Clinical Kidney Journal, 2023, vol. 16, no. 11, 2108–2128

https://doi.org/10.1093/ckj/sfad170
Advance Access Publication Date: 18 July 2023
Original Article

ORIGINAL ARTICLE

Measurement properties of performance-based measures to assess physical function in chronic kidney disease: recommendations from a COSMIN systematic review

Jennifer M. MacRae[1], Oksana Harasemiw[2,3], Courtney J. Lightfoot[4], Stephanie Thompson[5], Kathryn Wytsma-Fisher[6], Pelagia Koufaki[7], Clara Bohm[2,3] and Thomas J. Wilkinson[8]

[1]Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada, [2]Department of Internal Medicine, University of Manitoba, Winnipeg, Manitoba, Canada, [3]Chronic Disease Innovation Centre, Winnipeg, Manitoba, Canada, [4]Department of Health Sciences, University of Leicester, Leicester, UK, [5]Department of Medicine, University of Alberta, Edmonton, Alberta, Canada, [6]Alberta Kidney Care South, Alberta Health Services, Calgary, Alberta, Canada, [7]School of Health Sciences, Centre for Health, Activity and Rehabilitation Research, Queen Margaret University, Edinburgh, UK and [8]Leicester Diabetes Centre, University of Leicester, Leicester, UK

Correspondence to: Jennifer M. MacRae; E-mail: jmmacrae@ucalgary.ca

ABSTRACT

Background. There is wide heterogeneity in physical function tests available for clinical and research use, hindering our ability to synthesize evidence. The aim of this review was to identify and evaluate physical function measures that could be recommended for standardized use in chronic kidney disease (CKD).

Methods. MEDLINE, EMBASE, Cochrane Central Register of Controlled Trials, Cochrane Database of Systematic Reviews, CINAHL, Scopus and Web of Science were searched from inception to March 2022, identifying studies that evaluated a clinimetric property (validity, reliability, measurement error and/or responsiveness) of an objectively measured performance-based physical function outcomes using the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology and Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) based recommendations. Studies with individuals of all ages and of any stage of CKD were included.

Results. In total, 50 studies with 21 315 participants were included. Clinimetric properties were reported for 22 different physical function tests. The short physical performance battery (SPPB), Timed-up-and-go (TUG) test and Sit-to-stand tests (STS-5 and STS-60) had favorable properties to support their use in CKD and should be integrated into routine use. However, the majority of studies were conducted in the hemodialysis population, and very few provided information regarding validity or reliability.

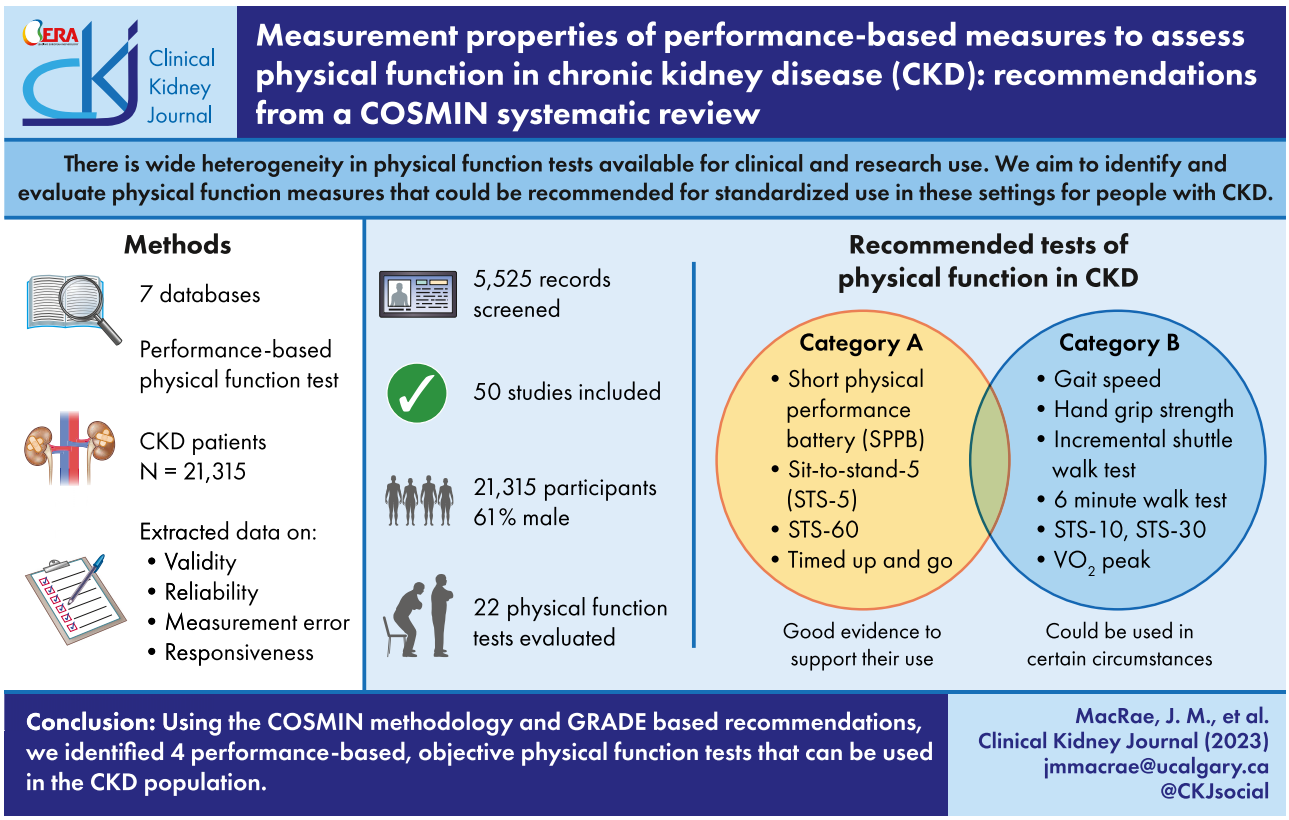
Received: 16.2.2023; Editorial decision: 26.6.2023

Conclusion. The SPPB demonstrated the highest quality of evidence for reliability, measurement error and construct validity amongst transplant, CKD and dialysis patients. This review is an important step towards standardizing a core outcome set of tools to measure physical function in research and clinical settings for the CKD population.

LAY SUMMARY

Poor physical function is very common among people living with chronic kidney disease (CKD), gets worse as the disease progresses, and is linked with reduced survival and poor quality of life. Routine assessment of physical function can help identify frailty and sarcopenia (loss of muscle mass) early on, allowing for timely introduction of therapies such as exercise rehabilitation. In this study, we reviewed the current literature to identify which physical function measures should be used to ensure that physical function is captured accurately and consistently, both in clinical practice and in research. We found that the short physical performance battery, Timed-up-and-go, Sit-to-stand-5 and Sit-to-stand-60 had good evidence to support their use in CKD. Our review and recommendations are the initial steps toward introducing a set of tools to measure physical function that can be used as part of routine CKD care.

GRAPHICAL ABSTRACT



Keywords: chronic kidney disease, clinimetric properties, measurement tools, physical function, systematic review

INTRODUCTION

Physical function is a key determinant of health and independent living, and can be assessed on a spectrum of outcomes ranging from self-care to activities that require varying degrees of mobility, balance, strength or endurance [1, 2]. In people living with chronic kidney disease (CKD), the effects of multimorbidity, physical inactivity and sedentary lifestyle result in functional limitations, which in turn negatively impact quality of life, patient-relevant outcomes such as falls and hospitalization, and survival [3–5]. Routine assessment of physical function may help clinicians monitor physical health status [6], and allow for the identification of people at risk of frailty and sarcopenia for timely implementation of therapeutic interventions, such as exercise, to prevent or delay disability and loss of independence [7].

There are a plethora of assessment tools reported in the CKD literature, ranging from objective measures of performance-based capacity, such as gait speed, chair-stand ability and hand-grip strength (HGS), to self-reported measures of physical ability [5]. A review by Jegatheesan *et al.* (2021) demonstrated that clinical and research practice seem to be varied and inconsistent

in their approach to measuring physical function; in 111 trials reporting physical fitness outcomes in adults with CKD, 87 different physical function tests/measurements were used to evaluate 30 different outcome measures [8]. Recommendations on best physical function assessment tools should be based on criteria of good measurement properties such as validity, reliability, responsiveness and interpretability, as well as safety and feasibility characteristics. We have therefore performed a systematic review to summarize the clinimetric measurement properties of performance-based physical function measures in CKD, in order to support evidence informed recommendations for use in clinical and research settings.

MATERIALS AND METHODS

This review follows recommendations from the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative [9, 10] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses protocols (PRISMA 2020) [11, 12]. The review was registered with PROSPERO (CRD42020182207). The PRISMA 2020 checklist was used for reporting of results (Supplementary Material 8).

Literature search

The following electronic databases were searched from their date of establishment to 22 March 2020: MEDLINE (Ovid), EMBASE (Ovid), Cochrane Central Register of Controlled Trials (Ovid), Cochrane Database of Systematic Reviews (Ovid), CINAHL, Scopus and Web of Science. Reference lists of key review articles and studies selected for inclusion were hand-searched. The searches were re-run prior to final analysis (March 2022). The full search strategy is in Supplementary Material 1. Results were exported into the Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia) for deduplication and screening. The list of studies excluded after full text review, and reasons for exclusion, are listed in Supplementary Material 9.

Inclusion criteria

The studies included people (of all ages) living with CKD across all stages and treatment modalities. Any study where the aim was to evaluate a clinimetric property of a performance-based physical function outcome, or any study that reported on the effect of an intervention using an objectively measured performance-based physical function outcome and provided a clinimetric property for the tool, was included.

Exclusion criteria

Studies with no data on clinimetric properties, or used tools that are highly specialized or that are not routinely available to measure physical function, were excluded. In addition, subjective questionnaires that assess physical function were excluded. Studies published in a language other than English were excluded due to resource constraints associated with translation [13, 14].

Study selection

Reviewers (J.M.M., O.H., C.J.L., S.T., K.W.-F., C.B., T.J.W.) working in pairs independently screened abstracts, performed full text adjudication and extracted data. Any disagreements were resolved by discussing discrepancies and reaching consensus, mediated by a third person if necessary.

Data extraction

The data extracted from the studies is shown in Table 1. Objective measures of physical function were included: tests of cardiorespiratory fitness (e.g. measured or estimated peak oxygen uptake, anaerobic threshold), exercise capacity (e.g. peak power output, time on test, distance travelled) or neuromuscular fitness (e.g. strength, power, balance) as defined by Jegatheesan *et al.* [8].

Measurement properties

The measurement properties were defined using COSMIN classification [9] but modified for objective measurement tools and included: validity by construct (hypothesis testing) or criterion (correlation with gold standard), reliability [intraclass correlation (ICC)], responsiveness [area under the curve (AUC)] and measurement error [standard error of the mean (SEM) and minimal detectable change (MDC)] [15]. Minimally clinically important difference (MCID) was considered to be an important clinimetric property that, although not part of the modified COSMIN checklist, was included.

Assessing the methodologic quality of each study per measurement properties

Methodologic quality was assessed using the modified COSMIN Risk of Bias (RoB) checklist [16], which is a checklist developed for assessing the methodological quality of studies on outcome measurement tools, outlining minimum standards for measurement properties (including content validity, structural validity, internal consistency, cross-cultural validity\measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity and responsiveness). Reviewers working in pairs ranked each measurement property (very good/adequate/doubtful/inadequate) and reached consensus on each item in the checklist, as it applied to the individual study; the rankings derived from this then informed the overall RoB score using the COSMIN RoB tool [15, 17] (Supplementary Material 3 and 4).

Assessing the criteria for good measurement properties using COSMIN quality criteria

The ratings of the results for each study are based on the criteria for good measurement properties defined in Supplementary Material 2 [9].

Data synthesis

To summarize multiple studies for a physical function measure, the results were first reviewed for consistency across the studies, qualitatively summarized and the overall result compared with the criteria for good measurement properties. Therefore, each physical function test was rated, as per the description listed in Supplementary Material 3. A (+) ranking denotes a strong level of evidence; (–) limited with only one study of fair methodological quality; (±) inconsistent/conflicting findings; and (?) denotes an unknown level of evidence (no evidence available or only studies of poor methodological quality).

Table 1: Summary of included study characteristics.

1st author (year), country	Physical function test reported	Population	Study design	Sample size (*n*)	Age, years (SD)	Male, %	BMI, kg/m^2 (SD)	DM, %	CAD, %
Abdulan 2020, Romania [20]	SPPB; HGS	CKD	Cross sectional	80	76.7 (5.9)	35	NR	36	NR
Bohannon 1995, USA [52]	Knee extension; STS-10; Gait speed	Mixed CKD	Cross sectional	110	45.0 (11.6)	69	NR	35	NR
Brar 2021, Canada [43]	SPPB	CKD 4/5	Prospective cohort	603	66 (54–73) NF; 73 (65–82) F	59.0	NR	57.0	20.4
Chakkera 2018, USA [38]	VO_{2peak}	Mixed CKD	Clinical prediction rule	637	56.6 (NR)	61	NR	52	21
Chakkerra 2022, USA [39]	VO_{2peak}	Tx	Prospective cohort	293	57.6 (12.8)	61.4	NR	48.8	15.8
Cheng 2020, USA [32]	6MWT; STS-60	Dialysis Tx waitlist	Prospective cohort	305	NR	53.5	NR	48.9	24.6
Collado-Mateo 2019, Portugal [68]	Knee extension strength; Knee flexion strength; Elbow extension strength; Elbow flexion strength	PD	Cross sectional	31	48.5 (13.4)	61	24.4 (3.6)	NR	NR
da Silva 2022, Brazil [69]	Glittre ADL test	HD	Test–retest	30	52 (12)	63.3	26 (3.59)	13	NR
de Villar 2018, Spain [47]	SPPB; OLST; TUG	HD	Test–retest	71	61.7 (16.4)	59	25.9 (6.2)	52	21
Figueiredo 2022, Brazil [70]	Glittre ADL test	HD	Test–retest	91	52.4 (16.5)	59.3	23.8 (4.4)	22.0	NR
Figueiredo 2021, Brazil [51]	STS-30	HD	Test–retest	63	48.3 (44.6–51.9)	66.7	24.1 (22.9–25.4)	20.6	NR
Greenwood 2019, UK [57]	ISWT	Mixed CKD and dialysis	Prospective cohort	757	56.1(12.4)	54	30.8 (7.0)	59	NR
Hadjiioannou 2020, UK [56]	ISWT; STS-60; TUG	Mixed CKD	Test–retest	40	56.7 (11.0)	70	32.1 (6.7)	35	71
Haugen 2020, USA [44]	SPPB	Potential KT candidates	Prospective cohort	3255	54 (SD 14, range 18–89)	60	NR	NR	NR
Hsieh 2006, Taiwan [36]	$VO_{2peak;}$ 6MWT	HD	Cross sectional	27	60.6 (9.2)	70	23.1 (3.0)	22	NR
Huang 2018, UK [50]	STS-5; STS-60; Quadriceps muscle strength	Tx	Test–retest (same day)	56	50 (13)	59	27.7 (4)	NR	NR
Isoyama 2014, Sweden [21]	HGS	HD	Cross sectional	330	53 (13)	62	NR	NR	NR
Jácome 2018, Portugal [71]	BESTest (36); Mini; BESTest (14); Brief-BESTest (2)	Mixed dialysis	Test–retest	74	63.9 (15.1)	66	25.1 (4.3)	NR	NR
Jamal 2006, Canada [31]	HGS; Functional reach; TUG; 6MWT	HD	Cross sectional	52	66 (9)	71	NR	31	NR
Jiménez 2022, Spain [29]	SPPB, 4 m gait speed, STS-10, STS-60, 6MWT, OLST, TUG, HGS	HD	Test–retest	30	66.4 (16.3)	66.7	NR	23.3	NR
Johnstone 2022, USA [48]	SPPB, STS-5, gait speed, 2-min walk test, 90-s balance test, knee extensor strength	CKD and HD	Test–retest	21	58.72 (13.56)	50.0	29.74 (6.42)	40.9	50.0
Kittiskulnam 2017, USA [25]	HGS; Gait speed	HD	Prospective cohort	645	56.7 (14.5)	59	28.1 (6.9)	44	9
Kohl 2012, Brazil [34]	6MWT; VO_{2peak}	HD	Prospective cohort	89	36 (11)	54	22 (4)	NR	NR
Koufaki 2001, UK [40]	$VO_{2\ peak}$	Mixed dialysis	Test–retest	12	46.8 (18.5)	92	NR	17	17
Koufaki 2002, UK [42]	VO_{2peak}; STS-5; STS-60; NSRI Walk	Mixed dialysis	Randomized controlled trial	48	57.3 (14.3)	73	25.9 (3.6)	16	45
Lopes 2018, Brazil [26]	HGS	HD	Prospective cohort	413	50.4 (15)	60	22.0 (6.5)	38	8

Table 1: Continued.

1st author (year), country	Physical function test reported	Population	Study design	Sample size (*n*)	Age, years (SD)	Male, %	BMI, kg/m^2 (SD)	DM, %	CAD, %
Lorenz 2017, USA [72]	SPPB	Mixed CKD	Prospective cohort	140	51.2 (15.1)	61	27.2 (5.0)	21	7
Macagnan 2019, Brazil [27]	HGS	HD	Randomized crossover trial	21	57.6 (8.7)	67	24.3 (3.74)	7	NR
Macdonald 2005, UK [73]	Knee extension strength; STS-30	HD	Test–retest	10	48.4 (5.3)	78	24.8 (1.5)	NR	NR
Matos 2014, Brazil [22]	HGS	HD	Prospective cohort	443	46.6 (14.1)	62	NR	22	5
Mercer 1998, UK [74]	NSRI Walk	Mixed dialysis	Test–retest	25	V: 58.8 (11.0); R: 61.8 (13.0)	V: 71; R: 72	NR	NR	NR
Mesquita 2013, Netherlands [54]	TUG	HD	Test–retest (same day)	72	64 (IQR 54–75)	63	24.3 (IQR 21.8–26.6)	NR	NR
Nastasi 2018, USA [46]	SPPB	Tx	Prospective cohort	652	51.8 (14.1)	62	NR	NR	NR
Nastasi 2018, USA [45]	SPPB; Balance; Gait speed; STS-5	Tx	Prospective cohort	719	51.6 (14.2)	62	26.7 (5.4)	32	NR
Nixon 2019, UK [19]	HGS; Gait speed; SPPB	Mixed CKD	Cross sectional	90	F 68 (13); NF 73 (11)	F: 58; NF: 21	F: 29 (6); NF: 28 (6)	NR	NR
Overend 2010, Canada [37]	6MWT; STS-30	HD	Test–retest	25	67.2 (14.2)	56	NR	NR	NR
Padilla 2008, USA [35]	$VO_{2peak;}$ 6MWT; Gait speed; STS-10	CKD	Cross sectional	32	57.4 (14.8)	84	28.1 (5.0)	NR	NR
Segura-Ortí 2011, Spain [28]	STS-10; STS-60; 6MWT; OLHR; HGS	HD	Test–retest	39	60.3 (15.8)	82	22.0 (3.3)	NR	NR
Shi 2017, China [30]	6MWT	PD	Prospective cohort	145	L: 51 (11); S: 59 (13)	L: 43; S: 44	L: 22.3 (3.35); S: 22.56 (3.33)	L: 11.1; S: 23.3	6.9
Souweine 2017, France [75]	Knee extension strength	HD	Test–retest	21	72.4 (13.3)	67	26.7 (NR)	33.3	NR
Sutcliffe 2018, Australia [55]	STS-30; TUG (8 ft)	HD	Prospective cohort	228	67.5 (13.2)	61	NR	39	NR
Torino 2014, Italy [33]	6MWT	Mixed dialysis	Prospective cohort	317	65 (13)	68	25 (5)	21	NR
Watanabe[a] 2016, Brazil [18]	6MWT	Mixed CKD with Tx	Cross sectional	38	11.2 (6.5–16)	63	17.5 (13.4–27.6)	NR	NR
Watson 2020, UK [58]	ISWT	CKD stages 3b-5 non-dialysis	Prospective cohort	89	62.8 (11.0)	55.1	NR	24.7	22.5
Wilkinson 2019, UK [76]	STS-5; STS-60; Gait speed; ISWT; SPPB	CKD	Cross sectional	30	57 (17.8)	53	29.5 (4.8)	30	NR
Wilkinson 2019, UK [53]	ISWT; STS-5; STS-60; Knee extension strength; VO_{2peak}	CKD	Test–retest	26	61.4 (13.7)	44	29.5 (6)	27	NR
Wilkinson 2019, UK [41]	ISWT; STS-5; STS-60; VO_{2peak}	CKD	Test–retest	41	62 (11)	44	30.1 (5.7)	22	NR
Wilkinson 2021, UK [23]	HGS	CKD	Prospective cohort	8767	62.8 (5.8)	46	29.3 (5.2)	NR	NR
Xu 2020, China [24]	HGS	PD	Prospective cohort	1089	56.4 (11.1)	50.8	23.2 (3.8)	38.9	39.5
Zanotto 2020, UK [49]	HGS; TUG; Gait speed; STS-5	HD	Prospective cohort	93	61.7 (13.3)	55	29.1 (6.4)	25	NR

Numerical data presented as frequency (%) or mean ± SD as reported. Mixed CKD includes all stages of CKD and dialysis.

[a]Conducted in pediatric population.

L: long 6MWD and S: short 6MWD groups; C: control; E: exercise groups; V: validity; R: reliability groups; F: frail; NF: non-frail; OLHR: One leg heel rise; NR: not reported; SD: standard deviation; DM: diabetes mellitus; CAD: coronary artery disease; BMI: body mass index; Tx: transplant; OLST: One leg stand test; IQR: interquartile range.

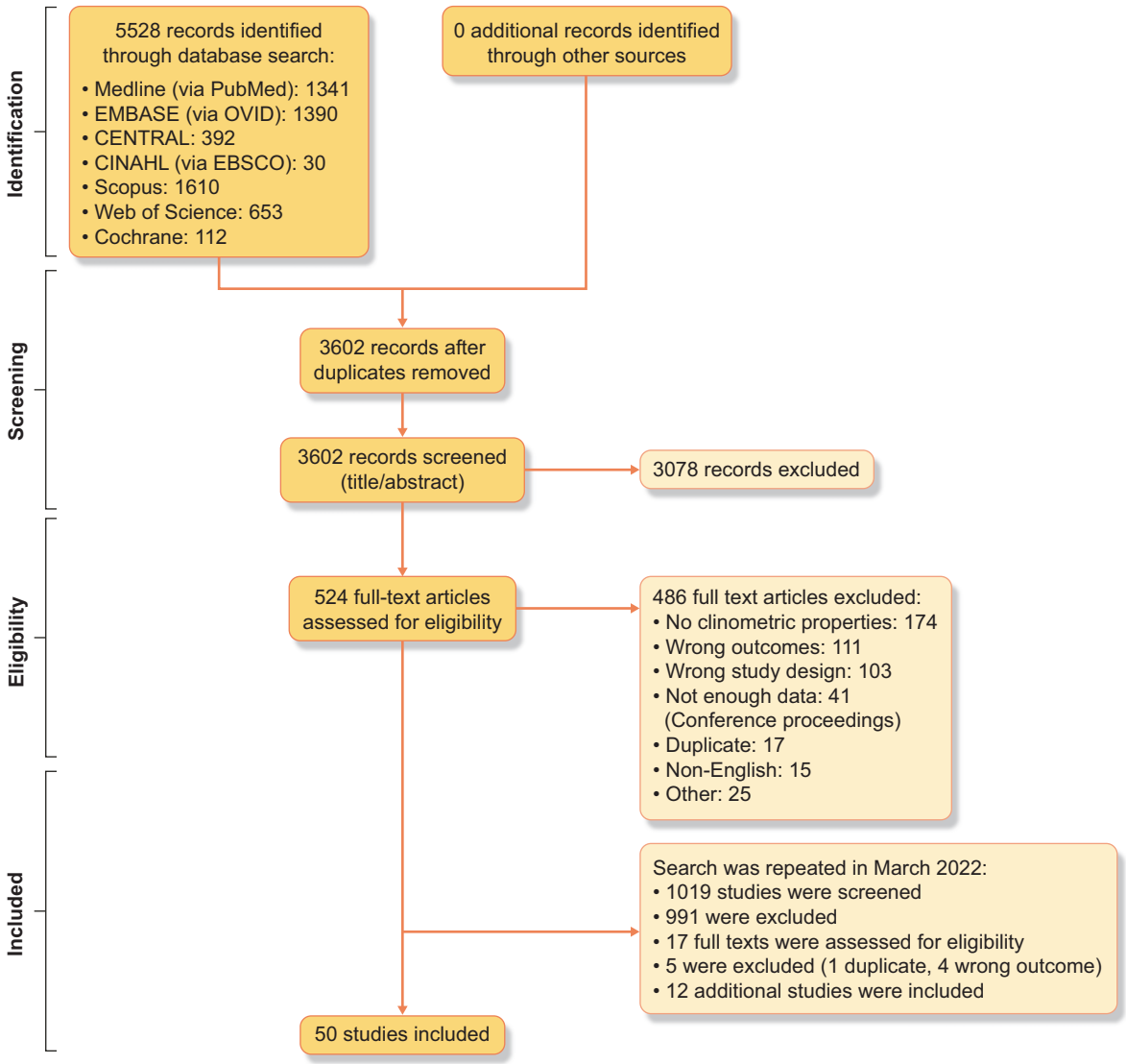


Figure 1: PRISMA flow diagram.

Due to heterogeneity of studies and populations for the measures we evaluated, a meta-analysis was not conducted.

GRADE evaluation of the quality of the evidence

The overall quality of the evidence for each measurement property was graded (Supplementary Material 4) using a modified Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) approach taking into consideration the number, methodologic quality, heterogeneity and consistency of the results from the studies [15].

For each measurement property, the overall quality of evidence was given an initial evidence grading of "high quality," which was then downgraded in a step-wise fashion from moderate, to low, to very low, if there were any emerging concerns regarding overall risk of bias, inconsistency of results, imprecision or indirectness. The overall quality of evidence could also be downgraded by three levels to "very low," if there was only one inadequate study present [17].

Final recommendations for physical function tools

Based on COSMIN criteria recommendations, physical function tools were classified as: Category A—recommended; Category B—promising; or Category C—insufficient [15]. Category A was defined as an instrument that had at least moderate quality for reliability and validity AND at least low evidence for measurement error or responsiveness. Category C was defined as an instrument with insufficient or low evidence for measurement property. Category B was an instrument that was not categorized as A or C.

RESULTS

Summary of included study characteristics

The initial literature search identified 5528 records of which 524 were selected for full text review and 38 were included in the review. An updated search (March 2022) revealed an additional 12 studies. In total, 50 studies were included (Fig. 1). Ongoing trials

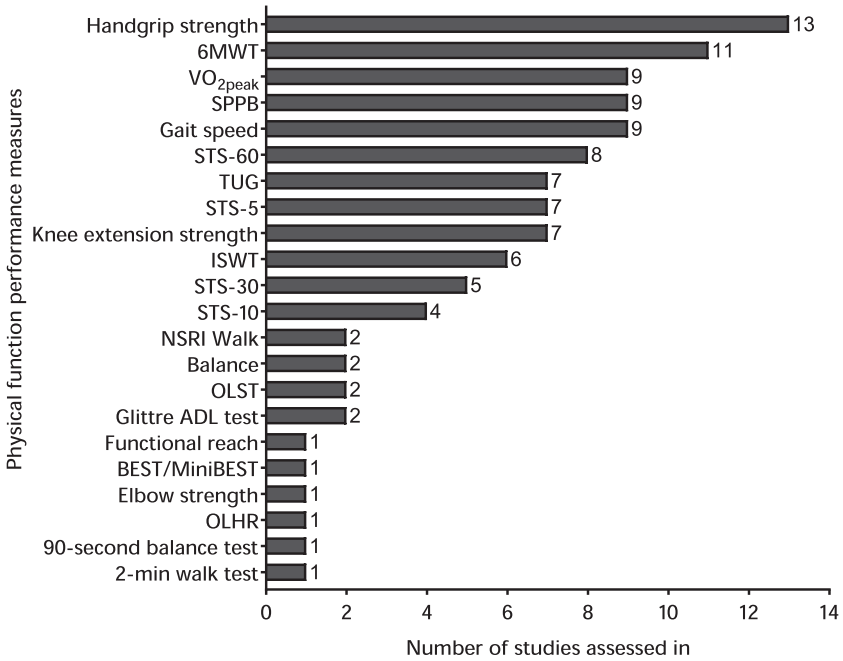


Figure 2: Frequency of physical function performance measures (in order of most frequently used). OLST: One leg stand test; ADL: Activity of daily living; OLHR: One leg heel raise. Knee extension strength assessed using either handheld dynamometer, Biodex System or estimated 1-repetiition maximum on a leg extension machine.

and conference proceedings/published abstracts were excluded during screening due to limited data that were available in these types of publications. The majority of studies were conducted in adults, whilst one was conducted in pediatric patients on either hemodialysis (HD) or peritoneal dialysis (PD) [18]. HD was the most common population studied ($n = 18$). There were two PD studies, and five combined both PD and HD patients. Seven studies were conducted exclusively in non-dialysis CKD and five in transplant recipients. The remaining studies included a combination of CKD stages. Study designs included were $n = 18$ test–retest, $n = 1$ clinical prediction rule, $n = 10$ cross-sectional, $n = 19$ prospective cohort, $n = 1$ randomized controlled trial and $n = 1$ randomized crossover trial.

Studies were published between 1995 and 2022 and included a range of study designs (Table 1). 21 315 participants were included. Study sample sizes ranged from 10 to 8767. The mean age was 57.3 years (range 11.2–76.7 years) and 61.0% of participants were male.

Clinimetric properties were reported for 22 different physical function tests (Fig. 2). Data on the tests that were most commonly assessed (four or more studies) are shown in Tables 2 and 3 and Fig. 2. Tests not commonly reported in the literature (in two or fewer studies) or tests with a lot of heterogeneity in the way they are measured (e.g. knee extension strength) are described in Supplementary Material 5 and 6.

Measurement properties for physical function measures

Tables 2 and 3 summarizes results on the clinimetric properties of physical function tests.

Handgrip strength

Of 13 studies, 10 reported on construct validity. Low HGS was associated with frailty [19], malnutrition [20] and an increased risk of mortality in six studies [21–26]. Reliability was reported in three studies and showed excellent ICC values (0.89–0.96) [27–29]. Two studies reported on measurement error with a SEM ranging from 1.5 to 2.3 kg in people on HD [28, 29]. The minimal detectable change, 90% confidence level (MDC90) ranged between 3.4 and 5.5 kg. The summarized quality of the evidence for HGS was moderate for measurement error and not assessed for responsiveness. The quality was low for reliability and validity (Supplementary Material 4).

Six-minute walk test

Of 11 studies, 7 reported on validity [30–36]. In people with CKD and receiving HD, data showed that the Six-minute walk test (6MWT) correlated with maximum oxygen uptake (VO_{2max}; r between 0.62 and 0.70) [35, 36], and low score was predictive of mortality [32–34], hospitalizations [33], cardiovascular events [33], technique failure amongst PD patients [30], increased fracture risk [31] and lower chance of transplantation [32]. Four studies showed excellent reliability [18, 28, 29, 37]. Measurement error was consistent (28.0–29.5 m) in an adult HD population [28, 29, 37] and amongst a mixed pediatric group of CKD, HD and transplant patients (21.8 m) [18]. The summarized quality of evidence is high for reliability and measurement error, low in the validity and very low for responsiveness studies (Supplementary Material 4).

Peak oxygen uptake

Of nine studies, five reported on validity [34–36, 38, 39]. Greater peak oxygen uptake (VO_{2peak}) values predicted lower risk of future cardiovascular events and mortality amongst mixed CKD patients [38] and a reduced risk of mortality amongst transplant patients [39], but not in HD [34]. Reliability properties were measured in two studies [40, 41], and three explored measurement error (SEM between 1.00 and 1.01 mL/kg/min) in different populations [40–42]. MCID, reported in one study, ranged from 0.3 to

Table 2: Summary of results on criterion validity, construct validity and responsiveness for physical function measures.

		Construct validity properties			Responsiveness properties		
Physical function measure	1st author, year	Comparison with other measures; predictive validity; correlation coefficients	Quality criteria rating[a]	COSMIN score	Criterion or construct approach	Quality criteria rating[a]	COSMIN score
Gait speed	Bohannon 1995 [52]	Usual [r/rs = 0.385 (95% CI 0.21–0.53)] and maximum gait speed (r/rs = 0.538) correlated to normalized knee extension force; data also presented for non-dominant and non-normalized associations	?	Inadequate	N/A	N/A	N/A
	Kittiskulnam 2017 [25]	Gait speed associated with death [adjusted C statistic: 0.66; NRI%: 18.8 (95% CI –12.5 to 49.9); HR 0.74 (95% CI 0.54–1.03), Harrell C: 0.74] Slowness (defined as gait speed ≤0.8 m/s) associated with death [C statistic: 0.68; NRI%: 50.5 (95% CI 24.3–73.0); HR 2.25 (95% CI 1.36–3.74), Harrell C: 0.74]	+	Inadequate	N/A	N/A	N/A
	Nastasi 2018 [45]	1 point decrease in gait speed produces 1.21 fold ↑ risk death (95% CI 0.89–1.65)	+	Adequate	N/A	N/A	N/A
	Zanotto 2020 [49]	Gait speed did not predict falls over 12 months [rate ratio 0.26 (95% CI 0.03–2.08)]	+	Adequate	N/A	N/A	N/A
	Nixon 2019 [19]	15 ft gait speed correlates with frailty [correlation coefficient: 0.70 (95% CI 0.55–0.80)]	+	Very good	N/A	N/A	N/A
	Brar 2021 [43]	Gait speed associated with mortality HR 1.82 (95% CI 1.36–2.43)	+	Doubtful	N/A	N/A	N/A
HGS	Zanotto 2020 [49]	HGS did not predict falls over 12 months [rate ratio 0.97 (95% CI 0.92–1.02)]	+	Adequate	N/A	N/A	N/A
	Isoyama 2014 [21]	HGS was inversely associated with mortality [HR 0.32 (95% CI 0.18–0.57)]	?	Inadequate	N/A	N/A	N/A
	Jamal 2006 [31]	HGS was not associated with spine and non-spine fractures [OR 0.64 (95% CI 0.30–1.30)]	+	Inadequate	N/A	N/A	N/A
	Abdulan 2020 [20]	HGS predicted malnutrition [AUC 0.766 (95% CI 0.658–0.854)]	?	Inadequate	N/A	N/A	N/A
	Nixon 2019 [19]	HGS correlated with frailty [correlation coefficient: –0.62 (95% CI –0.73 to –0.48)]	+	Very good	N/A	N/A	N/A

Table 2: Continued.

		Construct validity properties			Responsiveness properties		
Physical function measure	1st author, year	Comparison with other measures; predictive validity; correlation coefficients	Quality criteria rating[a]	COSMIN score	Criterion or construct approach	Quality criteria rating[a]	COSMIN score
	Matos 2014 [22]	Low HGS associated with mortality [HR 2.68 (95% CI 1.53–4.71)] in total cohort. Also in males [<28.3 kg: HR 3.35 (95% CI 1.67–6.73)] and females [21.5 kg: HR 1.10 (95% CI 0.49–2.46)]	+	Adequate	N/A	N/A	N/A
	Wilkinson 2021 [23]	Low HGS (based on EWGSOP criteria) associated with all-cause mortality [HR 1.33 (1.07–1.66)] and with ESKD [HR 2.08 (1.53–2.82)]	+	Inadequate	N/A	N/A	N/A
	Xu 2020 [24]	1 kg increase in HGS associated with lower risk of all-cause mortality [HR 0.97 (0.96–0.99)]; HGS <24.5 kg associated with increases risk of mortality [HR 1.96 (1.35–2.84)]	?	Inadequate	N/A	N/A	N/A
	Kittiskulnam 2017 [25]	HGS associated with mortality (C statistic: 0.68) NRI%: 32.0 (95% CI 6.8–59.2); HR 0.67 (95% CI 0.47–0.94, C: 0.74) Weakness (defined as HGS <26 and <16 kg in males and females) associated with mortality [C statistic: 0.66, NRI%: 33.7 (95% CI 9.8–62.7); HR 1.68 (95% CI 1.01–2.79), C: 0.73]	?	Inadequate	N/A	N/A	N/A
	Lopes 2018 [26]	HGS associated with mortality [HR 2.58 (95% CI 1.73–3.85)]	?	Inadequate	N/A	N/A	N/A
SPPB	Nastasi 2018 [45]	SPPB score ≤10 predicts 5-year all-cause mortality [OR 2.30 (95% CI 1.12–4.74)] 1-year mortality was 4.1% in those with SPPB score ≤10 (vs 1.5% in >10); 5-year mortality was 20.6% in those with SPPB score ≤10 (vs 4.5% in >10); C statistic: 0.78 1 point ↓ in SPPB associated with 1.19 fold ↑ risk of post Tx mortality (95% CI 1.09–1.30)	+	Adequate	N/A	N/A	N/A
	Nastasi 2018 [46]	SPPB score ≤10 vs >10 predicts length of stay >14 days at time of Tx [OR 1.90 (95% CI 1.23–2.94)]	+	Adequate	N/A	N/A	N/A
	Nixon 2019 [19]	SPPB correlated with frailty [r = –0.66 (95% CI –0.78 to –0.51)]	+	Very good	N/A	N/A	N/A

Table 2: Continued.

Physical function measure	1st author, year	Construct validity properties: Comparison with other measures; predictive validity; correlation coefficients	Construct validity properties: Quality criteria rating[a]	Construct validity properties: COSMIN score	Responsiveness properties: Criterion or construct approach	Responsiveness properties: Quality criteria rating[a]	Responsiveness properties: COSMIN score
	Wilkinson 2019 [76]	Postural stability correlated with SPPB (r = –0.674)	+	Doubtful	N/A	N/A	N/A
	Abdulan 2020 [20]	SPPB associated with malnutrition (via MNA) [AUC 0.668 (95% CI 0.554–0.770)]	?	Inadequate	N/A	N/A	N/A
	Brar 2021 [43]	Agreement between SPPB and frailty (Fried) is weak Cohen $\kappa = 0.43$ SPPB is not associated with choice of dialysis modality OR 1.86 (95% CI 0.99–3.5) Associated with mortality [HR 1.96 (95% CI 1.42–2.70)]	+	Doubtful	N/A	N/A	N/A
	Haugen 2020 [44]	SPPB impaired less chance of Tx listing vs nonimpaired [aHR 0.70 (95% CI 0.64–0.77), $P < .001$] After adjustment for participant demographic and health factors, SPPB impaired participants had nearly 1.6-fold higher risk of waitlist mortality [aSHR 1.56 (95% CI 1.18–2.06), $P = .002$] compared with nonimpaired Participants After adjustment for participant demographic and health factors, impaired participants were 16% less likely to undergo KT compared with nonimpaired [aIRR 0.84 (95% CI 0.73–0.98), $P = .02$]	+	Adequate	N/A	N/A	N/A
STS-5	Wilkinson 2019 [41]	STS-5 correlation with ISWT at repeated time points (r = 0.55 and r = 0.74) STS-5 poorly correlated with e1RM at repeated time points (r = 0.14 and r = 0.47)	+	Very Good	N/A	N/A	N/A
	Huang 2018 [50]	STS-5 correlates with STS-60, r = –0.90, $P = .001$ STS-5 does not correlate with isometric quadriceps muscle force r = –0.04 or torque r = –0.04	?	Doubtful	N/A	N/A	N/A
	Nastasi 2018 [45]	1 point ↓ in chair stand produced ↑ 1.28 fold risk of death (95% CI 1.02–1.60)	+	Adequate	N/A	N/A	N/A
	Zanotto 2020 [49]	STS-5 did not predict falls over 12 months [rate ratio 0.98 (95% CI 0.94–1.02)]	+	Adequate	N/A	N/A	N/A

Table 2: Continued.

Physical function measure	1st author, year	Construct validity properties: Comparison with other measures; predictive validity; correlation coefficients	Construct validity properties: Quality criteria rating[a]	Construct validity properties: COSMIN score	Responsiveness properties: Criterion or construct approach	Responsiveness properties: Quality criteria rating[a]	Responsiveness properties: COSMIN score
	Hsieh 2006 [36]	VO_{2max} negatively correlates with STS-5; regression coefficient –0.239, $P = .03$	+	Adequate	N/A	N/A	N/A
	Brar 2021 [43]	Associated with mortality [HR 2.52 (95% CI 1.61–3.94)]; associated with choice of in-centre HD [OR 2.61 (95% CI 1.35–5.09)]	+	Doubtful	N/A	N/A	N/A
STS-10	Bohannon 1995 [52]	STS-10 were correlated with the LEMP measures [0.568 (95% CI 0.43–0.68)]	?	Inadequate	N/A	N/A	N/A
STS-30	Figueiredo 2021 [51]	STS-30 and exercise capacity ($r = 0.724$; $P < .001$) (R^2 adjusted for sex = 0.33; b = 40.5 ± 4.7; $P < .001$)	+	Adequate	N/A	N/A	N/A
STS-60	Cheng 2020 [32]	STS-60 correlates with 6MWT, $R^2 = 0.49$ Per 5 reps fewer, waitlist removal or death HR 1.53 (95% CI 1.43–1.75) Per 5 reps lower, death HR 1.23 (95% CI 0.92–1.66) Per 5 reps fewer, transplantation HR 0.80 (95% CI 0.71–0.89)	+	Adequate	N/A	N/A	N/A
TUG	Jamal 2006 [31]	TUG associated with fractures [OR 0.14 (95% CI 0.11–0.23)]	+	Inadequate	N/A	N/A	N/A
	Zanotto 2020 [49]	TUG did not predict falls over 12 months [rate ratio 1.08 (95% CI 0.96–1.21)]	+	Adequate	N/A	N/A	N/A
VO_{2peak}	Kohl 2012 [34]	VO_{2peak} not predictive of mortality [OR 1.04 (95% CI 0.197–1.13)]	?	Inadequate	N/A	N/A	N/A
	Chakkera 2018 [38]	VO_{2peak} >17 mL/kg/min has a NPV of 95% ($P = .0481$) for future cardiovascular events and NPV 88.5% (0.0496) mortality in CKD patients evaluated for kidney transplant	+	Inadequate	N/A	N/A	N/A
	Chakkera 2022 [39]	Per one unit increase get 9% reduction mortality HR 0.91 (95% CI 0.81–1.02); $P = .105$ Amongst low risk Tx patients, VO_{2peak} <12 mL/kg/min higher risk mortality vs >15 mL/kg/min, $P = .0445$.	+	Adequate	N/A	N/A	N/A
	Hsieh 2006 [36]	VO_{2max} correlation with 6MWT, $r = 0.62$, $P = 0.008$	+	Adequate	N/A	N/A	N/A
	Padilla 2008 [35]	VO_{2peak} was correlated with 6MWT ($r = 0.703$)	+	Inadequate	N/A	N/A	N/A

Table 2: Continued.

Physical function measure	1st author, year	Construct validity properties: Comparison with other measures; predictive validity; correlation coefficients	Construct validity properties: Quality criteria rating[a]	Construct validity properties: COSMIN score	Responsiveness properties: Criterion or construct approach	Responsiveness properties: Quality criteria rating[a]	Responsiveness properties: COSMIN score
6MWT	Hsieh 2006 [36]	N/A	N/A	N/A	N/A	N/A	N/A
	Padilla 2008 [35]	VO_{2peak} was correlated with 6MWT ($r = 0.703$)	+	Inadequate	N/A	N/A	N/A
	Kohl 2012 [34]	6MWT predictive of mortality [OR 0.53 (95% CI 0.37–0.74)]	?	Inadequate	N/A	N/A	N/A
	Torino 2014 [33]	6MWT predictive of a composite end point consisting of death/hospitalization/CV events [HR 0.94 (95% CI 0.91–0.98)], all-cause mortality [HR 0.89 (95% CI 0.84–0.94)], hospitalizations [HR 0.96 (95% CI 0.92–0.99)], CV events [HR 0.96 (95% CI 0.91–1.01)] An ↑ 20 m led ↓ all-cause mortality by 12% and reduced risk hospitalization by 4% over 3.3 years	?	Very good	N/A	N/A	N/A
	Jamal 2006 [31]	6MWT associated with fracture risk [OR per 1 SD increase: 0.10 (95% CI 0.03–0.36)] Discrimination between patients with and without fractures [AUC 0.84 (95% CI 0.71–0.97)]	+	Inadequate	N/A	N/A	N/A
	Shi 2017 [30]	6MWT ≤396 m associated with technique failure from peritonitis ($P = .036$)	+	Inadequate	6MWT ≤396 m associated with mortality (log rank 4.983, $P = .0026$)	+	Inadequate
	Cheng 2020 [32]	Per 50 m lower, waitlist removal or death HR 1.42 (95% CI 1.30–1.56) Per 50 m lower, death HR 1.29 (95% CI 1.06–1.58) Per 50 m lower, transplantation HR 0.80 (95% CI 0.72–0.89)	+	Adequate	N/A	N/A	N/A
ISWT	Wilkinson 2019 [41]	N/A	N/A	N/A	Correlates with VO_{2peak}, $r = 0.73$, $P < .001$	+	Very good
	Greenwood 2019 [57]	Change in ISWT >50 m with rehab associated with reduced mortality and CV morbidity [HR 0.6 (95% CI 0.36–0.98), $P = .04$]	+	Inadequate	N/A	N/A	N/A
	Watson 2020 [58]	Development of ESKD: 1.02 (1.02–1.05) Time to first hospitalization: 0.99 (0.98–1.01) All-cause mortality: 1.00 (0.99–1.00)	+	Adequate	N/A	N/A	N/A

[a] Supplementary Material 2 provides a summary as to how this ranking was obtained.

N/A: non-applicable; 95% CI: 95% confidence interval; HR: hazard ratio; aHR: adjusted hazard ratio; aSHR: adjusted subhazard ratio; aIRR: adjusted incidence rate ratio; EWGSOP: European Working Group on Sarcopenia in Older People; ESKD: end-stage kidney disease; MNA: Mini Nutritional Assessment; OR: odds ratio; Tx: transplant; CV: cardiovascular; NPV: negative predictive value; KT: kidney transplant; NRI%: net reclassification index; LEMP; lower extremity muscle performance.

Table 3: Summary of results on reliability and measurement error for physical function measures.

Physical function measure	1st author, year	Reliability properties			Measurement error properties		
		ICC (Kappa)	Quality criteria rating[a]	COSMIN score	SEM%, LOA, MDC90/MDC95, CV%	Quality criteria rating[a]	COSMIN score
Gait speed	Johnstone 2021 [48]	ICC 0.886 (0.717 to 0.954)	+	Adequate	N/A	N/A	N/A
	Jimenez 2022 [29]	ICC 0.863 (0.733–0.933)	+	Adequate	SEM 0.1; MDC90 0.3	?	Adequate
	Macagnan 2019 [27]	ICC 0.89–0.95	+	Inadequate	N/A	N/A	N/A
HGS	Jimenez 2022 [29]	ICC 0.945 (0.887–0.973)	+	Adequate	SEM 2.3; MDC90 5.5	?	Adequate
	Segura-Orti 2011 [28]	ICC 0.96 (95%CI 0.88-0.99)	+	Very good	SEM 1.5 kg; MDC90 3.4 kg	?	Very good
ISWT	Hadjiioannou 2020 [56]	ICC 0.95 (95% CI 0.90–0.97) Subgroup analysis: pre-dialysis: ICC 0.97 (95% CI 0.91–0.99); dialysis: ICC 0.80 (95% CI 0.35–0.95); Tx: ICC 0.98 (95% CI 0.90–1.00)	+	Adequate	SEM 34.09 m Subgroup analysis: pre-dialysis 32.06 m; dialysis 47.26 m; Tx: 20.82 m MDC90: 79.6 m Subgroup analysis: pre-dialysis: 74.81 m; dialysis: 110.27 m; Tx: 48.59 m	–	Adequate
	Wilkinson 2019 [41]	ICC 0.973 (95% CI 0.950–0.986)	+	Doubtful	SEM 7.1 m; SEM% 1.8%; MDC individual: 20 m; MDC group: 3 m	+	Doubtful
6MWT	Overend 2010 [37]	ICC 0.93 (95% CI 0.82–0.97)	+	Adequate	SEM 28 m; SEM% 21%; MDC95 77 m	?	Adequate
	Jimenez 2022 [29]	ICC 0.932 (0.861–0.963)	+	Adequate	SEM 29.5; MDC90 68.8	?	Adequate
	Watanabe 2011 [18]	Overall ICC 0.91; post-Tx ICC 0.89; dialysis ICC 0.90	+	Doubtful	SEM 21.8 m; MDC 60.5 m	?	Doubtful
	Segura-Orti 2011 [28]	ICC 0.94 (95% CI 0.89–0.97)	+	Very good	SEM 28.4 m; MDC90 66.3m	?	Very good
SPPB	De Villar 2018 [47]	ICC 0.94 (95% CI 0.91– 0.97)	+	Adequate	SEM 0.72 points (95% CI 0.56–0.91); MDC90 1.7 points (95% CI 1.3–2.1)	+	Adequate
	Johnstone 2021 [48]	ICC 0.906 (0.772–0.961)	+	Adequate	N/A	N/A	N/A
	Jimenez 2022 [29]	ICC 0.947 (0.891–0.974)	+	Adequate	SEM 0.4; MDC90 0.9	?	Adequate
STS-5	Wilkinson 2019 [41]	ICC 0.676 (95% CI 0.468–0.813)	–	Doubtful	SEM 2.7 s; SEM% 24.5%; MDC individual: 7.5 s; MDC group: 1.2 s	?	Doubtful
	Johnstone 2021 [48]	ICC 0.952 (0.882–0.981)	+	Adequate	N/A	N/A	N/A
	Huang 2018 [50]	ICC 0.997 (95% CI 0.994–0.998) inter-tester ICC 0.970 (0.948–0.983) intra-tester	+	Inadequate	Inter-tester SEM 0.3 s; MDC95 0.8 s Intra-tester SEM 0.9 s; MDC95 2.5 s	+	Inadequate

Table 3: Continued.

Physical function measure	1st author, year	Reliability properties			Measurement error properties		
		ICC (Kappa)	Quality criteria rating[a]	COSMIN score	SEM%, LOA, MDC90/MDC95, CV%	Quality criteria rating[a]	COSMIN score
	Koufaki 2002 [42]	N/A	N/A	N/A	CV% 15.1%	?	Inadequate
STS-10	Bohannon 1995 [52]	ICC 0.843 (95% CI 0.68–0.93)	+	Very good	N/A	N/A	N/A
	Jimenez 2022 [29]	ICC 0.861(0.729–0.931)	+	Adequate	SEM 3.6; MDC90 8.5	?	Adequate
	Segura-Orti 2011 [28]	ICC 0.88 (95%CI 0.78–0.94)	+	Very good	SEM 3.6 s; MDC90 8.4s	?	Very good
STS-30	Macdonald 2005 [73]	N/A	N/A	N/A	SEM 0.47 reps; CV% 5.5%	?	Doubtful
	Overend 2010 [37]	ICC 0.93 (95% CI 0.58–0.98)	+	Adequate	SEM 0.9 reps; SEM% 10.3%; MDC95 2.6 reps	?	Adequate
	Sutcliffe 2018 [55]	↓ in the difference of RR: baseline to week 12 [0.83 (95% CI 0.81–0.86)]; baseline to Week 24 [0.82 (95% CI 0.80–0.85)]	?	Inadequate	N/A	N/A	N/A
	Figueiredo 2021 [51]	ICC 0.93 (0.86–0.96)	+	Adequate	SEM 0.91; MDC90 2.1; MDC95 2.5;	?	Adequate
STS-60	Wilkinson 2019 [53]	N/A	N/A	N/A	N/A	N/A	N/A
	Wilkinson 2019 [41]	ICC 0.927 (95% CI 0.866–0.961)	+	Doubtful	SEM 1.3 reps; SEM% 5%; MDC individual: 4 reps; MDC group: 1 rep	?	Doubtful
	Hadjiioannou 2020 [56]	ICC 0.89 (95% CI 0.80–0.95) Subgroup analysis: pre-dialysis: ICC 0.90 (95% CI 0.75–0.96); dialysis: ICC 0.84 (95% CI 0.48–0.96); Tx: ICC 0.88 (95% CI 0.62–0.97)	+	Adequate	SEM 14.1% Subgroup analysis: pre-dialysis: 15.3%; dialysis: 11.9%; Tx: 12.7% MDC90 7.0 reps Subgroup analysis: pre-dialysis 8.32 reps; dialysis 5.67 reps; Tx 5.74 reps	?	Adequate
	Koufaki 2002 [42]	N/A	N/A	N/A	CV% 12.8%	?	Inadequate
	Segura-Orti 2011 [28]	ICC 0.97 (95% CI 0.94–0.98)	+	Very good	SEM 1.7 rep; MDC90 4 rep	?	Very good
	Jimenez 2022 [29]	ICC 0.925 (95% CI 0.848–0.963)	+	Adequate	SEM 2.3; MDC90 5.4	?	Adequate

Table 3: Continued.

Physical function measure	1st author, year	Reliability properties			Measurement error properties		
		ICC (Kappa)	Quality criteria rating[a]	COSMIN score	SEM%, LOA, MDC90/MDC95, CV%	Quality criteria rating[a]	COSMIN score
Timed Up & Go	Mesquita 2013 [54]	Between 1st and 2nd trial: ICC 0.91 (95% CI 0.78–0.86); Kappa: 0.85	+	Very good	Between 1st and 2nd trial: SEM 1.24 s; LOA –3.82 to 2.05; MDC95 3.44 s (29%)	?	Very good
		Between 2nd and 3rd trial: ICC 0.96 (95% CI 0.94–0.98); Kappa: 0.49	+	Very good	Between 2nd and 3rd trial: SEM 0.78 s; LOA –2.25 to 1.94; MDC95 2.16 s (19%)	?	Very good
	Sutcliffe 2018 [55]	Rate deterioration: baseline to week 12 [HR 0.71 (95% CI 0.53–0.95)]; baseline to Week 24 [HR 0.56 (95% CI 0.39–0.80)]	?	Inadequate	N/A	N/A	N/A
	De Villar 2018 [47]	ICC 0.96 (95% CI 0.94–0.98)	+	Adequate	SEM 1.24 s (95% CI 0.56–0.91); MDC90 2.9 s (95% CI 2.2–3.7)	?	Adequate
	Hadjiioannou 2020 [56]	ICC 0.96 (95% CI 0.92–0.98) Subgroup analysis: pre-dialysis: ICC 0.99 (0.96–0.99); dialysis: ICC 0.71 (0.20–0.92); Tx: ICC 0.98 (0.92–0.99)	+	Adequate	SEM 1.24 s; MDC90 2.9 s Subgroup analysis: pre-dialysis: SEM 0.82 s; MDC90 1.92 s; Tx: SEM 2.26 s; MDC90 5.28 s	?	Adequate
	Jimenez 2022 [29]	ICC 0.945 (0.887–0.973)	+	Adequate	SEM 0.9; MDC90 2.1	?	Adequate
VO_{2peak}	Koufaki 2002 [42]	N/A	N/A	N/A	CV% 4.7%; VO_2 at VT: CV% 6.6%	?	Inadequate
	Koufaki 2001 [40]	ICC 0.99 (95% CI 0.93–0.99)	+	Adequate	SEM 1.01 mL/kg/min; CV% 5%; LOA ± 2.42	?	Very good
	Wilkinson 2019 [53]	N/A	N/A	N/A	N/A	N/A	N/A
	Wilkinson 2019 [41]	ICC 0.866 (95% CI 0.755–0.929)	+	Doubtful	SEM 1 mL/kg/min; SEM% 5.1%; MDC individual: 2.8 mL/kg/min; MDC group: 0.5 mL/kg/min	?	Doubtful

LOA: limits of agreement; CV: coefficient of variation; N/A: non-applicable; Tx: transplant; HR: hazard ratio; VT: ventilatory threshold.
[a]Supplementary Material 2 provides a summary as to how this ranking was obtained.

1.8 mL/kg/min [41]. The summarized quality of evidence is low for reliability and measurement error. The quality of evidence was low for construct validity (Supplementary Material 4).

Short physical performance battery

Of 10 studies, 7 reported on validity. Lower Short physical performance battery (SPPB) performance was predictive of mortality in CKD [43, 44] and post-transplantation [45], of lower likelihood of being transplanted [44], a longer length of stay at time of transplant [46], frailty in a mixed population of CKD and HD [19], and malnutrition in CKD [20]. The SPPB showed excellent reliability (ICC values 0.906–0.947) in two studies in HD and CKD [29, 47, 48]. In two studies, the SEM ranged between 0.4 and 0.72 in HD patients [29, 47] and the MDC90 was 1.7 [47]. The summarized quality of evidence was high for reliability, measurement error and construct validity. Responsiveness was not assessed (Supplementary Material 4).

Gait speed

The predictive validity of gait speed was reported in six studies. Slower gait speed was associated with an increased risk of mortality in people on dialysis [25, 43], but not post-transplant [45]. Gait speed was associated with frailty in a mixed CKD population [19]. Gait speed was not predictive of falls in HD patients [49]. The reliability was examined in two studies ($n = 51$) [29, 48], and showed consistent reliability properties (ICC values between 0.863 and 0.886). One study reported measurement error with an SEM of 0.1 m/s [29]. The summarized quality of evidence is moderate for validity and reliability but very low for measurement error (Supplementary Material 4).

Sit-to-stand tests

Four variations of the Sit-to-stand (STS) test were reported. The construct validity of STS-5, reported in six studies, demonstrated that it was correlated with other physical function measures [Incremental shuttle walk test (ISWT) [41], STS-60 [50] and VO_{2peak} [36]] and mortality [43, 45], but not falls [49]. STS-60 was correlated with 6MWT and transplant wait list removal, as well as death [32]. STS-30 correlated with exercise capacity [51] and STS-10 with lower extremity muscle power [52]. Reliability was reported in four studies for STS-5, five studies for STS-10, six studies for STS-30 and six studies for STS-60. The MCID for STS-5 ranged between –4.2 and –2.5 s. The MCID for STS-60 was three repetitions [53]. The quality of evidence for reliability of STS-5 was moderate, low for measurement error and high for construct validity. The STS-10 had high quality of evidence for reliability, moderate for measurement error and very low for construct validity. The STS-30 had high quality evidence for reliability, moderate for measurement error and low for construct validity. The STS-60 had high quality evidence for reliability, moderate for measurement error and moderate for construct validity. The clinimetric property of responsiveness was not reported in any of the STS variations (Supplementary Material 4).

Timed-up-and-go

The validity of the Timed-up-and-go (TUG) was assessed in two HD studies [31, 49]. The TUG was predictive of fractures in HD patients but not falls over 12 months. Reliability was reported in five studies with consistent SEM ranging from 0.78 to 1.24 s [29, 47, 54–56]. The quality of evidence was high for reliability and measurement error and moderate for validity due to bias (Supplementary Material 4).

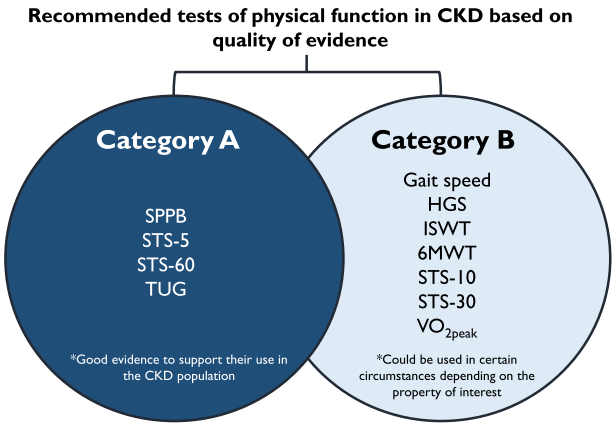


Figure 3: Recommended Category A and B tests of physical function.

Incremental shuttle walk test

Validity was assessed in three studies, one of which reported that ISWT correlated with the gold standard VO_{2peak} [41]. Improvements in the distance walked during the ISWT were associated with reduced mortality and cardiovascular morbidity risk [57] whereas worsening of ISWT performance was not associated with development of end-stage kidney disease [58]. Two small studies reported on the reliability (ICC values ranging from 0.950 to 0.973) and measurement error (SEM ranging from 7.1 to 34.1 m) of the ISWT [41, 56]. The MDC was determined between 20.0 and 79.6 m. The quality of evidence was low for both reliability and measurement error, and moderate validity and criterion (Supplementary Material 4).

Best evidence synthesis: levels of evidence

A summary of best evidence synthesis was derived from information in Table 2 and 3 for each of the most frequent performance tests shown in Table 4 (and tests, all graded B or C, not included in main synthesis in Supplementary Material 7). Given the large variety of performance-based measures, results were rarely combined. Based on the quality of evidence across the studies SPPB, STS-5, STS-60 and TUG are recommended as Category A tests with good evidence to support their use. Gait speed, HGS, ISWT, 6MWT, STS-10, STS-30,and VO_{2peak} are Category B and could be used in certain circumstances depending on the property of interest (Fig. 3).

DISCUSSION

In this review, we systematically reviewed the literature to evaluate the clinimetric properties of physical function tools amongst the CKD population. This represents the initial step in standardizing a set of tools available to measure physical function in research and clinical settings for CKD. Based on our findings, the SPPB, TUG, STS-5 and STS-60 have good evidence to support their use in CKD. It is important to note that most studies included in this review were conducted in the HD population, and of the many studies exploring objective measures of physical function, very few provided information regarding validity or reliability.

Table 4: Summary of the quality of the evidence for clinimetric properties of the physical function measures.

	Reliability		Measurement error		Validity		Responsiveness		
Physical function measure	Level	GRADE	Level	GRADE	Level	GRADE	Level	GRADE	Category
Gait speed	+	Mod	?	V.low	+	Mod	?	N/A	B
HGS	–	Low	+	Mod	–	Low	?	N/A	B
ISWT	–	Low	–	Low	+	Mod	?	N/A	B
6MWT	+	High	+	High	–	Low	–	V.low	B
SPPB	**+**	**High**	**+**	**High**	**+**	**High**	**?**	**N/A**	**A**
STS-5	**–**	**Mod**	**?**	**Low**	**+**	**High**	**?**	**N/A**	**A**
STS-10	+	High	–	Mod	?	V.low	?	N/A	B
STS-30	+	High	–	Mod	?	Low	?	N/A	B
STS-60	**+**	**High**	**±**	**Mod**	**–**	**Mod**	**?**	**N/A**	**A**
TUG	**+**	**High**	**+**	**High**	**–**	**Mod**	**?**	**N/A**	**A**
VO_{2peak}	?	Low	?	Low	–	Low	?	Low	B

The Category A physical function measures are highlighted in bold.
Category criteria adapted from COSMIN manual (step 9): A: outcome measures with evidence for at least moderate quality for reliability and validity AND at least low evidence for measurement error or responsiveness; B: outcome measures categorized not in A or C; and C:outcome measures with only indeterminate or unknown evidence for properties.
Criteria outlining the evaluation of quality of results across studies is presented in Supplementary Material 3; briefly: (+) denotes a strong level of evidence; (–) limited with only 1 study of fair methodological quality; (±) inconsistent/conflicting findings; and (?) denotes an unknown level of evidence (no evidence available or only studies of poor methodological quality).

The SPPB demonstrated high quality of evidence for reliability, measurement error and construct validity amongst transplant, CKD and dialysis patients. SPPB score predicted mortality [43, 44], frailty [19] and a lower likelihood of being transplanted [44]. The finding that SPPB predicts mortality is consistent with other chronic disease populations (e.g. myocardial infarction [59], chronic lung disease [60] and elderly populations [61]). One limitation of the SPPB is its ceiling effect in well-functioning CKD patients, and evidence on measurement error is poorly reported. Whilst not reported here, the SPPB can also have large floor effects in very poor functioning patients [62]. Floor and ceiling effects may limit how sensitive the test is to change and mean efforts to improve SPPB performance are often difficult. The SPPB is based on a summary score of each of its three components: standing balance, gait speed and STS-5. The latter two tests are commonly reported in CKD and the continuous nature may allow for better discrimination in patients with greater functional ability. Indeed, individually, both gait speed and STS-5 performance are associated with mortality across CKD [25, 43, 45], and both tests demonstrated moderate to high levels of evidence for reliability and validity.

Other well-performing tests reviewed include the TUG and STS-60. The TUG evaluates an individual's dynamic balance and mobility and is recommended for its good reliability and measurement error properties, in people on HD and with CKD [56]. The STS-60 showed a high level of quality evidence for reliability, moderate for measurement error and moderate construct validity showing correlation with 6MWT and mortality in a large group of mixed dialysis patients [32]. We found that several tests, such as the Category B tests, could be used in certain circumstances depending on the property of interest. Tests of gait speed, VO_{2peak}, 6MWT, STS-10, STS-30, ISWT and HGS all demonstrated a lower and varied quality level of evidence for clinimetric properties. For example, the ISWT demonstrated good validity when compared with VO_{2peak} (the gold standard of cardiorespiratory fitness), however the quality of evidence was low for both reliability and measurement error due to small sample size and risk of bias.

Whilst there is one other review which highlighted the heterogeneity of physical function tests in CKD [8], there are no other systematic reviews to explore clinimetric and measurement properties of physical function tools to compare our findings. However, a review of older community-dwelling persons (>60 years) found the SPPB was the measurement with the best reliability, validity and responsiveness [63]. Similar support for SPPB measurement properties is present in older adults during hospitalization [64]. A review of the clinimetric properties of muscle function tests in individuals with cystic fibrosis found good support for STS-60, although it lacked validity against quadriceps muscle strength [65]. STS tests and TUG had good level of support in a COSMIN review of performance-based measures in hip and knee osteoarthritis [66]. A COSMIN review found STS-5 and TUG to be promising for patients with chronic low back pain [67]. However, more research on the measurement error and responsiveness of these tests in CKD is needed to be able to fully recommend them as outcome measures in research and clinical practice.

Strengths and limitations

We used a systematic approach (i.e. the COSMIN methodology) for assessing the quality of the included articles. We compiled results from a heterogeneous population across all stages of CKD. This allowed us to identify suitable and stable measures for use across CKD stages; this is highly relevant in clinical practice and research settings for monitoring and classification purposes. However, most studies were limited by small sample sizes and lack of an *a priori* established hypothesis when exploring validity properties and responsiveness, thus reducing the quality of evidence that can lead to robust recommendations. We were unable to provide a meta-analysis due to the heterogeneity and the low quality of the studies. Whilst COSMIN provides a comprehensive, step-by-step and standardized framework to assess measurement properties, it is not exhaustive. The COSMIN tool has not yet developed a rating scale to evaluate interpretability or feasibility, but we believe these are important clinical considerations for functional testing. By adhering to COSMIN's guidance lowest scores method, included studies are evaluated with perhaps overly stringent criteria, particularly in relation to construct validity. Reduced scores because of unreported or unclear information, combined with instructions that

"lowest score counts" as the overall score, led to many tests being reported as either "doubtful" or "inadequate." The tools recommended as Category A or B are ones that do not require extensive training or specialty equipment (with the exception of HGS and VO_{2peak}). However, our recommendations do not take into consideration additional criteria beyond the methodological scoring we have applied, such as population acceptability and equity—future primary research in this area should explore these considerations.

The SPPB, STS-5, STS-60 and TUG demonstrate the best properties across the spectrum of CKD. However, knowledge gaps regarding measurement properties remain for many tests. To improve the quality of evidence for these measures, COSMIN guidelines should be followed for the design and reporting of studies investigating measurement properties of physical function outcomes in people with CKD.

SUPPLEMENTARY DATA

Supplementary data are available at *ckj* online.

ACKNOWLEDGEMENTS

The authors thank Helen Lee Robertson, MLIS from the Health Sciences Library, University of Calgary for providing the search results.

FUNDING

The authors acknowledge Satellite Healthcare, Inc., San Jose, California for their unrestricted grant to support the Global Renal Exercise Network (GREX).

AUTHORS' CONTRIBUTIONS

J.M.M.: conceptualization; abstract screening, full text extraction; methodology; writing—original draft. O.H.: abstract screening, full text extraction; methodology; writing—original draft. C.J.L.: abstract screening, full text extraction; methodology; writing—editing and approval of final draft. S.T.: conceptualization; abstract screening, full text extraction; methodology; writing—editing and approval of final draft. K.W.-F.: abstract screening, full text extraction; methodology; writing—editing and approval of final draft. P.K.: abstract screening, full text extraction; methodology; writing—editing and approval of final draft. C.B.: conceptualization; abstract screening, full text extraction; methodology; writing—editing and approval of final draft. T.W.: abstract screening, full text extraction; methodology; writing—original draft.

DATA AVAILABILITY STATEMENT

No new data were generated or analysed in support of this research.

CONFLICT OF INTEREST STATEMENT

T.J.W. is supported by the NIHR Applied Research Collaboration East Midlands. C.J.L. is salary supported by the Stoneygate Trust as part of the Leicester Kidney Lifestyle Research Programme Grant and is supported by the National Institute for Health Research Leicester Biomedical Research Centre (BRC).

REFERENCES

1. Koufaki P, Mercer T. Assessment and monitoring of physical function for people with CKD. *Adv Chronic Kidney Dis* 2009;**16**:410–9. https://doi.org/10.1053/j.ackd.2009.08.010
2. Wilkinson TJ, Macrae JM, Thompson S *et al*. Making the case for standardized outcome measures in exercise and physical activity research in chronic kidney disease. *Kidney Dial* 2023;**3**:219–28. https://doi.org/10.3390/kidneydial3020020
3. MacKinnon HJ, Wilkinson TJ, Clarke AL *et al*. The association of physical function and physical activity with all-cause mortality and adverse clinical outcomes in nondialysis chronic kidney disease: a systematic review. *Ther Adv Chronic Dis* 2018;**9**:209–26. https://doi.org/10.1177/2040622318785575
4. Koufaki P. Assessment of function limitations in people with chronic kidney disease for implementation in clinical practice. *Kidney Dial* 2022;**2**:234–44. https://doi.org/10.3390/kidneydial2020024
5. Iman Y, Harasemiw O, Tangri N. Assessing physical function in chronic kidney disease. *Curr Opin Nephrol Hypertens* 2020;**29**:346–50. https://doi.org/10.1097/MNH.0000000000000594
6. Bakker EA, Zoccali C, Dekker FW *et al*. Assessing physical activity and function in patients with chronic kidney disease: a narrative review. *Clin Kidney J* 2021;**14**:768–79. https://doi.org/10.1093/ckj/sfaa156
7. Baker LA, March DS, Wilkinson TJ *et al*. Clinical practice guideline exercise and lifestyle in chronic kidney disease. *BMC Nephrol* 2022;**23**:1–36. https://doi.org/10.1186/s12882-021-02618-1
8. Jegatheesan DK, Modderman R, Krishnasamy R *et al*. A systematic review of scope and consistency of outcome measures for physical fitness in chronic kidney disease trials. *Kidney Int Rep* 2021;**6**:1280–8. https://doi.org/10.1016/j.ekir.2021.02.010
9. Prinsen CAC, Mokkink LB, Bouter LM *et al*. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;**27**:1147–57. https://doi.org/10.1007/s11136-018-1798-3
10. Mokkink LB, Prinsen CAC, Bouter LM *et al*. The COnsensus-based standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther* 2016;**20**:105–13. https://doi.org/10.1590/bjpt-rbf.2014.0143
11. Page MJ, McKenzie JE, Bossuyt PM *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;**372**.
12. Moher D, Liberati A, Tetzlaff J *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg* 2010;**8**:336–41. https://doi.org/10.1016/j.ijsu.2010.02.007
13. Morrison A, Polisena J, Husereau D *et al*. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *Int J Technol Assess Health Care* 2012;**28**:138–44. https://doi.org/10.1017/S0266462312000086
14. Dobrescu AI, Nussbaumer-Streit B, Klerings I *et al*. Restricting evidence syntheses of interventions to English-language publications is a viable methodological shortcut for most medical topics: a systematic review. *J Clin Epidemiol* 2021;**137**:209–17. https://doi.org/10.1016/j.jclinepi.2021.04.012
15. Mokkink LB, Boers M, van der Vleuten CPM *et al*. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement

instruments: a Delphi study. *BMC Med Res Method* 2020;**20**:1–13. https://doi.org/10.1186/s12874-020-01179-5
16. Mokkink LB, de Vet HCW, Prinsen CAC *et al*. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* 2018;**27**:1171–9. https://doi.org/10.1007/s11136-017-1765-4
17. Mokkink LB, Prinsen CA, Patrick DL *et al*. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User Manual* [Internet] 2018. Available from: https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf (27 April 2023, date last accessed).
18. Watanabe FT, Koch VHK, Juliani RCTP *et al*. Six-minute walk test in children and adolescents with renal diseases: tolerance, reproducibility and comparison with healthy subjects. *Clinics* 2016;**71**:22–7. https://doi.org/10.6061/clinics/2016(01)05
19. Nixon AC, Bampouras TM, Pendleton N *et al*. Diagnostic accuracy of frailty screening methods in advanced chronic kidney disease. *Nephron* 2019;**141**:147–55. https://doi.org/10.1159/000494223
20. Abdulan IM, Stefaniu R, Mastaleru A *et al*. Cut-off values of anthropometric measurements, handgrip strength, physical activity and geriatric scores for the malnutrition risk among older patients with chronic kidney disease. *Rev Med Chir Soc Med Nat lasi* 2020;**124**:19–26.
21. Isoyama N, Qureshi AR, Avesani CM *et al*. Comparative associations of muscle mass and muscle strength with mortality in dialysis patients. *Clin J Am Soc Nephrol* 2014;**9**:1720–8. https://doi.org/10.2215/CJN.10261013
22. Matos CM, Silva LF, Santana LD *et al*. Handgrip strength at baseline and mortality risk in a cohort of women and men on hemodialysis: a 4-year study. *J Ren Nutr* 2014;**24**:157–62. https://doi.org/10.1053/j.jrn.2013.12.005
23. Wilkinson TJ, Miksza J, Yates T *et al*. Association of sarcopenia with mortality and end-stage renal disease in those with chronic kidney disease: a UK Biobank study. *J Cachexia Sarcopenia Muscle* 2021;**12**:586–98. https://doi.org/10.1002/jcsm.12705
24. Xu X, Yang Z, Ma T *et al*. The cut-off values of handgrip strength and lean mass index for sarcopenia among patients on peritoneal dialysis. *Nutr Metab* 2020;**17**:1–11.
25. Kittiskulnam P, Chertow GM, Carrero JJ *et al*. Sarcopenia and its individual criteria are associated, in part, with mortality among patients on hemodialysis. *Kidney Int* 2017;**92**:238–47. https://doi.org/10.1016/j.kint.2017.01.024
26. Lopes MB, Silva LF, Dantas MAS *et al*. Sex-age-specific handgrip strength and mortality in an incident hemodialysis cohort: the risk explained by nutrition and comorbidities. *Int J Artif Organs* 2018;**41**:825–32. https://doi.org/10.1177/0391398818793088
27. Macagnan FE, Baroni BM, Cristofoli EZ *et al*. Acute effect of photobiomodulation therapy on handgrip strength of chronic kidney disease patients during hemodialysis. *Lasers Med Sci* 2019;**24**:835–40. https://doi.org/10.1007/s10103-018-2593-7
28. Segura-Ortí E, Martínez-Olmos F. Test-retest reliability and minimal detectable change scores for sit-to- stand-to-sit tests, the six-minute walk test, the one-leg heel-rise test, and handgrip strength in people undergoing hemodialysis. *Phys Ther* 2011;**91**:1244–52. https://doi.org/10.2522/ptj.20100141
29. Jiménez AJ, Bernabeu ET, Périz LA *et al*. Impact of measurement timing on reproducibility of testing among haemodialysis patients. *Sci Rep* 2022;**12**:1–12.
30. Shi Y, Zheng D, Zhang L *et al*. Six-minute walk test predicts all-cause mortality and technique failure in ambulatory peritoneal dialysis patients. *Nephrology* 2017;**22**:118–24. https://doi.org/10.1111/nep.12726
31. Jamal SA, Leiter RE, Jassal V *et al*. Impaired muscle strength is associated with fractures in hemodialysis patients. *Osteoporos Int* 2006;**17**:1390–7. https://doi.org/10.1007/s00198-006-0133-y
32. Cheng XS, Myers J, Han J *et al*. Physical performance testing in kidney transplant candidates at the top of the waitlist. *Am J Kidney Dis* 2020;**76**:815–25. https://doi.org/10.1053/j.ajkd.2020.04.009
33. Torino C, Manfredini F, Bolignano D *et al*. Physical performance and clinical outcomes in dialysis patients: a secondary analysis of the excite trial EXCITE working group. *Kidney Blood Press Res* 2014;**39**:205–11. https://doi.org/10.1159/000355798
34. Kohl L de M, Signori LU, Ribeiro RA *et al*. Prognostic value of the six-minute walk test in endstage renal disease life expectancy: a prospective cohort study. *Clinics* 2012;**67**:581–6. https://doi.org/10.6061/clinics/2012(06)06
35. Padilla J, Krasnoff J, DaSilva M *et al*. Physical functioning in patients with chronic kidney disease. *J Nephrol* 2008;**21**:550–9.
36. Hsieh RL, Lee WC, Chang CH. Maximal cardiovascular fitness and its correlates in ambulatory hemodialysis patients. *Am J Kidney Dis* 2006;**48**:21–7. https://doi.org/10.1053/j.ajkd.2006.03.081
37. Overend T, Anderson C, Sawant A *et al*. Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiother Can* 2010;**62**:122–8. https://doi.org/10.3138/physio.62.2.122
38. Chakkera HA, Angadi SS, Heilman RL *et al*. Cardiorespiratory fitness (peak oxygen uptake): safe and effective measure for cardiovascular screening before kidney transplant. *J Am Heart Assoc* 2018;**7**:1–6. https://doi.org/10.1161/JAHA.118.008662
39. Chakkera HA, Kaplan B, Budhiraja P *et al*. VO2peak prior to transplant differentiates survival post kidney transplant. *Clin Transplant* 2022;**36**:e14517. https://doi.org/10.1111/ctr.14517
40. Koufaki P, Naish PF, Mercer TH. Reproducibility of exercise tolerance in patients with end-stage renal disease. *Arch Phys Med Rehabil* 2001;**82**:1421–4. https://doi.org/10.1053/apmr.2001.26076
41. Wilkinson TJ, Xenophontos S, Gould DW *et al*. Test–retest reliability, validation, and "minimal detectable change" scores for frequently reported tests of objective physical function in patients with non-dialysis chronic kidney disease. *Phyiother Theory Pract* 2019;**35**:565–76. https://doi.org/10.1080/09593985.2018.1455249
42. Koufaki P, Mercer TH, Naish PF. Effects of exercise training on aerobic and functional capacity of end-stage renal disease patients. *Clin Physiol Funct Imaging* 2002;**22**:115–24. https://doi.org/10.1046/j.1365-2281.2002.00405.x
43. Brar RS, Whitlock RH, Komenda PVJ *et al*. Provider perception of frailty is associated with dialysis decision making in patients with advanced CKD. *Clin J Am Soc Nephrol* 2021;**16**:552–9. https://doi.org/10.2215/CJN.12480720
44. Haugen CE, Agoons D, Chu NM *et al*. Physical impairment and access to kidney transplantation. *Transplantation* 2020;**104**:367–73. https://doi.org/10.1097/TP.0000000000002778
45. Nastasi AJ, McAdams-DeMarco MA, Schrack J *et al*. Pre-kidney transplant lower extremity impairment and

post-kidney transplant mortality. *Am J Transplant* 2018;**18**:189–96. https://doi.org/10.1111/ajt.14430
46. Nastasi AJ, Bryant TS, Le JT *et al*. Pre-kidney transplant lower extremity impairment and transplant length of stay: a time-to-discharge analysis of a prospective cohort study. *BMC Geriatr* 2018;**18**:246–55. https://doi.org/10.1186/s12877-018-0940-y
47. de Villar LOP, Martínez-Olmos FJ, Junqué-Jiménez A *et al*. Test-retest reliability and minimal detectable change scores for the short physical performance battery, one-legged standing test and timed up and go test in patients undergoing hemodialysis. *PLoS One* 2018;**13**:1–16.
48. Johnstone LM, Roshanravan B, Rundell SD *et al*. Instrumented and standard measures of physical performance in adults with chronic kidney disease. *J Acute Care Phys Ther* 2022;**13**:110–8. https://doi.org/10.1097/JAT.0000000000000179
49. Zanotto T, Mercer TH, Van Der Linden ML *et al*. The relative importance of frailty, physical and cardiovascular function as exercise-modifiable predictors of falls in haemodialysis patients: a prospective cohort study. *BMC Nephrol* 2020;**21**:99. https://doi.org/10.1186/s12882-020-01759-z
50. Huang K, O'Connor E, Tuffnell R *et al*. Reliability and validity of the five-repetition sit-to-stand test in adult kidney transplant recipients. *Int J Ther Rehabil* 2018;**25**:158–66. https://doi.org/10.12968/ijtr.2018.25.4.158
51. Figueiredo PHS, de S Veloso LR, Lima MMO *et al*. The reliability and validity of the 30-seconds sit-to-stand test and its capacity for assessment of the functional status of hemodialysis patients. *J Bodyw Mov Ther* 2021;**27**:157–64. https://doi.org/10.1016/j.jbmt.2021.02.020
52. Bohannon RW, Smith J, Hull D *et al*. Deficits in lower extremity muscle and gait performance among renal transplant candidates. *Arch Phys Med Rehabil* 1995;**76**:547–51.
53. Wilkinson TJ, Watson EL, Xenophontos S *et al*. The "minimum clinically important difference" in frequently reported objective physical function tests after a 12-week renal rehabilitation exercise intervention in nondialysis chronic kidney disease. *Am J Phys Med Rehabil* 2019;**98**:431–7.
54. Mesquita R, Janssen DJA, Wouters EFM *et al*. Within-day test-retest reliability of the Timed Up & Go test in patients with advanced chronic organ failure. *Arch Phys Med Rehabil* 2013;**94**:2131–8.
55. Sutcliffe BK, Bennett PN, Fraser SF *et al*. The deterioration in physical function of hemodialysis patients. *Hemodial Int* 2018;**22**:245–53. https://doi.org/10.1111/hdi.12570
56. Hadjiioannou I, Wong K, Lindup H *et al*. Test-retest reliability for physical function measures in patients with chronic kidney disease. *J Renal Care* 2020;**46**:25–34. https://doi.org/10.1111/jorc.12310
57. Greenwood SA, Castle E, Lindup H *et al*. Mortality and morbidity following exercise-based renal rehabilitation in patients with chronic kidney disease: the effect of programme completion and change in exercise capacity. *Nephrol Dial Transplant* 2019;**34**:618–25. https://doi.org/10.1093/ndt/gfy351
58. Watson EL, Major RW, Wilkinson TJ *et al*. The association of muscle size, strength and exercise capacity with all-cause mortality in non-dialysis-dependent CKD patients. *Clin Physiol Funct Imaging* 2020;**40**:399–406. https://doi.org/10.1111/cpf.12655
59. Tonet E, Ariza-Solé A, Serenelli M *et al*. The impact of sex and physical performance on long-term mortality in older patients with myocardial infarction. *BMC Med* 2022;**20**:15. https://doi.org/10.1186/s12916-021-02211-1
60. Zhang D, Tang W, Dou LY *et al*. Four different frailty models predict health outcomes in older patients with stable chronic obstructive pulmonary disease. *BMC Geriatr* 2022;**22**:57.
61. Mutambudzi M, Chen NW, Howrey B *et al*. Physical performance trajectories and mortality among older Mexican Americans. *J Gerontol A Biol Sci Med Sci* 2019;**74**:233–9. https://doi.org/10.1093/gerona/gly013
62. Parry SM, Denehy L, Beach LJ *et al*. Functional outcomes in ICU – what should we be using? – an observational study. *Crit Care* 2015;**19**:1–9. https://doi.org/10.1186/s13054-015-0829-5
63. Freiberger E, De vreede P, Schoene D *et al*. Performance-based physical function in older community-dwelling persons: a systematic review of instruments. *Age Ageing* 2012;**41**:712–21. https://doi.org/10.1093/ageing/afs099
64. Menezes KVRS, Auger C, de Souza Menezes WR *et al*. Instruments to evaluate mobility capacity of older adults during hospitalization: a systematic review. *Arch Gerontol Geriatr* 2017;**72**:67–79. https://doi.org/10.1016/j.archger.2017.05.009
65. Combret Y, Medrinal C, Bonnevie T *et al*. Clinimetric evaluation of muscle function tests for individuals with cystic fibrosis: a systematic review. *J Cyst Fibros* 2020;**19**:981–95. https://doi.org/10.1016/j.jcf.2020.05.014
66. Dobson F, Hinman RS, Hall M *et al*. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. *Osteoarthritis Cartilage* 2012;**20**:1548–62. https://doi.org/10.1016/j.joca.2012.08.015
67. Jakobsson M, Gutke A, Mokkink L *et al*. Level of evidence for reliability, validity, and responsiveness of physical capacity tasks designed to assess functioning in patients with low back pain: a systematic review using the COSMIN standards. *Phys Ther* 2019;**99**:457–77. https://doi.org/10.1093/ptj/pzy159
68. Collado-Mateo D, Dominguez-Munoz FJ, Charrua Z *et al*. Isokinetic strength in peritoneal dialysis patients: a reliability study. *Appl Sci (Basel)* 2019;**9**:3542–50.
69. da Silva KB, Leal DV, da Rocha JL *et al*. Glittre activities of daily living test is reliable and valid in hemodialysis patients. *Disabil Rehabil* 2022;**45**:542–8.
70. Figueiredo PHS, da Silva ACR, Costa HS *et al*. The Glittre activities of daily living as a potential test for functional evaluation of patients on hemodialysis: a validation study. *Disabil Rehabil* 2022;**44**:2083–90. https://doi.org/10.1080/09638288.2020.1812121
71. Jácome C, Flores I, Martins F *et al*. Validity, reliability and minimal detectable change of the balance evaluation systems test (BESTest), mini-BESTest and brief-BESTest in patients with end-stage renal disease. *Disabil Rehabil* 2018;**40**:3171–6. https://doi.org/10.1080/09638288.2017.1375034
72. Lorenz EC, Cheville AL, Amer H *et al*. Relationship between pre-transplant physical function and outcomes after kidney transplant. *Clin Transplant* 2017;**31**:1–19. https://doi.org/10.1111/ctr.12952

73. Macdonald JH, Marcora SM, Jibani M *et al*. Intradialytic exercise as anabolic therapy in haemodialysis patients—a pilot study. *Clin Physiol Funct Imag* 2005;**25**:113–8.
74. Mercer TH, Naish PF, Gleeson NP *et al*. Development of a walking test for the assessment of functional capacity in non-anaemic maintenance dialysis patients. *Nephrol Dial Transplant* 1998;**13**:2023–6. https://doi.org/10.1093/ndt/13.8.2023
75. Souweine JS, Boudet A, Chenine L *et al*. Standardized method to measure muscle force at the bedside in hemodialysis patients. *J Ren Nutr* 2017;**27**:194–200. https://doi.org/10.1053/j.jrn.2017.01.017
76. Wilkinson TJ, Nixon DGD, Smith AC. Postural stability during standing and its association with physical and cognitive functions in non-dialysis chronic kidney disease patients. *Int Urol Nephrol* 2019;**51**:1407–14.

Received: 16.2.2023; Editorial decision: 26.6.2023