ORIGINAL ARTICLE

Nephrol Dial Transplant (2023) 38: 1158–1169
https://doi.org/10.1093/ndt/gfac231
Advance Access publication date 1 August 2022

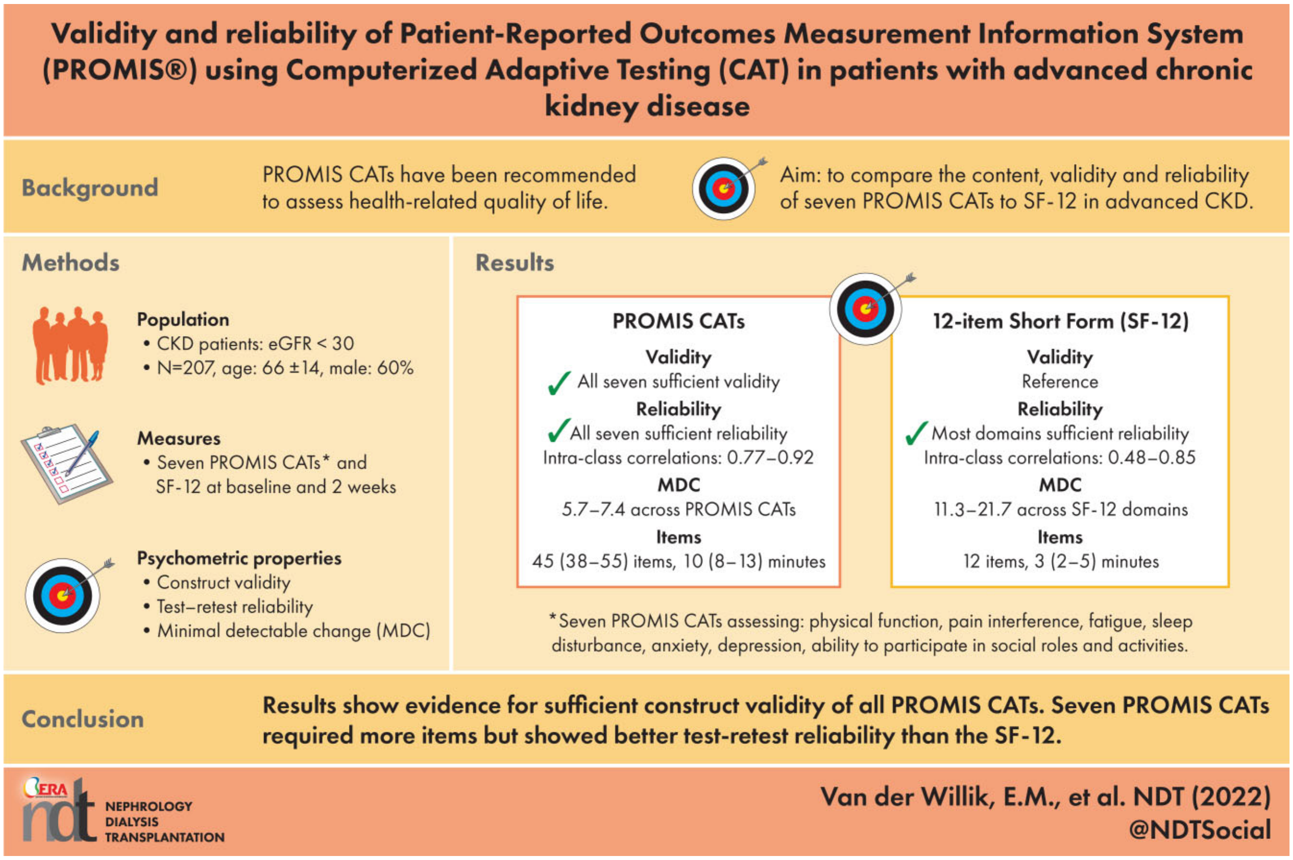
Validity and reliability of the Patient-Reported Outcomes Measurement Information System (PROMIS®) using computerized adaptive testing in patients with advanced chronic kidney disease

Esmee M. van der Willik[1,2], Fenna van Breda[3], Brigit C. van Jaarsveld[3], Marlon van de Putte[3], Isabelle W. Jetten[3], Friedo W. Dekker[2], Yvette Meuleman[2], Frans J. van Ittersum[3] and Caroline B. Terwee[1]

[1]Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health research institute, Amsterdam, The Netherlands , [2]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands and [3]Department of Nephrology, Amsterdam University Medical Centers, Amsterdam, The Netherlands

Correspondence to: Esmee M. van der Willik; E-mail: e.m.van_der_willik@lumc.nl

GRAPHICAL ABSTRACT



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

Background

PROMIS CATs have been recommended to assess health-related quality of life.

Aim: to compare the content, validity and reliability of seven PROMIS CATs to SF-12 in advanced CKD.

Methods

Population
- CKD patients: eGFR < 30
- N=207, age: 66 ±14, male: 60%

Measures
- Seven PROMIS CATs* and SF-12 at baseline and 2 weeks

Psychometric properties
- Construct validity
- Test–retest reliability
- Minimal detectable change (MDC)

Results

PROMIS CATs

Validity
✓ All seven sufficient validity
Reliability
✓ All seven sufficient reliability
Intra-class correlations: 0.77–0.92
MDC
5.7–7.4 across PROMIS CATs
Items
45 (38–55) items, 10 (8–13) minutes

12-item Short Form (SF-12)

Validity
Reference
Reliability
✓ Most domains sufficient reliability
Intra-class correlations: 0.48–0.85
MDC
11.3–21.7 across SF-12 domains
Items
12 items, 3 (2–5) minutes

*Seven PROMIS CATs assessing: physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, ability to participate in social roles and activities.

Conclusion

Results show evidence for sufficient construct validity of all PROMIS CATs. Seven PROMIS CATs required more items but showed better test-retest reliability than the SF-12.

NEPHROLOGY DIALYSIS TRANSPLANTATION

Van der Willik, E.M., et al. NDT (2022)
@NDTSocial

ABSTRACT

Background. The Patient-Reported Outcomes Measurement Information System (PROMIS®) has been recommended for computerized adaptive testing (CAT) of health-related quality of life. This study compared the content, validity, and reliability of seven PROMIS CATs to the 12-item Short-Form Health

KEY LEARNING POINTS

What is already known about this subject?

- Patient-reported outcome measures (PROMs) are increasingly being used in nephrology care, but there is no consensus on the preferred PROMs because of lack of knowledge of psychometric properties.
- The Patient-Reported Outcomes Measurement Information System (PROMIS®) using computerized adaptive testing (CAT) is one of the proposed PROMs to measure generic health-related quality of life (HRQOL) in patients with advanced chronic kidney disease (CKD).
- PROMIS CAT is a relatively novel measurement method in health care and has several advantages compared with traditional, nonadaptive PROMs, but it has not yet been validated in patients with CKD.

What this study adds?

- Psychometric properties of seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and the ability to participate in social roles and activities) are compared with the 12-item Short-Form Health Survey (SF-12), which is a validated and commonly used PROM and currently used in Dutch routine nephrology care.
- Results show evidence for sufficient construct validity and test-retest reliability of all PROMIS CATs.
- Seven PROMIS CATs required more items but showed better reliability than the SF-12.

What impact this may have on practice or policy?

- This study provides valuable information about the psychometric properties of seven PROMIS CATs compared with a commonly used PROM (i.e. the SF-12) to assess HRQOL in patients with CKD.
- Content comparison and reliability parameters, such as the minimal detectable change, are informative in the interpretation of PROM scores in routine nephrology care.
- Knowledge of validity and reliability can support considerations about which PROMs best fit routine nephrology care.

Survey (SF-12) in patients with advanced chronic kidney disease.

Methods. Adult patients with chronic kidney disease and an estimated glomerular filtration rate under 30 mL/min/1.73 m^2 who were not receiving dialysis treatment completed seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and the ability to participate in social roles and activities), the SF-12, and the PROMIS Pain Intensity single item and Dialysis Symptom Index at inclusion and 2 weeks. A content comparison was performed between PROMIS CATs and the SF-12. Construct validity of PROMIS CATs was assessed using Pearson's correlations. We assessed the test-retest reliability of all patient-reported outcome measures by calculating the intraclass correlation coefficient and minimal detectable change.

Results. In total, 207 patients participated in the study. A median of 45 items (10 minutes) were completed for PROMIS CATs. All PROMIS CATs showed evidence of sufficient construct validity. PROMIS CATs, most SF-12 domains and summary scores, and Dialysis Symptom Index showed sufficient test-retest reliability (intraclass correlation coefficient ≥ 0.70). PROMIS CATs had a lower minimal detectable change compared with the SF-12 (range, 5.7–7.4 compared with 11.3–21.7 across domains, respectively).

Conclusion. PROMIS CATs showed sufficient construct validity and test-retest reliability in patients with advanced chronic kidney disease. 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.

Keywords: chronic kidney disease (CKD), minimal detectable change, patient-reported outcome measures (PROMs), reliability, validity

INTRODUCTION

Patients with advanced chronic kidney disease (CKD) experience numerous physical and emotional disease-related symptoms that 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 health care professionals alike [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 doing so 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 tests (CATs). The use of CATs is relatively novel in health care and has several advantages compared with traditional fixed (i.e. nonadaptive) 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 that patient. In addition, on average, fewer questions will be required to obtain similar or even more precise measurements compared with 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 [MDC]) of seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and the ability to participate in social roles and activities) with the SF-12 in patients with advanced CKD. Additionally, we assessed the 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.

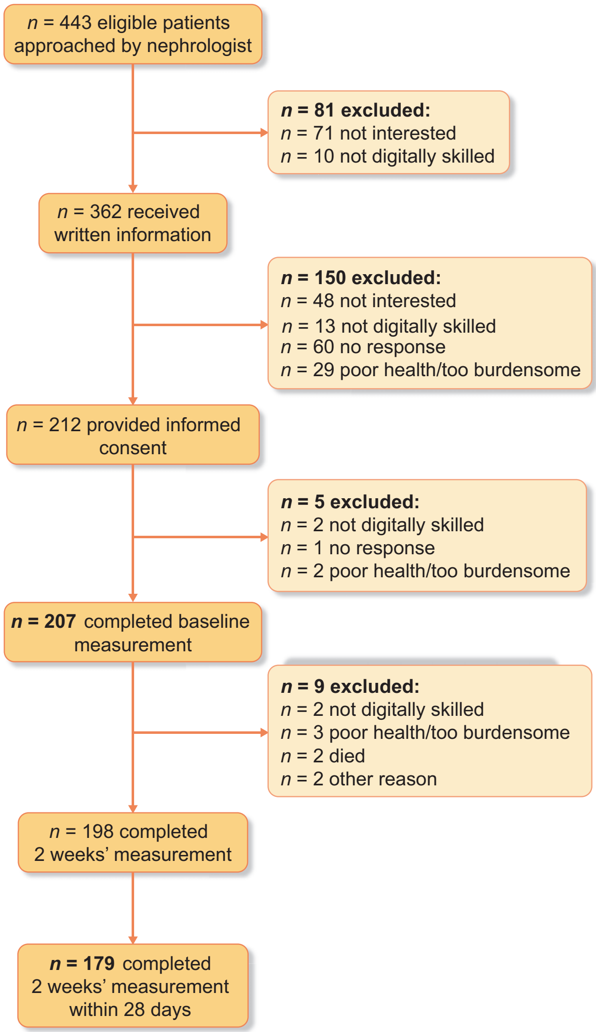

n = 443 eligible patients approached by nephrologist
***n* = 81 excluded:**
n = 71 not interested
n = 10 not digitally skilled
n = 362 received written information
***n* = 150 excluded:**
n = 48 not interested
n = 13 not digitally skilled
n = 60 no response
n = 29 poor health/too burdensome
n = 212 provided informed consent
***n* = 5 excluded:**
n = 2 not digitally skilled
n = 1 no response
n = 2 poor health/too burdensome
***n* = 207** completed baseline measurement
***n* = 9 excluded:**
n = 2 not digitally skilled
n = 3 poor health/too burdensome
n = 2 died
n = 2 other reason
n = 198 completed 2 weeks' measurement
***n* = 179** completed 2 weeks' measurement within 28 days

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

MATERIALS AND METHODS

Study design and population

This observational study included adult patients with advanced CKD and an estimated glomerular filtration rate (eGFR) less than 30 mL/min/1.73 m^2 not receiving dialysis treatment. Exclusion criteria were kidney replacement therapy (KRT—i.e. dialysis or kidney transplantation) planned within 4 weeks, rapid deterioration of kidney function (i.e. decrease in eGFR >20 mL/min/1.73 m^2 during the past 6 months), not able to complete PROMs because of cognitive impairment, poor knowledge of the Dutch language, and 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 in the Kwaliteit van Leven In Kaart (KLIK; www.hetklikt.nu) research platform at inclusion (i.e. baseline), after 2 weeks, and after 6 months. If necessary, two reminders were sent by email or patients were contacted by telephone. Patients without access to an electronic device with an internet connection could participate by telephone. This study used the baseline and 2-week measurements (Fig. 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 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 [CVD], lung disease, liver disease, and malignancy), as defined by ICHOM [11], eGFR (mL/min/1.73 m^2), KRT in medical history, start of KRT and death during follow-up were collected from medical records. Educational level and ethnocultural 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: version 1.2 Physical Function, version 1.1 Pain Interference, version 1.0 Fatigue, version 1.0 Sleep Disturbance, version 1.0 Anxiety, version 1.0 Depression, and version 2.0 Ability to Participate in Social Roles and Activities. All items have five response options, 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 greater than 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 having difficulties doing 2 hours of physical labor (first item). Then, the second item will be an 'easier' 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 5 miles) that he or she 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, patients must complete fewer questions. (See Supplement A for a visual illustration of a 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 an 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 had been administered. We used a lower SE compared with the standard stopping rule (i.e. SE, 3.0) [13] because a higher reliability may be preferable for routine care; by using this setting, the optimal performance of PROMIS CATs could be investigated. PROMIS CATs were administered using CAT software from the Dutch-Flemish Assessment Center, part of the Dutch-Flemish PROMIS National Center [27].

The SF-12, version 2 [28, 29] is a 12-item generic PROM that assesses 8 domains of HRQOL: physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health. Additionally, a physical component summary (PCS) score (including physical functioning, role-physical, bodily pain, and general health) and a mental component summary (MCS) 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 was used as reference, with an average score of 50 (SD, 10). Higher scores indicate a better HRQOL.

The PROMIS item version 1.0 Numerical Rating Scale Pain Intensity 1a is a single-item with a 0 to 10 scale, with higher scores indicating more pain.

The DSI [21] is a 30-item disease-specific PROM for assessing 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: (i) total number of symptoms (0–30 symptoms) and (ii) 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 because 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 the SF-12, we compared their content by providing (i) an overview of the PROM characteristics (e.g. domains, number of items, recall period, scoring, and interpretation) and (ii) 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

We assessed the construct validity of PROMIS CATs by using Pearson's correlations. Hypotheses were formulated *a priori* about the expected correlations between PROMIS CATs and the SF-12 and DSI based on literature [15–18] and expert judgement (E.vdW. and C.T.). 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% or more of the results were in accordance with the hypotheses.

Test-retest reliability

We assessed the test-retest reliability of PROMIS CATs, SF-12, PROMIS Pain Intensity single item, and DSI by calculating the intraclass correlation coefficient (ICC) in patients with valid baseline and 2-week measurements (Fig. 1). We calculated the ICC using a two-way random-effects model for

Table 1. Hypotheses for construct validity

PROMIS CAT	Strong correlation: Pearson's $r \geq 0.7$	Moderate correlation: Pearson's $r \geq 0.5$–0.7	No strong correlation: Pearson's $r \leq 0.6$
Physical Function	SF-12 physical functioning SF-12 PCS[a]	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 PCS	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)[b]		All other SF-12 domains DSI total number of symptoms and symptom burden score
Anxiety	SF-12 mental health SF-12 MCS[a] 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 MCS[a]		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

[a] SF-12 PCS includes the domains physical functioning, role-physical, bodily pain, and general health; SF-12 MCS includes the domains vitality, social functioning, role-emotional, and mental health.

[b] DSI Sleep problems were defined as trouble falling asleep and/or trouble staying asleep.

absolute agreement: $ICC\,agreement = \frac{\sigma_p^2}{\sigma_p^2+\sigma_m^2+\sigma_e^2}$, whereby σ_p^2 is the variation between patients, σ_m^2 is the variation between measurements, and σ_e^2 is random error variance. ICC ≥ 0.70 was considered sufficient [33].

We computed the ICC for each PROMIS CAT and SF-12 domain separately. Additionally, we calculated the ICC 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 health care, and insight into the reliability of these scores is therefore of clinical relevance.

The 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, which 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 following formula: $1.96 * \sqrt{2} * SEM$, whereby SEM was calculated as $\sqrt{\sigma_m^2 + \sigma_e^2}$.

The MDC, based on IRT, of each PROMIS CAT varies by 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 SE of the T-score at baseline and SE_2 at the 2-week measurement. A mean MDC of each PROMIS CAT was subsequently calculated for the whole group.

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

RESULTS

Study participants

Almost half of the patients 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-week measurement within 28 days and were eligible for reliability analyses (Fig. 1). The average time between the baseline and 2-week measurement was 14.1 (SD, 3.7) days. Eleven patients participated by telephone. Sociodemographic and clinical characteristics of the participants at baseline and 2-week measurements are shown in Table 2. The baseline and 2-week study samples were comparable. About 60% of patients were male, the mean (SD) age was 65.5 (13.8) years, and the majority (85%) had a Dutch ethnocultural background. Mean (SD) eGFR was 21.4 (6.7), and 17% of patients had had KRT in the past.

Content comparison

Table 3 shows the similarities and differences in characteristics of PROMIS CATs and the SF-12. Although assessing the same patient-relevant outcome (i.e. generic HRQOL), PROMIS CATs and the 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 patients with advanced CKD completed a median (interquartile range [IQR]) of 45 (38–55) items for all seven PROMIS CATs, which took them a median

Table 2. Characteristics of study sample at baseline and 2-week measurement

Characteristic	Study sample at baseline[a] ($n = 207$)	Study sample at 2 weeks[a] ($n = 179$)
Sex, male, *n* (%)	124 (59.9)	107 (59.8)
Age (years), mean (SD)	65.5 (13.8)	66.1 (13.1)
Ethnocultural group[b], Dutch, *n* (%)	176 (85.0)	152 (84.9)
Educational level[c], *n* (%)		
Low	85 (41.0)	74 (41.3)
Middle	49 (23.7)	43 (24.0)
High	73 (35.3)	62 (34.6)
Primary kidney disease, *n* (%)		
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), mean (SD)	21.4 (6.7)	21.6 (6.6)
KRT in medical history[d], yes, *n* (%)	35 (17.0)	30 (16.9)
BMI, mean (SD)	26.8 (5.2)	26.9 (5.2)
Smoking, *n* (%)		
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, *n* (%)		
Hypertension, yes	164 (79.2)	140 (78.2)
Diabetes mellitus, yes	62 (30.0)	53 (29.6)
CVD, 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)

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%).

[a] Study sample at baseline was used for validity analyses. Study sample at 2-week measurement was used for reliability analyses.

[b] Self-reported ethnocultural group: 'What ethnic group do you consider yourself to belong to?'

[c] Educational level according to International Standard Classification of Education levels 2011, classified as low (primary, lower secondary, or lower vocational education), middle (upper secondary or upper vocational education), and high (tertiary education [college/university]).

[d] 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 mL/min/1.73 m^2 and did not require dialysis treatment, in accordance with inclusion criteria.

(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 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 with SF-12 domains and summary scores. Overall, PROMIS CATs showed 'better' (toward the green area) HRQOL scores compared with 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 with 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 PCS [39.2 ± 10.7]).

Construct validity

All PROMIS CATs showed evidence of sufficient construct validity because 75% or more 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 of 15 hypotheses were met. For Fatigue and Ability to Participate in Social Roles and Activities, 13 of 15 correlations were met, and for Anxiety, 12 of 15 correlations were in accordance with the hypotheses.

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 PCS and MCS 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 (ICCs of 0.85 and 0.88, respectively).

Table 3. Content comparison of PROMIS CAT with SF-12[a]

	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*[b]: PCS MCS
Number of items	All PROMIS domains[c] 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*[b]: PCS 6 items MCS 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 (SD), 50 (10)	General US population: mean (SD), 50 (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 favorable 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 (IRT)	Classical Test Theory (CTT)
Completion options	Electronic only	Electronic or paper based
Time to complete[d]	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.

[a] The 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[43].

[b] SF-12 PCS includes the domains physical functioning, role-physical, bodily pain, and general health; SF-12 MCS includes the domains vitality, social functioning, role-emotional, and mental health.

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

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

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

Category	*n* (%)[a]	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
PCS[b]	204 (98.6)	39.2 (10.7)	11.1–61.4
MCS[b]	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)[c]	203 (98.1)	2.0 (1.6)	0–5
Sleep problems (0–10)[c,d]	203 (98.1)	2 (0–3)	0–10
Feeling anxious (0–5)[c]	203 (98.1)	0 (0–0)	0–5

[a] In total, four people did not finish the measurement and completed only part of the PROMs.

[b] SF-12 PCS includes the domains physical functioning, role-physical, bodily pain, and general health; SF-12 MCS includes the domains vitality, social functioning, role-emotional, and mental health.

[c] Prevalence of feeling tired or lack of energy: 70.0%; sleep problems: 52.7%; feeling anxious: 18.7%.

[d] Sleep problems were defined as trouble falling asleep and/or trouble staying asleep.

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 11.3 to 21.7, respectively, across domains.

DISCUSSION

This study examined the validity and reliability of seven PROMIS CATs compared with the SF-12 in patients with advanced CKD in the Netherlands. This study is the first to investigate 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 with the SF-12.

The observed average HRQOL scores are in line with scores that would be expected from existing literature in patients with CKD for both the PROMIS CATs [18, 36, 37] and the SF-12 [2, 3, 38, 39]. Comparison of the domain scores, however, revealed a slightly better HRQOL in patients with advanced CKD based on PROMIS CATs compared with the SF-12. This finding 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 [SD], 50 [10]; US reference population). This difference 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 [40], which is reflected in the smaller SDs for PROMIS CATs compared with SF-12. By means of linking [41], the scores of comparable PROMIS CAT and SF-12 domains could be converted into each other; this has been done for many other PROMs [42] and would be a valuable next step because it facilitates harmonization of data across studies or health care organizations (e.g. when both instruments are used across different health care specialties) and comparison with historical data if one changes from one PROM to the other [41].

All PROMIS CATs showed sufficient test-retest reliability, with better ICCs and small MDCs compared with 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 [43]. For PROMIS, the MIC has been estimated at 2 to 6 points [26], which is slightly smaller but close to the MDC of 6 to 7 points. Information about the MIC for SF-12 domains is limited, which makes it difficult to say to what extent the SF-12 can distinguish important changes from measurement error [43]. 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 result 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-

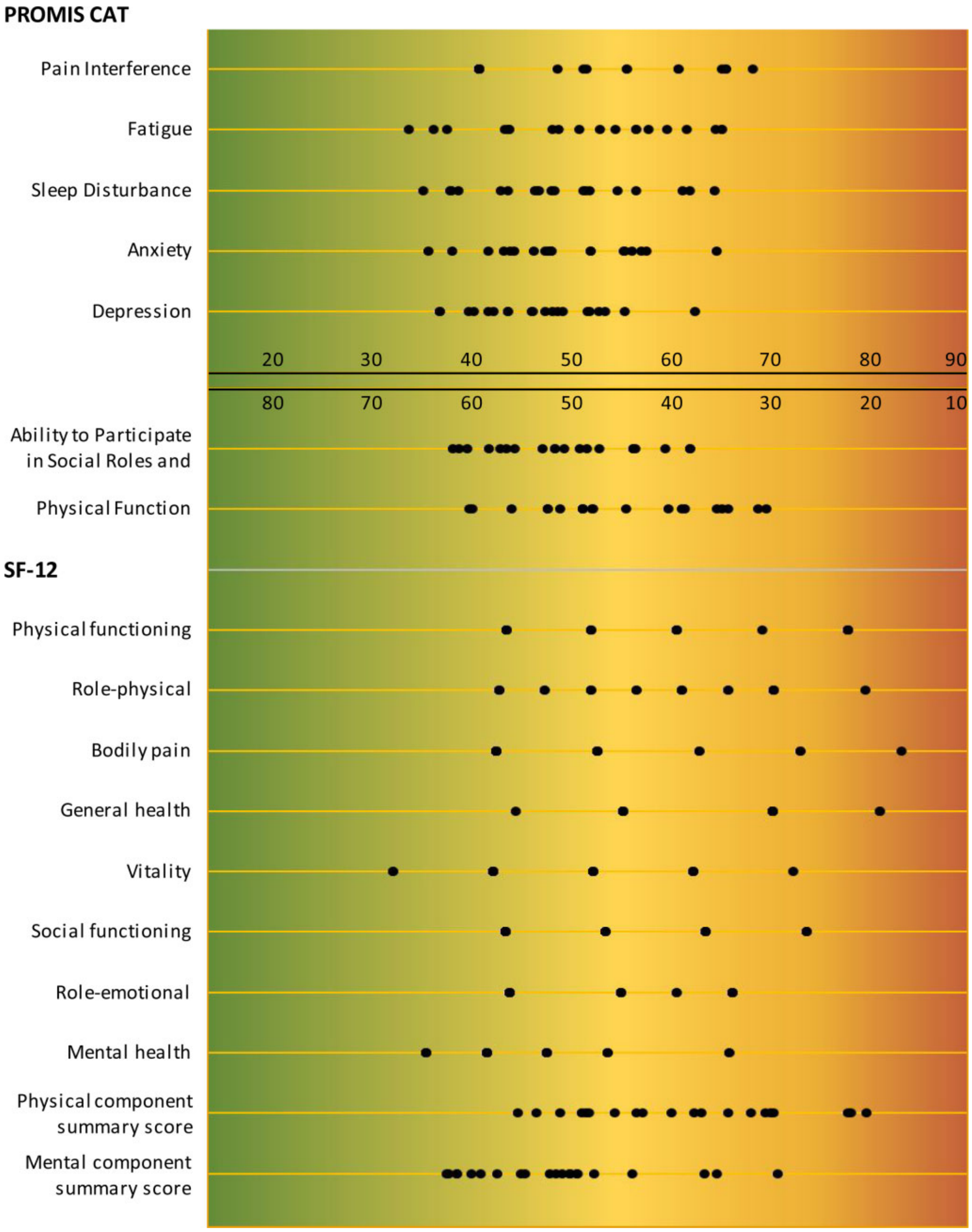


FIGURE 2: Score distributions of PROMIS CATs and SF-12 domains and summary scores. The figure's background color gives an indication of the interpretation of scores, ranging from good (green) to worse (red) HRQOL [31]. Note that the first five PROMIS CATs use a reverse scale compared with the other PROMIS CATs and the SF-12.

precision stopping rule is the relatively large number of 45 questions asked (i.e. 6–7 items per domain and 3–4 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 rules could be considered but with a detriment to precision. In this study, we applied a stopping rule with a smaller SE of 2.2 compared with the standard stopping rule (SE, 3.0) [13] to investigate the optimal performance of the PROMIS CATs. We expect that application of the standard stopping rule will result in 36 to 43 items in total (5–6 items per CAT), with a minimum of 28 items because the standard stopping rule requires 4 items per CAT and fewer than 45 items because of the higher SE compared with this study. Other alternative stopping rules to consider are 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 cutoff 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 optimal use of PROMIS CATs in routine nephrology care in close collaboration with patients and health care professionals.

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

		PROMIS CATs						
PROM	Domain	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
	PCS score[a]	**0.80**	*−0.63*	−0.55	−0.24	−0.13	−0.20	0.58
	MCS score[a]	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[b]	−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).

[a] SF-12 PCS includes the domains physical functioning, role-physical, bodily pain, and general health; SF-12 MCS includes the domains vitality, social functioning, role-emotional, and mental health.

[b] DSI sleep problems were defined as trouble falling asleep and/or trouble staying asleep.

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

Tool and domain	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
PCS score[a]	0.85 (0.81–0.89)	4.07	11.29
MCS score[a]	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

[a] SF-12 PCS includes the domains physical functioning, role-physical, bodily pain, and general health; SF-12 MCS includes the domains vitality, social functioning, role-emotional, and mental health. CI, confidence interval.

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 55 years of age or older is sufficiently digitally skilled [44], but in many countries—including European countries—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, other methods could be considered, as well, such as 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 the patient more likely considers relevant. As a result, the PROM may be perceived as less burdensome. In contrast, items may vary over time, meaning that progression of individual items cannot easily be monitored over time, which is in contrast to how the SF-12 (and DSI) is 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 its use in routine nephrology care.

Furthermore, it is important to mention that patients selected the SF-12 for use in routine nephrology care partly because of the low number of items. In addition, 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 the SF-12 and how they complement other PROMs, should thus be taken into account when considering which PROM fits routine nephrology care to measure HRQOL.

An important strength of this study is that the PROMIS CATs were compared with the PROM 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 health care, less focus is often paid to individual SF-12 domains. To expand on current findings, future research could investigate the validity of PROMIS CATs compared with the SF-36 [11, 29].

CONCLUSION

All seven PROMIS CATs (assessing physical function, pain interference, fatigue, sleep disturbance, anxiety, depression, and the 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 in routine nephrology care.

SUPPLEMENTARY DATA

Supplementary data are available at *ndt* online.

ACKNOWLEDGEMENTS

The authors are grateful to all the patients and health care professionals for their participation in and contributions to this study. The authors thank KLIK and the Dutch-Flemish PROMIS National Center for facilitating the collection of PROMs.

FUNDING

This study was supported by an unrestricted grant from the Dutch Kidney Foundation (18SWO03). The funding organization had no role in the study design; collection, analysis, and interpretation of the data; writing the report; and the decision to submit the report for publication.

AUTHORS' CONTRIBUTIONS

The authors E.vdW., F.vB., B.vJ., F.vI., and C.T. designed the study. Authors E.vdW. and C.T. conducted the data analysis and drafted the manuscript. Y.M., F.vI., and C.T. provided supervision and mentorship. All authors (E.vdW., F.vB., B.vJ., M.vdP., I.J., F.D., Y.M., F.vI., and C.T.) supported the interpretation of results, provided important intellectual content, and revised the final version of the manuscript. All authors provided final approval of the version to be published.

DATA AVAILABILITY STATEMENT

The data used for this research is available upon request. Contact information: Caroline B. Terwee, cb.terwee@amsterdamumc.nl.

CONFLICT OF INTEREST STATEMENT

There are no financial or other conflicts of interest to declare. The results presented in this article have not been published previously in whole or part, except in abstract format.

REFERENCES

1. Almutary H, Bonner A, Douglas C. Symptom burden in chronic kidney disease: a review of recent literature. *J Ren Care* 2013;**39**:140–50. https://doi.org.10.1111/j.1755-6686.2013.12022.x
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**:1707–15. https://doi.org/10.1093/ndt/gfy167
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**:1535–44. https://doi.org/10.1093/ckj/sfz192
4. Raj R, Ahuja KD, Frandsen M *et al.* Symptoms and their recognition in adult haemodialysis patients: interactions with quality of life. *Nephrology (Carlton)* 2017;**22**:228–33. https://doi.org/10.1111/nep.12754
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**:1813–21. https://doi.org/10.2215/CJN.01610214
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**:44–54. https://doi.org/10.1053/j.ajkd.2016.02.037
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**:960–7. https://doi.org/10.2215/CJN.00990207
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**:1504–13. https://doi.org/10.1093/ckj/sfab008
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**:e0179733. https://doi.org/10.1371/journal.pone.0179733
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**:344. https://doi.org.10.1186/s12882-019-1521-9

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. *Am J Kidney Dis* 2019;**73**:372–84. https://doi.org/10.1053/j.ajkd.2018.10.007
12. Terwee CB, Vonkeman HE, Zuidgeest M. *Het Menu van Generieke PROMs: Advies.* https://www.linnean.nl/inspiratie/bibliotheek/handlerdownloadfiles.ashx?idnv=1501426 (30 November 2021, date last accessed).
13. HealthMeasures. *Intro to PROMIS®*. https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis (22 November 2021, date last accessed).
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. *J Clin Epidemiol* 2010;**63**:1179–94. https://doi.org/10.1016/j.jclinepi.2010.04.011
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**:2687–701. https://doi.org/10.1111/jth.15454
16. Rose AJ, Bayliss E, Huang W *et al.* Evaluating the PROMIS-29 v2.0 for use among older adults with multiple chronic conditions. *Qual Life Res* 2018;**27**:2935–44. https://doi.org/10.1007/s11136-018-1958-5
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 Res (Hoboken)* 2017;**69**:1312–21. https://doi.org/10.1002/acr.23183
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. *Qual Life Res* 2019;**28**:815–27. https://doi.org/10.1007/s11136-018-2058-2
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**:2347–56. https://doi.org/10.1007/s00467-014-2858-8
20. European Renal Association—European Dialysis and Transplant Association. *Annual report 2019*. Department of Medical Informatics, Amsterdam UMC, Amsterdam, the Netherlands, 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**:226–40. https://doi.org/10.1016/j.jpainsymman.2003.07.004
22. Pakpour AH, Nourozi S, Molsted S *et al.* Validity and reliability of Short Form-12 questionnaire in Iranian hemodialysis patients. *Iran J Kidney Dis* 2011;**5**: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**:1170–6. https://doi.org/10.1093/ndt/gfv066
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. https://doi.org/10.1186/1477-7525-10-46
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**:1733–41. https://doi.org/10.1007/s11136-013-0611-6
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. *Qual Life Res* 2021;**30**:2729–54. https://doi.org/10.1007/s11136-021-02925-y
27. Dutch-Flemish PROMIS. *Het Dutch–Flemish PROMIS National Center*. http://www.dutchflemishpromis.nl/ (22 November 2021, date last accessed).
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**:220–33. https://doi.org/10.1097/000005650-199603000-0003
29. Ware JE Jr SF-36 Health Survey update. *Spine* 2000;**25**: 3130–9. https://doi.org.10.1097/00007632-200012150-00008
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**:1057–64. https://doi.org/10.2215/CJN.00430109
31. HealthMeasures. *PROMIS® score cut points.* https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points (22 November 2021, date last accessed).
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**:55. https://doi.org/10.1186/s41687-021-00324-y
33. De Vet HC, Terwee CB, Mokkink LB *et al. Measurement in Medicine: A Practical Guide.* Cambridge: 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**:559–72. https://doi.org/10.1177/0146621616664046
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**:781–92. https://doi.org/10.1007/s11336-021-09774-1
36. 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**:7. https://doi.org/10.1186/s12882-019-1671-9
37. Sturgill DA, Bal N, Nagavally S *et al.* The relationship between dialysis metrics and patient-reported cognition, fatigue, and physical function. *Kidney Dis (Basel)* 2020;**6**:364–70. https://doi.org/10.1159/000508919
38. Gorodetskaya I, Zenios S, McCulloch CE *et al.* Health-related quality of life and estimates of utility in chronic kidney disease. *Kidney Int* 2005;**68**:2801–8. https://doi.org/10.1111/j.1523-1755.2005.00752.x
39. 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**:2799–809. https://doi.org/10.1007/s11136-016-1313-7
40. 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. https://doi.org/10.1016/j.jclinepi.2021.01.011
41. Schalet BD, Lim S, Cella D *et al.* Linking scores with patient-reported health outcome instruments: a validation study and comparison of three linking methods. *Psychometrika* 2021;**86**:717–46. https://doi.org/10.1007/s11336-021-09776-z
42. PROsetta Stone®. *What is PROsetta Stone*? https://www.prosettastone.org/ (22 November 2021, date last accessed).
43. 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**:391–9. https://doi.org/10.1111/nep.13843
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 Rev* 2021;**39**:324–39. https://doi.org/10.1111/dpr.12479
46. Eurostat. Individuals' level of digital skills (from 2021 onwards). https://ec.europa.eu/eurostat/databrowser/view/ISOC_SK_DSKL_I21__custom_2397093/bookmark/table?lang=en&bookmarkId=dc481686-c938-4e07-b03c-8e039f532857 (June 10, 2022, date last accessed).

Received: 7.4.2022; Editorial decision: 13.7.2022