HHS Public Access

Author manuscript

J Geriatr Oncol. Author manuscript; available in PMC 2023 June 01.

Published in final edited form as:

J Geriatr Oncol. 2022 June ; 13(5): 691–697. doi:10.1016/j.jgo.2022.02.002.

Remote administration of physical performance tests among persons with and without a cancer history: Establishing reliability and agreement with in-person assessment

Carolyn Guidarelli, MPH[1],

Colin Lipps, BS[1],

Sydnee Stoyles, MBST, MAT[2],

Nathan F. Dieckmann, PhD[2,3],

Kerri M. Winters-Stone, PhD[1,2,*]

[1]Division of Oncological Sciences, Knight Cancer Institute, Oregon Health & Science University, Portland OR, USA

[2]School of Nursing, Oregon Health & Science University, Portland, OR, USA

[3]Department of Psychiatry, School of Medicine, Oregon Health & Science University, Portland, OR, USA

Abstract

Objectives: To assess the reliability of using videoconference technology to remotely administer the Short Physical Performance Battery (SPPB), including the 5-time sit-to-stand (5XSTS) and usual 4-meter walk (4mWT), and the Timed Up and Go (TUG) tests and agreement with in-person administration among adults with and without cancer.

Methods: Participants from two ongoing clinical exercise trials in cancer survivors, one that included partners without cancer, comprised the available sample (n=176; mean age 62.5 ± 11.5 years.). Remote tests were administered on two separate days by either the same or a different assessor to determine intra-rater and inter-rater reliability, respectively. We also compared tests conducted remotely and in-person using the same assessor and the same participant. Intraclass correlation coefficients (ICC) and 95% confidence intervals (95% CI) were used for

*Corresponding author at: Oregon Health & Science University, KR-CPC, 3181 SW Sam Jackson Park Road, Portland, OR 97239. wintersk@ohsu.edu.

Author's contributions

Conception and design: KWS, CG; Collection and assembly of data: CG, CL; Data analysis and interpretation: SS, NF, CG, KWS; Writing- original draft: CG, KWS; Writing- review and editing: All authors; Final approval of manuscript: All authors

Declaration of Competing Interests

The authors declare no conflict of interest.

DECLARATIONS

Ethics Approval and Consent to Participate

The protocols involved in this ancillary study were approved by an Institutional Review Board and all participants gave informed consent before participating.

all comparisons, except for the SPPB score, which used Cohen's kappa and Krippendorf's alpha for intra- and inter-rater reliability, respectively.

Results: Remote assessment of the TUG test had excellent intra-rater reliability (0.98, 95% CI 0.93–0.99), inter-rater reliability (ICC = 0.96, 95% CI 0.90–0.99), and good agreement with in-person tests (ICC = 0.88, 95% CI 0.74–0.94). The 5XSTS and 4mWT showed excellent (ICC = 0.92, 95% CI 0.84–0.96) and good (ICC = 0.87, 95% CI 0.71–0.94) intra-rater reliability, respectively, but somewhat lower inter-rater reliability (5XSTS: ICC = 0.65, 95% CI 0.34–0.83 and 4mWT: ICC = 0.62, 95% CI 0.30–0.81). Remote 5XSTS had moderate agreement (ICC = 0.72, 95% CI 0.62–0.80) and 4mWT had poor agreement (ICC = 0.48, 95% CI −0.07–0.76) with in-person tests.

Conclusions: Remote assessment of common physical function tests in older adults, including those who have cancer, is feasible and highly reliable when using the same assessor. TUG may be the most methodologically robust measure for remote assessment because it is also highly reliable when using different assessors and correlates strongly with in-person testing. Adapting administration of objective measures of physical function for the remote environment could significantly expand the reach of research and clinical practice to assess populations at risk of functional decline.

Keywords

Physical function; cancer survivors; older adults; COVID-19; video technology; reliability; measurement

INTRODUCTION

Patient-reported physical functioning assesses a person's perception of their physical abilities can be obtained through questionnaires or interviews no matter where patients live (1). Patient-reported physical functioning can independently predict short-term mortality and nursing home admissions in older adults (2). While patient-reported measures are highly practical, objective assessment of physical functioning is more reliable, detects changes earlier and with more sensitivity, and less prone to bias (2–4). For example, in a study comparing older breast cancer survivors to age-matched healthy controls, survivors reported physical function scores about 10% lower than their peers, whereas scores on objective function tests were up to 25% lower in survivors than controls (5).

The Short Physical Performance Battery (SPPB) is a test of physical functioning designed to objectively assess lower-body physical function in older adults and is comprised of a sum score of three subtests: standing balance, usual walk speed (4mWT), and 5 time sit-to-stand (5XSTS). Low SPPB scores correlate with disability (6, 7), declines in mobility (8) activities of daily living (ADLs), hospitalization, and mortality (8, 9). The SPPB is typically administered in-person and has acceptable internal consistency (2), ability to detect clinically meaningful change (10), and versatility as it can be administered in clinic or home settings (9). The Timed Up and Go (TUG) test is another objective and reliable measure of functional mobility that is also feasible in a clinic setting (11) and predicts health outcomes, such as fall risk in community-dwelling adults (12). In cancer survivors,

poor TUG and SPPB scores correlate with treatment-related complications, higher rates of functional decline, and decreased survival (13).

A notable limitation with objective measures is that in-person assessment requires travel by the patient to a facility or by the assessor to the patient's home (14, 15). In early 2020, the COVID-19 pandemic disrupted clinical trials across the world, forcing studies to forgo in-person data collection (e.g., biologic samples, physical assessments) or to adapt protocols for remote delivery using approaches such as videoconferencing. Our laboratory suspended in-person study activities in two large NIH-funded exercise trials, including assessments of objective physical function. In turn, we adapted protocols for both exercise training (16) and for assessment of physical function to remote formats using videoconference technology.

While reliability estimates for in-person administration of SPPB and TUG have been established (17–19), there are very few published estimates of reliability of remote administration (20) or data comparing scores collected in a remote setting to scores collected in-person (21, 22). One recent study in a veteran population assessed inter-rater reliability of remote assessments using two assessors who observed a single participant on the same videoconference for repetition-based functional tests (i.e., # arm curls completed in 30 seconds), but did not measure intra-rater reliability or compare remote assessments to laboratory based tests (20). Another study conducted during COVID-19 in a small sample of otherwise healthy adults compared remote assessment of functional tests performed in the home to assessments repeated in an outdoor setting, but did not include reliability assessments of either type of delivery setting (21). To date there is no study that has assessed both inter-rater and intra-rater reliability of remote functional tests along with comparison of remote testing to traditional laboratory-based testing and included participants of varying ranges and health status. Thus, we seized an opportunity to use participants from our ongoing trials to conduct an ancillary study to estimate the measurement properties of physical function tests conducted by remote assessment using videoconferencing technology. The purpose of this study was to estimate the intra- and inter-rater reliability of remotely assessed SPPB, 5XSTS, 4mWT, and TUG tests in a sample of middle-aged and older adults with and without cancer. Eventual resumption of in-person visits allowed us to also estimate the agreement between tests administered remotely to those administered in-person.

METHODS

Participants

We used data collected in two on-going National Institutes of Health-funded randomized controlled exercise trials, details of which are described in detail elsewhere (23, 24). Study 1 (NCT03630354) enrolls breast and prostate cancer survivors within three years of diagnosis plus their intimate partner and Study 2 (NCT03741335) enrolls prostate cancer survivors treated with androgen deprivation therapy (ADT). All participants involved completed two testing assessments on separate days. Sixty-five participated in repeat remote testing appointments and 114 participated in both in-person and remote testing. Three of the participants that completed intra-rater reliability also participated in in-person testing and therefore their data was used in both the intra-rater reliability analysis as well as the

analysis that compared remote testing to in-person testing. Testing was completed during a combination of baseline and follow-up visits concordant with participants' timeline in the ongoing trials. Both studies included the SPPB as an outcome measure, while Study 2 also included the TUG. Both in-person and remote study protocols were approved by the OHSU Institutional Review Board and all participants provided informed consent prior to participation in testing. Participants with upcoming study visits were asked if they were willing to participate in an additional remote testing session for intra- or inter-rater reliability testing, whereas participation in remote and in-person testing became standard testing procedure for scheduled study visits from August 2020 onward.

Procedures

Assessors—Seven experienced assessors conducted the assessments. Only two assessors were involved in any given reliability or agreement assessment and conducted their assessment on separate days. Assessors followed Standard Operating Procedures (SOP) to better ensure tests were administered the same way regardless of delivery format. Prior to initiating remote testing in the studies, a project director observed up to 5 testing sessions per assessor to ensure uniformity across assessors and trials.

Test sessions—Videoconference-based testing was conducted securely via Cisco Webex Meetings with both assessors and participants connecting from their home residence. Participants had to have the following resources for remote assessments: internet access; computer or tablet with video capabilities; 16' space; armless, non-rolling, straight-backed, standard height chair; and a measuring tape ≥ 14'. If a participant lacked any resources, items were mailed or delivered to them. If space was inadequate for a given test, data collection did not occur for only that measure. If poor video quality interfered with test administration (i.e., lags or skips) and could not be remediated the session was rescheduled. For reliability, each participant completed a testing assessment with their regularly appointed assessor. A second remote testing assessment was scheduled to be completed on a different day by either the same assessor (intra-rater) or by a different assessor (inter-rater) from each assessor's home and then scores were compared. For inter-rater reliability the second assessor was picked based on availability and assessor pairs and the order they tested varied. Five assessors were involved in the intra-rater reliability testing. Six assessors were involved in the inter-rater reliability testing: five assessors with 10 testing pair combinations for TUG and five assessors with nine testing pair combinations for SPPB. For agreement estimates, in-person tests were conducted in our laboratory following established SOPs (2, 25) after OHSU permitted the resumption of in-person research visits. Six assessors were involved in agreement testing and the order of testing was not formally randomized because this study was an add-on to ongoing trials and tests were scheduled based on participant convenience. Of the 114 agreement assessments, 82% were conducted in-person first and 18% conducted remote testing first. Repeat assessment visits were not completed on the same day to minimize any confounding from fatigue, but were to be completed within a timeframe where we would expect participant performance to be stable (e.g., 1–14 days) and to accommodate participant schedules. Adverse events (AEs) during testing were routinely tracked in both trials and used to assess safety of remote testing (23, 24).

Test Protocols—The SPPB consists of three timed assessments: standing balance, 4mWT, and 5XSTS. Each test is scored from 0 to 4, then scores are summed (SPPB sum) for a range of possible scores of 0–12 (2). Higher scores indicate better physical function. We used the SPPB subtest and sum scores and the continuous values of the 5XSTS (sec) and 4mWT (m/s) for analysis. For the TUG, the time it takes a participant to rise from a chair, walk at their usual pace for 3-meters, turn around, and return to a seated position in the chair was used for analysis (26). We adapted SOPs for remote administration [Supplement 1] to stay as closely as possible to in-person administration. The supplement includes detailed outlines of equipment, videoconferencing setup, and SOPs for each subtest of the SPPB and the TUG with modified language for instructing participants to setup optimal camera placement and photo examples. Key adaptations for remote administration are summarized in Fig. 1. Safety precautions included ensuring walk courses were free of obstacles, chairs were stationed against a fixed object, balance activities took place near a fixed object, another person in the participant's home was on call nearby and/or an emergency contact number was on hand. Raters used the same stopwatch that they use for in-person tests to time each assessment.

Statistical Analysis

Descriptive statistics were calculated for the subsamples used for each analysis. Intraclass correlations (ICC) were used to estimate intra- and inter-rater reliability for all continuous outcomes (5XSTS, 4mWT, and TUG). Decisions about ICC type (single assessor versus average of multiple assessors in future applications), model (one-way, two-way, random vs mixed effects), and reliability estimate (absolute agreement versus consistency) followed Koo & Li's (27) recommendations and flow chart. All ICCs used single type which assumes future assessments would be conducted by one assessor and absolute agreement as tests are used to measure change over time in longitudinal studies. Intra-rater reliability used two-way mixed models to account for multiple scores from the same assessor while inter-rater reliability used one-way random models as assessors were not the same for all participants. Agreement between remote and in-person tests was determined using ICC two-way mixed-effects models with absolute agreement since each measurement from the subject was rated by the same assessor. Interpretations of the ICCs were based on Koo & Li's (27) guidelines: ICC values <0.5 = poor reliability; 0.5–0.75 = moderate reliability; 0.75–0.9 = good reliability; and >0.9 = excellent reliability.

The SPPB sum and subtests scores and the standing balance test itself are discrete interval measures that do not fit the assumptions required for ICC. Intra-rater reliability and agreement of SPPB sum and standing balance between the two settings were estimated using Cohen's kappa. This statistic is most appropriate for a single assessor performing two evaluations and is not ideal for multiple raters as in this study. However, multiple raters increase the error and decreases the magnitude of kappa, thus estimates were conservative. Inter-rater reliability of SPPB sum and standing balance was estimated using Krippendorf's alpha with quadratic weighting, which allows for multiple assessors who do not need to rate all participants (28). An alpha of 0 suggests agreement equivalent to chance, while an alpha of 1 indicates perfect agreement. The interval rating provides a weighting structure to adjust for ordinal discrete variables (29). Krippendorf tentatively suggested an alpha of 0.667 as the lowest conceivable limit for equivalence, but stressed that there is no single

valid cutoff (30, 31). Additionally, we calculated the joint probability of agreement for SPPB and its subtests where participants had exactly the same score within the intra-rater (agreement within assessors), inter-rater (agreement between assessors), and agreement testing conditions (agreement between remote and in-person). All analyses were performed using R (R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) and the irr package (Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr).

RESULTS

Participant Characteristics

Overall, data from 118 participants in Study 1 and 58 participants in Study 2 were used. The number of cases for each analysis varied, thus characteristics of each subsample is provided (Table 1). In brief, 32 participants completed intra-rater reliability testing, 33 completed inter-rater reliability testing, and 114 completed remote and in-person testing. On average, the entire sample were relatively healthy (Charlson Comorbidity Index score: 2.2 ± 2.2) older adults (62.5 ± 11.5 years). The study sample for the TUG test averaged about a decade older than the sample for SPPB, but the latter includes a broader age range and both genders. Participants in the study were primarily in Oregon and 25% lived in rural counties, but the sample included several participants living in one of 3 nearby states (Washington, Idaho, Nevada). The median number of days between any two testing appointments across all test conditions was seven. For the SPPB agreement analysis, there were 18 cases where remote and in-person testing occurred more than 14 days apart. A sensitivity analysis removing these cases did not substantively change the results so these cases were retained. During our transition from in-person to remote testing, 99% of our sample were able to complete remote testing instead of the originally scheduled in-person testing with only one participant who had insufficient internet to do any tests and five participants who did not have enough space to do walk tests. Of the 176 participants in our ancillary study, we had to deliver equipment to 15 of them (<9%), including the following items: One tape measure, three chairs, two webcams, seven tablet stands, one laptop, and one cellular-enabled tablet. Internet lag or disruptions occurred periodically, but only led to a complete reschedule of a testing appointment in two sessions (1% of all tests). There were no AEs during any remote assessment.

Intra-rater reliability

Using the same assessor to administer remote tests (Table 2), TUG and 5XSTS had excellent reliability (0.96, 95% CI 0.90–0.99 and 0.92, 95% CI 0.84–0.96, respectively) while reliability for 4mWT was good (0.87, 95% CI 0.70–0.94). SPPB sum had a Cohen's kappa of 0.78 and standing balance had a kappa of 0.82, both of which imply substantial agreement between repeat assessments (32).

Inter-rater reliability

Using different assessors to administer remote tests (Table 3), TUG showed excellent reliability (0.98, 95% CI 0.93–0.99). Reliability for both 5XSTS and 4mWT was moderate (0.65, 95% CI 0.34–0.83 and 0.62, 95% CI 0.30–0.81, respectively). Krippendorf's alpha was 0.59 for SPPB sum, which implies unacceptable agreement between different assessors. For standing balance nearly all participants scored a 4 on successive tests, providing a skewed distribution that would not generate a reliable alpha when using multiple assessors.

Agreement between testing formats

Remote administration of TUG (Table 4) had good agreement (0.88, 95% CI 0.74–0.94), while 5XSTS had moderate agreement (0.72, 95% CI 0.62–0.80) and 4mWT had poor agreement (0.48, 95% CI −0.07–0.76) with in-person tests. SPPB sum and standing balance had slight to moderate agreement between test formats (kappa = 0.38 and 0.46, respectively). Performance on all tests were better (or faster) when participants performed tests in-person compared to home, with the exception of standing balance. The average observed difference for TUG was 0.14 (95% CI −0.40, 0.67), SPPB sum = 0.25 (95% CI 0.06, 0.43), 5XSTS = 0.20 (−0.27, 0.68), and 4mWT = 0.14 (95% CI 0.12, 0.16). Standing balance, which trended in the other direction, had an observed mean difference of −0.11 (95% CI −0.20, 0.68), but these scores had very little variability and demonstrated a ceiling effect making reliability estimates less robust.

Agreement of SPPB scores across comparisons

Using the same guidelines for interpretation as used for ICC and the discrete scores for the SPPB, standing balance scores had the best absolute agreement (Table 5) across all types of re-tests with agreement for intra-rater reliability = 92.6% (excellent), inter-rater reliability = 87.5% (good), and agreement = 86.0% (good). 4mWT scores had good absolute agreement for intra- (81.5%) and inter-rater (87.5%) reliability and moderate agreement for comparison between remote and in-person testing (65.8%). 5XSTS scores had moderate agreement for intra-rater reliability = 74.1%, inter-rater reliability = 58.3% and agreement = 60.5%. SPPB sum scores had poor agreement across all types of re-tests with intra-rater reliability = 59.3%, inter-rater reliability = 50.0%, and agreement = 43.9%.

DISCUSSION

In an effort to sustain our research during COVID-19, we successfully adapted established objective measures of physical function used in clinical care and research in older adults and clinical populations at risk of functional decline (4, 9). Remote assessment was both feasible and safe as 99% of our sample originally slated for in-person testing completed assessments remotely, instead, and without injury. To our knowledge, we are among the first to assess reliability of two widely-used functional tests, the SPPB and TUG, that both have high utility and predictive validity. We are also among the first to make comparisons between remote and in-person administration. The TUG test appears the most robust when assessed remotely, showing excellent intra- and inter-rater reliability and good agreement with in-person assessment. Compared to the TUG, the intra-rater reliability of the SPPB and the individual tests of 5XSTS and 4mWT were also very strong, but these tests had

somewhat lower inter-rater reliability when assessors administered remote visits out of different homes and was less consistent with in-person tests. Test scores were generally better in-person than under remote conditions in the home, a finding consistent with other studies comparing walk tests conducted in a lab to tests conducted in the home (33–35). Even though under both conditions in our study participants were observed by an assessor, the laboratory setting may induce more of a “white coat” effect that triggers the sympathetic nervous system and results in higher effort (36). Thus, laboratory-based tests may measure a person’s best performance, whereas a home assessment may be a more ecologically valid measure of daily functioning in a person’s home environment.

Of the two tests, the TUG was most robust. The consistency of the TUG test across assessors and settings might be attributable to the elements of the test that make it simpler to administer using videoconferencing technology. Specifically, the slower pace (i.e., usual vs. fast) and easier video angle to capture the end of the test (i.e., return to upright sitting position) might allow for better consistency and accuracy of timing, despite varying technologies between homes, like internet speed. Remote administration of the subtests of the SPPB was very consistent within assessors, but may be more susceptible to differences between assessors due to the ways these tests are administered and the impact of varying technology between homes. Chair stands are completed for speed and thus may be more susceptible to differences in video quality between assessor homes that leads to subtle, but variable transmission speeds. Further, it may be more difficult for assessors to ensure quality of test performance (i.e., full stand and full sit) using only verbal feedback during remote tests compared to verbal and tactile feedback that is used in-person, potentially adding another source of error. For the 4mWT, in addition to the variation in video quality between assessor homes, accurate timing is also dependent upon the video angle setup to capture the foot crossing the finish line and could introduce variability between assessors in how cameras are positioned and when foot crossing is timed. To date, only one other recently published study has reported on remotely administered functional tests. They reported high inter-rater reliability on repetition-based functional tests (i.e., 30-second arm curl and chair stand tests and 2-minute step test), but two assessors simultaneously timed a single participant on the same video conference (20), which differed from how we conducted our inter-rater reliability assessments.

In most cases, our reliability estimates for remote administration are comparable to those reported for the same measures conducted only in-person. Our remote intra-and inter-rater reliability for TUG is consistent with what is found in the literature for in-person testing (18, 19). Among pulmonary hypertension patients in an out-patient clinic setting, TUG intra-rater reliability (0.96, 95% CI 0.93, 0.98) (18) was nearly identical to that of our remote assessments (0.96, 95% CI 0.90, 0.99). Inter-rater reliability of the TUG test in a small sample (n=11) of a geriatric out-patient center (0.98, 95% CI 0.93, 1.00) (19) was also nearly identical to our remote TUG inter-rater reliability (0.98, 95% CI 0.93, 0.99). Our remote 4mWT intra- and inter-rater reliability is also comparable to reliability estimates reported for in-person administration in people with and without asthma which reported good intra-rater reliability (0.86, 95% CI 0.73, 0.92) and moderate inter-rater reliability (ICC = 0.58 (95% CI 0.26, 0.76) (17). For 5XSTS our intra-rater reliability for remote administration (ICC = 0.92) exceeds that reported in a meta-analysis (n=779 participants)

for in-person (adjusted mean ICC of 0.81) (37). However, our inter-rater reliability estimates for the 5XSTS were lower than that reported in another meta-analysis in 400 participants of mixed health status (ICC = 0.947, 95% CI 0.88, 0.97; range 0.74 to 0.99) (38). Some of the differences between in-person reliability estimates and ours on certain timed tests could be related to varying video quality in the home setup of different assessors. We could not avoid these differences during COVID-19 where staff were under stay-at-home orders. However, homogenizing the videoconferencing setup across assessors by having them use the same hardware and software in a central location, as opposed to their different computers in different homes, might remove a source of error and improve reliability across assessors.

Scoring for the SPPB converts timed scores to discrete values and likely affected our reliability estimates for this test. Even with high interrater reliability in the timed tests used for the SPPB, minor differences in timing between assessors can result in different discrete SPPB scores. In addition, potential ceiling effects when testing younger and/or or high functioning adults, such as what we observed for the standing balance subtest, can further limit variability needed for robust reliability estimates. These limitations introduced by SPPB scoring fundamentally influence the degree of agreement that is possible when comparing remote to in-person formats. These preliminary findings suggest that studies should have a single assessment format and not interchange remote and in-person administration of the SPPB.

Use of videoconferencing technology introduces potential unique sources of error, otherwise not previously considered when trying to maximize reliability of repeat assessments. Our staff worked with the OHSU Information Technology Group (ITG) to try to uncover new sources of error when conducting remote assessments of physical functioning. Inspection of recorded repeat assessments and additional mock testing sessions helped to identify technology-specific sources of error, such as internet connectivity, bandwidth speed, and audio and visual quality (39). Though we could not specifically quantify error, OHSU ITG provided guidelines for both assessors and participants to maximize call quality (Fig. 2). A recently published study on remote assessments of other types of physical performance tests noted similar technical difficulties (20), highlighting a consistent limitation with technology.

Our ancillary study had strengths and limitations. We successfully adapted well established objective measures of physical functioning for remote assessment using videoconference technology and in a sample that included older adults and cancer survivors in whom these tests are most relevant. Older adults are sometimes assumed to have limited internet and computer skills and there may be concerns about their safety when doing performance-based tests alone in a home setting. However, nearly our entire sample was able to easily complete remote assessment and only a few needed additional resources, which were easily mailed, and none experienced an adverse event during testing. A quarter of our sample also lived in rural areas, which is another presumed limitation to remote delivery approaches. Thus, the age, geographic and clinical diversity of our sample strongly suggests that our findings could generalize to a broader sample of older adults, rural/urban settings and chronic illnesses. It is noteworthy, though, that our sample consisted of trial participants who were willing and able to participate in moderate-intensity exercise training and thus feasibility and safety of remote assessment in more frail or medically complex adults needs to be established. An

alternative option could be the virtual SPPB (vSPPB) where a participant self-assesses their performance against computer animated test performances (40). Though the virtual SPPB is not recommended as a substitute for a person's actual performance it could be a reasonable alternative for testing older, frail populations due to safety concerns and/or space limitations in the home. Due to the logistics of implementing this ancillary study amidst state and institutional pandemic-related restrictions, we were not able to randomize the order of our remote and in-person testing. The size of our sample also varied across the different types of comparisons, thus further confirmation of the measurement properties of remote assessments of physical function in larger samples is needed. Our sample also lacked racial and ethnic diversity thus it is unclear whether or not the measurement properties of remote assessment of physical function tests is similar in non-white adults. Finally, we only used a single type of videoconferencing software, Cisco Webex Meetings, to conduct our remote tests and thus our results may not generalize when using other software.

In summary, our ancillary findings suggest that two very well-established standard tests of physical functioning, the SPPB and TUG tests, that have always been conducted in-person, can be successfully adapted for remote administration through video conferencing technology with evidence for feasibility, safety, reliability and agreement with in-person assessments. Remote administration of the TUG test, in particular, was highly consistent between assessors and concordant with tests administered in-person suggesting it could be used both in large, longitudinal studies but also possibly to screen for fall and disability risk in individuals. Remote administration of the SPPB may also be useful in longitudinal studies, but may require more standardization when using multiple assessors for remote assessments. Certainly, more work in this area will be forthcoming as the incorporation of video conferencing in the conduct of research and practice have accelerated during the COVID-19 pandemic. The added value of adapting functional tests for remote administration is the widened potential to reach a broader population who may have been less able to participate in in-person tests because of distance, (e.g., rural areas), who are otherwise unable to travel on their own (e.g., homebound elderly, without transportation) or who are unwilling to come to a central facility. Better representation of the aging and cancer survivor population in research that aims to reduce rates of disability and improve outcomes will ultimately lead to better population health.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank the participants in this study, for their willingness to complete additional testing visits. We would also like to acknowledge the research staff who conducted the testing assessments: Joanna Chadd, Christopher Chalmers, Pablo Herrera-Fuentes, Colin Lipps, Ashley Lyons, Jade Moon, and Christopher Palmer. Special thanks to Ramyar Eslami for working with OHSU ITG to identify strategies to improve video quality during remote assessments.

Funding

Funding for this study was provided by NIH 1R01CA218093 and 1R01CA222605 to Dr. Winters-Stone. Dr. Winters-Stone is also funded in part by Cancer Center Support Grant 5P30CA069533-23.

Availability of Data and Materials

The author's state that they have full control of all primary data and that they agree to allow the journal to review their data if requested.

REFERENCES

1. Reuben DB, Seeman TE, Keeler E, Hayes RP, Bowman L, Sewall A, et al. Refining the Categorization of Physical Functional Status: The Added Value of Combining Self-Reported and Performance-Based Measures. The Journals of Gerontology: Series A. 2004;59(10):M1056–M61.
2. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, 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(2):M85–94. [PubMed: 8126356]
3. Brach JS, VanSwearingen JM, Newman AB, Kriska AM. Identifying early decline of physical function in community-dwelling older women: performance-based and self-report measures. Phys Ther. 2002;82(4):320–8. [PubMed: 11922849]
4. Latham NK, Mehta V, Nguyen AM, Jette AM, Olarsch S, Papanicolaou D, et al. Performance-based or self-report measures of physical function: which should be used in clinical trials of hip fracture patients? Arch Phys Med Rehabil. 2008;89(11):2146–55. [PubMed: 18996244]
5. Winters-Stone KM, Medysky ME, Savin MA. Patient-reported and objectively measured physical function in older breast cancer survivors and cancer-free controls. J Geriatr Oncol. 2019;10(2):311–6. [PubMed: 30344000]
6. Guralnik JM, Ferrucci L, Pieper CF, Leveille SG, Markides KS, Ostir GV, et al. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. J Gerontol A Biol Sci Med Sci. 2000;55(4):M221–31. [PubMed: 10811152]
7. Ostir GV, Markides KS, Black SA, Goodwin JS. Lower body functioning as a predictor of subsequent disability among older Mexican Americans. J Gerontol A Biol Sci Med Sci. 1998;53(6):M491–5. [PubMed: 9823755]
8. Perera S, Studenski S, Newman A, Simonsick E, Harris T, Schwartz A, et al. Are estimates of meaningful decline in mobility performance consistent among clinically important subgroups? (Health ABC study). J Gerontol A Biol Sci Med Sci. 2014;69(10):1260–8. [PubMed: 24615070]
9. Studenski S, Perera S, Wallace D, Chandler JM, Duncan PW, Rooney E, et al. Physical performance measures in the clinical setting. J Am Geriatr Soc. 2003;51(3):314–22. [PubMed: 12588574]
10. Gill TM. Assessment of function and disability in longitudinal studies. J Am Geriatr Soc. 2010;58 Suppl 2:S308–12. [PubMed: 21029059]
11. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc. 1991;39(2):142–8. [PubMed: 1991946]
12. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test. Physical Therapy. 2000;80(9):896–903. [PubMed: 10960937]
13. Verweij NM, Schiphorst AHW, Pronk A, van den Bos F, Hamaker ME. Physical performance measures for predicting outcome in cancer patients: a systematic review. Acta Oncologica. 2016;55(12):1386–91. [PubMed: 27718777]
14. Dibartolo MC, McCrone S. Recruitment of rural community-dwelling older adults: Barriers, challenges, and strategies. Aging & Mental Health. 2003;7(2):75–82. [PubMed: 12745386]
15. Hensen B, Mackworth-Young CRS, Simwinga M, Abdelmagid N, Banda J, Mavodza C, et al. Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. Health Policy Plan. 2021;36(3):360–8. [PubMed: 33881138]
16. Winters-Stone KM, Boisvert C, Li F, Lyons KS, Beer TM, Mitri Z, et al. Delivering Exercise Medicine to Cancer Survivors: Has COVID-19 shifted the landscape for how and who can be reached with supervised group exercise? Journal Supportive Care Cancer. 2021.

17. Oliveira JMd, Spositon T, Cerci Neto A, Soares FMC, Pitta F, Furlanetto KC. Functional tests for adults with asthma: validity, reliability, minimal detectable change, and feasibility. Journal of Asthma. 2020:1–9.
18. Ozcan Kahraman B, Ozsoy I, Akdeniz B, Ozpelit E, Sevinc C, Acar S, et al. Test-retest reliability and validity of the timed up and go test and 30-second sit to stand test in patients with pulmonary hypertension. International Journal of Cardiology. 2020;304:159–63. [PubMed: 31980271]
19. Kristensen MT, Bloch ML, Jønsson LR, Jakobsen TL. Interrater reliability of the standardized Timed Up and Go Test when used in hospitalized and community-dwelling older individuals. Physiotherapy Research International. 2019;24(2):e1769. [PubMed: 30657232]
20. Ogawa EF, Harris R, Dufour AB, Morey MC, Bean J. Reliability of Virtual Physical Performance Assessments in Veterans During the COVID-19 Pandemic. Archives of Rehabilitation Research and Clinical Translation. 2021;3(3):100146. [PubMed: 34589696]
21. 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. The journal of nutrition, health & aging. 2021.
22. Holland AE, Malaguti C, Hoffman M, Lahham A, Burge AT, Dowman L, et al. Home-based or remote exercise testing in chronic respiratory disease, during the COVID-19 pandemic and beyond: A rapid review. Chron Respir Dis. 2020;17:1479973120952418. [PubMed: 32840385]
23. Winters-Stone K, li F, Horak F, Dieckmann N, Hung A, Amling C, et al. Protocol for GET FIT Prostate: A Randomized, Controlled Trial of Group Exercise Training for Fall Prevention and Functional Improvements During and After Treatment for Prostate Cancer2021.
24. Winters-Stone KM, Lyons KS, Dieckmann NF, Lee CS, Mitri Z, Beer TM. Study protocol for the Exercising Together© trial: a randomized, controlled trial of partnered exercise for couples coping with cancer. Trials. 2021;22(1):579. [PubMed: 34461975]
25. Guralnik JM. SHORT physical PERFORMANCE battery (sppb) guide 2017 [Available from: https://sppbguide.com/.
26. Thompson M, Medley A. Performance of Community Dwelling Elderly on the Timed Up and Go Test. Physical & Occupational Therapy In Geriatrics. 1995;13(3):17–30.
27. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–63. [PubMed: 27330520]
28. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement. 1973;33(3):613–9.
29. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy. 2005;85(3):257–68. [PubMed: 15733050]
30. Krippendorff K Content analysis: An introduction to its methodology 2nd ed2004.
31. Krippendorff K Reliability in content analysis: Some common misconceptions and recommendations. Human communication research 2004.
32. Glen S Cohen's Kappa Statistic Statistics How To: StatisticsHowTo.com: Elementary Statistics for the rest of us!; 2014 [StatisticsHowTo.com: Elementary Statistics for the rest of us!]. Available from: https://www.statisticshowto.com/cohens-kappa-statistic/.
33. Portegijs E, Karavirta L, Saajanaho M, Rantalainen T, Rantanen T. Assessing physical performance and physical activity in large population-based aging studies: home-based assessments or visits to the research center? BMC Public Health. 2019;19(1):1570. [PubMed: 31775684]
34. Brodie MAD, Coppens MJM, Lord SR, Lovell NH, Gschwind YJ, Redmond SJ, et al. Wearable pendant device monitoring using new wavelet-based methods shows daily life and laboratory gaits are different. Medical & Biological Engineering & Computing. 2016;54(4):663–74. [PubMed: 26245255]
35. Zampieri C, Salarian A, Carlson-Kuhta P, Nutt JG, Horak FB. Assessing mobility at home in people with early Parkinson's disease using an instrumented Timed Up and Go test. Parkinsonism Relat Disord. 2011;17(4):277–80. [PubMed: 20801706]
36. Pioli MR, Ritter AM, de Faria AP, Modolo R. White coat syndrome and its variations: differences and clinical impact. Integr Blood Press Control. 2018;11:73–9. [PubMed: 30519088]

37. Bohannon RW. Test-Retest Reliability of the Five-Repetition Sit-to-Stand Test: A Systematic Review of the Literature Involving Adults. The Journal of Strength & Conditioning Research. 2011;25(11):3205–7. [PubMed: 21904240]

38. Muñoz-Bermejo L, Adsuar JC, Mendoza-Muñoz M, Barrios-Fernández S, Garcia-Gordillo MA, Pérez-Gómez J, et al. Test-Retest Reliability of Five Times Sit to Stand Test (FTSST) in Adults: A Systematic Review and Meta-Analysis. Biology. 2021;10(6):510. [PubMed: 34207604]

39. Eslami R Remote Data Collection: How Broadband Connectivity Affected the Conversion of Physical Performance Testing to the Remote Setting During COVID-19: A Research Assistant's Assessment: Portland State University; 2021.

40. Marsh AP, Wrights AP, Haakonssen EH, Dobrosielski MA, Chmelo EA, Barnard RT, et al. The Virtual Short Physical Performance Battery. The Journals of Gerontology: Series A. 2015;70(10):1233–41.

- Raters gave verbal instructions to the participants and showed visual demonstrations by sharing their screen
- Participants set up their own testing space, including measuring out their own walking courses
- Participants were instructed on how to setup optimal video angles for viewing

Fig. 1.
Summary of key adaptations from in-person to remote administration of Timed Up and Go and Short Physical Performance Battery Standard Operating Procedures

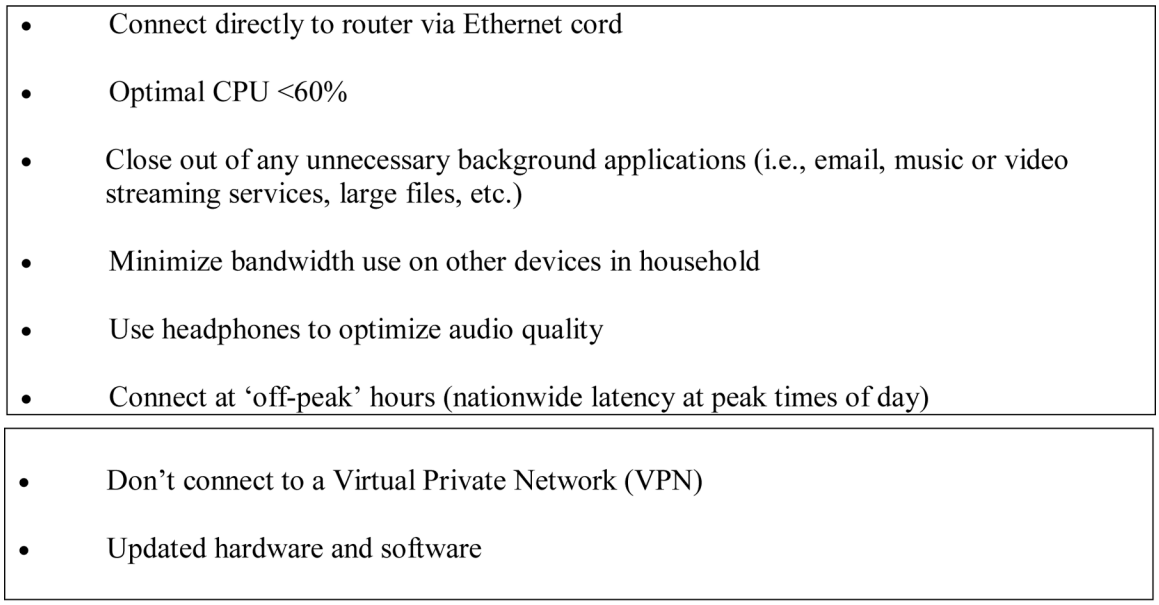

Fig. 2.
Remote Videoconferencing Tips
Abbreviations: CPU- Central Processing Unit

Table 1

Sample characteristics by assessment type and measure

	Intra-rater		Inter-rater		Remote to In-Person Agreement	
Characteristics	**TUG n=18**	**SPPB n=27**	**TUG n=15**	**SPPB n=24**	**TUG n=25**	**SPPB n=114**
Age, (years)	72.2 ± 7.7	63.7 ± 9.8	75.0 ± 6.8	59.6 ± 13.0	71.0 ± 6.8	61.7 ± 11.0
Gender						
Female	0 (0.0%)	7 (25.9%)	0 (0.0%)	9 (37.5%)	0 (0.0%)	47 (41.2%)
Male	18 (100%)	20 (74.1%)	15 (100%)	15 (62.5%)	25 (100%)	67 (58.8%)
Cancer survivors	18 (100%)	20 (74.1%)	15 (100%)	15 (62.5%)	25 (100%)	70 (61.4%)
Cancer Type						
Breast	0 (0.0%)	6 (22.2%)	0 (0.0%)	7 (29.2%)	0 (0.0%)	43 (37.7%)
Prostate	18 (100%)	14 (51.9%)	15 (100%)	8 (33.3%)	25 (100%)	27 (23.7%)
Comorbidities	2.5 ± 2.4	1.9 ± 2.1	3.1 ± 2.1	1.8 ± 1.9	3.6 ± 2.8	2.3 ± 2.2
Difference in days between testing	5.2 ± 1.6	5.9 ± 1.3	4.5 ± 2.2	5.3 ± 1.8	6.3 ± 2.5	11.2 ± 11.1
Difference in days (median (range))	6 (2–7)	6 (2–7)	5 (1–7)	5.5 (2–7)	6 (2–14)	7(2–51)

Abbreviations: TUG- Timed Up and Go; SPPB- Short Physical Performance Battery

[a]Measured using the Charlson Comorbidity Index

Table 2

Intra-rater reliability of remote administration of the Timed Up and Go and Short Physical Performance Battery tests

	n[a]	ICC (95% CI)	Cohen's kappa
TUG (sec)	18	0.96 (0.90, 0.99)	
SPPB sum	27		0.78
Standing balance	27		0.82
5XSTS (sec)	27	0.92 (0.84, 0.96)	
4mWT (m/s)	27	0.87 (0.71, 0.94)	

Abbreviations: ICC- Intraclass Correlation Coefficient; TUG- Timed Up and Go; SPPB- Short Physical Performance Battery; 5XSTS- 5-time sit-to-stand; 4mWT- 4-meter walk test

[a] n = number of participants tested for each measure

Table 3

Inter-rater reliability of remote administration of the Timed Up and Go and Short Physical Performance Battery tests[a]

Measure	n[b]	ICC (95% CI)	Krippendorf's alpha
TUG (sec)	15	0.98 (0.93, 0.99)	
SPPB sum	24		0.59
5XSTS (sec)	24	0.65 (0.34, 0.83)	
4mWT (m/s)	24	0.62 (0.30, 0.81)	

Abbreviations: ICC- Intraclass Correlation Coefficient; TUG- Timed Up and Go; SPPB- Short Physical Performance Battery; 5XSTS- 5-time sit-to-stand; 4mWT- 4-meter walk test

[a] Unable to calculate alpha for standing balance

[b] n = number of participants tested for each measure

Table 4

Agreement of remote administration of the Timed Up and Go and Short Physical Performance Battery against in-person administration

	n[a]	ICC (95% CI)	Cohen's kappa	Difference Mean[b] (95% CI)
TUG (sec)	25	0.88 (0.74, 0.94)		0.14 (−0.40, 0.67)
SPPB sum	114		0.38	0.25 (0.06, 0.43)
Standing balance	114		0.46	−0.11 (−0.20, −0.02)
5XSTS (sec)	114	0.72 (0.62, 0.80)		0.20 (−0.27, 0.68)
4mWT (m/s)	114	0.48 (−0.07, 0.76)		0.14 (0.12, 0.16)

Abbreviations: ICC- Intraclass Correlation Coefficient; TUG- Timed Up and Go; SPPB- Short Physical Performance Battery; 5XSTS- 5-time sit-to-stand; 4mWT- 4-meter walk test

[a] n = number of participants tested for each measure

[b] Difference (In-person – Remote)

Table 5

Agreement of the Short Physical Performance Battery and subtest scores across test conditions

Agreement Type	SPPB sum[a]	Standing Balance[b]	4mWT[b]	5XSTS[b]
Agreement within assessors (intra-rater)	59.3%	92.6%	81.5%	74.1%
Agreement between assessors (inter-rater)	50.0%	87.5%	87.5%	58.3%
Agreement between remote and in-person testing	43.9%	86.0%	65.8%	60.5%

Abbreviations: SPPB- Short Physical Performance Battery; 4mWT- 4-meter walk test; 5XSTS- 5-time sit-to-stand

[a] SPPB sum score ranges 0–12

[b] SPPB, Standing Balance, 4mWT, and 5XSTS score ranges 0–4.