individuals aged over 50 years who participate in clinical studies or contact a clinician. Additionally, existing normative demographic data aligned with the collected populations of the respective countries.

Third, we collected data during a period of lockdown owing to coronavirus-19 worldwide (April 12, 2021, to April 28, 2021), which might have influenced PF and PI. We have no way of knowing the extent to which our measured constructs were affected by the pandemic; however, this highlights the importance of timely and context-specific reference data.

Fourth, we provided patient-specific reference values for only three countries, which limits worldwide application. However, until further country-specific reference values become available, our patient-specific reference model could be adapted by matching the percentiles of another country's sample with our age- and gender-specific values because those were universal for the three countries we evaluated.



1. Country

Select the patient`s country of residence:

○ Germany
● United Kingdom
○ United States

2. Age

Select the patient`s age (between 50 and 100 years): 65

3. Sex

Select the patient`s sex: Male

4. PROMIS Measures

Note: Input the patient`s *T-score* for the relevant PROMIS measure.

PROMIS Physical Functioning *T*-Score: 50

PROMIS Upper Extremities *T*-Score: 50

PROMIS Pain Interference *T*-Score: 50

Note: These tables show the quantiles from 1% to 99%, ranging from lowest to highest possible score for each domain, based on patients with similar age, sex, and country of residence.

PROMIS Physical Functioning

1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
25.3 [22.8; 26.2]	29.4 [27.2; 31.6]	34.6 [32.6; 36.6]	41.0 [39.4; 42.3]	44.6 [43.4; 45.5]	47.4 [46.4; 48.3]	50.0 [48.7; 50.8]	51.8 [51.2; 52.7]	54.2 [53.3; 55.5]	57.6 [56.9; 58.8]	62.4 [60.9; 63.4]	65.1 [63.6; 67.4]	70.3 [68.4; 84.1]

PROMIS Upper Extremities

1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
23.9 [12.2; 26.1]	29.5 [27.0; 32.6]	34.2 [31.8; 36.3]	40.1 [37.7; 41.5]	43.5 [41.6; 45.3]	47.0 [45.0; 48.1]	49.4 [48.1; 50.6]	51.9 [50.6; 53.7]	55.3 [53.9; 56.5]	58.3 [57.8; 59.4]	62.1 [60.8; 63.3]	65.5 [63.6; 67.2]	70.6 [67.3; 76.8]

PROMIS Pain Interference

1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
27.5 [24.4; 30.9]	34.4 [31.8; 35.2]	36.3 [35.2; 37.3]	39.6 [38.8; 40.6]	43.2 [41.2; 44.8]	46.1 [44.8; 47.3]	48.8 [46.8; 50.2]	51.3 [50.0; 52.7]	53.7 [52.6; 54.8]	57.0 [54.8; 59.8]	63.3 [60.2; 66.9]	68.1 [64.5; 75.2]	77.3 [75.2; 80.3]

Fig. 1 This screenshot shows the Shiny Web application for patient-specific reference data.

Discussion of Key Findings

We found notable differences in PROMIS PF, UE, and PI scores across the United States, United Kingdom, and Germany. Individuals from Germany generally had higher PF, UE, and PI scores across most percentiles than those from the United States, while participants from the United Kingdom tended to have lower PF and PI scores. However, this trend reversed in the higher percentiles, particularly after the 90th percentile. Our findings are comparable to those of the European Organization for Research and Treatment of Cancer computerized adaptive testing core norm data study, which found larger country-related effect sizes for the physical function domain than for pain [20]. Our results provide valuable information for clinicians and researchers in these countries, because our study illustrates how patient-specific factors might influence PROs. It is also beneficial for multinational studies, allowing for more accurate and informed interpretation of data. For instance, a clinician treating a patient in Germany can compare their patient's scores with the German-specific reference group, providing a more accurate representation of their patient's health status compared with the general population. Similarly, researchers conducting cross-nation studies can account for these differences when designing their studies and interpreting their results, leading to more meaningful and applicable findings.

Our study reveals an association of age and gender with all PROMIS domains. As expected, aging was related to a decrease in PROMIS PF scores and an increase in PROMIS PI scores, affirming the well-established connection between advancing age and deteriorating physical function as well as elevated pain levels [22, 24]. Additionally, we observed a gender-based difference, with women generally reporting lower PROMIS PF scores and higher PROMIS PI scores than men, aligning with previous studies highlighting gender disparities in physical function and pain perception [24, 34]. In clinical practice, these findings underscore the importance of interpreting PROs in the context of a patient's age and gender. It is crucial that clinicians avoid one-size-fits-all interpretations of PROMIS scores, but instead, compare patient scores to age-specific and gender-specific reference values, recognizing that what is considered "normal" or "average" may vary greatly depending on these factors. In research, recognizing these differences can lead to more sophisticated study designs that account for the potential confounding effects of age and gender, enhancing the validity and generalizability of our findings.

The use of plausible value imputation to model reference scores had some clear benefits. Because we accounted for measurement error, we modeled PF, UE, and PI on the PRO metric that was independent from the specific measure used in this study. We can hence use the same reference values for PROMIS computerized adaptive tests and other short forms. Furthermore, our reference values are applicable to other measures that can be scored using the PROMIS metric [15]; for instance, those that have been already linked in the PROsetta Stone project (https://www.prosettastone.org). This project links the PROMIS scales with other related instruments (such as the Short Form Health Survey, Brief Pain Inventory; Center for Epidemiologic Studies Depression Scale; Mood and Anxiety Symptom Questionnaire; or Functional Assessment of Chronic Illness Therapy-Fatigue scales) to expand the range of PRO assessment options in a common metric. Additionally, we not only modeled mean values but also provided quantiles for the entire distribution of the relevant sample to allow fine-grained comparisons.

Conclusion

Our study provides new insights into variations in PROMIS PF, UE, and PI scores across different ages, genders, and countries. It highlights the importance of using specific reference values to accurately interpret PROs. These findings can inform clinicians, researchers, healthcare policymakers, and developers of PROMs, offering stratified reference data that can aid in more personalized healthcare delivery. Specifically, these reference values could allow clinicians to understand their patients' health status more accurately and potentially adjust their treatment plans accordingly. For researchers, these findings may serve as a reference point for future studies examining PROs. Healthcare policymakers can use these data to form a more comprehensive view of the healthcare needs of different populations, especially those older than 50 years, helping to shape informed policies. Moving forward, we suggest future studies should expand on our work by considering a more complex modeling of patient-specific reference values. This could entail the inclusion of factors such as comorbidities and treatment variables, which could offer a more comprehensive understanding of patient health. However, conducting such studies requires careful design and thoughtful consideration of the methodologies used. The goal is to extend the use of precise, patient-specific reference values across the medical field, helping to improve our interpretation of PROs and ultimately contribute to more effective patient care.

References

1. Amtmann D, Cook KF, Jensen MP, et al. Development of a PROMIS item bank to measure pain interference. *Pain*. 2010; 150:173-182.
2. Amtmann D, Kim J, Chung H, Askew R, Park R, Cook K. Minimally important differences for Patient Reported Outcomes Measurement Information System Pain Interference for individuals with back pain. *J Pain Res*. 2016:251-255.
3. Bates AT, Divino C. Laparoscopic surgery in the elderly: a review of the literature. *Aging Dis*. 2015;6:149-155.
4. Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007;16:133-141.
5. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol*. 2010;63:1179-1194.
6. Chang W, Cheng J, Allaire J, et al. Shiny: Web Application Framework for R. Available at: https://CRAN.R-project.org/package=shiny. Accessed September 18, 2022.
7. Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *J Stat Softw*. 2011;39:1-30.
8. Cook KF, Jensen SE, Schalet BD, et al. PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. *J Clin Epidemiol*. 2016;73:89-102.
9. Embretson SE, Reise SP. *Item Response Theory*. Psychology Press; 2013.
10. Fischer F, Gibbons C, Coste J, Malderas JM, Rose M, Leplege A. Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany. *Qual Life Res*. 2018;27:999-1014.
11. Fischer F, Rose M. Scoring depression on a common metric: a comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivar Behav Res*. 2019;54:85-99.
12. Hays RD, Schalet BD, Spritzer KL, Cella D. Two-item PROMIS® global physical and mental health scales. *J Patient Rep Outcomes*. 2017;1:2.
13. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. *Ann Rheum Dis*. 2015;74:104-107.
14. Kaat AJ, Buckenmaier CT, Cook KF, et al. The expansion and validation of a new upper extremity item bank for the Patient-Reported Outcomes Measurement Information System® (PROMIS). *J Patient Rep Outcomes*. 2019;3:69.
15. Kaat AJ, Schalet BD, Rutsohn J, Jensen RE, Cella D. Physical function metric over measure: an illustration with the Patient-Reported Outcomes Measurement Information System (PROMIS) and the Functional Assessment of Cancer Therapy (FACT). *Cancer*. 2018;124:153-160.
16. Karhade AV, Bernstein DN, Desai V, et al. What is the clinical benefit of common orthopaedic procedures as assessed by the PROMIS versus other validated outcomes tools? *Clin Orthop Relat Res*. 2022;480:1672-1681.
17. Kisala PA, Boulton AJ, Cohen ML, et al. Interviewer- versus self-administration of PROMIS® measures for adults with traumatic injury. *Health Psychol*. 2019;38:435-444.
18. Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. *J Bone Joint Surg Am*. 2007;89:780-785.

19. Liegl G, Gandek B, Fischer HF, et al. Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. *Arthritis Res Ther*. 2017; 19:66.
20. Liegl G, Petersen MA, Groenvold M, et al. Establishing the European norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core. *Eur J Cancer*. 2019;107:133-141.
21. Magnus BE, Liu Y, He J, et al. Mode effects between computer self-administration and telephone interviewer-administration of the PROMIS® pediatric measures, self- and proxy report. *Qual Life Res*. 2016;25:1655-1665.
22. Marengoni A, Angleman S, Melis R, et al. Aging with multimorbidity: a systematic review of the literature. *Ageing Res Rev*. 2011;10:430-439.
23. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9:57.
24. Okabe T, Suzuki M, Goto H, et al. Sex differences in age-related physical changes among community-dwelling adults. *J Clin Med*. 2021;10:4800.
25. Piccinin C, Basch E, Bhatnagar V, et al. Recommendations on the use of item libraries for patient-reported outcome measurement in oncology trials: findings from an international, multidisciplinary working group. *Lancet Oncol*. 2023;24:e86-e95.
26. Rapp K, Becker C, Cameron ID, König H-H, Büchele G. Epidemiology of falls in residential aged care: analysis of more than 70,000 falls from residents of Bavarian nursing homes. *J Am Med Dir Assoc*. 2012;13:187.e1-6.
27. Remillard ML, Mazor KM, Cutrona SL, Gurwitz JH, Tjia J. Systematic review of the use of online questionnaires among the geriatric population. *J Am Geriatr Soc*. 2014;62:696-705.
28. Revicki DA, Kawata AK, Harnam N, Chen W-H, Hays RD, Cella D. Predicting EuroQol (EQ-5D) scores from the patient-reported outcomes measurement information system (PROMIS) global items and domain item banks in a United States sample. *Qual Life Res*. 2009;18:783-791.
29. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol*. 2014;67:516-526.
30. Sandvall B, Okoroafor UC, Gerull W, Guattery J, Calfee RP. Minimal clinically important difference for PROMIS Physical Function in patients with distal radius fractures. *J Hand Surg Am*. 2019;44:454-459.e1.
31. Schalet BD, Revicki DA, Cook KF, Krishnan E, Fries JF, Cella D. Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS® Physical Function. *J Gen Intern Med*. 2015;30:1517-1523.
32. Schonlau M, van Soest A, Kapteyn A, Couper M. Selection bias in web surveys and the use of propensity scores. *Sociol Methods Res*. 2009;37:291-318.
33. Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual Life Outcomes*. 2010;8:81.
34. Sialino LD, Picavet HSJ, Wijnhoven HAH, et al. Exploring the difference between men and women in physical functioning: how do sociodemographic, lifestyle- and health-related determinants contribute? *BMC Geriatr*. 2022;22:610.
35. Tang X, Schalet BD, Hung M, Brodke DS, Saltzman CL, Cella D. Linking Oswestry Disability Index to the PROMIS pain interference CAT with equipercentile methods. *Spine J*. 2021;21: 1185-1192.
36. Terwee CB, Roorda LD, de Vet HCW, et al. Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Qual Life Res*. 2014;23:1733-1741.
37. Von Davier M, Gonzalez E, Mislevy R. What are plausible values and why are they useful. *IERI Monogr Ser*. 2009;2:9-36.