Age and Ageing 2023; **52:** 1–11
https://doi.org/10.1093/ageing/afac327

SYSTEMATIC REVIEW

Remote collection of physical performance measures for older people: a systematic review

Philip A. Heslop, Christopher Hurst, Avan A. Sayer, Miles D. Witham

AGE Research Group, NIHR Newcastle Biomedical Research Centre, Newcastle University, Newcastle upon Tyne Hospitals NHS Foundation Trust and Cumbria, Northumberland, Tyne and Wear NHS Foundation Trust, Newcastle upon Tyne NE4 5PL, UK

Address correspondence to: Miles D. Witham, NIHR Newcastle Biomedical Research Centre, Newcastle University Campus for Ageing and Vitality, Newcastle upon Tyne, NE4 5PL, UK. Tel: (+44) 191 208 1317. Email: Miles.Witham@newcastle.ac.uk

Abstract

Remotely collected physical performance measures could improve inclusion of under-served groups in clinical research as well as enabling continuation of research in pandemic conditions. It is unclear whether remote collection is feasible and acceptable to older patients, or whether results are comparable to face-to-face measures. We conducted a systematic review according to a prespecified protocol. We included studies with mean participant age ≥ 60 years, with no language restriction. Studies examining the gait speed, Short Physical Performance Battery, distance walk tests, grip strength, Tinetti score, Berg balance test, sit-to-stand test and timed up and go were included. Reports of feasibility, acceptability, correlation between remote and face-to-face assessments and absolute differences between remote and face-to-face assessments were sought. Data were synthesised using Synthesis Without Meta-analysis methodology; 30 analyses from 17 publications were included. Study size ranged from 10 to 300 participants, with a mean age ranging from 61 to >80 years. Studies included a broad range of participants and conditions. Most studies had a moderate or high risk of bias. Only two studies undertook assessment of acceptability or feasibility, reporting good results. Correlation between face-to-face and remote measures was variable across studies, with no measure showing consistently good correlation. Only nine studies examined the accuracy of remote measures; in six studies, accuracy was rated as good (<5% mean difference between face-to-face and remote measures). There is a lack of robust evidence that remote collection of physical performance measures is acceptable to patients, feasible or provides comparable results to face-to-face measures.

Keywords: systematic review, remote outcomes, physical performance, older people

Key Points

- Remote assessment of outcomes for research and clinical practice has become common during the COVID-19 pandemic.
- Evidence for the acceptability and feasibility of remote assessment of physical performance for older people is lacking.
- Remotely measured physical performance outcomes are not always comparable to face-to-face measures.
- Better-designed comparison studies are needed before the remote assessment of physical performance outcomes could be recommended for research or clinical practice.

Introduction

Recruitment to clinical research studies for older people is often challenging [1]. Limited mobility, social isolation and transportation barriers may all make it difficult for older people to attend research centres for study visits, contributing to the low recruitment and retention rates in clinical studies. Remote delivery (defined as any non-face-to-face method, including telephone, video or postal delivery) of trial processes [2] provides a way to broaden the inclusion of older people into research. The recent COVID-19 pandemic has forced many clinical research studies to stop, or to conduct their processes remotely. While this has enabled some studies to continue [3], recent guidance has highlighted the lack of research testing the acceptability, feasibility and validity of remote research delivery approaches [4]. Remote delivery may generate new barriers to inclusion, particularly for older people with sensory impairment, cognitive impairment or

who lack digital infrastructure or training [5]. This area has been highlighted as a priority area for future research [6], both to enable research delivery for older people in pandemic situations but also to widen participation after the pandemic.

Measurements of physical performance are important outcomes in many research studies for older people, and improving physical performance is highly prioritised by older people as an aim of treatment and care [7]. If such measures are to be of value in clinical research, it is essential that remote versions of measures of physical performance are robustly evaluated. Ensuring that remotely assessed outcome measures are acceptable to participants and are feasible for research teams to deliver is necessary to minimise missing data and participation bias. It is necessary to ensure that remote assessments are comparable with face-to-face assessments if different ways of measuring an outcome are to be combined in analyses. It is also necessary to ensure that the results of studies using remotely conducted assessments are comparable to studies using face-to-face measures [8].

To date, there has been no attempt to review and integrate the literature on remote assessment of commonly used measures of physical performance for older people. We therefore sought to (i) systematically review the evidence on the acceptability to participants of conducting physical performance measures remotely; (ii) systematically review the evidence on the feasibility of conducting physical performance measures remotely and (iii) systematically review the evidence that remote assessment of physical performance measures are comparable with face-to-face assessment.

Methods

Search strategy

We conducted a systematic review according to a prespecified protocol, which was registered on the PROSPERO database (CRD42020219855). A summary of methods and search strategies is given in the Supplementary material. We searched six electronic databases (MEDLINE, CINAHL, Embase, Cochrane Central Register of Controlled Trials, Controlled Clinical Trials.com and NHS elibrary) from inception up to the end of April 2022. No language restrictions were used, and papers that were available as preprints or online ahead of print were eligible for inclusion. Reference lists of included papers were searched for further potentially eligible studies. A series of physical performance measures commonly used in research with older people (mean age of participants ≥60 years) were preselected by the investigators, and a separate search was conducted for each measure. Searches were conducted for the following measures: Short Physical Performance Battery (SPPB) [9], Gait speed [10], Timed walk test [11, 12], Tinetti gait and balance score [13], Berg balance scale [14], timed up and go (TUAG) test [15], sit-to-stand (STS) test [16], Handgrip strength [17] and shuttle walk test and step test.

Inclusion criteria

We included studies involving human participants with a mean age of ≥60 years. Studies were required to include the remote assessment of one or more of the listed physical performance measures by a human assessor. In order to avoid duplicating other planned and ongoing work [18] that we were aware of, we excluded studies using sensor-based remote monitoring approaches. Standalone observational studies or studies nested within clinical trials were eligible for inclusion. We considered face-to-face assessment as the gold standard and included studies that either assessed the feasibility or acceptability of conducting remote assessments of physical performance, or compared the accuracy of remote assessments of physical performance with to face-to-face delivery of the same assessment. We did not seek to study the reliability, responsiveness or external validity of remote measures or of the face-to-face measures in this review, as these should have already been assessed in the validation studies of face-to-face measures.

Study selection and data extraction

Following the removal of duplicates, two reviewers (P.A.H. and M.D.W.) screened all titles with those identified as being potentially eligible for inclusion having abstracts retrieved. Following this, both reviewers screened the retrieved abstracts, and abstracts flagged as being potentially eligible for inclusion by either reviewer had full text papers retrieved. Papers agreed as eligible by both reviewers were included in the review. Data were extracted using a standard, piloted form. One reviewer (P.A.H.) extracted data which were then checked by M.D.W. Discrepancies were resolved by discussion until consensus was reached.

We extracted baseline data on trial populations (including age, sex, functional status and level of cognition). We identified the physical performance measures that met the criteria of being remotely assessed with some form of clinician involvement. We extracted descriptive data on acceptability or feasibility (e.g. acceptability questionnaire results and the percentage of participants successfully completing the remote assessment). We extracted the statistical tests used for comparing the remote and face-to-face performance measure (e.g. comparison of means and correlation between measures) and the results of those analyses.

Assessment of methodological quality

Risk of bias for each trial was independently assessed using a modified version of the QUADAS-2 tool [19]. Four domains were assessed: Patient Selection, Index Tests, Reference Standard and Flow and Timing. We considered an appropriate face-to-face measurement as the reference standard. We followed the QUADAS-2 guidelines to assess the processes used, how participants were recruited, what (if any) randomisation occurred, if index testing occurred without knowledge of the reference testing, if the testing adequately captured the remote assessment by a clinician and

whether the timing of the testing was suitable (i.e. minimal delay between face-to-face and remote testing). These factors were condensed into a risk of bias for each domain.

Data synthesis

Given the high degree of heterogeneity in the results, we did not attempt to conduct meta-analysis. Instead, data were synthesised in summary tables, with a narrative synthesis conducted according to the principles of Synthesis Without Meta-Analysis methodology [20]. We grouped studies by the physical performance measure. We present a description of each study population and identify which physical performance measure was compared between the remote and clinical applications. We also present each physical performance measure in separate tables, indicating the comparison method used, the statistical test involved and the results of those tests. All studies were included in the summary and synthesis; we did not attempt to impose a standardised metric or transformation method as the included data were too heterogeneous. We summarised the available correlation (correlation coefficient or R^2 value from regression) and accuracy data (percentage difference compared to face-to-face measures) by using arbitrary categorisations for ease of comparison. Categories used for correlation were: Good ($r > 0.8$ or $R^2 > 0.65$), Moderate ($r = 0.6$–0.8 or $R^2 = 0.40$–0.65) or Poor ($r \leq 0.6$ or $R^2 < 0.40$). Categories used for accuracy were: Good (≤5% mean difference compared to face-to-face measure), Moderate (5–10% mean difference) and Poor (>10% mean difference). We chose to use the face-to-face method as the gold standard for each comparison as this is the method currently used in research and practice, and the focus of our review was on whether the remote performance of measures could substitute for face-to-face measurement rather than which method was more consistent or responsive. Given the range of measures and heterogeneity of measurement, we did not attempt to formally evaluate the certainty of the evidence.

Results

Our searches yielded a total of 30 analyses contained in 17 publications [21–37]. A summary PRISMA flow diagram is shown in Figure 1, and the details of the individual searches are given in Supplementary Table 1. Four studies used the SPPB [22, 26, 29, 33], four used gait speed [32, 33, 36, 37], three used timed walk [28, 29, 31], one used the Tinetti gait and balance score [34] and three used the Berg balance score [23, 35, 36]. Six studies used the TUAG test [21, 23, 27, 31, 36, 37], four used the STS test [25, 27, 33, 37] and four used handgrip strength [24, 30–32]. One study [23] used the step test, and no studies used the shuttle walk test. A summary of included studies is shown in Table 1. Study size ranged from 10 to 300 participants; the most common country of study was the USA. The mean age of included participants ranged from 61 to >80, with most studies having a mean age between 70 and 80 years. Most studies included a broad range of older people, unconstrained by a specific disease condition.

Quality assessment

Table 2 shows the results from the QUADAS-2 risk of bias assessment. Most studies had a moderate or high risk of bias for patient selection due to non-consecutive patients being included, or highly selected populations studied. There was insufficient information present in most papers to assess the risk of bias in the domains of either the index test or the reference test. Published descriptions of participant flow through the included studies were adequate to assess the risk of bias in most cases; most studies demonstrated a low risk of bias in this domain.

Acceptability and feasibility

Only three of the included studies undertook quantitative assessment of acceptability or feasibility. Simpson et al. [33] designed a remotely monitored exercise application to assess gait speed. They reported that the monitoring application was feasible to use, with participants performing >100% of the prescribed exercise sessions completed and had good participant satisfaction (rating the system usability (78%), enjoyment (70%) and system benefit (80%) as high). Gillespie et al. [35] conducted an acceptability survey with their participants after the study, with >60% self-assessing that they were be able to use the video conferencing software for the study. One other study (Hwang et al. [31]) employed the system usability scale, a validated measurement tool, that rates a user's experience of technology on 10-question, 5-point Likert Scale administered to the participant at the end of the telerehabilitation assessment. The mean system usability scale total score in this study was 85 (SD 15) out of 100.

Comparability of remote measures with face-to-face measures

Measures of comparability for each study are shown in Supplementary Tables 2–10, grouped by physical performance measure; we summarise and discuss the results for each measure in narrative synthesis below. Most, but not all, included studies that compared remote and face-to-face measures; those that did not are still included in the review as they include data on feasibility and acceptability. Table 3 summarises the findings for both correlation between face-to-face and remote measures and for the degree of difference between face-to-face and remote measures (accuracy).

Short Physical Performance Battery

Four studies compared remote assessment with face-to-face SPPB (Supplementary Table 2). One of the four studies [32] used a remotely monitored exercise application. Three of the four studies studied video-prompted patient self-report as the remote outcome measure. The Mobility Assessment Tool (MAT-sf) tool was used in two of these studies

Table 1. Details of included studies

Paper	Year	Country	*N*	Mean age (years)	Age range (years)	% women	Inclusion criteria (ages, physical/cognitive function)	Physical performance measure								
								SPPB	Gait speed	Timed walk test/ 400 m walk	Tinetti gait and balance score	Berg balance scale	TUAG test	STS test	Handgrip strength	Step test
Botolfsen et al. [21]	2008	Norway	28	80	NA	82%	Aged 75 +. Impaired mobility (fallen 1+ times in past year or previously used walking aid) BUT able to walk 10 m unaided. No stroke in past year. MMSE ≥ 22						•			
Rejeski et al. [22]	2010	USA	234	82	NA	71%	Aged 65–90 years. Able to walk without help. MMSE ≥ 23.	•								
Russell et al. [23]	2013	Australia	12	66	45–76	50%	PD severity rating between Stages I and IV on the Hoehn and Yahr Score. Independently mobile. Communication and cognitive status adequate to complete assessment tasks. No comorbid conditions that would prevent them from safely completing assessment tasks.					•	•			•
Tseng et al. [24]	2013	Taiwan	101	N/A	55–85	64%	Checked for basic function/could complete all four assessments.								•	
Banerjee et al. [25]	2014	USA	10	N/A	83–97	NA	Aged 83–97.							•		
Guerra et al. [26]	2014	Brazil/ Colombia	300	69	NA	59%	Aged 65–74 years. Able to walk without help + SPPB < 10. MMSE ≥ 21.	•								
Verheyden et al. [27]	2014	UK	*s*: 38 *c*: 19	*s*: 69 (8) *c*: 68 (0)	47-88 52–85	*s*: 39% *c*: 68%	Study group: diagnosed PD. MMSE ≥ 24.						•	•		
Holland et al. [28]	2015	Australia	19	71	56–86	42%	Mean age = 71. COPD diagnosis. Normal exercise response (6MWT).			•						

(*Continued*)

Table 1. Continued

Paper	Year	Country	*N*	Mean age (years)	Age range (years)	% women	Inclusion criteria (ages, physical/cognitive function)	Physical performance measure								
								SPPB	Gait speed	Timed walk test/ 400 m walk	Tinetti gait and balance score	Berg balance scale	TUAG test	STS test	Handgrip strength	Step test
Marsh et al. [29]	2015	USA	110	81	65–94	73%	Aged 65–94 years. Able to walk without help + SPPB < 10. MMSE ≥ 21.	•		•						
Blomkvist et al. [30]	2016	Switzerland	30	69	NA	NA	Aged 65 +. Able to pass custom dementia screening (year, month and president). No surgery on limbs in last 6 months. No neurological disease.								•	
Hwang et al. [31]	2017	Australia	17	69	39–87	12%	Mean age = 69. Stable CHF.			•			•		•	
Chkeir et al. [32]	2019	France	194	79	NA	60%	Aged 70+ years. Able to stand without help. MMSE ≥ 10.		•						•	
Simpson et al. [33]	2020	Australia	10	74	58–88	40%	Aged 18 +. Had stroke within 2 years. Able to stand independently.	•	•					•		
Venkataraman et al. [34]	2020	USA	42	61	NA	19%	Used a cane/orthosis/prosthesis at least once per week for walking. Scored 6 of 10 or higher on the Short Portable Mental Status Questionnaire.				•					
Gillespie et al. [35]	2021	Canada	20	72	NA	45%	Aged 18 +. Had stroke diagnosis. FIM expression score > 3. FIM memory and comprehension score > 3.					•				
Pelicioni et al. [36]	2022	New Zealand	15	71	64–78	53%	Able to walk 30 m (including with aids) Able to complete tests safely. MoCA score ≥ 21		•			•	•			
Peyrusqué et al. [37]	2022	Canada	15	69	NA	60%	Aged 60 +. Live independently in community. Able to perform physical tests. Able to consent.		•				•	•		

NA, not available; 6MWT, 6-minute walk test; CHF, chronic heart failure; COPD, chronic obstructive pulmonary disease; FIM, Functional Independence Measure; MMSE, Mini-Mental State Examination; MoCA, Montreal Cognitive Assessment; PD, Parkinson's disease.

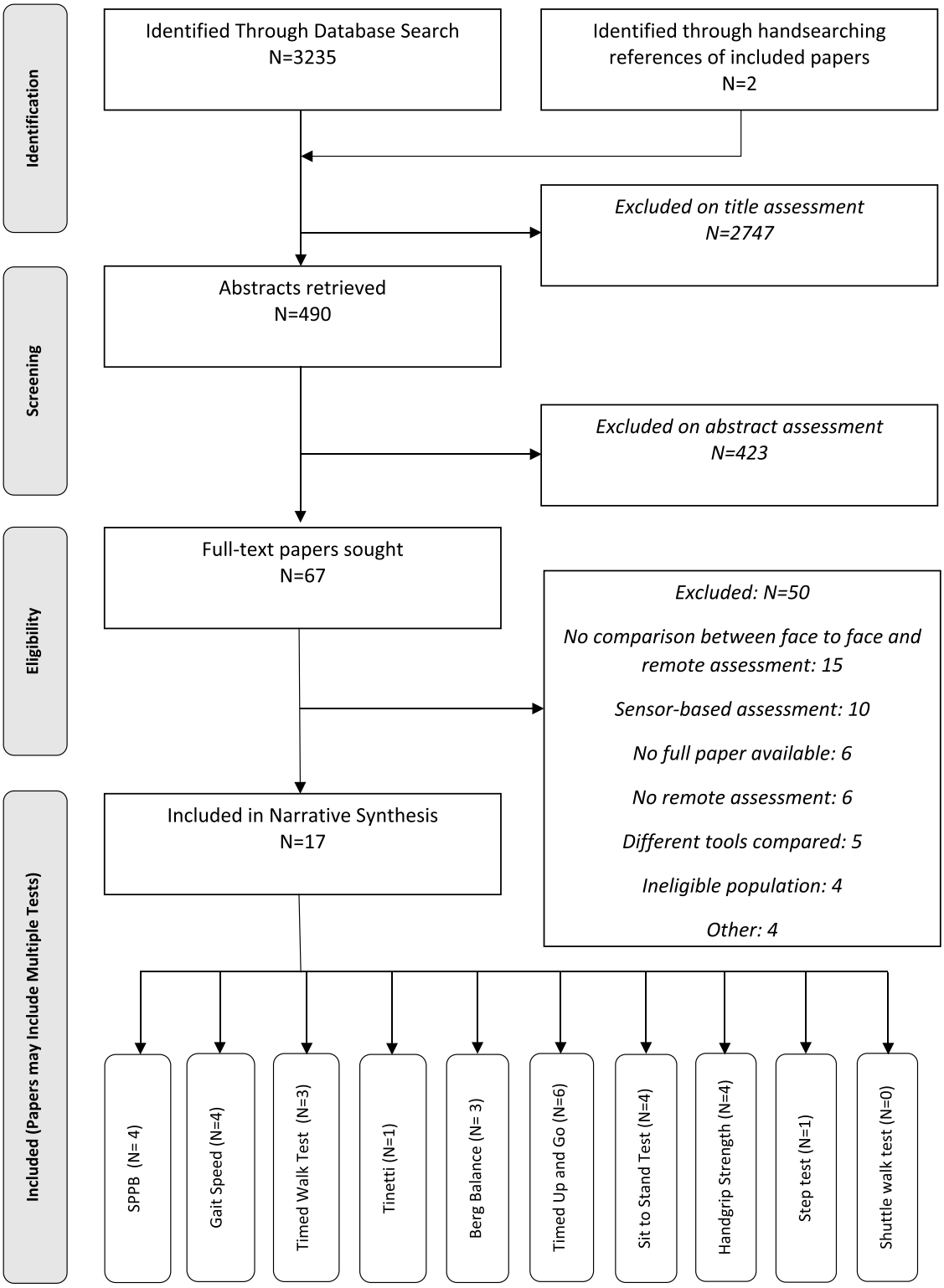

Figure 1. Summary PRISMA flow diagram for all searches combined.

[22, 26], and the virtual Short Physical Performance Battery (vSPPB), was used in one [29]. In both cases, participants rate their own physical capabilities based on viewing a video or animation of graded physical tasks, and the results of this self-assessment were compared with face-to-face SPPB measurements. No study used a clinician observing the SPPB over a video link as the remote assessment, and no data were available assessing the difference in scores between the vSPPB and the in-person SPPB. One study showed only a weak association between the MAT-sf and the SPPB

Table 2. Risk of bias based on QUADAS-2 domains

Study	QUADAS-2 domain			
	Patient selection	Index tests	Reference tests	Flow and timing
Botolfsen et al. 2008 [21]				
Rejeski et al. 2010 [22]				
Russell et al. 2013 [23]				
Tseng et al. 2013 [24]				
Banerjee et al. 2014 [25]				
Guerra et al. 2014 [26]				
Verheyden et al. 2014 [27]				
Holland et al. 2015 [28]				
Marsh et al. 2015 [29]				
Blomkvist et al. 2016 [30]				
Hwang et al. 2017 [31]				
Chkeir et al. 2019 [32]				
Simpson et al. 2020 [33]				
Venkataraman et al. 2020 [34]				
Gillespie et al. 2021 [35]				
Pelicioni et al. 2022 [36]				
Peyrusqué et al. 2022 [37]				

Red, high risk of bias; orange, moderate risk of bias; green, low risk of bias, grey, unclear risk of bias.

Table 3. Summary table for accuracy of remotely assessed physical performance measures

	SPPB	Gait speed	Walk tests	Tinetti	Berg	Step test	TUAG	STS	Grip strength
Number of studies with data on accuracy	3	3	3	1	3	1	6	2	4
Correlation between face-to-face and remote measures	Good: 0/3 Moderate: 2/3 Poor: 1/3	Good: 2/3 Moderate: 1/3 Poor: 0/3	Good: 0/3 Moderate: 0/3 Poor: 1/3 No data: 2/3	Good: 0/1 Moderate: 1/1 Poor: 0/1	Good: 3/3 Moderate: 0/3 Poor: 0/3	No data	Good: 3/6 Moderate: 0/6 Poor: 0/6 No data: 3/6	Good: 1/2 Moderate: 0/2 Poor: 0/2 No data: 1/2	Good: 2/4 Moderate: 0/4 Poor: 1/4 No data: 1/4
Accuracy of remote measures	No data	Good: 0/3 Moderate: 0/3 Poor: 1/3 No data: 2/3	Good: 1/3 Moderate: 1/3 Poor: 0/3 No data: 1/3	Good: 0/1 Moderate: 1/1 Poor: 0/1	No data	Good: 1/1	Good: 2/6 Moderate: 0/6 Poor: 0/6 No data: 4/6	Good: 1/2 Moderate: 0/2 Poor: 0/2 No data: 1/2	Good: 1/4 Moderate: 0/4 Poor: 0/4 No data: 3/4

Correlation: good: $r > 0.8$ or $R^2 > 0.65$; moderate: $r = 0.6$–0.8 or $R^2 = 0.40$–0.65; poor: $r \leq 0.6$ or $R^2 < 0.40$. Accuracy: good: ≤5% mean difference compared to face to face; moderate: 5–10% mean difference; poor: >10% mean difference.

(R^2 = 0.24–0.29); another showed a moderate association (R^2 = 0.40). The vSPPB showed moderate correlation with the in-person SPPB (r = 0.60).

Gait speed

Four studies evaluated Gait Speed (Supplementary Table 3) of which three compared face-to-face and remote assessment. Chkeir et al. [32] compared a radar-based gait speed measure with clinician timed walk speed. They reported a correlation of 0.62 by linear regression, but gait speed by radar was only 77% of that measured by face-to-face timing. The study did not use a video link for remote clinical assessment. Pelicioni et al. [36] compared face-to-face with both live and recorded video assessments. Both interrater and intrarater reliabilities (measured by intraclass correlation coefficient) were good. Peyrusqué et al. [37] compared remote assessment via video link with a face-to-face assessment conducted later. Face-to-face assessments were performed outside by the same evaluator within a week of the remote assessment and intraclass correlation was moderate to good between the two measures (ICC = 0.62–0.77).

Timed walk/400 m walk

Three studies used a timed walk or 400 m walk as a physical performance measure (Supplementary Table 4). Marsh et al. [29] compared the vSPPB, (where patients self-report their

ability levels using reference videos), with face-to-face measurement of 400-m walk time. The two measures were not highly correlated ($r = 0.54$); the face-to-face SPPB showed much stronger correlation with the face-to-face 400 m walk ($r = 0.82$). Hwang et al. [31] compared 6-minute walk distance performed in a hospital clinic and assessed via video link with 6-minute walk distance assessed face to face. They found good average agreement but considerable individual variation between the modes of assessment (mean difference = −4 m and limits of agreement −84 to 76 m). Holland et al. [28] also measured 6-minute walk distance but compared self-reported home measurements (i.e. outdoors, without a dedicated, measured track) with in-clinic measurements (i.e. a controlled, measured environment). This study found consistently lower walk distance at home, with wide individual variation (mean difference = −30 m, limits of agreement = −167 to 102 m). In this study, track length at home showed a significant correlation with the discrepancy between home and hospital measures ($r = 0.58$; $P < 0.001$); shorter home tracks were associated with shorter home walk distances relative to the hospital walk distance.

Tinetti gait and balance score

One study [34] compared Tinetti score assessed from video recordings with face-to-face assessment (Supplementary Table 5). This study found higher scores on in-person assessment than via video, with moderate agreement between in-person and slow-motion video on regression analysis ($B = 0.62$; 95% CI = 0.37–0.87).

Berg balance scale

Three studies evaluated the Berg Balance Scale by comparing face-to-face assessment with video (Supplementary Table 6). Russell et al. [23] compared face-to-face assessment with simultaneous remote assessment and found a high level of agreement. Although percent exact agreement (%EA) scores were low, percent agreement within one point was high, suggesting that differences in scoring were minor. Gillespie et al. [35] used an interrater reliability study design to evaluate the reliability between two individual raters by using Krippendorff's alpha reliability estimate, producing a result of 0.97 (0.96–0.99 CI). They also conducted an acceptability survey with 19/20 of the participants, with 60% self-assessing that they were be able to use the video conferencing software for the study. Pelicioni et al. [36] also found good intrarater results (ICC 95% CI) were: 0.82 (0.442–0.940) versus realtime video and 0.78 (0.339–0.927) versus recorded video.

TUAG test

Six studies used the TUAG test to measure physical performance (Supplementary Table 7), and comparative data were available from five of these studies. Russell et al. [23] analysed the limits of agreement (with a clinically acceptable limit of 5.0 seconds). They found differences in the range of −1.25 to 1.24, with a mean difference of −0.01 (SD 0.63) and a mean absolute difference of 0.47 seconds. Hwang et al. [31] compared assessment via video link with a separate face-to-face assessment and found good average agreement but some individual variation between the modes of assessment (mean difference = 0.2 s, limits of agreement = −2.8 to 3.3 s). Botolfsen et al. [21] compared an extended timed up and go, where each part of the test, such as turning, was timed separately, with a standard TUAG, showing good correlation ($r = 0.85$, $P < 0.001$); no data comparing difference in time taken to complete the two tests were shown. Both Pelicioni et al. [36] and Peyrusqué [37] also found a high degree of correlation between remote and face-to-face TUG using ICC measures.

STS test

Four studies compared remote with face-to-face assessment, two of which compared remote assessment with face-to-face assessment (Supplementary Table 8). Banerjee et al. [25] compared a video algorithm with clinician stopwatch timings, reporting only minor mean differences between stopwatch timings and timing from each of three versions of their algorithm (0.17, 0.094 and 0.011 seconds, respectively). Limits of agreement were not given. They also concluded that the video angle and chair choice were important factors in their video processing algorithm accuracy. Peyrusqué et al. [37] found very high correlation between remote and face-to-face measurement (ICC = 0.96 [95% CI = 0.89–0.99]) where face-to-face assessment was conducted within 7 days of the remote assessment.

Grip strength

Four papers studied grip strength (Supplementary Table 9). Chkeir et al. [32] provided a home testing mechanism for measuring grip strength and reported good correlations with clinic-based measurements ($R^2 = 0.80$ by linear regression, with remote measurement by grip-ball being on average 10% lower than by face-to-face measurement by Jamar dynamometer). Hwang et al. [31] compared remote grip strength assessment via video link with a separate face-to-face assessment and found good average agreement but considerable individual variation between the modes of assessment (right grip: mean difference = 0.2 kg, limits of agreement = −6.5 to 6.8 kg; left grip: mean difference = 0.3 kg, limits of agreement = −5.6 to 6.1 kg). Tseng et al. [24] also compared clinic-based grip measurements with a remote home-based system; only a weak correlation was noted ($r = 0.29$) and app-based grip strength was approximately half that recorded by clinic-based measures. Much better results were seen by Blomkvist et al. [30] who compared squeezing a Nintendo Wii balance board with standard grip strength measuring equipment. This showed good correlation between the two methods ($r = 0.86$ to 0.87) but much lower estimates of grip strength by the Wii balance board (mean difference = 15.4 kg dominant hand, 11.9 kg non-dominant hand).

Step test

Russell et al. [23] showed high levels of agreement between the remote and face-to-face assessments of the step test using weighted kappa, %EA and percent agreement within one point on the ordinal scale. The results are shown in Supplementary Table 10.

Discussion

Summary of evidence

Our results suggest that great care is needed in interpreting results from remote assessments of physical performance for older people. Current evidence is insufficient to give confidence in either the feasibility or acceptability of remote assessment or the interchangeability of remote and face-to-face assessment results. The studies included in our review did not have comparison of remote and face-to-face assessments as their primary goal. We found variable correlation between remote and face-to-face measures, and even where correlation is good, systematic over or underestimation of results by remote assessment may mean that remote and face-to-face results are not interchangeable without recourse to correction factors. Few studies have been performed that sought to formally test the feasibility, acceptability and accuracy of remote assessment of physical performance measures. For most studies included in this review, the data reported were not adequate to enable conclusions to be drawn about the robustness of remote assessment. Formal testing of acceptability to participants, and feasibility (i.e. the proportion of participants for whom a remote measure could be completed) were lacking from the majority of studies. Most studies did not attempt to replicate the face-to-face measurement remotely, either via self-report or via video-linked assessment. Differences in the type of assessment used remotely, or the way that an assessment was deployed, are likely to contribute to the lack of remote test accuracy noted above. In addition, most studies did not report on contextual factors or attempt to highlight how study processes (e.g. who was conducting assessments and how they were trained) might differ from the real-world research or clinical practice. Many studies relied on correlation rather than including more rigorous tests of accuracy and precision such as Bland–Altman plots [38].

Study designs

Many of the included studies used designs that were not well suited to comparing face-to-face and remote physical performance measures. Our adapted QUADAS-2 analysis shows there are likely biases across the studies. Populations were not always representative and were often selected from larger studies. In particular, few studies included patients living with frailty and the feasibility, and the acceptability of conducting remote assessments in this group is therefore uncertain. Socioeconomic data and other population descriptors were not reported in most studies, and we were thus unable to summarise these study characteristics in this review. It was unclear in most studies if the remote assessment measurements and face-to-face measurements were performed without knowledge of each other, and face-to-face as a 'gold standard' was not always performed. Studies that employed video footage (either live or recorded) did not represent the likely scenario for a real-world deployment in that they used cameras, tripods and other equipment that are beyond the means of the general public. Some studies used physical performance measures that do not lend themselves to assessment by observation, for example, grip strength, and are thus less suitable for video link remote assessment. For such measures, provision of a grip strength device for use unobserved in the home, with measurements reported back to the research team by the participant, may be an alternative, but this approach requires further study. Only one study reported a significant difference between the remote and clinical measurements, and this was only due to the specialised environment set up in clinic (i.e. a measured walking track) that might not be available in all clinics, much less in people's homes.

Self- report versus remote assessment

Several of the studies use guided or coached self-report (using example video clips or via an application) instead of direct remote assessment via video link. In these cases, the studies reported moderate to good correlation with face-to face measures but without a thorough comparison with alternative assessment methods. This is particularly necessary when a measurement is calculated from a device or sensor rather than observation, i.e. grip strength. These findings suggest there may be value in a study comparing clinical, direct remote and guided or coached self-report assessments, or even a mixed approach. This future work should ensure that sound testing practices, including the standardisation of verbal instructions and participant familiarisation, are incorporated as well as a thorough quantification of reliability [39].

Limitations

There are a number of limitations to our analysis. As with any systematic review, it is possible that we have omitted relevant literature, although the use of a broad search strategy, multiple reviewers and hand searching reduced the chances of missing relevant literature. The scope of our review excluded populations <60 years of age and studies that used sensors or applications to remotely monitor (i.e. without clinical involvement). We deliberately chose to focus on older populations for which the chosen outcome measures are commonly used, and we also wished to avoid overlap with other published or planned work. The heterogeneity of the included study populations, study methods and analysis methods did not permit meta-analysis. The heterogeneity of included study populations and reported measures also did not allow the generation of meaningful funnel plots as a way to test for possible publication bias. Such bias cannot be excluded, and it is important to note that if the

publication bias favours results that support the validity of remote studies, the real state of the evidence will be even less robust than our summary presented in this paper. We did not study reliability, responsiveness or external validity of remotely collected measures in this review. If a measure has been shown to have good reliability and responsiveness face to face, and the remotely collected results are similar, we assume that the remote measure will therefore also be reliable and responsive. Empirical data to support this assumption are lacking, however, and would be a useful focus for future study.

Implications for practice and research

This review has found a lack of robust evidence to support the use of remotely collected physical performance measures in older people. Although the use of remote measures is a practical response to the restrictions on social and health care contact imposed by the COVID-19 pandemic, it is unclear whether the remote collection of physical performance measures is acceptable to patients, feasible or provides comparable results to face-to-face measures. Using remotely collected measures, therefore, risks problems of missing data, inaccurate or biased data and imprecision, leading to greater dispersion of measurements. Before using remotely collected physical performance measures more widely, validation studies are required that are specifically designed to compare remote and face-to-face collections of the same outcome, in representative populations, using study designs that adhere to STARD guidelines [40]. In the meantime, caution is needed in the interpretation of studies or clinical practice measurements that are collected remotely, and we recommend avoiding combining and comparing data from face-to-face and remote assessments of physical performance in the same study.

Acknowledgements: All the authors acknowledge support from the NIHR Newcastle Biomedical Research Centre.

Supplementary Data: Supplementary data mentioned in the text are available to subscribers in *Age and Ageing* online.

Declaration of Conflicts of Interest: None.

Declaration of Sources of Funding: None.

References

1. McMurdo ME, Roberts H, Parker S *et al.* Improving recruitment of older people to research through good practice. Age Ageing 2011; 40: 659–65.
2. Trials@Home – Centre of Excellence Remote and Decentralised Clinical Trials. https://trialsathome.com/ (30th November 2021, date last accessed).
3. Witham MD, Anderson E, Carroll CB *et al.* Ensuring that COVID-19 research is inclusive: guidance from the NIHR INCLUDE project. BMJ Open 2020; 10: e043634. https://doi.org/10.1136/bmjopen-2020-043634.
4. NIHR Remote Trial Delivery Working Group: Preliminary Guidance. https://sites.google.com/nihr.ac.uk/remotetrialdelivery/home (30 November 2021, date last accessed).
5. Hewitt J, Pennington A, Smith A *et al.* A multi-centre, UK-based, non-inferiority randomised controlled trial of 4 follow-up assessment methods in stroke survivors. BMC Med 2019; 17: 111. https://doi.org/10.1186/s12916-019-1350-5.
6. Richardson SJ, Carroll CB, Close J *et al.* Research with older people in a world with COVID-19: identification of current and future priorities, challenges and opportunities. Age Ageing 2020; 49: 901–6.
7. Roberts H, Khee TS, Philp I. Setting priorities for measures of performance for geriatric medical services. Age Ageing 1994; 23: 154–7.
8. Goldsack JC, Coravos A, Bakker JP *et al.* Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020; 3: 55. https://doi.org/10.1038/s41746-020-0260-4.
9. Guralnik JM, Simonsick EM, Ferrucci L *et al.* A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994; 49: M85–94.
10. Graham JE, Ostir GV, Fisher SR, Ottenbacher KJ. Assessing walking speed in clinical research: a systematic review. J Eval Clin Pract 2008; 14: 552–62.
11. Guyatt GH, Sullivan MJ, Thompson PJ *et al.* The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. Can Med Assoc J 1985; 132: 919–23.
12. Simonsick EM, Montgomery PS, Newman AB, Bauer DC, Harris T. Measuring fitness in healthy older adults: the health ABC long distance corridor walk. J Am Geriatr Soc 2001; 49: 1544–8.
13. Tinetti ME. Performance-oriented assessment of mobility problems in elderly patients. J Am Geriatr Soc 1986; 34: 119–26.
14. Berg KO, Wood-Dauphinee SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992; 83: S7–11.
15. Podsiadlo D, Richardson S. The timed "up & go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991; 39: 142–8.
16. Bohannon RW. Reference values for the timed up and go test: a descriptive meta-analysis. J Geriatr Phys Ther 2006; 29: 64–8.
17. Roberts HC, Denison HJ, Martin HJ *et al.* A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach. Age Ageing 2011; 40: 423–9.
18. Polhemus AM, Bergquist R, de Basea MB *et al.* Walking-related digital mobility outcomes as clinical trial endpoint measures: protocol for a scoping review. BMJ Open 2020; 10: e038704. https://doi.org/10.1136/bmjopen-2020-038704.
19. Whiting PF, Rutjes AWS, Westwood ME *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155: 529–36.
20. Campbell M, McKenzie JE, Sowden A *et al.* Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020; 368: l6890.
21. Botolfsen P, Helbostad JL, Moe-Nilssen R, Wall JC. Reliability and concurrent validity of the expanded timed up-and-go

test in older people with impaired mobility. Physiother Res Int 2008; 13: 94–106.
22. Rejeski WJ, Ip EH, Marsh AP, Barnard RT. Development and validation of a video-animated tool for assessing mobility. J Gerontol A Biol Sci Med Sci 2010; 65: 664–71.
23. Russell TG, Hoffmann TC, Nelson M, Thompson L, Vincent A. Internet-based physical assessment of people with Parkinson disease is accurate and reliable: a pilot study. J Rehabil Res Dev 2013; 50: 643–50.
24. Tseng KC, Wong AM, Hsu CL, Tsai TH, Han CM, Lee MR. The iFit: an integrated physical fitness testing system to evaluate the degree of physical fitness of the elderly. IEEE Trans Biomed Eng 2013; 60: 184–8.
25. Banerjee T, Skubic M, Keller JM, Abbott C. Sit-to-stand measurement for in-home monitoring using voxel analysis. IEEE J Biomed Health Inform 2014; 18: 1502–9.
26. Guerra RO, Oliveira BS, Alvarado BE *et al.* Validity and applicability of a video-based animated tool to assess mobility in elderly Latin American populations. Geriatr Gerontol Int 2014; 14: 864–73.
27. Verheyden G, Kampshoff CS, Burnett ME *et al.* Psychometric properties of 3 functional mobility tests for people with Parkinson disease. Phys Ther 2014; 94: 230–9.
28. Holland AE, Rasekaba T, Fiore JF Jr, Burge AT, Lee AL. The 6-minute walk distance cannot be accurately assessed at home in people with COPD. Disabil Rehabil 2015; 37: 1102–6.
29. Marsh AP, Wrights AP, Haakonssen EH *et al.* The virtual short physical performance battery. J Gerontol A Biol Sci Med Sci 2015; 70: 1233–41.
30. Blomkvist AW, Andersen S, de Bruin ED, Jorgensen MG. Isometric hand grip strength measured by the Nintendo Wii Balance Board - a reliable new method. BMC Musculoskelet Disord 2016; 17: 56. https://doi.org/10.1186/s12891-016-0907-0.
31. Hwang R, Mandrusiak A, Morris NR, Peters R, Korczyk D, Russell T. Assessing functional exercise capacity using telehealth: is it valid and reliable in patients with chronic heart failure? J Telemed Telecare 2017; 23: 225–32.
32. Chkeir A, Novella JL, Dramé M, Bera D, Collart M, Duchêne J. In-home physical frailty monitoring: relevance with respect to clinical tests. BMC Geriatr 2019; 19: 34. https://doi.org/10.1186/s12877-019-1048-8.
33. Simpson DB, Bird ML, English C *et al.* Connecting patients and therapists remotely using technology is feasible and facilitates exercise adherence after stroke. Top Stroke Rehabil 2020; 27: 93–102.
34. Venkataraman K, Amis K, Landerman LR, Caves K, Koh GC, Hoenig H. Teleassessment of gait and gait aids: validity and interrater reliability. Phys Ther 2020; 100: 708–17.
35. Gillespie D, MacLellan C, Ferguson-Pell M, Taeger A, Manns PJ. Balancing access with technology: comparing in-person and telerehabilitation Berg balance scale scores among stroke survivors. Physiother Can 2021; 73: 276–85.
36. Pelicioni PH, Waters DL, Still A, Hale L. A pilot investigation of reliability and validity of balance and gait assessments using telehealth with healthy older adults. Exp Gerontol 2022; 162: 111747. https://doi.org/10.1016/j.exger.2022.111747.
37. Peyrusqué E, Granet J, Pageaux B, Buckinx F, Aubertin-Leheudre M. Assessing physical performance in older adults during isolation or lockdown periods: web-based video conferencing as a solution. J Nutr Health Aging 2022; 26: 52–6.
38. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Int J Nurs Stud 2010; 47: 931–6.
39. Hurst C, Batterham AM, Weston KL, Weston M. Short-and long-term reliability of leg extensor power measurement in middle-aged and older adults. J Sports Sci 2018; 36: 970–7.
40. Bossuyt PM, Reitsma JB, Bruns DE *et al.* STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015; 351: h5527. https://doi.org/10.1136/bmj.h5527.

Received 2 December 2021; editorial decision 2 November 2022