Measurement properties and interpretability of the PROMIS item banks in stroke patients: a systematic review

Oosterveer, D.M.; Arwert, H.; Terwee, C.B.; Schoones, J.W.; Vlieland, T.P.M.V.

Quality of Life Research (2022) 31:3305–3315
https://doi.org/10.1007/s11136-022-03149-4

REVIEW

Measurement properties and interpretability of the PROMIS item banks in stroke patients: a systematic review

Daniëlla M. Oosterveer[1] · Henk Arwert[1,2] · Caroline B. Terwee[3,4] · Jan W. Schoones[5] · Thea P. M. Vliet Vlieland[1,6]

Accepted: 22 April 2022 / Published online: 14 May 2022

Abstract

Purpose Both the International Consortium for Health Outcomes Measurement and the National Institutes of Health recommend the use of Patient-Reported Outcomes Measurement Information System (PROMIS®) measures in clinical care and research for stroke patients. This study aimed to systematically review the literature on the measurement properties and interpretability of PROMIS measures in stroke patients.

Methods Nine databases were searched from January 1st, 2007 till April 12th, 2021 for studies concerning the measurement properties and interpretability of PROMIS measures in stroke patients. The findings of these studies were analyzed according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline for systematic reviews of Patient-Reported Outcome Measures (PROMs).

Results Ten studies were included. The PROMIS Global Health was studied the most: its two subscales had sufficient structural validity in one study of very good quality, sufficient construct validity with > 75% of hypotheses tested confirmed (high GRADE rating), sufficient internal consistency, i.e. $\alpha \geq 0.70$ in two studies (high GRADE rating), sufficient reliability, i.e. ICC ≥ 0.70 in one study of doubtful quality, and indeterminate responsiveness in one study of inadequate quality. For other PROMIS measures, the measurement properties and interpretability were limitedly studied.

Conclusion The PROMIS Global Health showed sufficient structural and construct validity and internal consistency in stroke patients. There is a need for further research on content validity, structural validity, and measurement invariance of PROMIS measures in stroke patients.

Trial Registration Information: CRD42020203044 (PROSPERO).

Keywords Stroke · PROMIS · Review · Measurement properties · Interpretability · Psychometrics · Patient-reported outcome measures

✉ Daniëlla M. Oosterveer
d.oosterveer@basaltrevalidatie.nl

1 Basalt Rehabilitation, Wassenaarseweg 501, 2333 AL Leiden, The Netherlands

2 Department of Rehabilitation, Haaglanden Medical Center, The Hague, The Netherlands

3 Epidemiology and Data Science, Amsterdam University Medical Center Location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

4 Methology, Amsterdam Public Health, Amsterdam, The Netherlands

5 Directorate of Research Policy (Formerly: Walaeus Library), Leiden University Medical Center, Leiden, The Netherlands

6 Department of Orthopaedics, Rehabilitation and Physical Therapy, Leiden University Medical Center, Leiden, The Netherlands

Introduction

Patient-reported outcome measures (PROMs) are nowadays acknowledged as essential tools to evaluate health-related quality of life (HR-QoL) and other outcomes in stroke patients for both clinical care and research. Their value lies in the fact that PROMs reflect the patient's opinion with no external influence or interpretation by others [1]. The selection of a PROM is complex: it involves not only defining the construct of interest (i.e. the 'subject' you want to measure), but also the consideration of the burden for patients, its costs, and its measurement properties and interpretability (i.e. this is not a measurement property, but the degree to which one can assign qualitative meaning to scores or change in scores) [2]. Currently, a large number of different PROMs are used for the evaluation of stroke care and in

research, even measuring the same construct [2]. This variation hampers the comparison of outcomes of stroke care among institutions and of research results.

To overcome this problem, the International Consortium for Health Outcomes Measurement (ICHOM) has introduced a Standard Set for Stroke in 2015 [3]. This set includes a relatively new measure for HR-QoL, namely the Patient-Reported Outcomes Measurement Information System (PROMIS) Global Health. The inclusion of this PROMIS measure is in line with the overall recommendation of the National Institutes of Health (NIH) to use PROMIS measures in clinical care and in their funded research to enhance and standardize the use of PROMs in clinical care and research [4].[1]

In a previous review by Arwert et al. [5] it was shown that PROMIS measures are not yet widely used in stroke research, despite the abovementioned recommendations of the ICHOM and the NIH [3, 4]. A possible barrier for the use of PROMIS measures may be the lack of knowledge of the measurement properties and interpretability of these measures in stroke populations. The aim of our study was therefore to systematically review the literature on the measurement properties and interpretability of all PROMIS measures in stroke patients. This will assist clinicians and researchers in using these PROMIS measures in these patients.

Methods

Design

This study comprises a systematic review to summarize, evaluate, and compare the current literature on the measurement properties and interpretability of all PROMIS measures in stroke patients. The study protocol was registered in PROSPERO (https://www.crd.york.ac.uk/prospero/; CRD42020203044) and was based on the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline for systematic reviews of Patient-Reported Outcome Measures (PROMs). [6]

[1] PROMIS measures assess physical, mental, and social aspects of health and are available both as computer-adaptive tests (CAT) and as traditional "paper and pencil" instruments (called short forms or scales). Raw scores of each PROMIS measure are converted to an item response theory (IRT)-based T score. A T score of 50 is the average for the USA general population with a standard deviation (SD) of 10.[4] PROMIS short forms and scales are freely available in different languages at the PROMIS website (www.healthmeasures.net).

Step 1: search

PubMed, MEDLINE (OVID version), Embase (OVID version), Emcare, PsycINFO (EbscoHOST version), Google Scholar, Academic Search Premier, Web of Science, and Cochrane Library were searched for studies published between January 1st, 2007 and August 14th, 2020. This search was updated to include studies published till April 12th, 2021. In cooperation with a trained librarian (JS), a detailed search strategy was composed. The query consisted of the combination of the following two concepts: (1) Stroke and (2) PROMIS. Filters were used to exclude meeting abstracts and the results were limited to the following languages: English, Dutch, French, and German. Full details of the search strategy can be found in Appendix 1.

Step 2: selection of studies

Prior to the search, the inclusion criteria for this review were defined as follows: (1) patients were diagnosed with ischemic or hemorrhagic stroke and were 18 years or older; (2) a PROMIS measure should be used; (3) the study was published in English, Dutch, French, or German; and (4) the aim of the study was the evaluation of one or more measurement properties or interpretability of one or more PROMIS measures.

Studies were excluded when (1) only patients with subarachnoid hemorrhages were studied as these patients have a distinct clinical course; (2) no separate information on stroke patients was provided in case patients with various medical conditions were included; and (3) the study was a meeting abstract, protocol description, a letter to the editor, or a review. References of the reviews were screened for additional studies.

Two reviewers (HA and DO) independently screened all titles and abstracts using Rayyan (2016). The full-text papers of relevant studies were retrieved based on abovementioned criteria. If a study seemed relevant by at least one reviewer based on the title and abstract or in case of doubt, the full-text paper was retrieved. These full-text studies were read and included independently by the two reviewers. In case of disagreement on the final selection the two reviewers discussed their views. If agreement was not reached, a third reviewer (TV) was consulted.

Step 3: data extraction

The two reviewers independently extracted data of each selected study regarding: (1) the PROMIS measure used; (2) patient characteristics, i.e., number (N), age, and percentage (%) of female participants; (3) stroke characteristics, i.e., type of stroke, severity; (4) details of instrument

administration, i.e., population (general, hospital, or rehabilitation population; outpatients or inpatients), country, and timing after stroke; (5) the measurement properties that were evaluated; and (6) interpretability.

Step 4: sorting of the results and the application of criteria for good measurement properties

Measurement properties and interpretability results were sorted according to the COSMIN taxonomy. The results of each study on a measurement property were rated against the criteria for good measurement properties as 'sufficient,' 'insufficient,' or 'indeterminate' [6]. The criteria for good measurement properties are described below.

Validity – Validity is defined by COSMIN as 'the degree to which an instrument truly measures the construct(s) it purports to measure' [9]. In the COSMIN taxonomy three types of validity are distinguished:

(1) Content validity (i.e., does the content of the PROMIS measure corresponds with the construct one intents to measure in terms of relevance, comprehensiveness, and comprehensibility). This was considered sufficient if the items were relevant, comprehensive, and comprehensible with respect to the construct and stroke populations. Content validity is important: if there is high-quality evidence that the content validity of a PROMIS measure is insufficient in a stroke population or other population, the measure should not be recommended. There is a separate COSMIN manual available to evaluate this measurement property. [10]
(2) Criterion validity (i.e., how well the PROMIS measure agrees with the scores on a gold standard). This was considered sufficient when the correlation with a gold standard was ≥ 0.70 or when the area under the receiver operator characteristic curve (AUC) was ≥ 0.70.
(3) Construct validity (i.e., are the PROMIS scores consistent with hypotheses, e.g., with regard to internal relationships, relations with scores of other instruments, or differences between relevant groups). This can be further categorized into structural validity, hypotheses testing, and cross-cultural validity. Structural validity was considered sufficient when the expected dimensionality of a scale was demonstrated in a confirmatory factor analysis criteria (Comparative fit index (CFI) or Tucker–Lewis index or comparable measure > 0.95 OR Root Mean Square Error of Approximation (RMSEA) < 0.06 OR Standardized Root Mean Residuals (SRMR) < 0.08) or an IRT or in a Rasch model there was no violation of unidimensionality, local independence, and monotricity, and there was a good model fit [6]. Hypotheses testing was considered sufficient when at least 75% of the results were in accordance with the predefined hypotheses [6]. The hypotheses of the original studies were used; only when none was described explicitly, we have defined hypotheses based on the methods used in that study. Cross-cultural validity was considered sufficient when no important differences were found in the probability of giving a certain answer to items between relevant groups with similar levels of the studied construct. This is the case when no important differences are seen between group factors in multiple group factor analysis or no important differential item functioning (DIF) for group factors (McFadden's $R^2 < 0.02$). [6]

Reliability—Reliability is defined by COSMIN as 'the degree to which the measurement is free from measurement error' and can be subdivided into

(1) Internal consistency. This was considered sufficient when there was evidence for sufficient structural validity (in a stroke population or when not available in another population) and Cronbach's alpha was ≥ 0.70 for each unidimensional scale or subscale or when mean of the Standard Error of T-score was ≤ 3.3 using IRT analyses. [6, 11]
(2) Reliability. This was considered sufficient when the intraclass correlation coefficient or weighted Kappa was ≥ 0.70. [6]
(3) Measurement error. This was considered sufficient when the smallest detectable change (SDC, i.e., the minimal change that can be distinguished from measurement error in an individual patient with 95% confidence) or the limits of agreement were smaller than the minimal important change (MIC, i.e., the smallest change in score that patients perceive as important). [6]

Responsiveness—Responsiveness is defined by COSMIN as 'the ability of an instrument to detect change over time in the construct to be measured' [9]. It can be seen as an aspect of validity, but referring to the validity of a change score instead of a single score. Responsiveness was considered sufficient when at least 75% of the results were in accordance with a priori-formulated hypotheses or if the AUC > 0.70 [6]. The hypotheses of the original studies were used and only when none was described explicitly, we have defined hypotheses based on the methods used in that study.

Interpretability—Interpretability is defined by COSMIN as 'the degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores' [9]. It is not considered a measurement property, but an important aspect in the selection of an instrument [6]. The distribution of scores in the study population, floor and ceiling effects, clinically relevant differences in scores

between subgroups, and the MIC or minimal important difference (MID) were extracted from each study, if reported. When > 15% of patients scored 0 or 100 on a PROMIS measure, a floor or ceiling effect was considered present [12].

Step 5: rating of the methodological quality of each study

For each measurement property separately, the methodological quality of each study was rated according to the four-point rating system of the COSMIN Risk of Bias checklist [7]. For each measurement property COSMIN has provided several standards with design requirements and preferred statistical methods on measurement properties. Each standard of this checklist was rated as 'very good,' 'adequate,' 'doubtful,' or 'inadequate.' The lowest rating of all standards per measurement property was used to rate the overall methodological quality of a particular study for a specific measurement property. This reflects to whether the results of that specific measurement property of that study are trustworthy.

Step 6: summarizing the results for each measurement property

First, the consistency of the results of step 4 (i.e., whether or not the findings of each study fulfilled the criteria for good measurement properties) across studies was determined for each measurement property when two or more studies were available: overall, the measurement property of a measure was rated as 'sufficient,' 'insufficient,' 'inconsistent,' or 'indeterminate.' If all results of a measurement property were consistent, the results were summarized and overall rating was given: 'sufficient' or 'insufficient.' If the results were inconsistent, the results were not summarized and step 7 was not taken. [6]

Step 7: grading the quality of the evidence

Finally, the quality of the evidence was graded when two or more studies assessing a measurement property were available, using a modified Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [8]: the quality of the evidence was graded by taking into account the risk of bias (i.e., the methodological quality of the studies), inconsistency (i.e., unexplained inconsistency of results across studies), imprecision (i.e., total sample size of the available studies), and indirectness (i.e., evidence from different populations than the population of interest in the review). Because of lack of registry of studies on measurement properties, publication bias could not be taken into account [6]. The grading was done by two reviewers (DO en HA) independently. A high-quality level was defined as 'We are very confident that the true measurement property lies close to that of the (pooled) estimate of the measurement property,' moderate as 'We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different,' low as 'Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property,' and very low-quality level as 'We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property.' [6]

Results

Search

The search resulted in 174 studies, of which 50 were selected for full-text review after screening the titles and abstracts. After reading the full-text studies, ten studies were included. Backward reference tracking of the 15 reviews in the original yield of the search did not result in additional studies. A flowchart of the study selection is depicted in Fig. 1.

Characteristics of the study populations in the ten included studies are shown in Table 1. All study populations were hospital-based and concerned outpatients. The median stroke severity was mostly mild (NIHSS 0 and mRS 1–2) if reported [13–17]. Four studies reported on the same study population [18–21]. All PROMIS measures used English versions with the exception of two studies [16, 22] using the Dutch version of the PROMIS Global Health.

Measurement properties

A summary of the measurement properties is given in Table 2. Content validity was not evaluated for any of the PROMIS measures. Because there was no high-quality evidence that the content validity of a PROMIS measure was insufficient, we continued the review process. In addition, none of the studies valuated criterion validity and cross-cultural validity/measurement invariance.

In eAppendix 2 the hypotheses for construct validity and responsiveness are shown. There was only one study in which hypotheses for responsiveness were not predefined and clearly formulated [17]. Details concerning the rating of the quality of each study for each measurement property are given in eAppendix 3.

The PROMIS Global Health subscales showed sufficient structural validity with SRMR of 0.043 in one study of very good quality [17], sufficient construct validity measured with hypotheses testing, i.e., > 75% of hypotheses was confirmed, with a high GRADE rating [16, 17, 22],

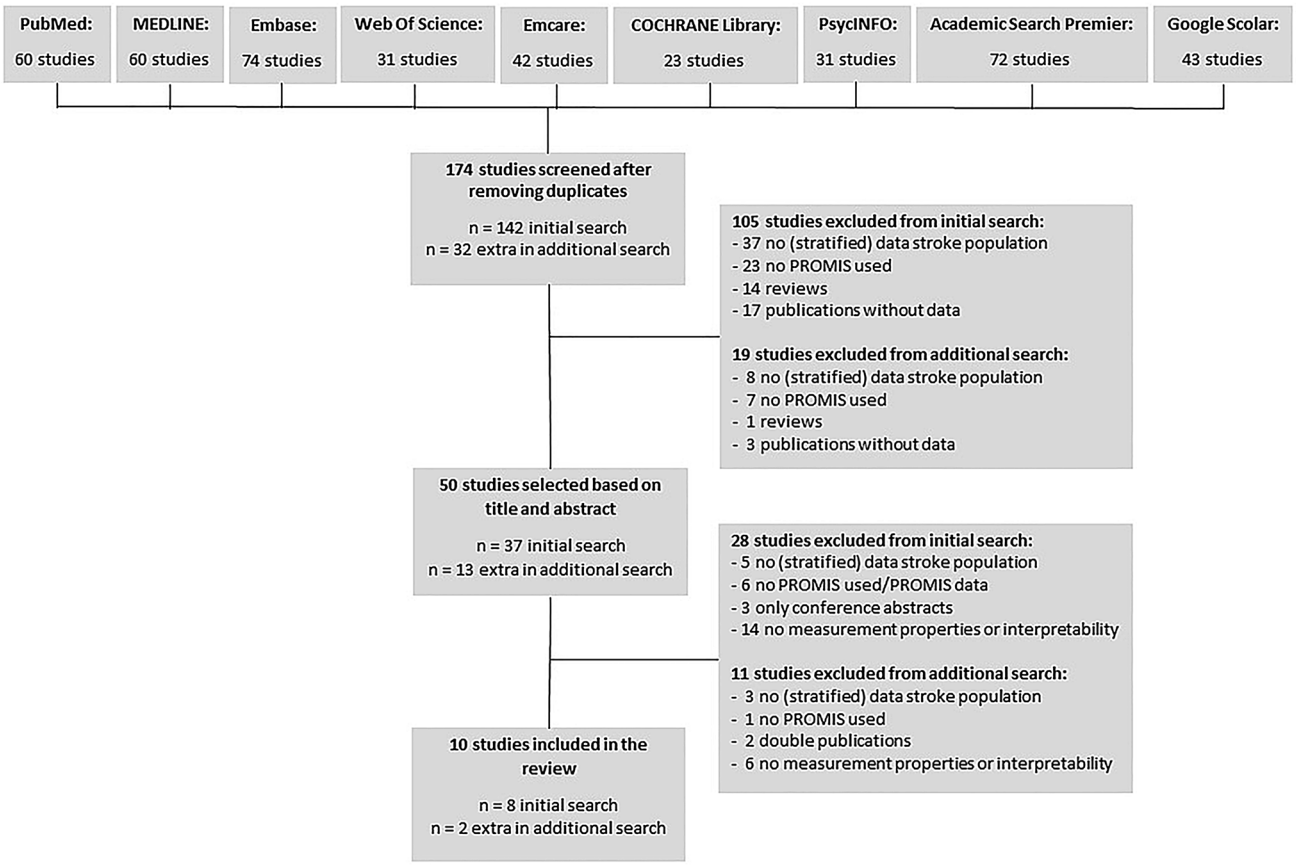

Fig. 1 Flowchart of the study selection

sufficient internal consistency, i.e., $\alpha \geq 0.70$ in two studies, with a high GRADE rating [16, 17], sufficient reliability, i.e., ICC ≥ 0.70, in one study of doubtful quality [17], and indeterminate responsiveness in one study of inadequate quality. [17].

Structural validity of other PROMIS measures was not assessed in any of the studies and reliability was only assessed in one study: a test–retest correlation of 0.82 for the PROMIS Physical Function v1.0 CAT was found in 36 stroke patients [14]. In addition, this study demonstrated that this PROMIS measure had a sufficient construct validity in stroke patients, as all hypotheses were confirmed [14]. In contrast, one hypothesis formulated for responsiveness of the PROMIS Physical Function v1.0 CAT was not confirmed in 739 stroke patients. [15].

One study reported on measurement error of several PROMIS measures (i.e., Physical functioning v1.0, Satisfaction with social roles, Fatigue, and Anxiety): for all measures the SDC was found to be larger than the MIC. [13].

Responsiveness of four Pain Interference and four Depression short forms in stroke patients was found low, i.e., AUC < 0.70; however these results were found in studies of doubtful quality. [18, 21].

Interpretability

Seven studies reported on interpretability of PROMIS measures (Table3) [13–15, 18–20, 22]. Floor and ceiling effects of PROMIS item banks were low (< 15%) for the PROMIS Physical Function CAT and Fatigue CAT [14]. Katzan et al. [14, 15] reported that the PROMIS Physical Function CAT had a lower ceiling effect (0.68–1.3%), than the Stroke Impact Scale-16 (15.4–19.6%), EuroQoL 5Dimensions (EQ5D, 17.2–23.3%), Patient Health Questionnaire-9 (14.0–19.9%), NIHSS (52.9–55.8%), and modified Rankin Scale (20.9–21.8%). De Graaf et al. [22] reported that the PROMIS Global Health had no floor or ceiling effects as opposed to the EQ5D. For the PROMIS Pain Interference 31–36% of patients had the lowest score possible. [18].

MIC values based on an anchor question varied between 2.4 for PROMIS Fatigue v1.0 and 6.2 for PROMIS Satisfaction for social roles v1.0. [13].

The percentage of missing items and total scores were reported in only two studies. Lam et al. [16] reported missing values in 3 (8.1%) out of 37 patients who completed the PROMIS Global Health in the paper-and-pencil group, and Katzan et al. [17] mentioned that the pain item of the PROMIS

Table 1 Characteristics of the included study populations

Reference	PROMIS	Population			Stroke characteristics		Instrument administration		
		N	Age mean (SD, range) yr	Sex % female	% ischemic	Severity median (IQR)	Population	Country	Timing after stroke
Lapin et al. [13]	Physical Function v1.0 Satisfaction with Social Roles v1.0; Fatigue v1.0; Anxiety v1.0	337	61 ± 14 SD	56	89	NIHSS 0 (0–2); mRS 1 (1–2)	Hospital outpatients	USA	< 1 (T1) and 6 (T2) months
Katzan et al. [14]	Physical Function v1.0 CAT	1946	63 ± 14 SD	46	100	NIHSS 0 (0–2); mRS 1 (1–2)	Hospital outpatients	USA	Median 75 days (IQR 35–277)
Katzan et al. [15]	Physical Function v1.0 CAT; Fatigue v1.0 CAT	3283	64 ± 14 SD		100	NIHSS 0 (0–2); mRS 2 (1–3)	Hospital outpatients	USA	Median 58 days (IQR 32–258)
Katzan et al. [17]	Global Health	1102	61 ± 15 SD	46	87	NIHSS 0 (0–1); mRS 1 (0–2)	Hospital outpatients	USA	Median 134 days (IQR 47–540)
Lam et al. [16]	Global Health	75	69 ± 11 SD	32	100	No neurological deficit after 3 days	Hospital outpatients	Netherlands	$Mean_{paper}$ 375 days (SD 60); $Mean_{telephone}$ 376 (SD 31)
De Graaf et al. [22]	Global Health SF 10	360	71 (17)*	40	93	NIHSS 3 (3)	Hospital outpatients	Netherlands	3 months
Chen et al. [18–19]	Pain interference SF 4a; Pain interference SF 6a; Pain interference SF 6b; Pain interference SF 8a	258	62 ± 11 SD	19			Hospital outpatients	USA	Baseline and 3 months
Kroenke et al. [20–21]	Depression SF 4a; Depression SF 6a; Depression SF 8a; Depression SF 8b	258	62 ± 11 SD	19			Hospital outpatients	USA	Baseline and 3 months

*Median (IQR)

CAT computer-adaptive testing; IQR interquartile range; mRS modified Rankin Scale; NIHSS National Institutes of Health Stroke Scale; SD standard deviation; SF short form; USA the United States of America; yr years; v version

Table 2 Summary of measurement properties of PROMIS measures

PROMIS	Validity							
	Structural validity				Hypothesis testing for construct validity			
	n	Result	Meth qual study	Overall result (GRADE)	*n*	Result	Meth qual study	Overall result (GRADE)
Global Health	1102 [17]	*Sufficient;* RMSEA 0.114, 95%CI 0.103–0.126; CFI 0.940; SRMR 0.043	Very good		1102 [17]	*Sufficient;* Results in line with 3 hypo and not in line with 1 hypo	adequate	*Sufficient;* (HIGH GRADE)
					75 [16]	*Sufficient;* Results in line with 2 hypo		
					360 [22]	*Sufficient;* Results in line with 2 hypo		
Physical Function v1.0 CAT					1946 [14]	*Sufficient;* Results in line with 3 hypo		

PROMIS	Reliability											
	Internal consistency				Reliability				Measurement error			
	n	Result	Meth qual study	Overall result (GRADE)	*n*	Result	Meth qual study	Overall result (GRADE)	*n*	Result	Meth qual study	Overall result (GRADE)
Global health	1102 [17]	*Sufficient;* Ordinal α GMH 0.875 / GPH 0.823	Doubtful	*Sufficient;* (HIGH GRADE)	195 [17]	*Sufficient;* ICC GMH 0.86/ GPH 0.88	Doubtful					
	75 [16]	*Sufficient;* Cronbach's α GMH 0.83 and GPH 0.79	Very good									
	360 [22]	*Sufficient;* Cronbach's α 0.90	Inadequate									
Physical function v1.0 CAT	1946 [14]	*Sufficient;* SEM 2.4 (SD 0.46)/ ≥ 3.3 in 98.7% of patients*	Very good		36 [14]	*Sufficient;* Test–retest correlation 0.82	Doubtful					
Physical function v1.0									337 [13]	*Insufficient;* SDC 6.51 > MIC 3.98	Inadequate	
Satisfaction with social roles v1.0									337 [13]	*Insufficient;* SDC 7.26 > MIC 6.20	Inadequate	

Table 2 (continued)

PROMIS	Reliability											
	Internal consistency				Reliability				Measurement error			
	n	Result	Meth qual study	Overall result (GRADE)	*n*	Result	Meth qual study	Overall result (GRADE)	*n*	Result	Meth qual study	Overall result (GRADE)
Fatigue v1.0									337 [13]	*Insufficient;* SDC 6.27 > MIC 2.41	Inadequate	
Anxiety v1.0									337 [13]	*Insufficient;* SDC 8.17 > MIC 3.51	Inadequate	

PROMIS	Responsiveness			
	n	Result	Meth qual study	Overall result (GRADE)
Global health	195 [17]	*Indeterminate;* Results partially in line with 3 hypo	Inadequate	
Physical function v1.0 CAT	739 [15]	*Insufficient;* Results not in line with 1 hypo	Adequate	
Fatigue v1.0 CAT	720 [15]	*Insufficient;* Results not in line with 1 hypo	Adequate	
Pain interference SF 4a	258 [18]	*Insufficient;* AUC 0.59	Doubtful	
Pain interference SF 6a	258 [18]	*Insufficient;* AUC 0.56	Doubtful	
Pain interference SF 6b	258 [18]	*Insufficient;* AUC 0.55	Doubtful	
Pain interference SF 8a	258 [18]	*Insufficient;* AUC 0.56	Doubtful	
Depression SF 4a	258 [21]	*Insufficient;* AUC 0.55–0.66	Doubtful	
Depression SF 6a	258 [21]	*Insufficient;* AUC 0.56–0.66	Doubtful	
Depression SF 8a	258 [21]	*Insufficient;* AUC 0.56–0.69	Doubtful	
Depression SF 8b	258 [21]	*Insufficient;* AUC 0.56–0.68	Doubtful	

*Structural validity was not tested in a stroke population, but unidimensionality was demonstrated in other populations described by Abma et al. [27]

CFI Comparative fit index; GMH Global Mental Health; GPH Global Physical Health; hypo hypothesis/hypotheses; GRADE Grading of Recommendations Assessment, Development, and Evaluation; ISI insomnia severity index; Meth qual methodological quality; MIC minimal important change; n number; ref reference; RMSEA Root mean square error of approximation; SDC smallest detectable change; SF short form; SRMR standardized root mean square residual; hypotheses that were tested for construct validity and responsiveness are described in eAppendix 2. The rating of quality of each study according to the four-point rating system of the COSMIN Risk of Bias checklist is described in eAppendix 3. The GRADE rating of the quality of the evidence was done when two or more studies were available on a measurement property of a PROMIS measure, details of this rating are described in eAppendix 4

Global Health was skipped most (in 3.2%) and that < 1% had missing items across 9 of the 10. The Global Mental Health (GMH) was complete for 99.6% of patients and the Global Physical Health (GPH) in 96.4% [17]. In the other studies patients with incomplete scores were excluded from the analyses. [13, 15, 18–20].

Proxy help was required by 28.4–30.1% of the patients for completing a PROMIS measure in three studies. [13, 14, 17].

Discussion

The HR-QoL measure, the PROMIS Global Health, has been studied most, demonstrating sufficient structural validity and sufficient internal consistency with high-quality evidence (i.e., there is high confidence that this finding is true in stroke populations). This is supported by findings in large general US and Dutch populations, demonstrating evidence for the unidimensionality and internal consistency of the two subscales [23–25]. We also found

Table 3 Summary of interpretability of PROMIS item banks

PROMIS item bank [ref]	Distribution of scores in the study population in mean ± SD	Floor and ceiling effects	Scores and/or change scores available for relevant (sub)groups	Minimal important change (MIC) or minimal important difference (MID)
Global Health SF10 [22]	54.3 ± 18.5	1.9% floor or ceiling	By mRS	
Physical Function v1.0 [13]	42.1 ± 10.8		By sex, race and mRS	MID_{anchor} 4.0
Physical Function v1.0 CAT [14]	41.4 ± 11.5	1.77% floor 0.68% ceiling		
Physical Function v1.0 CAT [15]	40.9 (33.1–48.8)*	1.3% ceiling	By sex, age, race, marital status, income, hypertension, CAD	
Satisfaction with social roles v1.0 [13]	45.2 ± 11.7		By sex, race and mRS	MID_{anchor} 6.2
Pain interference SF 4a; Pain interference SF 6a; Pain interference SF 6b; Pain interference SF 8a [18–19]	53.2 ± 10.4; 53.1 ± 10.6; 53.2 ± 10.3; 53.1 ± 10.6	31–36% floor		$MID_{BPI\text{-}I\ 1\ point}$ 2.8–2.9
Fatigue v1.0 [13]	51.5 ± 10.3		By sex, race and mRS	MID_{anchor} 2.4
Fatigue v1.0 CAT [15]	52.2 (46.2–60.3)*	1.3% ceiling	By sex, age, race, marital status, income, hypertension, CAD	
Anxiety v1.0 [13]	49.9 ± 10.4		By sex, race and mRS	MID_{anchor} 3.5
Depression SF 4a; Depression SF 6a; Depression SF 8a; Depression SF 8b [22]	51.3 ± 9.2; 50.5 ± 10.0; 50.3 ± 9.9; 50.0 ± 10.3			$MID_{PHQ\text{-}9\ 3\ points}$ 3.4–3.6

BPI-I Brief Pain Inventory Inference Scale; CAD coronary artery disease; CAT computer-adaptive testing; ES effect size; IQR interquartile range; MDC minimal detectable change; MID minimal important difference; mRS modified Rankin Scale; PGC Prospective global change; PHQ-9 patient health questionnaire; RGC Retrospective global change; SEM standard error of measurement; SD standard deviation; SF short form; v version. *Median with interquartile range

sufficient construct validity with high-quality evidence (i.e., there is high confidence that this finding is true in stroke populations), sufficient reliability was found in one study of doubtful quality, and PROMIS General Health showed no floor or ceiling effects in contrast to the frequently used HR-QoL measure, the EQ5D. No conclusions could be drawn on the other measurement properties and interpretability.

Although the knowledge on the measurement properties of the PROMIS Global Health in stroke populations might seem limited, one should take into account that this knowledge is also limited for other frequently used measures for HR-QoL measures, such as EQ5D and SF-36 [26]. In a review, both the EQ5D and SF-36 have 'limited' to 'moderate'-positive ratings for test–retest reliability and construct validity. Their content validity, intra-rater reliability, inter-rater reliability, and measurement error are unknown in stroke populations. The EQ5D had 'limited' to 'moderate' positive ratings for responsiveness but nothing was known about structural validity, while for the SF-36 nothing was known about responsiveness and even strong negative evidence was found for structural validity [26]. In this light, the PROMIS Global Health seems a good alternative HR-QoL measure to use in clinical care and research for stroke patients.

Although no firm conclusions could be drawn for the other PROMIS measures in stroke patients, there is an increasing amount of evidence on the promising measurement properties of PROMIS measures in other populations. Examples are high-quality evidence for sufficient structural validity and measurement precision of the Dutch-Flemish PROMIS Physical function item bank and the upper extremity subdomain in a review [27] and sufficient construct validity and reliability of PROMIS item banks and short forms in the US general population and clinical groups. [23].

The range of MICs of 2.4–6.2 found in stroke patients is in line with a recent systematic review of MIC values of all PROMIS measures in all populations. This review concluded that a MIC value of 2–6 T score points seems reasonable to assume [28]. When investigated, floor and ceiling effects were found in less than 2% of stroke patients. Only on the PROMIS pain interference short forms, 31–36% had the lowest score, similar to the 26% found in adults with hemophilia reported by Kuijlaars et al. [29], but this does not have to indicate lack of validity because these patients may just have no pain.

No studies were found that examined content validity, criterion validity and cross-cultural validity/measurement invariance in stroke populations. Content validity is considered the most important measurement property of a PROM, however we were not able to evaluate this measurement property because of lack of studies including this information [10]. An explanation for this limitation is that PROMIS measures are generic measures, and perhaps therefore content validity is not studied in every patient population. On the other hand, content validity of especially the short forms (i.e., a selection of questions of an item bank) might differ between populations. For the PROMIS Global Health, authors of the ICHOM Standard Set for Stroke stated that 'the PROMIS-10 covers the majority of the outcome domains considered most important by the expert panel', but they added additional questions about for example feeding and communication, because these domains are not included in the PROMIS Global Health [3]. In addition, our review demonstrated that missing values were seen [16, 17] and proxy help was needed [13, 14, 17], which might be due to diminished comprehensibility in stroke populations. These findings suggest that content validity research in stroke patients is still needed.

We expected that research into criterion validity would be scarce, because a gold standard for the majority of constructs studied by PROMIS measures does not exist.

Besides our findings that not all measurement properties were studied, the included studies concerned a quite similar subgroup of stroke patients. All studies included an outpatient hospital-based patient sample, mainly from the USA. When reported, the studies included patients that were relatively mildly impaired as measured by the NIHSS and mRS. The lack of cultural diversity and of more severely impaired stroke populations such as seen in rehabilitation may limit the generalizability of the results to all stroke populations.

The strength of our study is that this is the first study summarizing the measurement properties and interpretability of PROMIS measures in stroke patients, which are increasingly used in clinical care and research with the support of important institutes as the ICHOM and NIH. In addition, we have used the standardized approach of the COSMIN methodology for systematic reviews of PROMs to summarize, evaluate, and compare the data [6]. This approach aims to improve the quality of systematic reviews of measurement properties of PROMs. The criteria of good measurement properties are clearly defined; however, some standards of the COSMIN Risk of Bias checklist have room for personal interpretations, for example: "Were there any other important flaws in the design or statistical methods of the study?" It is to the interpretation of the reviewer whether a flaw is a minor flaw or an important flaw. Because almost all studies had clear predefined hypotheses, we choose to use these instead of redefining all hypotheses as described in COSMIN methodology. To optimize transparency of our ratings, we added Appendices, where all hypotheses and ratings are given (eAppendix 2–4).

In conclusion, although the ICHOM Standard Set for Stroke and the NIH advise the use of PROMIS measures in clinical care and research, the measurement properties and interpretability of PROMIS measures are not yet fully established in stroke populations. The first studies show promising results in line with the additional evidence for sufficient measurement properties of PROMIS from other populations. Nevertheless, there is a need for further research on content validity, structural validity, and measurement invariance of PROMIS measures in stroke patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-022-03149-4.

Funding No funding was received for conducting this study.

Declarations

Conflict of interest All authors declare that there are no conflicts of interest.

References

1. Reeves, M., Lisabeth, L., Williams, L., Katzan, I., Kapral, M., Deutsch, A., & Prvu-Bettger, J. (2018). Patient-Reported Outcome Measures (PROMs) for acute stroke: Rationale. *Methods and Future Directions. Stroke., 49*(6), 1549–1556. https://doi.org/10.1161/STROKEAHA.117.018912
2. Mokkink, L. B., Prinsen, C. A., Bouter, L. M., Vet, H. C., & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Brazilian Journal of Physical Therapy, 20*(2), 105–113. https://doi.org/10.1590/bjpt-rbf.2014.0143
3. Salinas, J., Sprinkhuizen, S. M., Ackerson, T., Bernhardt, J., Davie, C., George, M. G., Gething, S., Kelly, A. G., Lindsay, P., Liu, L., Martins, S. C., Morgan, L., Norrving, B., Ribbers, G. M., Silver, F. L., Smith, E. E., Williams, L. S., & Schwamm, L. H. (2016). An international standard set of patient-centered outcome measures after stroke. *Stroke, 47*(1), 180–186. https://doi.org/10.1161/STROKEAHA.115.010898
4. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Rose, M., PROMIS Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care., 45*(5), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55
5. Arwert, H., Oosterveer, D. M., Schoones, J. W., Terwee, C. G., & Vliet Vlieland, T. P. M. A systematic review on the current use of PROMIS item banks as outcome measurement in stroke patients. Archives of Rehabilitation Research & Clinical Translation. In press. https://doi.org/10.1016/j.arrct.2022.100191.
6. Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures.

Quality of Life Research, 27, 1147–1157. https://doi.org/10.1007/s11136-018-1798-3
7. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*, 1171–1179. https://doi.org/10.1007/s11136-017-1765-4
8. Schünemann, H., Brożek, J., Guyatt, G., Oxman, A. (2013). GRADE handbook—Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated in 2013. Retrieved from: https://gdt.gradepro.org/app/handbook/handbook.html
9. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006
10. Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0
11. Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum.
12. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012
13. Lapin, B., Thompson, N. R., Schuster, A., & Katzan, I. L. (2019). Clinical utility of Patient-Reported Outcome Measurement information system domain scales. *Circulation Cardiovascular Quality and Outcomes, 12*(1), e004753. https://doi.org/10.1161/CIRCOUTCOMES.118.004753
14. Katzan, I. L., Fan, Y., Uchino, K., & Griffith, S. D. (2016). The PROMIS physical function scale: A promising scale for use in patients with ischemic stroke. *Neurology, 86*(19), 1801–1807. https://doi.org/10.1212/WNL.0000000000002652
15. Katzan, I. L., Thompson, N. R., Lapin, B., & Uchino, K. (2017). Added value of Patient-Reported Outcome Measures in stroke clinical practice. *Journal of the American Heart Association, 6*(7), e005356. https://doi.org/10.1161/JAHA.116.005356
16. Lam, K. H., & Kwa, V. I. H. (2018). Validity of the PROMIS-10 Global Health assessed by telephone and on paper in minor stroke and transient ischaemic attack in the Netherlands. *British Medical Journal Open, 8*(7), e019919. https://doi.org/10.1136/bmjopen-2017-019919
17. Katzan, I. L., & Lapin, B. (2018). PROMIS GH (Patient-Reported Outcomes Measurement Information System Global Health) scale in stroke: A validation study. *Stroke, 49*(1), 147–154. https://doi.org/10.1161/STROKEAHA.117.018766
18. Chen, C. X., Kroenke, K., Stump, T., Kean, J., Krebs, E. E., Bair, M. J., Damush, T., & Monahan, P. O. (2019). Comparative responsiveness of the PROMIS pain interference short forms with legacy pain measures: Results from three randomized clinical trials. *The Journal of Pain, 20*(6), 664–675. https://doi.org/10.1016/j.jpain.2018.11.010
19. Chen, C. X., Kroenke, K., Stump, T. E., Kean, J., Carpenter, J. S., Krebs, E. E., Bair, M. J., Damush, T. M., & Monahan, P. O. (2018). Estimating minimally important differences for the PROMIS pain interference scales: Results from 3 randomized clinical trials. *Pain, 159*(4), 775–782. https://doi.org/10.1097/j.pain.0000000000001121
20. Kroenke, K., Stump, T. E., Chen, C. X., Kean, J., Bair, M. J., Damush, T. M., Krebs, E. E., & Monahan, P. O. (2020). Minimally important differences and severity thresholds are estimated for the PROMIS depression scales from three randomized clinical trials. *Journal of Affective Disorders, 266*, 100–108. https://doi.org/10.1016/j.jad.2020.01.101
21. Kroenke, K., Stump, T. E., Chen, C. X., Kean, J., Damush, T. M., Bair, M. J., Krebs, E. E., & Monahan, P. O. (2021). Responsiveness of PROMIS and Patient Health Questionnaire (PHQ) depression scales in three clinical trials. *Health and Quality of Life Outcomes, 19*(1), 41. https://doi.org/10.1186/s12955-021-01674-3
22. de Graaf, J. A., Visser-Meily, J. M., Schepers, V. P., Baars, A., Kappelle, L. J., Passier, P. E., Wermer, M. J., de Wit, D. C., & Post, M. W. (2021). Comparison between EQ-5D-5L and PROMIS-10 to evaluate health-related quality of life 3 months after stroke: A cross-sectional multicenter study. *European Journal of Physical and Rehabilitation Medicine, 57*(3), 337–346. https://doi.org/10.23736/S1973-9087.21.06335-8
23. Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. *Quality of Life Research, 18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9
24. Pellicciari, L., Chiarotto, A., Giusti, E., Crins, M. H. P., Roorda, L. D., & Terwee, C. B. (2021). Psychometric properties of the Patient-Reported Outcomes Measurement Information System scale v1.2: Global Health (PROMIS-GH) in a Dutch general population. *Health and Quality of Life Outcomes, 19*(1), 226. https://doi.org/10.1186/s12955-021-01855-0
25. Alcantara, J., Whetten, A., Zabriskie, C., & Jones, S. (2021). Exploratory factor analysis of PROMIS-29 V1.0, PROMIS Global Health and the RAND SF-36 from chiropractic responders attending care in a practice-based research network. *Health and Quality of Life Outcomes, 19*(1), 82. https://doi.org/10.1186/s12955-021-01725-9
26. Cameron, L. J., Wales, K., Casey, A., Pike, S., Jolliffe, L., Schneider, E. J., Christie, L. J., Ratcliffe, J., & Lannin, N. A. (2021). Self-reported quality of life following stroke: A systematic review of instruments with a focus on their psychometric properties. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02944-9
27. Abma, I. L., Butje, B. J. D., Ten Klooster, P. M., & van der Wees, P. J. (2021). Measurement properties of the Dutch-Flemish Patient-Reported Outcomes Measurement Information System (PROMIS) physical function item bank and instruments: A systematic review. *Health and Quality of Life Outcomes, 19*(1), 62. https://doi.org/10.1186/s12955-020-01647-y
28. Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J. S., Terluin, B., Cella, D., Griffith, P., & Mokkink, L. B. (2021). Minimal important change (MIC): A conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Quality of Life Research*. https://doi.org/10.1007/s11136-021-02925-y
29. Kuijlaars, I. A. R., Teela, L., van Vulpen, L. F. D., Timmer, M. A., Coppens, M., Gouw, S. C., Peters, M., Kruip, M. J. H. A., Cnossen, M. H., Muis, J. J., van Hoorn, E. S., Haverman, L., & Fischer, K. (2021). Generic PROMIS item banks in adults with hemophilia for patient-reported outcome assessment: Feasibility, measurement properties, and relevance. *Research and Practice in Thrombosis and Haemostasis, 5*(8), e12621. https://doi.org/10.1002/rth2.12621