Title

Validity and Responsiveness of a 10-Item Patient-Reported Measure of Physical Function in a Rheumatoid Arthritis Clinic Population

Journal

Arthritis Care & Research, 69(3)

ISSN

2151-464X

Authors

Wahl, Elizabeth
Gross, Andrew
Chernitskiy, Vladimir
et al.

Publication Date

2017-03-01

DOI

10.1002/acr.22956

Peer reviewed

HHS Public Access

Author manuscript

Arthritis Care Res (Hoboken). Author manuscript; available in PMC 2019 February 11.

Published in final edited form as:

Arthritis Care Res (Hoboken). 2017 March ; 69(3): 338–346. doi:10.1002/acr.22956.

Validity and Responsiveness of a 10-Item Patient-Reported Measure of Physical Function, PROMIS PF-10a, in a Rheumatoid Arthritis Clinic Population

Elizabeth Wahl, MD,
Division of Rheumatology, VA Quality Scholars Program, San Francisco Veterans Affairs Medical Center, San Francisco, CA

Andrew Gross, MD,
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Vladimir Chernitsky, BA,
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Laura Trupin, MPH,
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Lianne S. Gensler, MD,
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Krishna Chaganti, MD MPH,
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Kaleb Michaud, PhD,
Division of Rheumatology and Immunology, University of Nebraska Medical Center, Omaha, NE, National Data Bank for Rheumatic Diseases, Wichita, KS

Patricia Katz, PhD, and
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Jinoos Yazdany, MD MPH
Division of Rheumatology, University of California, San Francisco, San Francisco, CA

Abstract

Objective—We assessed implementation of the 10-item PROMIS physical function form (PF-10a) in routine practice in a racial/ethnically diverse population with rheumatoid arthritis (RA). Objectives were to determine feasibility of implementing PF-10a in the electronic health record (EHR) and PF-10a validity and longitudinal responsiveness.

Corresponding Author: Elizabeth Wahl, MD, Quality Scholars Fellow, Division of Rheumatology, Department of Internal Medicine, San Francisco Veterans Affairs Medical Center, 4150 Clement Street, Building 1, room 207, #181G, San Francisco, CA 94121, Tel (415) 221-4810 ext. 22546, Fax (415) 750-6641, elizabeth.wahl@va.gov.

COI/ statement

Dr. Wahl, Dr. Gross, Mr. Chernitsky, Ms. Trupin, Dr. Gensler, Dr. Chaganti, Dr. Michaud, Dr. Katz, and Dr. Yazdany have no commercial or financial conflicts of interest to disclose.

Methods—Clinical and demographic data were abstracted from EHRs for all RA patients seen at a university-based rheumatology clinic between February 2013 and February 2015. We evaluated floor and ceiling (edge) effects and construct validity of PF-10a in a subgroup of patients with HAQ scores (n=189). We used linear mixed effects models to assess responsiveness of PF-10a to longitudinal changes in the Clinical Disease Activity Index (CDAI), for patients in the entire clinical cohort with both scores recorded on at least two encounters (n=326).

Results—Half of patients were non-white, and 15% were non-English speakers. Over a two-year period, PF10a was successfully implemented; 97% of patients and 89% of encounters had at least one measurement performed. PF-10a had fewer ceiling (defined as best) effects than HAQ (8% vs. 22%) and convergent validity was high ($r = -0.85$). PF-10a was sensitive to expected differences (older versus younger patients, more versus less active disease). Longitudinal changes in PF-10a were highly associated with changes in CDAI score ($p<0.0001$).

Conclusion—The PF-10a was feasible to implement in a diverse RA population. It strongly correlates with HAQ but has fewer ceiling effects and is responsive to changes in RA disease activity, suggesting its validity for use in routine clinical practice.

Physical function is a strong predictor of clinical outcomes among patients with rheumatoid arthritis (RA) (1,2). RA patients rank physical function as a top concern, and physicians agree that functional status is an essential outcome in RA (3,4). Annual documentation of patient-reported physical function is now endorsed by the National Quality Forum and mandated by some insurance payers (5). However, the most widely used patient-reported physical function measure in rheumatology, the Health Assessment Questionnaire (HAQ) was designed to detect impairment among patients with greater disability than those seen today (6), has significant floor effects, and is not sensitive to changes in disease activity over time (7,8). Given increased recognition of the importance of measurement, documentation, and integration of patient-reported outcomes into routine clinical care of RA patients, identification of a self-reported physical function measure that is sensitive to a wide range of functional status and changes in disease activity and is feasible to implement in a variety of clinical settings is both essential and timely.

The Patient-Reported Outcome Measurement Information System (PROMIS), a NIH-initiative to develop patient-reported outcome instruments with improved validity and efficiency in experimental and observational studies, includes measures to assess patients' physical, social, and emotional functioning (www.nihpromis.org). These measures are efficient, sensitive to changes across a broad range of functioning, and are not disease specific, yielding scores that can be compared across different disease states (9,10). PROMIS Physical Function items can be administered by Computer Adaptive Test (CAT), as a stand-alone 6-, 8-, 10- or 20-item paper questionnaire (PF-10a and PF-20), or as part of a 29-item profile; mode of administration does not affect score precision (11,12).

Global quality of life measures, such as the EQ 5D, have been validated in RA (8,13), though these measures are not adequately sensitive to clinically important changes in function (14). Validity and responsiveness of the PF-10a and PF-20 to changes in RA disease activity have been demonstrated in a predominantly Caucasian, well-educated, research cohort (15–17). However, implementation of the PF-10a and validation of its

sensitivity to changes in disease activity have not been described in a real world clinical setting. A patient-reported measure of physical function that is responsive to changes over time, actionable, generic and interpretable across disease states, and readily incorporated into the EHR, is critically needed to advance care for individuals and populations with RA. In this paper, we describe the implementation of the PF-10a in a racially/ethnically diverse RA clinic population and evaluate its content validity, floor and ceiling effects, and sensitivity to change over time.

Patients and Methods

Data sources

Clinical and demographic data were extracted from the electronic health record (EHR) for all patients seen at the UCSF Rheumatology clinic with at least one face-to-face encounter with a rheumatologist that was associated with an ICD-9 code for RA between February 2013 and February 2015. All individuals with a Clinical Disease Activity Index (CDAI) score and a PF-10a score recorded at an encounter were included in the analysis. A subset of patients were participants in the multi-site longitudinal University of California San Francisco RA Cohort study (18), which collected additional clinical and demographic data.

The UCSF Committee on Human Research approved this study.

Measures

Physical function was measured using the PF-10a for all patients; the HAQ was also administered to the RA Cohort subgroup. The PF-10a is a 10-item questionnaire assessing current self-reported physical function. Raw scores range from 10 to 50 and can be translated into T-scores, with a mean of 50 and a standard deviation of 10, for comparison with the U.S. general population mean; for this study, all reported PF-10a scores are T-scores. A higher PROMIS-PF10a T-score represents better physical function. Chinese and Spanish PF-10a forms were obtained from www.nih.promis.gov and were utilized for patients who preferred these languages and were not accompanied by a trained interpreter. PF-10a can be scored if the participant answers at least 5 of the 10 items. All forms were scored and entered by clinic staff prior to the clinic visit. The HAQ is a 40-item validated questionnaire measuring functioning in RA; scores range from 0 to 3, with higher scores reflecting greater functional limitations (6).

RA disease activity was measured using the Clinical Disease Activity Index (CDAI) (19), a composite measure of Patient Global Assessment (visual analog scale from 0–100 mm), Evaluator Global Assessment (visual analog scale from 0–100 mm), and 28-Tender and -Swollen Joint counts. Scores range from 0–76 with higher values reflecting more severe disease.

All patients completed a visual analog scale (0–100 mm) for pain at each visit, and serum C-reactive protein (CRP) was measured at least every 3 months. Demographics (date of birth, sex, self-reported race/ ethnicity, preferred language, and insurance status) were retrieved from the EHR. Patients who were participants in the RA Cohort study had additional clinical and demographic information available: rheumatoid factor and anti-cyclic citrullinated

peptide serology, presence of erosive disease, date of RA diagnosis, and number of comorbidities present at diagnosis.

In 2013, the UCSF rheumatology clinic implemented the PF-10a in routine clinical practice. Several workflows were tested and optimized to achieve maximum patient participation and to collect information efficiently. The final workflow included identifying patients with RA through an 'appointment type' in our scheduling system, having front desk staff request that patients complete the PF-10a short form in the waiting room, having medical assistants score the PF-10a score and enter this information into a documentation flowsheet in our EHR. Clinic staff tallied and entered raw scores into the EHR prior to the clinical encounter, which are converted to T-scores for the clinician to view.

Statistical analysis

Floor and ceiling effects—We calculated the proportion of individuals with floor (defined as worst) and ceiling (defined as best) scores for PF-10a and HAQ and compared these proportions using Fisher's exact test.

Validity—Construct validity, the extent to which a test measures the concept or construct that it is intended to measure (20,21), was assessed by looking at convergent, discriminant, and known-groups validity in a cross-sectional analysis of all patients with at least one HAQ, one PF-10a, and one CDAI score. Convergent and discriminant validity were assessed by comparing correlation of PF-10a to that of HAQ, patient global RA assessments, pain, swollen- and tender- joint count, and CRP with Spearman's correlation coefficient, as not all scores were normally distributed. We expected that the PF-10a would correlate strongly ($r < -0.60$) with other measures of physical function (HAQ) and other measures of physical health (Patient Global Assessment VAS, Pain VAS), and moderately ($-0.30 < r < -0.60$) with clinical outcome measures (28 Tender and Swollen Joint Counts) (20).

Known-groups validity was investigated by evaluating differences in mean PF-10a scores among predefined groups of differing disease severity. PF-10a was hypothesized to show poorer scores in older patients (age ≥65 compared with age <50), non-whites (22), those with moderate or severe RA disease activity (CDAI score >10 compared with CDAI score <10), those with more severe disease history (RF/CCP positive, history of erosive disease, history of joint replacement), and those with more comorbid conditions (≥2 comorbidities compared to no comorbidities). T-tests were used to compare mean group differences, and Cohen's effect size (the difference in mean scores divided by the pooled standard deviation) was reported. Effect size values for dichotomous variables were categorized as small (<0.5), medium (0.5–0.8), or large (>0.8).

Responsiveness—Responsiveness was determined by correspondence of changes in PF-10a scores to changes in disease activity (CDAI). Clinically important change in the CDAI has been defined as a 12-point change (the minimal clinically important difference) (23). To estimate the standardized response mean (SRM), patients with PF-10a measures recorded on two occasions at least 1 month apart were divided into three groups: those with a 12-point decrease in CDAI (clinical improvement), those with a 12-point increase in CDAI (clinical worsening), and those with a <12 point change in CDAI (no change). ANOVA was

used to test for difference in mean score changes in these subgroups, and the ratio of the mean score change to the standard deviation of that change was calculated (SRM) (21). Values were categorized a small (<0.5), medium (0.5–0.8), and large (>0.8).

We then used multi-level mixed effects linear regression to assess the responsiveness of PF-10a to changes over time by modeling the relationship between PF-10a scores and changes in CDAI scores among all patients with at least two clinical encounters with recorded scores (24). This approach was selected for its ability to handle multiple levels of clustering by provider and by patient, account for different patient intercepts and trajectories, and handle missing data. We took a stepwise approach to model selection. A random effects model was fit, first allowing each subject to have his/her own starting intercept and disease trajectory. Next, since there may be systematic differences in how providers rate swollen and tender joints in the CDAI, we accounted for clustering by provider. Because different providers evaluated different sets of patients, a crossed random effects model was used. Because the association between change in CDAI score and change in PF-10a could be confounded by the magnitude of the initial CDAI score, we constructed a model separating out initial CDAI score and change in CDAI score over baselines. We then stratified by race (non-white/white) and language (non-English/English) to assess longitudinal changes in PF-10a in these subgroups. Time was included as a linear predictor; adequacy of this was evaluated for each model and addressed by including quadratic or cubic terms as appropriate. Goodness of fit was evaluated using the likelihood-ratio test. More complex models were considered to be statistically significantly better for p values <0.05. Model checking to evaluate for linearity of predictors, normality, constant variance and outliers was performed with sensitivity analyses and did not substantively affect results.

Analysis was performed using Stata (StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP).

Results

Baseline demographic and clinical characteristics of the cohort included in the cross-sectional validation study were similar to those included in the longitudinal responsiveness study (Table 1). The majority of patients were female with mean age of 59 years (SD); the group was racially/ethnically diverse (52% non-white), and the majority reported English as their preferred language. Most (85%) were seropositive for RF or CCP, median disease duration was 12 years (IQR 5–21), and about half had moderate or severe disease activity scores at baseline.

Data from 485 RA patients from 1,989 encounters (mean 3.2 visits, range 1–12) were extracted from the EHR. 472 (97%) of individuals had PROMIS scores recorded at 1780 (89%) encounters (Figure 1). The final dataset included 416 people who had both PF-10a and CDAI scores recorded on at least one encounter. Of these, 189 also had HAQ scores recorded and comprised the research subgroup used to evaluate construct validity. In the entire clinical cohort used to evaluate responsiveness, 326 individuals had both PF-10a and CDAI scores recorded on at least two encounters, in 880 encounters (mean 3.4 visits, range 2–11), over the study period. This included data from 20 providers (average of 44 RA

encounters per provider, range 1–530). Mean PF-10a score in the cross-sectional cohort was 42.4 (SD 10.2) (Figure 2), nearly a standard deviation lower than the overall US population.

A significant proportion (35/189, 19%) of HAQ scores clustered at the ceiling (highest level of functioning), compared to only 16 (8%) of PF-10a scores ($p<0.001$). Neither PF-10a nor HAQ had significant floor effects (one patient scored at the floor of the PF-10a score and none at the floor of the HAQ). To assess whether the correlation between HAQ and PF-10a was influenced by the substantial ceiling effect of the HAQ, we performed a sensitivity analysis excluding 35 individuals with a HAQ score of zero (N=154), and found no substantive change in the strength of the correlation (r=−0.81).

Validity

Convergent and discriminant validity—PF-10a scores were strongly correlated with HAQ scores, ($r = -0.874$) and patient global assessment of RA activity ($r = -0.720$), and moderately correlated with pain scores ($r = -0.631$) (Table 2). PF-10a did not correlate with swollen or tender joint counts ($r < -0.3$ for both). Correlation with CRP was modest (r = −0.446). Findings were similar for HAQ. Strength of correlation did not differ substantially when stratifying by race (white/non-white), language (English/non-English), or age (>65/<50).

Known-groups validity—Patients who were older, non-white, non-English speaking, had more active disease (CDAI>=10), longer disease duration, and more comorbidities had significantly lower PF-10a scores, as hypothesized (Table 3). Effect size (Cohen's *d*) was large in the group dichotomized by disease activity (0.93) and moderate in the group dichotomized by age (0.62) and language (0.59). There were no significant differences in groups based on RF or CCP status, a history of erosive disease, or a history of joint replacement.

Responsiveness

Of 326 patients with at least two clinical encounters, median (IQR) interval between visits was 104 (88–139) days. Patients with two encounters were divided into 3 subgroups based on whether they had a 12-point change in CDAI (clinical improvement, n=34; stable disease, n=273; and clinical deterioration, n=19). Mean PF-10a scores differed significantly between groups ($p < 0.001$). The standardized response mean was moderate in the improvement group: 0.73, and small in the groups with stable disease (−0.02) and clinical deterioration (−0.43).

Linear mixed effects modeling showed that changes in CDAI scores over time were significantly associated with changes in PF-10a scores over time ($p<0.001$), suggesting that PF-10a is responsive to changes in disease activity. The rate of increase in CDAI score was related to the rate of decline in PF-10a score (Table 4). In an unadjusted model (Model 1), in which individuals were allowed their own starting intercept and trajectory, we found that each 1-point increase in CDAI score was associated with a mean decrease of 0.30 points in PF-10a score, 95% CI (0.26 to 0.35). In order to assess whether there was clustering by provider, we tested a crossed mixed effects model; however this model did not converge and

was subsequently excluded. Model 2 isolated specific within-person changes by including a separate term for initial CDAI score. This modification resulted in improved model fit ($p<0.001$) by the likelihood ratio test when compared to Model 1, but the point estimate was similar to Model 1. Initial CDAI score was not significantly associated with changes in PF-10a score ($p=0.16$). A 12-point increase in CDAI, reflective of the minimally important difference in disease activity, was associated with a 3.3 (2.7–3.8) point decrease in PF-10a score. Stratifying by non-white race/ethnicity and non-English language did not affect the magnitude or significance of the point estimate.

Discussion

Demonstrating that a new, brief patient-reported measure is both valid and responsive to change is critical prior to its widespread uptake and implementation in clinical practice. Our study shows that PF-10a has strong psychometric properties and is responsive to clinically important changes in disease activity over time. Our work also suggests that the validity of PF-10a is not affected by non-White race/ethnicity or non-English language. In addition, we demonstrate that PF-10a can be collected efficiently and consistently over a prolonged period in a busy practice.

Physical function in our RA clinic population was relatively poor, about one standard deviation below that of the average US population. This is consistent with PROMIS physical function scores from other RA patients using PF-20 and CAT-based instruments (15), and notably, similar to those with other chronic diseases, such as cancer (mean 44.9) (25), chronic obstructive pulmonary disease (mean 40.6) (26), and congestive heart failure (mean 37.2) (27). In healthier or more highly-functioning individuals, and in settings where detecting the smallest changes among the highest functioning individuals is key, PROMIS-CAT may be more appropriate (9). However, given that the large majority of our patients did not score at the ceiling, PF-10a seems both practical and acceptable for use in a general practice setting. Impressively, all eligible patients had PF-10a scores recorded at each visit, even those with non-English language proficiency. And while the PF-CAT may reduce item response burden, implementation of a computer-based measure may not be feasible for all practices, due to lack of access to equipment, challenges integrating technology into clinical workflow, and limited computer literacy in certain patient populations (28).

In our evaluation of convergent and discriminant validity, PF-10a scores correlated strongly with HAQ, patient global assessment by VAS, and pain scores, but did not correlate with swollen or tender joint counts. The high correlation with HAQ strongly suggests that PF10a and HAQ are measuring the same underlying construct, so clinicians will still be able to monitor physical function as they have with HAQ. PF10a, however, is much more responsive to change. Several interpretations of the poor correlation with joint counts are possible. Tender and swollen joint counts are performed by a physician, and are not self-reported. Others have suggested that clinical outcome measures (tender and swollen joints) should be less strongly related to a patient-reported physical function scale (20); PF-10a may measure a truly distinct component of patient-reported physical function. Alternatively, the cross-sectional value may not be reflective of some features of disease activity (tender, swollen

joints), but the score may be more useful when the trajectory of a patient over time is considered.

Known group differences by demographic and clinical characteristics performed as hypothesized. PF-10a score differences were greatest, and had the greatest effect size, among those with moderate to severe disease activity (CDAI ≥ 10) compared to those with less active disease. This relationship has been noted in assessments of other PRO measures of physical function (8), but ours is the first study to report known-groups validity of PF-10a in an RA population. While significant differences in function were seen among Non-Hispanic whites compared to others, and among non-English speakers compared to others, this finding likely reflects racial/ethnic differences in physical function known to exist in our population and described previously (22).

Ours is the first study to longitudinally evaluate and validate the responsiveness of PF-10a scores to changes in disease activity measured by CDAI. Standardized response means suggest that PF-10a may be more responsive to clinical improvement than to clinical deterioration. This is consistent with previous studies of SF-36 physical function and role limitation subscales, though not of HAQ, which does not appear responsive to change (8,13,29). Better understanding of this phenomenon will be critical to help clinicians interpret changes in PF-10a scores and use this information to talk with patients about treatment decisions.

While a standard approach to evaluating responsiveness relies on patient self-reported change anchors obtained at a fixed time point, this was not feasible in our retrospective analysis of clinical data. Mixed effects modeling has been used previously to assess longitudinal responsiveness of a measure (24), and was used here to model the relationship between changes in CDAI and changes in PF-10a score. We found that for each 12-point increase in disease activity measured by CDAI (the MID), there is, on average, a 3-point worsening in physical function measured by PF-10a. Importantly, these findings are quantitatively consistent with Hays' evaluation of responsiveness of PF-20 in a large RA cohort, using a change anchor (15). Taken together, these results support the clinical significance of a 3-point change in PF-10a as the minimally important difference, though less than the half-standard deviation (5.2) proposed by others (30,31). Future research should examine responsiveness of PF-10a to patient-reported change using validated change anchors.

Given data suggesting cultural factors influence patients' selection of a survey response questions (32), and the fact that the PROMIS physical function item bank has been validated in a predominantly Caucasian population with high levels of education, it was especially important to validate the PF-10a in our multi-ethnic population. Stratifying by non-white race did not qualitatively affect the point estimate of the relationship between changes in CDAI and changes in PF-10a score, although subgroups were not large enough to specifically evaluate differences among Blacks, Hispanics, Asians and whites. Similarly, stratifying by non-English speakers did not affect the relationship between changes in CDAI and changes in PF-10a score, though we were limited by a relatively small non English-speaking population. Interestingly, no patients were missing PF-10a scores, regardless of

their language preferences, though 15% of eligible Chinese-speaking patients and 29% of eligible Spanish-speaking patients were missing HAQ scores. Given that HAQ forms become more cumbersome (more than 10 pages) in Chinese, and that non-English versions of the HAQ have a reading level rated as 'Difficult' (33), this is not surprising and suggests that use of the PF-10a in a multi-lingual clinical setting may be more appropriate.

Our study has several important limitations. First, for maximal efficiency, PF-10a score data was entered as a raw score into the EMR, and individual questions were not recorded, so we were not able to examine the item characteristics of the PF-10a. Second, we used real clinical data from a two-year period, from individuals with different time intervals between visits. Thus, although we were able to assess measure responsiveness, we were not able to calculate measure repeatability, though this has been tested previously (16,17). Additionally, patients with more frequent clinical encounters may have differed from those with less frequent visits, which may have affected the beta-estimates from our mixed effects model. However, in a sensitivity analysis, patients at different quartiles of visit frequency did not differ in their baseline characteristics or in their starting PF-10a score, suggesting this would not have affected the estimates we obtained. Finally, when evaluating responsiveness of PF-10a to changes in CDAI in the mixed effects model, CDAI score was treated as a continuous measure, and a 12-point minimally important difference in CDAI was used as a reference. More recent work suggests that the MID for CDAI differs depending on an individuals' starting score and the directionality of the change (34). While this may be true, we have nonetheless shown that PF-10a scores change in the expected direction and over time.

Several important questions remain regarding the relationship between changes in CDAI and PF-10a in RA patients. While our model suggests that PF-10a is responsive to changes in CDAI over time, the optimal frequency with which we should be measuring PF-10a remains unknown. Our work serves as the foundation for identifying those patients for whom changes in CDAI do not predict changes in PF-10a, and understanding why this is the case. Unraveling the contribution of other factors such as depression and fibromyalgia are also critical to understanding the responsiveness of this measure and its clinical utility, and future studies should focus on these factors. Most importantly, PF-10a is a generic instrument that could be used by health institutions in many clinical settings to understand physical function at the population level. Placing RA patients on a scale with population standards is an important step forward in rheumatology patient care.

Acknowledgments

Funding sources

KM is supported by Rheumatology Research Foundation Investigator Award and Innovative Research Grant. PK and LT are supported by NIAMS P60 AR053308-01. JY is supported by NIAMS K23 AR060259, the Russell/Engleman Rheumatology Research Center, and the Robert L. Kroc Endowed Chair in Rheumatic and Connective Tissue Diseases at the University of California, San Francisco. EW supported by a VA Quality Scholars Fellowship through the VA Office of Academic Affiliations.

References

1. Gabriel SE, Crowson CS, Kremers HM, Doran MF, Turesson C, O'Fallon WM, et al. Survival in rheumatoid arthritis: a population-based analysis of trends over 40 years. Arthritis Rheum. 2003; 48:54–58. [PubMed: 12528103]
2. Gabriel SE, Michaud K. Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases. Arthritis Res Ther. 2009; 11:229. [PubMed: 19519924]
3. Bartlett SJ, Hewlett S, Bingham CO, Woodworth TG, Alten R, Pohl C, et al. Identifying core domains to assess flare in rheumatoid arthritis: an OMERACT international patient and provider combined Delphi consensus. 712012; :1855–1860.
4. van Hulst LTC, Kievit W, van Bommel R, van Riel PLCM, Fraenkel L. Rheumatoid arthritis patients and rheumatologists approach the decision to escalate care differently: results of a maximum difference scaling experiment. Arthritis Care Res (Hoboken). 2011; 63:1407–1414. [PubMed: 21748861]
5. NQF-Endorsed Measures for Musculoskeletal Conditions. National Quality Forum Available at: http://www.qualityforum.org/ProjectMeasures.aspx?projectID=73845 Accessed December 17, 2015.
6. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980; 23:137–145. [PubMed: 7362664]
7. Sokka T, Kautiainen H, Hannonen P, Pincus T. Changes in Health Assessment Questionnaire disability scores over five years in patients with rheumatoid arthritis compared with the general population. Arthritis Rheum. 2006; 54:3113–3118. [PubMed: 17009231]
8. Linde L, Sørensen J, Ostergaard M, Hørslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D [corrected] RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol. 2008; 35:1528–1537. [PubMed: 18484697]
9. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: assessment of physical function. J Rheumatol. 2014; 41:153–158. [PubMed: 24241485]
10. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. 362009; :2061–2066.
11. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE. Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. J Clin Epidemiol. 2014; 67:108–113. [PubMed: 24262772]
12. Rose M, Bezjak A. Logistics of collecting patient-reported outcomes (PROs) in clinical practice: an overview and practical examples. Qual Life Res. 2009; 18:125–136. [PubMed: 19152119]
13. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol. 1997; 36:551–559. [PubMed: 9189057]
14. Kvien TK, Uhlig T. Quality of life in rheumatoid arthritis. Scand J Rheumatol. 2005; 34:333–341. [PubMed: 16234180]
15. 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. 2013
16. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011; 13:R147. [PubMed: 21914216]
17. Bartlett SJ, Orbai A-M, Duncan T, DeLeon E, Ruffing V, Clegg-Smith K, et al. Reliability and Validity of Selected PROMIS Measures in People with Rheumatoid Arthritis. PLoS ONE. 2015; 10:e0138543. [PubMed: 26379233]
18. Barton JL, Imboden J, Graf J, Glidden D, Yelin EH, Schillinger D. Patient-physician discordance in assessments of global disease severity in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010; 62:857–864. [PubMed: 20535797]

19. Aletaha D, Nell VPK, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005; 7:R796–806. [PubMed: 15987481]
20. Oude Voshaar MAH, Klooster Ten PM, Taal E, van de Laar MAFJ. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes. 2011; 9:99. [PubMed: 22059801]
21. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol. 1989; 42:403–408. [PubMed: 2659745]
22. Barton JL, Trupin L, Schillinger D, Gansky SA, Tonner C, Margaretten M, et al. Racial and ethnic disparities in disease activity and function among persons with rheumatoid arthritis from university-affiliated clinics. Arthritis Care Res (Hoboken). 2011; 63:1238–1246. [PubMed: 21671414]
23. Ward MM, Guthrie LC, Alba MI. Clinically important changes in individual and composite measures of rheumatoid arthritis activity: thresholds applicable in clinical trials. Ann Rheum Dis. 2014
24. Koster N, Knol DL, Uitdehaag BMJ, Scheltens P, Sikkes SAM. The sensitivity to change over time of the Amsterdam IADL Questionnaire(©). Alzheimers Dement. 2015; 11:1231–1240. [PubMed: 25598195]
25. Jensen RE, Potosky AL, Reeve BB, Hahn E, Cella D, Fries J, et al. Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. Qual Life Res. 2015; 24:2333–2344. [PubMed: 25935353]
26. Lin F-J, Pickard AS, Krishnan JA, Joo MJ, Au DH, Carson SS, et al. Measuring health-related quality of life in chronic obstructive pulmonary disease: properties of the EQ-5D-5L and PROMIS-43 short form. BMC Med Res Methodol. 2014; 14:78. [PubMed: 24934150]
27. Flynn KE, Dew MA, Lin L, Fawzy M, Graham FL, Hahn EA, et al. Reliability and construct validity of PROMIS(®) measures for patients with heart failure who undergo heart transplant. Qual Life Res. 2015; 24:2591–2599. [PubMed: 26038213]
28. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Qual Life Res. 2012; 21:1305–1314. [PubMed: 22048932]
29. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol. 1998; 37:425–436. [PubMed: 9619895]
30. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61:102–109. [PubMed: 18177782]
31. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003; 41:582–592. [PubMed: 12719681]
32. Katz PP, Barton J, Trupin L, Schmajuk G, Yazdany J, Ruiz PJ, et al. Poverty, depression, or lost in translation? Ethnic and language variation in patient-reported outcomes in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2015
33. Adams J, Chapman J, Bradley S, Ryan SJ. Literacy levels required to complete routinely used patient-reported outcome measures in rheumatology. Rheumatology (Oxford). 2013; 52:460–464. [PubMed: 23118412]
34. Curtis JR, Yang S, Chen L, Pope JE, Keystone EC, Haraoui B, et al. Determining the Minimally Important Difference in the Clinical Disease Activity Index for Improvement and Worsening in Early Rheumatoid Arthritis Patients. Arthritis Care Res (Hoboken). 2015; 67:1345–1353. [PubMed: 25988705]

Significance and Innovations

- Patients, providers, and payers value integration of patient-reported measures of physical function into rheumatoid arthritis (RA) care, but validity and responsiveness of these measures have not been well studied in a real-world clinic setting.
- We found that a brief measure of patient-reported physical function, PROMIS-PF10a (PF-10a), was feasible to implement in our RA clinic, with high rates of completion even among non-English speakers, minimal ceiling effects and robust construct validity.
- This is the first study to demonstrate that changes in PF-10a are sensitive to changes in RA disease activity measures over time, and that this association is preserved among non-White and non-English speaking patients.
- Incorporating a brief, responsive, patient-reported measure of physical function into routine clinical care of RA patients provides an opportunity to change how providers display information to patients and engage in treatment decisions, and importantly, how health systems understand the burden of RA relative to other diseases.

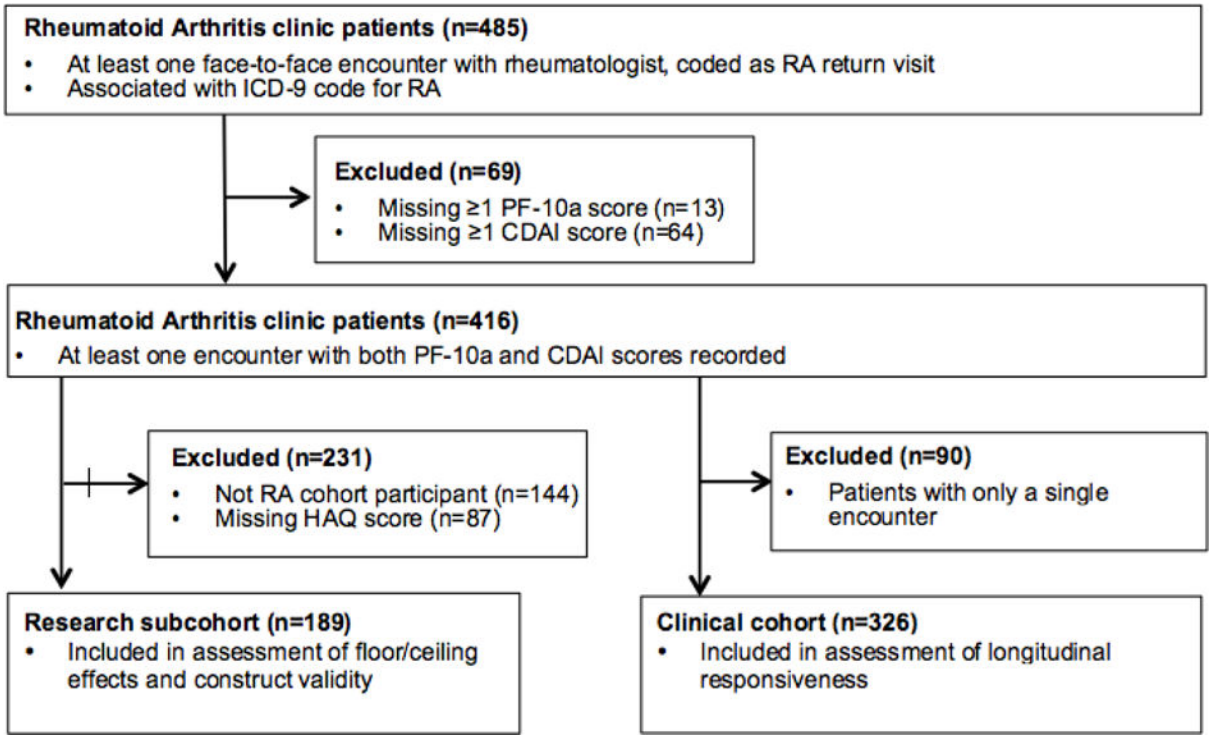

Figure 1.
Flow diagram of patients included in validation and responsiveness studies

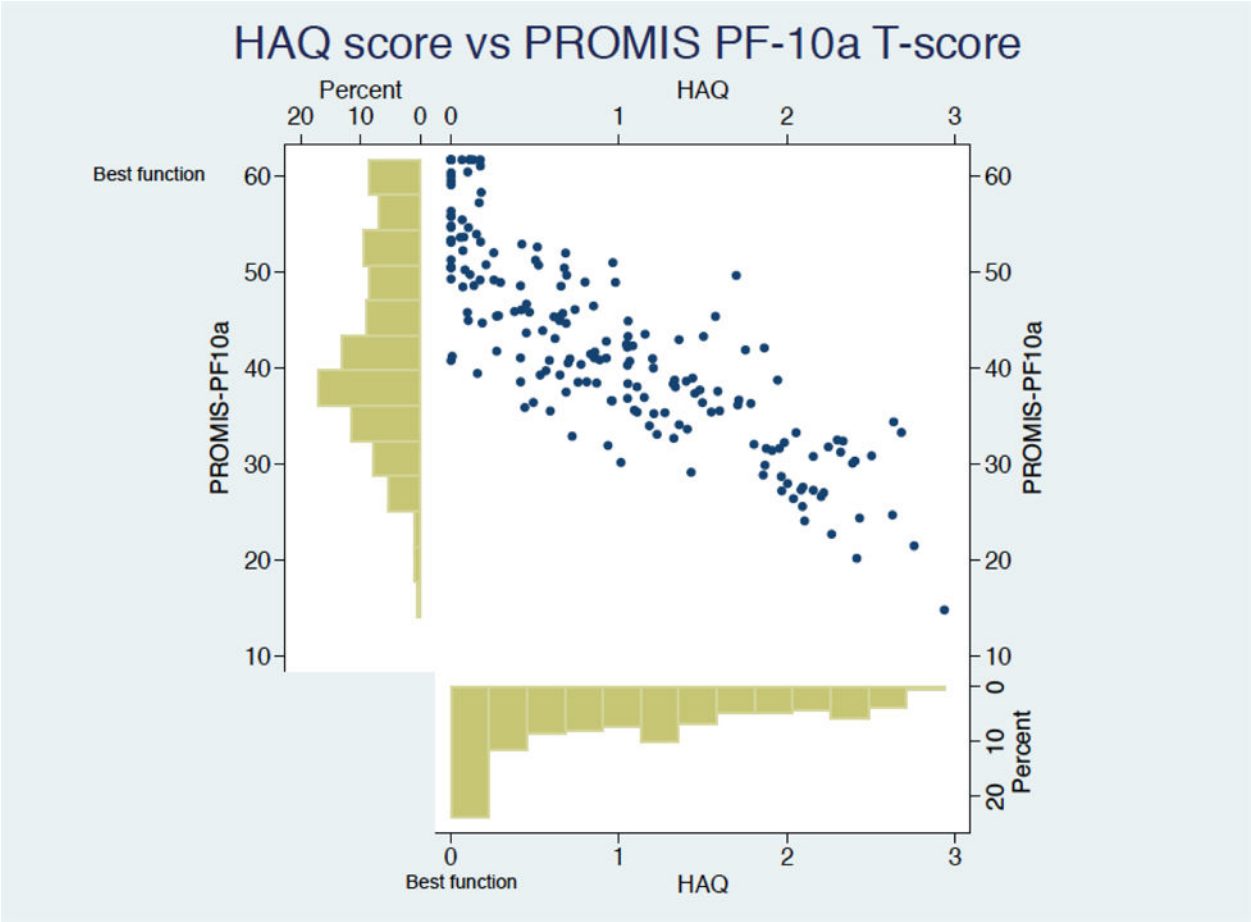

Figure 2.

Percentage distributions of HAQ and PROMIS PF-10a T-score and scatterplot of correlation between them ($r=-0.85$) in 189 RA clinic patients.

Table 1

Baseline characteristics of the RA clinic population

	Research subcohort N=189	Entire clinical cohort N=326
Age in years, mean ± SD	59 ± 14	59 ± 14
Female, %	81.5	81.6
Race/Ethnicity, %		
White	56	48
African American	5	8
Hispanic	16	15
Asian	14	18
Other	8	10
English preferred, %	87	82
Insurance type, %		
Private or state plan	54	52
Medicare	36	36
Medicaid	9	11
RA disease activity		
CDAI category, %		
Remission	16	14
Low	39	39
Moderate	30	27
Severe	14	20
Patient Global RA Assessment VAS, median (IQR)	29.0 (13.0 – 52.0)	35.0 (14.0 – 61.0)
Pain VAS, median (IQR)	27.0 (9.0 – 55.0)	34.0 (12.0 – 64.0)
Tender 28-joint count, median (IQR)	1.0 (0.0 – 4.0)	1.0 (0.0 – 5.0)
Swollen 28-joint count, median (IQR)	2.0 (0.0 – 4.0)	1.0 (0.0 – 5.0)
CRP mg/dl, median (IQR)	3.6 (1.5 – 9.7)	3.8 (1.9 – 9.8)
RA disease severity		
RF or CCP positive, n (%)	159 (85)	172 (87)
≥ 2 Comorbid conditions, n (%)	44 (23.5)	40 (20)
Disease duration in years, median (IQR)	13 (5 – 22)	12 (5 – 22)
History of erosive disease, %	58	61
History of joint replacement, %	10	12
PF10a, mean ± SD	42.4 ± 10.2	40.2 ± 10.5
HAQ, mean ± SD	0.9 ± 0.8	0.9 ± 0.8

Table 2

Convergent and discriminant validity for PF-10a scores reflected by Spearman's correlation coefficients in 189 RA clinic patients.

Domain and expected correlation	Validated measures	*r*
Strong correlation ($r < -0.6$)		
Patient-reported physical function	HAQ	**−0.874**
Patient-reported global RA activity	VAS global	**−0.720**
Patient-reported physical health	VAS pain	−0.631
Moderate correlation ($-0.3 < r < -0.6$)		
Clinical outcomes	Tender joint count	**−0.293**
	Swollen joint count	**−0.280**
Weak correlation ($r < -0.3$)		
Biological process measure	CRP (mg/dl)	−0.446

Bolded values indicate domains for which PF-10a has good convergent or discriminant validity.

RA = Rheumatoid Arthritis; HAQ = Health-Assessment Questionnaire; VAS = Visual Analog Scale; CRP = C-Reactive Protein

Table 3

Known-group validity for PF-10a in 189 RA clinic patients.

Known-group comparisons	n	Mean (SD)	Mean group difference (95% CI)	Effect size (Cohen's *d*)
Demographic characteristics				
Age				
≥65 years	78	40.6 (9.9)	**−6.0 (−9.7, −2.4)**	**0.62**
<50 years)	17	46.6 (9.7)		
Race/Ethnicity				
Non-white	83	39.8 (9.7)	−4.6 (−7.5, −1.7)	0.46
White	106	44.4 (10.2)		
Preferred language				
Non-English	24	37.2 (10.1)	−5.9 (−10.2, −1.6)	0.59
English	165	43.2 (10.1)		
Clinical characteristics at baseline				
More active disease				
CDAI ≥10	84	37.6 (9.0)	**−8.6 (−11.3, −6.0)**	**0.93**
CDAI <10	105	46.2 (9.5)		
Seropositive				
RF or CCP +	159	42.6 (10.1)	1.5 (−2.6, 5.6)	−0.15
RF and CCP −	28	41.1 (10.8)		
Disease duration				
≥ 12 years	119	41.1 (10.2)	−3.4 (−6.4, −0.4)	0.34
< 12 years	70	44.6 (9.9)		
Erosive disease				
Present	109	41.1(9.9)	−2.9 (−5.9, 0.02)	0.29
Absent	78	44.1 (10.2)		
Joint replacement				
Present	19	38.8 (9.5)	−3.9 (−8.7, 0.9)	0.39
Absent	168	42.8 (10.2)		
Comorbidities				
≥2 present	44	39.3 (9.8)	−3.9 (−0.2, −7.6)	0.39
None present	88	43.3 (10.3)		

Bolded values indicate those with a significant effect size, Cohen's $d > 0.6$.

Table 4

Linear mixed effects model coefficients (95% CI) representing the association between average change in CDAI score and change in physical function score (PF-10a) over time.

	β (95% CI)
Model 1	−0.30 (−0.35, −0.26)
Model 2 (baseline CDAI adjustment)	
Entire group, n=326	−0.27 [−0.32 – −0.23]
Stratified by race	
Non-white (n=169)	−0.28 [−0.34 – −0.21]
White (n=157)	−0.26 [−0.33 – −0.20]
Stratified by language	
Non-English (n=60)	−0.27 [−0.37 – −0.17]
English (n=266)	−0.27 [−0.32 – −0.22]

Adjusting for age did not affect the point estimate of the model.