Implementation and use of patient-reported outcome measures in routine nephrology care

Willik, E.M. van der

3

Validity and reliability of Patient-Reported Outcomes Measurement Information System (PROMIS®) using Computerized Adaptive Testing (CAT) in patients with advanced chronic kidney disease

Esmee M. van der Willik, Fenna van Breda, Brigit C. van Jaarsveld, Marlon van de Putte, Isabelle W. Jetten, Friedo W. Dekker, Yvette Meuleman, Frans J. van Ittersum, Caroline B. Terwee

Nephrology Dialysis Transplantation. 2022;gfac231.

Abstract

Background: The Patient-Reported Outcomes Measurement Information System (PROMIS®) has been recommended for computerized adaptive testing (CAT) of health-related quality of life (HRQOL). This study compared the content, validity and reliability of seven PROMIS CATs to the 12-item Short-Form Health Survey (SF-12) in patients with advanced chronic kidney disease (CKD).

Methods: Adult CKD patients with an eGFR<30 ml/min.1.73m^2 not receiving dialysis treatment completed seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and ability to participate in social roles and activities), the SF-12 and, additionally, the PROMIS Pain Intensity single item and Dialysis Symptom Index (DSI) at inclusion and 2-weeks. A content comparison was performed between PROMIS CATs and SF-12. Construct validity of PROMIS CATs was assessed using Pearson's correlations. Test-retest reliability of all patient-reported outcome measures (PROMs) was assessed by calculating the intra-class correlation coefficient (ICC) and minimal detectable change (MDC).

Results: In total, 207 patients participated in the study. A median of 45 items (10 minutes) was completed for PROMIS CATs. All PROMIS CATs showed evidence for sufficient construct validity. PROMIS CATs, most SF-12 domains and summary scores, and DSI showed sufficient test-retest reliability (ICC≥0.70). PROMIS CATs had a lower MDC compared to the SF-12 (5.7-7.4 compared to 11.2-21.7 across domains, respectively).

Conclusion: PROMIS CATs showed sufficient construct validity and test-retest reliability in patients with advanced CKD. PROMIS CATs required more items but showed better reliability than the SF-12. Future research is needed to investigate the feasibility of PROMIS CATs for routine nephrology care.

Introduction

Patients with advanced chronic kidney disease (CKD) experience numerous physical and emotional disease-related symptoms, which are associated with a decreased health-related quality of life (HRQOL).[1-4] Although several symptoms and the impact on physical, mental, and social functioning have been considered of great importance by patients and healthcare professionals[5, 6], these patient-relevant outcomes may still be regularly underrecognized and therefore insufficiently managed in routine nephrology care.[4, 7] Patient-reported outcome measures (PROMs) can be used to improve insight into these important outcomes. PROMs have been incorporated into Dutch routine dialysis care[3] and are now also being implemented into the care for Dutch patients with advanced CKD and kidney transplant recipients[8].

Many different generic and disease-specific PROMs are being used within and across countries.[9, 10] In Dutch nephrology care, the 12-item Short-Form Health Survey (SF-12) and the Dialysis Symptom Index (DSI) are used to assess generic HRQOL and disease-related symptom burden, respectively.[3] A major advantage of using the same PROMs is that this enables comparison and monitoring of outcomes across CKD stages and treatments.

Recently, the Patient-Reported Outcomes Measurement Information System (PROMIS®) was selected as one of the recommended PROMs to measure generic HRQOL in patients with CKD by a consensus group of the International Consortium of Health Outcomes Measurement (ICHOM).[11] Additionally, PROMIS was recommended by the Linnean initiative, a nationwide network of stakeholders in The Netherlands, for all patient populations, to standardize outcome measurement across medical conditions.[12] PROMIS consists of a collection of item banks (i.e. large sets of questions), developed to measure commonly relevant domains across patient conditions, such as physical function, fatigue and anxiety. Because PROMIS item banks were developed using item response theory (IRT) models, they can also be administered as Computerized Adaptive Test (CAT). The use of CAT is relatively novel in healthcare and has several advantages compared to traditional fixed (i.e. non-adaptive) PROMs. In a CAT, the computer selects questions from an item bank based on the answers to previous questions. With this method, the PROM is adapted to the patient, resulting in questions that are likely more relevant to the patient. In addition, on average less questions will be required to obtain similar or even more precise measurements compared to fixed PROMs.[13, 14] Sufficient validity and reliability of fixed PROMIS measures was found in several disease populations[15-17], including patients with CKD.[18, 19] However, the psychometric properties of PROMIS CATs have not yet been studied in patients with CKD.

Therefore, this study aimed to examine and compare the content, construct validity and test-retest reliability (including minimal detectable change) of seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and ability to participate in social roles and activities) with the SF-12 in patients with advanced CKD. Additionally, we assessed test-retest reliability of the PROMIS Pain Intensity single item and the DSI, as these PROMs are often used together with the PROMIS CATs and SF-12.

Methods

Study design and population

This observational study included adult patients with advanced CKD with an estimated glomerular filtration rate (eGFR) <30 ml/min.1.73m^2, not receiving dialysis treatment. Exclusion criteria were start with kidney replacement therapy (KRT; dialysis or kidney transplantation) planned within 4 weeks, rapid deterioration of kidney function (i.e. decrease in eGFR of >20 ml/min.1.73 m^2 during the last 6 months), not able to complete PROMs due to cognitive impairment, poor knowledge of the Dutch language, or no informed consent. Patients were recruited between November 2020 and August 2021 by their nephrologist at the outpatient clinics of Amsterdam University Medical Centre in Amsterdam and Niercentrum aan de Amstel in Amstelveen, the Netherlands. Eligible patients received written information by mail and were, if needed, approached by telephone after 2 weeks for further information. After providing written informed consent, patients were invited by e-mail to complete the PROMs digitally at the KLIK ('Kwaliteit van Leven In Kaart'; www.hetklikt.nu) research platform at inclusion (i.e. baseline), after 2 weeks and after 6 months. If necessary, two reminders were sent by e-mail or patients were contacted by telephone. Patients without access to an electronic device with internet connection could participate by telephone. In this study, the baseline and 2-weeks measurements were used (Figure 1).

The study was reviewed by the Medical Ethics Review Committee of VU University Medical Centre in the Netherlands, which confirmed that the Medical Research Involving Human Subjects Act (WMO) does not apply.

Measures

Demographic and clinical characteristics, including age, sex, primary kidney disease according to European Renal Association codes[20], body mass index (BMI), smoking status, comorbidities (hypertension, diabetes mellitus, cardiovascular disease, lung disease, liver disease and malignancy) as defined by ICHOM[11], eGFR

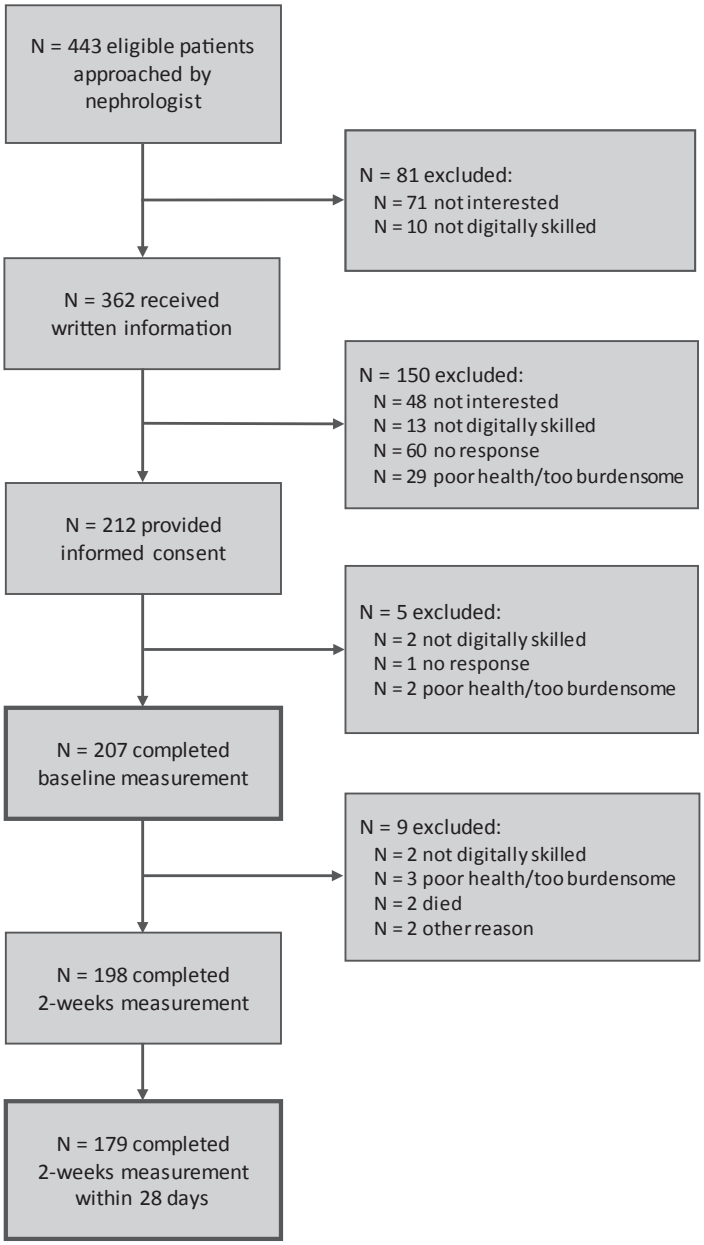

Figure 1: Flow diagram of patient inclusion for baseline and 2-weeks measurements.
All patients that completed the baseline measurement constitute the study sample for validity analyses. All patients that completed the 2-weeks measurement within 28 days after baseline are included for reliability analyses.
The reason for exclusion was indicated by the patient. Patients who were not digitally skilled were offered participation by telephone, but were not willing to participate in that manner.

(ml/min/1.73m^2), KRT in medical history, start of KRT and death during follow-up were collected from medical records. Educational level and ethno-cultural background were self-reported at baseline.

The PROMs included in this study are seven PROMIS CATs, the SF-12, one PROMIS single item and the DSI. The SF-12 and DSI have demonstrated validity within patients with CKD.[10, 21-24] PROMs were presented in random order across patients, but with fixed order within patients during follow-up. The research platform to complete PROMs did not allow for any missing values within a PROM.

Seven Dutch-Flemish PROMIS CATs[25] were administered: v1.2 Physical Function, v1.1 Pain Interference, v1.0 Fatigue, v1.0 Sleep Disturbance, v1.0 Anxiety, v1.0 Depression, and v2.0 Ability to Participate in Social Roles and Activities. All items have 5 response options (e.g. ranging from 'never' to 'always' or from 'not at all' to 'very much'). PROMIS CATs are presented as T-scores where 50 (SD: 10) represents the average score of the US general population. A difference of >2 points was considered relevant.[26] Higher scores indicate more of the construct (e.g. a higher Depression score means more depression, a higher Physical Function score means more [better] function). Within each PROMIS CAT, questions were selected one-by-one from an underlying item bank. The starting item is the item with the highest information value for the average level of the domain in the general population. The next items are subsequently selected from the item bank based on the respondent's answers to previous items. For example: a respondent reports to have difficulties with doing two hours of physical labor (first item). Then the second item will be a more 'easy' activity, e.g. a question about ability to do chores such as vacuuming. The respondent is not asked about more 'difficult' activities (e.g. running five miles) that (s)he is assumably not able to do. By tailoring the next item to the person's ability, questions are more often relevant to that person and on average less questions need to be completed. (See Supplement A for a visual illustration of CAT). After each item, the score and standard error (SE) are estimated based on all items completed so far. In this study, the CAT stopped when a SE of 2.2 on the T-score metric was reached (comparable to a reliability of approximately 0.95) or when a maximum of 12 items per CAT was administered. We used a lower SE compared to the standard stopping rule (i.e. SE: 3.0)[13], because a higher reliability may be preferable for routine care and by using this setting, the optimal performance of PROMIS CATs could be investigated. PROMIS CATs were administered using CAT software of the Dutch-Flemish Assessment Center, part of the Dutch-Flemish PROMIS National Center[27].

The SF-12 v2[28, 29] is a 12-item generic PROM assessing 8 domains of HRQOL: physical functioning, role-physical, bodily pain, general health, vitality, social function-

ing, role-emotional and mental health. Additionally, a physical component summary score (including physical functioning, role-physical, bodily pain and general health) and a mental component summary score (including vitality, social functioning, role-emotional and mental health) can be calculated. Domain and summary scores range from 0 to 100 and the US general population is used as reference with an average score of 50 (SD: 10). Higher scores indicate a better HRQOL.
The PROMIS item v1.0 Numerical Rating Scale Pain Intensity 1a is a single item with a 0-10 scale, with higher scores indicating more pain.
The DSI[21] is a 30-item disease-specific PROM to assess physical and emotional symptom burden. Patients report the presence of 30 symptoms (yes/no) during the past week and, if present, the burden of each symptom on a 5-point Likert scale ranging from 1 'not at all' to 5 'very much' bothersome. Two overall scores were calculated: 1) total number of symptoms (0-30 symptoms), and 2) total symptom burden score, which is the sum of burden on individual symptoms ranging from 0 (no symptoms) to 150 (all 30 symptoms are very much bothersome).[3, 30] The DSI items 'feeling tired or lack of energy', 'feeling anxious', 'trouble falling asleep' and 'trouble staying asleep' (hereafter combined as 'sleep problems') were used as comparison items in the construct validity analyses since these items intend to measure constructs comparable to the PROMIS CATs Fatigue, Anxiety and Sleep Disturbance.

Content comparison

To provide insight into the comparability of PROMIS CATs and SF-12, their content was compared by providing 1) an overview of the PROM characteristics (e.g. domains, number of items, recall period, scoring and interpretation), and 2) a visual comparison of the domain score distributions using an interpretative color indication (from green [better] to red [worse] HRQOL), in line with the use in routine care.[31, 32]

Construct validity

Construct validity of PROMIS CATs was assessed using Pearson's correlations. Hypotheses were formulated a priori about the expected correlations between PROMIS CATs and SF-12 and DSI, based on literature[15-18] and expert judgement (EvdW and CT). We expect strong correlations ($r \geq 0.7$) between PROMIS CATs and comparable SF-12 domains and similar DSI items, moderate correlations ($r = 0.5$-0.7) between PROMIS CATs and largely related SF-12 domains, and no strong correlations for other comparisons ($r \leq 0.6$) (see Table 1). Construct validity was considered sufficient if ≥75% of the results was in accordance with the hypotheses.

Table 1. Hypotheses for construct validity

PROMIS CAT	Strong correlation Pearson's r ≥ 0.7	Moderate correlation Pearson's r 0.5-0.7	No strong correlation Pearson's r ≤ 0.6
Physical Function	SF-12 physical functioning SF-12 physical component summary*	SF-12 general health SF-12 bodily pain	All other SF-12 domains DSI total number of symptoms and symptom burden score
Pain Interference	SF-12 bodily pain	SF-12 physical functioning SF-12 physical component summary	All other SF-12 domains DSI total number of symptoms and symptom burden score
Fatigue	SF-12 vitality DSI feeling tired or lack of energy (1 item)		All other SF-12 domains DSI total number of symptoms and symptom burden score
Sleep Disturbance	DSI sleep problems (2 items)#		All other SF-12 domains DSI total number of symptoms and symptom burden score
Anxiety	SF-12 mental health SF-12 mental component summary* DSI feeling anxious (1 item)		All other SF-12 domains DSI total number of symptoms and symptom burden score
Depression	SF-12 mental health SF-12 mental component summary*		All other SF-12 domains DSI total number of symptoms and symptom burden score
Ability to Participate in Social Roles and Activities	SF-12 social functioning	SF-12 role physical SF-12 role emotional	All other SF-12 domains DSI total number of symptoms and symptom burden score

** SF-12 physical component summary includes the domains physical functioning, role-physical, bodily pain and general health; SF-12 mental component summary includes the domains vitality, social functioning, role-emotional and mental health.*

 # DSI Sleep problems were defined as trouble falling asleep and/or trouble staying asleep.

Test-retest reliability

Test-retest reliability of PROMIS CATs, SF-12, PROMIS Pain Intensity single item, and DSI was assessed by calculating the intra-class correlation coefficient (ICC) in patients with valid baseline and 2-weeks measurements (Figure 1). The ICC was calculated using a two-way random-effects model for absolute agreement: *ICC agreement* = $\sigma^2_p / (\sigma^2_p + \sigma^2_m + \sigma^2_e)$, whereby σ^2_p is the variation between patients, σ^2_m is the variation between measurements and σ^2_e is random error variance. An ICC ≥0.70 was considered sufficient.[33]

The ICC was computed for each PROMIS CAT and SF-12 domain separately. Additionally, the ICC was calculated for the PROMIS Pain Intensity single item and for the DSI total number of symptoms and symptom burden score. Although the DSI was not designed to be interpreted as an overall score (as it measures 30 different symptoms), the total number of symptoms and symptom burden score are often used within healthcare, and insight into the reliability of these scores is therefore of clinical relevance.

The minimal detectable change (MDC) was also calculated for each domain of the PROMIS CATs and SF-12, the PROMIS Pain Intensity single item, and the DSI total number of symptoms and symptom burden score. The MDC is a parameter of measurement error and is defined as the "smallest change in score that can be detected beyond measurement error", with 95% confidence.[33] Two different methods were applied to calculate the MDC, in line with the underlying measurement theories, namely classical test theory (CTT) or IRT, that assume a constant or varying standard error of measurement (SEM) across the PROM-scale, respectively.[34, 35]

The MDC, based on CTT, of the SF-12 domains, PROMIS Pain Intensity single item, and the DSI total number of symptoms and symptom burden score was calculated using the formula: $1.96 * \sqrt{2} * SEM$, whereby SEM was calculated as: $\sqrt{(\sigma^2_m + \sigma^2_e)}$.

The MDC, based on IRT, of each PROMIS CAT varies per patient (because with IRT the SE of each score is different) and was calculated using the following formula: $1.96 * \sqrt{(SE_1^2 + SE_2^2)}$, whereby SE_1 is the patient's IRT estimated standard error of the T-score at baseline and SE_2 at the 2-weeks measurement. A mean MDC of each PROMIS CAT was subsequently calculated for the whole group.

Data analyses were performed using SPSS V.25.0 (IBM Corp., Armonk, NY, USA).

Table 2. Characteristics of study sample at baseline and 2-weeks measurements.

	Study sample at baseline* (n=207)	Study sample at 2 weeks* (n=179)
Sex, male	124 (59.9)	107 (59.8)
Age, years	65.5 (13.8)	66.1 (13.1)
Ethno-cultural group[$], Dutch	176 (85.0)	152 (84.9)
Educational level[#]		
Low	85 (41.0)	74 (41.3)
Middle	49 (23.7)	43 (24.0)
High	73 (35.3)	62 (34.6)
Primary kidney disease		
Glomerulonephritis/sclerosis	34 (16.6)	33 (18.6)
Pyelonephritis	7 (3.4)	7 (4.0)
Polycystic kidney disease	16 (7.8)	15 (8.5)
Other congenital/hereditary kidney diseases	15 (7.3)	13 (7.3)
Hypertension/renal vascular disease	46 (22.5)	42 (23.7)
Diabetes mellitus	14 (6.8)	12 (6.8)
Miscellaneous	63 (30.7)	49 (27.7)
Unknown	10 (4.9)	6 (3.4)
Kidney function, eGFR	21.4 (6.7)	21.6 (6.6)
KRT in medical history[£], yes	35 (17.0)	30 (16.9)
BMI	26.8 (5.2)	26.9 (5.2)
Smoking		
Yes	25 (13.2)	19 (11.7)
No, stopped	94 (49.7)	82 (50.6)
No, never smoked	70 (37.0)	61 (37.7)
Comorbidities		
Hypertension, yes	164 (79.2)	140 (78.2)
Diabetes mellitus, yes	62 (30.0)	53 (29.6)
Cardiovascular disease, yes	53 (25.6)	43 (24.0)
Lung disease, yes	30 (14.5)	28 (15.6)
Liver disease, yes	11 (5.3)	8 (4.5)
Malignancy, yes	50 (24.2)	43 (24.0)

Values are shown in n (%) or mean (SD).
** Study sample at baseline was used for validity analyses. Study sample at 2-weeks measurement was used for reliability analyses.*
$ Self-reported ethno-cultural group: "What ethnic group do you consider yourself to belong to?"
 # Educational level according to International Standard Classification of Education (ISCED) levels 2011, classified as low: primary, lower secondary or lower vocational education; middle: upper secondary or upper vocational education; high: tertiary education (college/university).
£ KRT in medical history includes patients who have undergone (temporary) dialysis treatment or a kidney transplant in the past. At study inclusion, all patients had an eGFR<30 and did not require dialysis treatment, in accordance with inclusion criteria.
Missing values at baseline: primary kidney disease: n=2 (1.0%); KRT in medical history: n=1 (0.5%); BMI: n=11 (5.3%); smoking: n=18 (8.7%). Missing values at 2 weeks: primary kidney disease: n=2 (1.1%); KRT in medical history: n=1 (0.6%); BMI: n=9 (5.0%); smoking: n=17 (9.5%).
Abbreviations: eGFR, estimated glomerular filtration rate; KRT, kidney replacement therapy; BMI, body mass index.

Results

Study participants

Almost half of the patients that were approached provided written informed consent. In total, 207 participants completed the baseline measurement and were included in current analyses. Of them, 179 (86.5%) participants completed the 2-weeks measurement within 28 days and were eligible for reliability analyses (Figure 1). The average time between the baseline and 2-weeks measurement was 14.1 (SD: 3.7) days. Eleven patients participated by telephone. Sociodemographic and clinical characteristics of the participants at baseline and 2-weeks measurements are shown in Table 2. The baseline and 2-weeks study samples were comparable. About 60% was male, mean (SD) age was 65.5 (13.8) years and the majority (85%) had a Dutch ethno-cultural background. Mean (SD) eGFR was 21.4 (6.7) and 17% has had KRT in the past.

Content comparison

Table 3 shows the similarities and differences in characteristics of PROMIS CATs and SF-12. Although assessing the same patient-relevant outcome (i.e. generic HRQOL), PROMIS CATs and SF-12 include related but slightly different domains. The PROMs have similarities in scoring (e.g. score range and US reference), but use a different underlying measurement method and score interpretation. In PROMIS CATs, the (number of) items varies from person to person, depending on the severity of symptoms or the function level on the domain being measured and the consistency of the answers. Our study sample of advanced CKD patients completed a median (IQR) of 45 (38-55) items for all seven PROMIS CATs, which took them a median (IQR) of 10.2 (8.3-12.6) minutes. The median (IQR) time to complete the SF-12 was 3.3 (2.4-4.6) minutes.

Table 3. Content comparison of PROMIS CAT with SF-12^

	PROMIS CAT	SF-12
Type of PROM	Generic	Generic
PRO	HRQOL	HRQOL
Domains	Physical Function Pain Interference Fatigue Sleep Disturbance Anxiety Depression Ability to Participate in Social Roles and Activities	Physical functioning Bodily pain Vitality Role-physical Role-emotional Mental health Social functioning General health *Composite summary scores*:* Physical component summary Mental component summary
Number of items	All PROMIS domains[$] median (IQR): 45 (38-55) items Physical Function median (IQR): 4 (3-6) items Pain Interference median (IQR): 4 (2-12) items Fatigue median (IQR): 5 (4-6) items Sleep Disturbance median (IQR): 10 (8-12) items Anxiety median (IQR): 7 (6-10) items Depression median (IQR): 8 (5-12) items Ability to Participate in Social Roles and Activities median (IQR): 5 (4-6) items	All SF-12 items 12 items Physical functioning 2 items Bodily pain 1 item Vitality 1 item Role-physical 2 items Role-emotional 2 items Mental health 2 items Social functioning 1 item General health 1 item *Composite summary scores*:* Physical component summary 6 items Mental component summary 6 items
Recall period	In general/1 week	In general/4 weeks
Rating scale	5-point scale	3- and 5-point scales
Score (range)	Norm-based scoring T-score (roughly 0–100)	Norm-based scoring (roughly 0-100)
Norm or reference standard	General US population: mean 50, SD 10	General US population: mean 50, SD 10

Score interpretation	Higher scores represent more of the HRQOL domain being measured. E.g. a higher score on fatigue means a worse fatigue, and a higher score on physical function means a better physical function.	Higher scores represent a more favourable HRQOL. E.g. a higher score on bodily pain means less bodily pain, and a higher score on physical functioning means a better physical functioning.
Measurement method	Item Response Theory	Classical Test Theory
Completion options	Electronic only	Electronic or paper-based
Time to complete[#]	All PROMIS CATs median (IQR): 10.2 (8.3-12.6) min. Physical Function median (IQR): 1.3 (0.8-1.7) min. Pain Interference median (IQR): 1.2 (0.8-1.8) min. Fatigue median (IQR): 1.3 (1.0-2.0) min. Sleep Disturbance median (IQR): 2.0 (1.5-2.6) min. Anxiety median (IQR): 1.4 (1.0-1.9) min. Depression median (IQR): 1.3 (1.0-1.8) min. Ability to Participate in Social Roles and Activities median (IQR): 1.2 (1.0-1.6) min.	All SF-12 items median (IQR): 3.3 (2.4-4.6) min.

[^] The Dialysis Symptom Index (DSI) aims to measure a different patient-relevant outcome and is therefore not included in this table. For characteristics of the DSI, see Weisbord 2004[21] and Van der Willik 2021[36].

[*] SF-12 physical component summary includes the domains physical functioning, role-physical, bodily pain and general health; SF-12 mental component summary includes the domains vitality, social functioning, role-emotional and mental health.

[$] Number of items used as observed in current study sample at baseline. Additional item details, including the top 3 most frequently used items of PROMIS CATs are provided in Supplement B.

[#] Time to complete the PROMs as observed in current study sample at baseline.

Abbreviations: PROM, patient-reported outcome measure; PROMIS, Patient-Reported Outcomes Measurement Information System; CAT, Computerized Adaptive Test; SF-12, 12-item Short Form Health Survey; PRO, patient-reported outcome; IQR, interquartile range; SD, standard deviation.

Table 4 and Figure 2 show the PROM-scores in our study sample of patients with advanced CKD. Less variation (i.e. lower SDs) was observed in PROMIS CATs compared to SF-12 domains and summary scores. Overall, PROMIS CATs showed 'better' (towards the green area) HRQOL scores compared to the SF-12; only two PROMIS CATs showed worse HRQOL scores than the general US population (Physical Function [mean ± SD: 43.4±8.3] and Fatigue [53.2±8.7]), compared to six SF-12 domains and one summary score (physical functioning [40.5±11.3], role-physical [40.1±10.3], bodily pain [46.9±11.3], general health [36.3±10.9], social functioning [43.4±12.1], role-emotional [44.2±11.3] and physical component summary [39.2±10.7]).

Construct validity

All PROMIS CATs showed evidence for sufficient construct validity as ≥75% of the results were in accordance with the hypotheses (Table 5). For Pain Interference, Sleep Disturbance and Depression, all correlations were in accordance with the hypotheses. For Physical Function 14 out of 15 hypotheses were met. For Fatigue and Ability to Participate in Social Roles and Activities, 13 out of 15 correlations, and for Anxiety, 12 out of 15 correlations were in accordance with the hypotheses.

Table 4. Baseline scores on PROMIS CATs, SF-12, PROMIS Pain Intensity, and DSI in patients with chronic kidney disease (n=207)

	N (%)[$]	Mean (SD) or Median (IQR)	Range (min-max)
PROMIS CATs			
Physical Function	205 (99.0)	43.4 (8.3)	24.1 - 67.6
Pain Interference	203 (98.1)	51.9 (9.1)	41.0 - 74.9
Fatigue	203 (98.1)	53.2 (8.7)	28.8 - 70.7
Sleep Disturbance	203 (98.1)	49.3 (7.9)	30.0 - 71.6
Anxiety	203 (98.1)	51.2 (7.7)	35.9 - 70.3
Depression	204 (98.6)	49.8 (7.5)	37.1 - 70.0
Ability to Participate in Social Roles and Activities	203 (98.1)	49.2 (8.6)	29.9 - 64.9
SF-12			
Physical functioning	204 (98.6)	40.5 (11.3)	22.1 - 56.5
Role-physical	204 (98.6)	40.1 (10.3)	20.3 - 57.2
Bodily pain	204 (98.6)	46.9 (11.3)	16.7 - 57.4
General health	204 (98.6)	36.3 (10.9)	18.9 - 62.0
Vitality	204 (98.6)	48.5 (10.2)	27.6 - 67.9

Social functioning	204 (98.6)	43.4 (12.1)	16.2 - 56.6
Role-emotional	204 (98.6)	44.2 (11.3)	11.3 - 56.1
Mental health	204 (98.6)	50.1 (9.3)	28.0 - 64.5
Physical component summary*	204 (98.6)	39.2 (10.7)	11.1 - 61.4
Mental component summary*	204 (98.6)	49.3 (9.7)	23.4 - 69.0
PROMIS single item			
Pain Intensity (0-10)	204 (98.6)	1 (0-5)	0 - 10
DSI			
Number of symptoms (0-30)	203 (98.1)	9.4 (5.6)	0 - 28
Symptom burden score (0-150)	203 (98.1)	22 (12-36)	0 - 96
Feeling tired or lack of energy (0-5)^	203 (98.1)	2.0 (1.6)	0 - 5
Sleep problems (0-10)^#	203 (98.1)	2 (0-3)	0 - 10
Feeling anxious (0-5)^	203 (98.1)	0 (0-0)	0 - 5

$ In total, four people did not finish the measurement and only completed part of the PROMs.
** SF-12 physical component summary includes the domains physical functioning, role-physical, bodily pain and general health; SF-12 mental component summary includes the domains vitality, social functioning, role-emotional and mental health.*
^ Prevalence of feeling tired or lack of energy: 70.0%, sleep problems: 52.7%, feeling anxious: 18.7%.
 # Sleep problems were defined as trouble falling asleep and/or trouble staying asleep.
Abbreviations: PROMIS, Patient-Reported Outcomes Measurement Information System; CAT, Computerized Adaptive Test; SF-12, 12-item Short Form Health Survey; DSI, Dialysis Symptom Index; SD, standard deviation; IQR, interquartile range.

Test-retest reliability

The reliability measures – ICC agreement, SEM and MDC – of the PROMIS CATs, SF-12, PROMIS Pain Intensity single item, and DSI are shown in Table 6. All PROMIS CATs showed sufficient test-retest reliability (ICCs between 0.77 and 0.92). The SF-12 domains physical functioning, role-physical, bodily pain, general health, mental health, and the physical and mental component summary scores also showed sufficient reliability (ICCs between 0.70 and 0.85). For SF-12 role-emotional, social functioning and vitality the ICC was between 0.48 and 0.67. The PROMIS Pain Intensity single item showed an ICC of 0.68. The DSI total number of symptoms and symptom burden score showed sufficient reliability (ICC 0.85 and 0.88, respectively).

The SEM and MDC of PROMIS CATs ranged from 2.1 to 2.7, and from 5.7 to 7.4, respectively, across domains. For the SF-12, the SEM and MDC ranged from 4.1 to 7.8, and from 11.2 to 21.7, respectively, across domains.

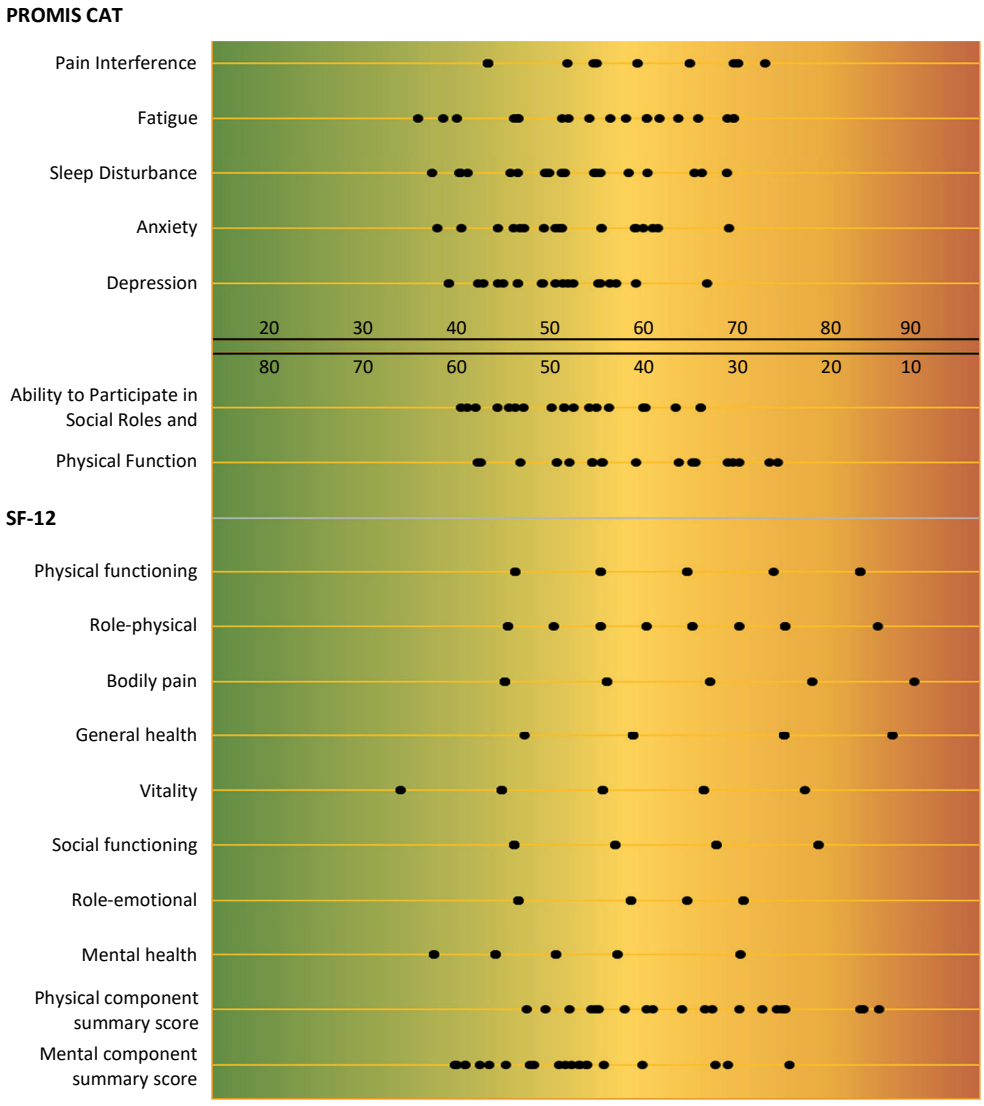

Figure 2. Score distributions of PROMIS CATs and SF-12 domains and summary scores.

The figure's background color gives an indication of the interpretation of scores, ranging from good (green) to worse (red) HRQOL.[31] *Note that the first five PROMIS CATs use a reverse scale compared to the other PROMIS CATs and SF-12.*

Table 5. Pearson's r for correlations between PROMIS CATs and SF-12 and DSI scores (n=207)

		PROMIS CATs						
		Physical Function	**Pain Interference**	**Fatigue**	**Sleep Disturbance**	**Anxiety**	**Depression**	**Ability to Participate**
SF-12	Physical functioning	**0.80**	*-0.52*	-0.49	-0.17	-0.19	-0.25	0.52
	Role-physical	0.65	-0.49	-0.59	-0.26	-0.24	-0.36	*0.59*
	Bodily pain	*0.59*	**-0.79**	-0.47	-0.35	-0.33	-0.33	0.47
	General health	*0.52*	-0.36	-0.53	-0.27	-0.24	-0.32	0.52
	Vitality	0.52	-0.39	**-0.66**	-0.31	-0.32	-0.43	0.59
	Social functioning	0.54	-0.49	-0.54	-0.34	-0.54	-0.58	**0.66**
	Role-emotional	0.30	-0.34	-0.41	-0.26	-0.40	-0.49	*0.39*
	Mental health	0.22	-0.33	-0.46	-0.33	**-0.66**	**-0.73**	0.40
	Physical component summary score*	**0.80**	*-0.63*	-0.55	-0.24	-0.13	-0.20	0.58
	Mental component summary score*	0.20	-0.29	-0.49	-0.35	**-0.64**	**-0.72**	0.47
DSI	Number of symptoms	-0.45	0.53	0.59	0.48	0.54	0.54	-0.48
	Symptom burden score	-0.48	0.55	0.60	0.51	0.51	0.52	-0.49
	Feeling tired or lack of energy	-0.41	0.41	**0.76**	0.35	0.36	0.46	-0.50
	Sleep problems#	-0.29	0.32	0.30	**0.79**	0.27	0.23	-0.27
	Feeling anxious	-0.04	0.16	0.16	0.25	**0.56**	0.48	-0.13
Hypotheses confirmed (%)		**93**	**100**	**87**	**100**	**80**	**100**	**87**

Correlations in **bold** *were expected to be strong (≥0.7 or ≤-0.7), correlations in italic were expected to be moderate (±0.5-0.7), other correlations were expected not to be strong (≤ 0.6 or ≥-0.6).*

** SF-12 physical component summary includes the domains physical functioning, role-physical, bodily pain and general health; SF-12 mental component summary includes the domains vitality, social functioning, role-emotional and mental health.*

 # DSI Sleep problems were defined as trouble falling asleep and/or trouble staying asleep.

Abbreviations: PROMIS, Patient-Reported Outcomes Measurement Information System; CAT, Computerized Adaptive Test; SF-12, 12-item Short Form Health Survey; DSI, Dialysis Symptom Index.

Discussion

This study examined the validity and reliability of seven PROMIS CATs in comparison to the SF-12 in patients with advanced CKD in The Netherlands. This is the first study investigating the psychometric performance of the Dutch-Flemish version of these PROMIS domains using CATs. All PROMIS CATs demonstrated evidence for sufficient construct validity and test-retest reliability. Overall, PROMIS CATs showed better reliability, with higher ICCs and lower MDCs, but required more items compared to the SF-12.

The observed average HRQOL scores are in line with scores that would be expected from existing literature in CKD patients for both the PROMIS CATs[18, 37, 38] and SF-12[2, 3, 39, 40]. However, comparison of the domain scores revealed a slightly better HRQOL in advanced CKD patients based on PROMIS CATs compared to SF-12. This demonstrates that the scores are not directly comparable in contrast to what one might intuitively expect based on the corresponding characteristics of both PROMs (0-100 scale, mean 50 with SD 10, US reference population). This can be explained by the fact that PROMIS CAT and SF-12 scores are on a different metric because they originate from different (calibration) samples[41], which is reflected in the smaller SDs for PROMIS CATs compared to SF-12. By means of linking[42], the scores of comparable PROMIS CAT and SF-12 domains could be converted into each other; this has been done for many other PROMs[43] and would be a valuable next step, as it facilitates harmonization of data across studies or healthcare organizations (e.g. when both instruments are used across different healthcare specialties) and comparison to historical data if one changes from one PROM to the other.[42]

All PROMIS CATs showed sufficient test-retest reliability with better ICCs and small MDCs compared to the SF-12. Small MDCs allow for small changes to be distinguished from measurement error with 95% confidence, and are therefore desirable especially when the minimal important change (MIC) is small.[36] For PROMIS, the MIC has been estimated at 2-6 points[26], which is slightly smaller but close to the MDC of 6-7 points. Information about the MIC for SF-12 domains is limited, which makes it difficult to say to what extent SF-12 can distinguish important changes from measurement error.[36] Our reliability results were better than results found in other research using PROMIS short forms (e.g. PROMIS-29 and -57, including 4 and 8 fixed items per domain, respectively).[18] This was expected given the underlying method of CAT and the stopping rule including a low SE to achieve high reliability. A downside of the higher precision stopping rule is the relatively large number of 45 questions asked (i.e. six to seven items per domain and three to four times the length of the SF-12). This number of items might raise some feasibility concerns for use in routine clinical practice. If fewer items are preferred, alternative stopping

Table 6. Reliability measures of PROMIS CAT, SF-12, PROMIS Pain Intensity, and DSI in patients with chronic kidney disease (n=179)

	ICC agreement (95%CI)	SEM	MDC
PROMIS CAT			
Physical Function	0.92 (0.89-0.94)	2.06	5.72
Pain Interference	0.78 (0.71-0.83)	2.65	7.43
Fatigue	0.81 (0.75-0.86)	2.06	5.71
Sleep Disturbance	0.84 (0.79-0.88)	2.22	6.15
Anxiety	0.78 (0.71-0.83)	2.29	6.36
Depression	0.81 (0.76-0.86)	2.35	6.53
Ability to Participate in Social Roles and Activities	0.77 (0.71-0.83)	2.09	5.80
SF-12			
Physical functioning	0.76 (0.69-0.82)	5.27	14.61
Role-physical	0.73 (0.65-0.79)	5.10	14.13
Bodily pain	0.70 (0.62-0.77)	6.02	16.67
General health	0.75 (0.68-0.81)	5.23	14.50
Vitality	0.67 (0.58-0.75)	5.72	15.85
Social functioning	0.64 (0.54-0.72)	7.20	19.96
Role-emotional	0.48 (0.36-0.58)	7.82	21.67
Mental health	0.78 (0.82-0.83)	4.32	11.98
Physical component summary score*	0.85 (0.81-0.89)	4.07	11.29
Mental component summary score*	0.72 (0.65-0.79)	5.09	14.11
PROMIS single item			
Pain Intensity (0-10)	0.68 (0.59-0.76)	1.53	4.24
DSI			
Number of symptoms (0-30)	0.85 (0.80-0.88)	2.12	5.87
Total symptom burden score (0-150)	0.88 (0.85-0.91)	5.75	15.94

* *SF-12 physical component summary includes the domains physical functioning, role-physical, bodily pain and general health; SF-12 mental component summary includes the domains vitality, social functioning, role-emotional and mental health.*

Abbreviations: PROMIS, Patient-Reported Outcomes Measurement Information System; CAT, Computerized Adaptive Test; SF-12, 12-item Short Form Health Survey; DSI, Dialysis Symptom Index; ICC, intra-class correlation coefficient; CI, confidence interval; SEM, standard error of measurement; MDC, minimal detectable change

rules could be considered but with detriment to precision. In this study, we applied a stopping rule with a smaller SE of 2.2 compared to the standard stopping rule (SE: 3.0)[13] to investigate optimal performance of the PROMIS CATs. We expect that application of the standard stopping rule will result in 36-43 items in total (5-6 items per CAT), with a minimum of 28 items since the standard stopping rule requires 4 items per CAT, and less than 45 items due to the higher SE in comparison to this study. Other alternative stopping rules to consider might be a lower maximum number of items (a maximum of 8 instead of 12 items per domain is currently being considered for the standard PROMIS CAT algorithms), stopping when the SE does not change much anymore (e.g. <0.1), or stopping when the score range is above or below a certain cut-off point on the scale (e.g. when the functionality or symptom burden is at such a level that it is probably not perceived as burdensome). The latter may be particularly beneficial for domains such as Pain Interference and Sleep Disturbance, to keep the number of items low for patients with no pain or sleep problems. Further research is needed to explore feasibility and the most optimal use of PROMIS CATs in routine nephrology care, in close collaboration with patients and healthcare professionals.

A limitation of PROMIS CATs is that they can only be completed digitally. Participants thus have to have access to an electronic device and be digitally skilled. In the Netherlands, approximately 80% of the population aged 55+ is sufficiently digitally skilled[44], but in many countries – also within Europe – citizens are less digitally skilled.[45, 46] Consequently, it may be challenging to reach the total advanced CKD population. In our study, we therefore enabled participation by telephone. For routine care, also other methods could be considered, for instance offering help or making tablets available on site.

An advantage of PROMIS CATs is that the PROM adapts to the patient, resulting in items that are more likely considered relevant by the patient. As a result, the PROM might be perceived as less burdensome. On the other hand, items may vary over time, meaning that progression of individual items cannot be easily monitored over time, which is in contrast to how the SF-12 (and DSI) is also being used in routine nephrology care.[3] In addition, the varying items and 'black box algorithm' (i.e. not a simple sum of scores) may also lead to patients and professionals finding it more difficult to interpret the scores. Qualitative research is needed to investigate patients' and professionals' preferences for use in routine nephrology care.

Furthermore, it may be important to mention that the SF-12 was selected by patients for use in routine nephrology care, partly because of the low number of items. Besides, the SF-12 was considered a good fit with the DSI to provide insight into both generic HRQOL and disease-specific symptom burden.[3, 10] Differences

in characteristics of the PROMIS CATs and SF-12, and how they complement other PRO(M)s, should thus be taken into account when considering which PROM fits routine nephrology care best to measure HRQOL.

An important strength of this study is that the PROMIS CATs were compared to the PROM that is currently being used in routine nephrology care to assess generic HRQOL (i.e. the SF-12). The findings from this study are therefore of clinical relevance and can contribute to considerations regarding which PROMs best fit routine practice to measure HRQOL. A disadvantage is that the SF-12 may not be the best comparator (i.e. 'golden standard') for the PROMIS CATs, for instance because of the low number of items per domain and the fact that in practice, both in research and in healthcare, less focus is often being paid towards individual SF-12 domains. To expand on current findings, future research could investigate the validity of PROMIS CATs in comparison to the SF-36.[11, 29]

Conclusion

All seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and ability to participate in social roles and activities) demonstrated evidence for sufficient construct validity and test-retest reliability in patients with advanced CKD in The Netherlands. PROMIS CATs required more items but showed better reliability than the SF-12. Future research is needed to investigate the optimal use of PROMIS CATs for routine nephrology care.

References

1. Almutary H, Bonner A, Douglas C. Symptom burden in chronic kidney disease: a review of recent literature. *J Ren Care* 2013;39(3):140-50

2. Voskamp PWM, van Diepen M, Evans M, et al. The impact of symptoms on health-related quality of life in elderly pre-dialysis patients: effect and importance in the EQUAL study. *Nephrol Dial Transplant* 2019;34(10):1707-15

3. van der Willik EM, Hemmelder MH, Bart HAJ, et al. Routinely measuring symptom burden and health-related quality of life in dialysis patients: first results from the Dutch registry of patient-reported outcome measures. *Clin Kidney J* 2021;14(6):1535-44

4. Raj R, Ahuja KD, Frandsen M, Jose M. Symptoms and their recognition in adult haemodialysis patients: Interactions with quality of life. *Nephrology (Carlton)* 2017;22(3):228-33

5. Manns B, Hemmelgarn B, Lillie E, et al. Setting research priorities for patients on or nearing dialysis. *Clin J Am Soc Nephrol* 2014;9(10):1813-21

6. Urquhart-Secord R, Craig JC, Hemmelgarn B, et al. Patient and Caregiver Priorities for Outcomes in Hemodialysis: An International Nominal Group Technique Study. *Am J Kidney Dis* 2016;68(3):444-54

7. Weisbord SD, Fried LF, Mor MK, et al. Renal provider recognition of symptoms in patients on maintenance hemodialysis. *Clin J Am Soc Nephrol* 2007;2(5):960-7

8. Wang Y, Snoep JD, Hemmelder MH, et al. Outcomes after kidney transplantation, let's focus on the patients' perspectives. *Clin Kidney J* 2021;14(6):1504-13

9. Aiyegbusi OL, Kyte D, Cockwell P, et al. Measurement properties of patient-reported outcome measures (PROMs) used in adult patients with chronic kidney disease: A systematic review. *PLoS One* 2017;12(6):e0179733

10. van der Willik EM, Meuleman Y, Prantl K, et al. Patient-reported outcome measures: selection of a valid questionnaire for routine symptom assessment in patients with advanced chronic kidney disease - a four-phase mixed methods study. *BMC Nephrol* 2019;20(1):344

11. Verberne WR, Das-Gupta Z, Allegretti AS, et al. Development of an International Standard Set of Value-Based Outcome Measures for Patients With Chronic Kidney Disease: A Report of the International Consortium for Health Outcomes Measurement (ICHOM) CKD Working Group. *American Journal of Kidney Diseases* 2019;73(3):372-84

12. Terwee CB, Vonkeman HE, Zuidgeest M: Het menu van generieke PROMs Advies, The Netherlands, Linnean initiatief, werkgroep PROMs en PREMs, 2019

13. HealthMeasures. Intro to PROMIS®, Northwestern University, 2021. https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis

14. 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. *Journal of Clinical Epidemiology* 2010;63(11):1179-94

15. van Balen EC, Haverman L, Hassan S, et al. Validation of PROMIS Profile-29 in adults with hemophilia in the Netherlands. *J Thromb Haemost* 2021;19(11):2687-701

16. Rose AJ, Bayliss E, Huang W, et al. Evaluating the PROMIS-29 v2.0 for use among

older adults with multiple chronic conditions. *Quality of Life Research* 2018;27(11):2935-44

17. Katz P, Pedro S, Michaud K. Performance of the Patient-Reported Outcomes Measurement Information System 29-Item Profile in Rheumatoid Arthritis, Osteoarthritis, Fibromyalgia, and Systemic Lupus Erythematosus. *Arthritis Care & Research* 2017;69(9):1312-21

18. Tang E, Ekundayo O, Peipert JD, et al. Validation of the Patient-Reported Outcomes Measurement Information System (PROMIS)-57 and -29 item short forms among kidney transplant recipients. *Quality of Life Research* 2019;28(3):815-27

19. Selewski DT, Massengill SF, Troost JP, et al. Gaining the Patient Reported Outcomes Measurement Information System (PROMIS) perspective in chronic kidney disease: a Midwest Pediatric Nephrology Consortium study. *Pediatr Nephrol* 2014;29(12):2347-56

20. ERA-EDTA Registry Annual Report 2019, Amsterdam UMC, location AMC, Department of Medical Informatics, Amsterdam, the Netherlands, European Renal Association, 2021

21. Weisbord SD, Fried LF, Arnold RM, et al. Development of a symptom assessment instrument for chronic hemodialysis patients: the Dialysis Symptom Index. *J Pain Symptom Manage* 2004;27(3):226-40

22. Pakpour AH, Nourozi S, Molsted S, Harrison AP, Nourozi K, Fridlund B. Validity and reliability of short form-12 questionnaire in Iranian hemodialysis patients. *Iran J Kidney Dis* 2011;5(3):175-81

23. Loosman WL, Hoekstra T, van Dijk S, et al. Short-Form 12 or Short-Form 36 to measure quality-of-life changes in dialysis patients? *Nephrol Dial Transplant* 2015;30(7):1170-6

24. Østhus TB, Preljevic VT, Sandvik L, et al. Mortality and health-related quality of life in prevalent dialysis patients: Comparison between 12-items and 36-items short-form health survey. *Health Qual Life Outcomes* 2012;10:46

25. Terwee CB, Roorda LD, de Vet HC, et al. Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Qual Life Res* 2014;23(6):1733-41

26. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Quality of Life Research* 2021;30(10):2729-54

27. Dutch-Flemisch PROMIS. http://www.dutchflemishpromis.nl/

28. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34(3):220-33

29. Ware JE, Jr. SF-36 health survey update. *Spine (Phila Pa 1976)* 2000;25(24):3130-9

30. Abdel-Kader K, Unruh ML, Weisbord SD. Symptom burden, depression, and quality of life in chronic and end-stage kidney disease. *Clin J Am Soc Nephrol* 2009;4(6):1057-64

31. HealthMeasures. PROMIS® Score Cut Points, Northwestern University, 2021. https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points

32. van Muilekom MM, Luijten MAJ, van Oers HA, et al. From statistics to clinics: the visual feedback of PROMIS® CATs. *J Patient Rep Outcomes* 2021;5(1):55

33. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge university press: 2011;

34. Jabrayilov R, Emons WHM, Sijtsma K. Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Appl Psychol Meas* 2016;40(8):559-72

35. Hays RD, Spritzer KL, Reise SP. Using Item Response Theory to Identify Responders to Treatment: Examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) Physical Function Scale and Emotional Distress Composite. *Psychometrika* 2021;86(3):781-92

36. van der Willik EM, Terwee CB, Bos WJW, et al. Patient-reported outcome measures (PROMs): making sense of individual PROM scores and changes in PROM scores over time. *Nephrology (Carlton)* 2021;26(5):391-9

37. Vanden Wyngaert K, Van Craenenbroeck AH, Eloot S, et al. Associations between the measures of physical function, risk of falls and the quality of life in haemodialysis patients: a cross-sectional study. *BMC Nephrol* 2020;21(1):7

38. Sturgill DA, Bal N, Nagavally S, Wolfgram DF. The Relationship between Dialysis Metrics and Patient-Reported Cognition, Fatigue, and Physical Function. *Kidney Dis (Basel)* 2020;6(5):364-70

39. Gorodetskaya I, Zenios S, McCulloch CE, et al. Health-related quality of life and estimates of utility in chronic kidney disease. *Kidney International* 2005;68(6):2801-8

40. Erez G, Selman L, Murtagh FE. Measuring health-related quality of life in patients with conservatively managed stage 5 chronic kidney disease: limitations of the Medical Outcomes Study Short Form 36: SF-36. *Qual Life Res* 2016;25(11):2799-809

41. Terwee CB, Crins MHP, Roorda LD, et al. International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. *J Clin Epidemiol* 2021;134:1-13

42. Schalet BD, Lim S, Cella D, Choi SW. Linking Scores with Patient-Reported Health Outcome Instruments: a Validation Study and Comparison of Three Linking Methods. *Psychometrika* 2021;86(3):717-46

43. PROsetta Stone® Linking Patient-Reported Outcome Measures. https://www.prosettastone.org/

44. van Deursen AJAM: Digitale ongelijkheid in Nederland: Internetgebruik van mensen van 55 jaar en ouder., Enschede, Nederland, Universiteit Twente, 2019

45. James J. Confronting the scarcity of digital skills among the poor in developing countries. *Development Policy Review* 2021;39(2):324-39

46. Eurostat. Individuals' level of digital skills (from 2021 onwards) 2022. https://ec.europa.eu/eurostat/databrowser/view/ISOC_SK_DSKL_I21__custom_2397093/bookmark/table?lang=en&bookmarkId=dc481686-c938-4e07-b03c-8e039f532857. Accessed June 10, 2022

Supplemental Material for Chapter 3

Supplement A – Visual illustration of Computerized Adaptive Testing (CAT)

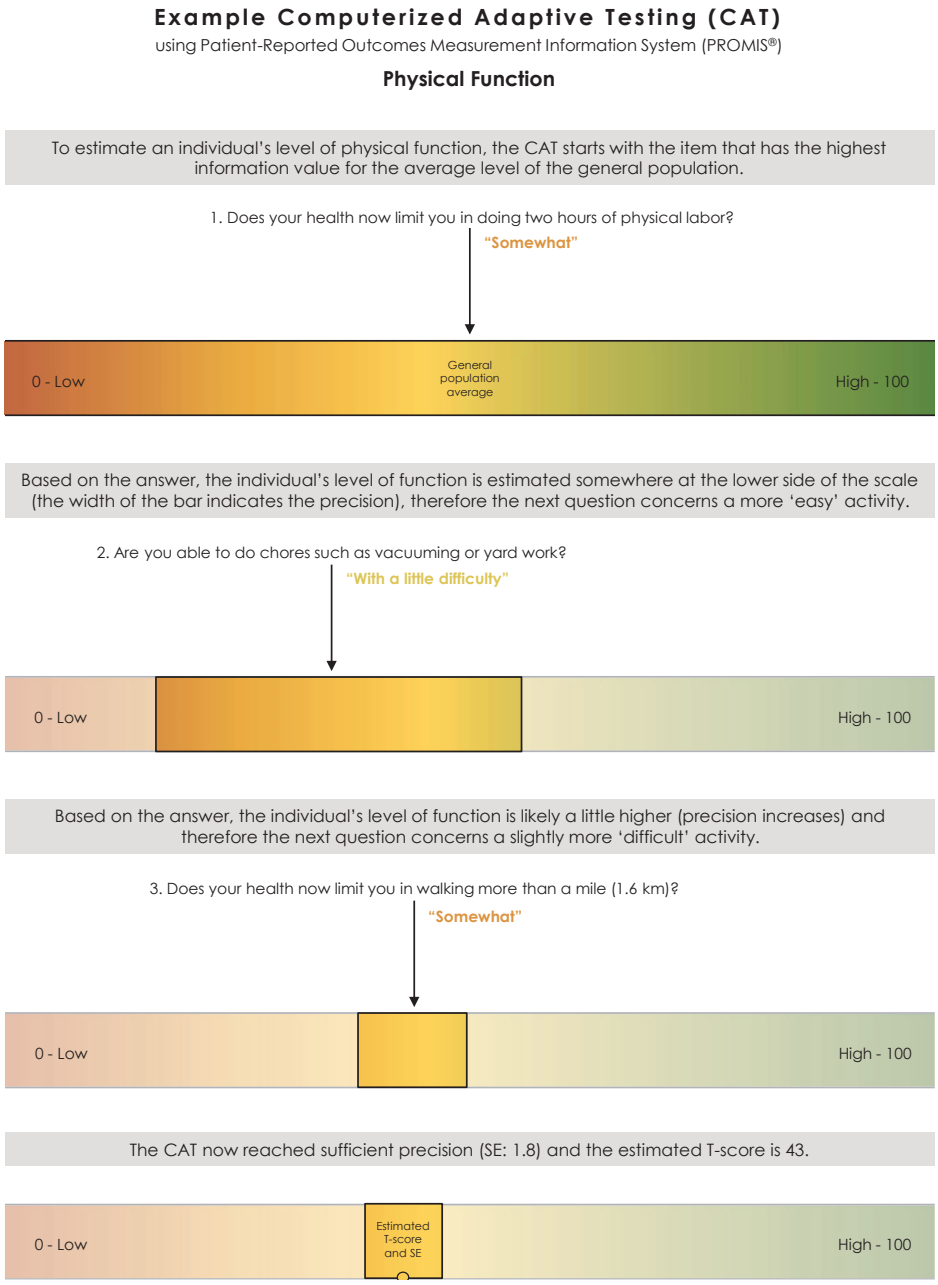

Figure S1. Visual illustration of Computerized Adaptive Testing (CAT) using Patient-Reported Outcomes Measurement Information System (PROMIS®) Physical Function

Supplement B – PROMIS CAT item characteristics

Table S1. Characteristics of items used in PROMIS CATs.

PROMIS CAT	Items used / total item bank	Top 3 items used	Items per patient
Physical Function	28 / 121	1. Does your health now limit you in doing two hours of physical labor? (n=205) 2. Are you able to do chores such as vacuuming or yard work? (n=122) 3. Does your health now limit you in walking more than a mile (1.6 km)? (n=119)	Median (IQR): 4 (3-6) Min-max: 3-12 12 items: n=8 (3.9%)
Pain Interference	24 / 40	1. How much did pain interfere with your day to day activities? (n=203) 2. How much did pain interfere with your ability to participate in social activities? (n=109) 3. How often was pain distressing to you? (n=95)	Median (IQR): 4 (2-12) Min-max: 2-12 12 items: n=75 (36.9%)
Fatigue	27 / 95	1. How often did you have to push yourself to get things done because of your fatigue? (n=203) 2. I have trouble starting things because I am tired. (n=180) 3. How exhausted were you on average? (n=33)	Median (IQR): 5 (4-6) Min-max: 4-12 12 items: n=6 (3.0%)
Sleep Disturbance	22 / 27	1. My sleep quality was… (n=203) 2. I had trouble sleeping. (n=203) 3. I had a problem with my sleep. (n=201)	Median (IQR): 10 (8-12) Min-max: 6-12 12 items: n=71 (35.0%)
Anxiety	21 / 29	1. I felt uneasy... (n=203) 2. I felt tense… (n=157) 3. I felt anxious and worried… (n=142)	Median (IQR): 7 (6-10) Min-max: 5-12 12 items: n=44 (21.7%)
Depression	21 / 28	1. I felt depressed. (n=204) 2. I felt unhappy. (n=144) 3. I felt discouraged about the future. (n=130)	Median (IQR): 8 (5-12) Min-max: 3-12 12 items: n=55 (27.0%)
Ability to Participate in Social Roles and Activities	23 / 35	1. I have trouble doing all of my regular leisure activities with others … (n=203) 2. I have trouble doing all of the activities with friends that I want to do... (n=132) 3. I have trouble doing all of the family activities that I want to do… (n=114)	Median (IQR): 5 (4-6) Min-max: 3-12 12 items: n=13 (6.4%)