International Journal of COPD

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Intra- and inter-rater reproducibility of the 6-minute walk test and the 30-second sit-to-stand test in patients with severe and very severe COPD

This article was published in the following Dove Press journal:
International Journal of COPD

Henrik Hansen[1]
Nina Beyer[2]
Anne Frølich[1]
Nina Godtfredsen[2,3]
Theresa Bieler[4]

[1]Research Unit of Chronic Diseases and Telemedicine – Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg University Hospital, Copenhagen, Denmark; [2]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [3]Department of Respiratory Medicine, Hvidovre University Hospital, Hvidovre, Denmark; [4]Department of Physical and Occupational Therapy, Bispebjerg and Frederiksberg Hospital, University of Copenhagen, Copenhagen, Denmark

Background: In patients with COPD, the 6-minute walk test (6MWT) and the 30-second sit-to-stand test (30sec-STS) are widely used as clinical outcome measures of walking capacity, lower limb muscle strength, and functional ability. Due to a documented learning effect, at least two trials are recommended for assessment. The aim of our study was to investigate the intra- and inter-rater reliability and agreement of the two tests in patients with severe and very severe COPD (FEV_1 <50%).

Patients and methods: Fifty patients (22 females; mean [SD]: age 67 [9] years, FEV_1 predicted 32 [9]%) were assessed with the 6MWT and the 30sec-STS twice by the same assessor on test-day 1 (T1) and by another assessor 7–10 days later on test-day 2 (T2).

Results: The 6MWT intra- and inter-rater reliability (intraclass correlation coefficient, $ICC_{1,1}$) was 0.98 (lower limit 95% CI: 0.94) and 0.96 (lower limit 95% CI: 0.94), respectively, and agreement (standard error of the measurement, SEM) was 14.8 and 20.5 m, respectively. The 30sec-STS intra- and inter-rater reliability and agreement results were, respectively, $ICC_{1,1}$ 0.94 (lower limit 95% CI: 0.90) and 0.92 (lower limit 95% CI: 0.86), with SEM of 0.97 and 1.14 repetitions. There was no difference (95% CI: −5.3; 8.1) between the 6MWT distances on T1, while the mean walking distance improved 7.9 m (0.0 m; 15.8 m) from T1 to T2. Improvement on the same test date was less likely (OR: 3.6 [95% CI: 1.1; 11.8], Fisher's exact test, P=0.047) in patients who walked less than 350 m in the 6MWT. We found no clinically relevant learning effect in the 30sec-STS.

Conclusion: In patients with severe and very severe COPD the 6MWT and the 30sec-STS showed excellent intra- and inter-rater reliability and acceptable agreement. No learning effect was documented for the tests when performed on the same day. Our data suggest that in clinical practice using different assessors is acceptable, and that a single test trial may be sufficient to assess patients with severe and very severe COPD.

Keywords: COPD, exercise test, outcome assessment, lower extremity, reproducibility of results

Background

Pulmonary rehabilitation is a key standard treatment of COPD in order to reduce COPD symptoms and improve functional ability and quality of life.[1–3] Functional performance tests designed to assess changes in the ability to perform specific tasks of daily activities should in particular be considered when evaluating pulmonary rehabilitation.[4,5] Feasible and reproducible tests for assessing functional ability are needed. Reproducibility parameters are population-specific and can be distinguished in reliability parameters that assess whether patients can be discriminated from each

Correspondence: Henrik Hansen
Research Unit of Chronic Diseases and Telemedicine – Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg University Hospital, Bispebjerg Bakke 23, 2400 Copenhagen NV, Denmark
Tel +45 2 894 6780
Email henrik.hansen.09@regionh.dk

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/COPD.S174248

other despite measurement errors, and agreement parameters that assess how close the results of repeated measures are by estimation of the error in repeated measurements.[6,7] Agreement parameters are preferable when the test is used for evaluating changes over time.[6–8]

The 6-minute walk test (6MWT) is recommended as one of the two gold-standard field tests to assess functional walking capacity and to evaluate treatment response from COPD rehabilitation programs.[9,10] The ability to get up from a seated position is crucial to a number of everyday activities necessary for autonomy; and in older persons the sit-to-stand (STS) tests are commonly used to measure lower extremity function.[4,5,11–16] Different STS tests are used in the evaluation of pulmonary rehabilitation including the time to perform five repetitions of STS (5STS), and the number of chair rises performed in 30 seconds, 60 seconds, or 3 minutes.[11,12] Most studies on patients with COPD have used the 60-second sit-to-stand test as a proxy for the 6MWT and as an indicator of functional capacity.[12,14] Fewer studies have used the shorter tests, for example, the 30sec-STS and the 5STS.[12] 5STS has a documented floor effect of 21% in elderly people from the general population and 15% in a cohort of patients with COPD.[13,16] While the 5STS has demonstrated high reliability, no reproducibility studies have been performed for the 30sec-STS.[11] The 30sec-STS has been shown to be a valid proxy measure for assessing lower limb muscle strength and function in patients with COPD.[15,17] Furthermore, several national health authorities recommend the 30sec-STS as a part of a generic screening tool for functional limitation, decline, and frailty in the elderly and persons with medical conditions.[18–20]

Numerous studies have reported the intra-rater reproducibility for 6MWT in groups with varying severity of COPD.[9,21–28] The Guidelines for Reporting Reliability and Agreement Studies (GRRAS) emphasize that inter-rater reproducibility is of high clinical relevance for multicenter studies and daily clinical practice because patients are often assessed by different assessors.[6,7] Yet, studies reporting the inter-rater reproducibility for 6MWT in patients with COPD are limited to one study[28] while no study has investigated the intra- and inter-rater reproducibility for 30sec-STS in patients with COPD.[11,12]

Results from studies on various patient groups show that patients typically perform better if the 6MWT and 30sec-STS are repeated, indicating that there is a systematic bias, commonly due to a learning effect.[29–34] Because learning effect has been an issue of concern in patients with COPD, the recommendation is to use a standardized instruction and to use the highest recorded value from at least two test trials for the 6MWT with a minimum of 30 minutes rest in between test trials.[9,10] This recommendation is primarily based on studies on heterogeneous groups of patients with varying severity of COPD (stage II–IV/moderate to very severe). Since physical capacity, activity of daily living, disease symptoms, and comorbidities differ substantially from mild to very severe COPD, this could influence test variability. Whether a learning effect is present in a homogenous group of patients with severe COPD is unknown. It could be hypothesized that if patients with very little or no reserve capacity perform two strenuous tests with a 30-minute pause between the trials a potential learning effect may be counteracted by fatigue or exhaustion.

The primary aim of this study was to investigate the intra- and inter-rater reliability and agreement of the 6MWT and 30sec-STS in patients with severe and very severe COPD (stage III–IV). Secondly, we intended to explore possible learning effects and other factors associated with repeated assessments.

Materials and methods

Study design

This intra- and inter-rater reproducibility study was part of a randomized controlled multicenter trial (RCT) (ClinicalTrials.gov identifier: NCT02667171) investigating the effect of online COPD rehabilitation in patients with severe and very severe (stage III–IV) COPD.[35] We followed the GRRAS.[7]

Participants

Eligible patients for the RCT were identified and recruited by respiratory nurses during outpatient COPD control visits from University Hospitals Amager, Hvidovre, Bispebjerg, Frederiksberg, Herlev, Gentofte, Frederikssund, and Hillerød. All patients provided written informed consent. The study was approved by the Ethics Committee of the Capital Region of Denmark (H-15019380) and the Danish Data Protection Agency approved the research database (j.nr.:2012-58-0004). All patients who agreed to participate in the RCT were consecutively asked to participate in the intra- and inter-rater reproducibility study that required an extra assessment visit prior to randomization and intervention. Recruitment for the reproducibility study commenced on March 18, 2016 and continued until 50 patients were recruited in March 20, 2017 (Figure S1 – flowchart). The sample size of 50 patients

was chosen based on the recommendation from COnsensus-based Standards for the selection of health Measurement INstruments.[8,36]

Inclusion criteria were[35] adults with a clinical diagnosis of COPD defined as the ratio of FEV_1 to FVC <0.70 and no history of asthma, and a FEV_1 <50% corresponding to severe or very severe airflow limitation; symptoms equivalent to the Medical Research Council dyspnea scale from 2 to 5; no participation in pulmonary rehabilitation within the last 6 months before start of the intervention; and medically cleared to participate in a COPD rehabilitation program. Inclusion and exclusion criteria corresponded with the criteria in routine COPD rehabilitation in the Capital Region of Copenhagen, Denmark.[35]

Study setting and raters

The assessments were conducted at the Respiratory and Physical Therapy Departments of five different University Hospitals (Hvidovre, Bispebjerg, Herlev, Gentofte, and Frederikssund) in Greater Copenhagen. All 10 raters underwent a 4-hour assessor course to ensure that they followed the same assessment protocol, and that testing procedures and the recording of results were standardized. In addition, the raters were required to pass four approved tests to obtain accreditation as blinded rater. All raters were familiar with the 6MWT and 30sec-STS from clinical practice. The median years of experience after graduation as therapist was 11.5 years (<10 years [n=3]; 10–20 years [n=4]; and >20 years [n=3]). The therapists worked in disease areas of geriatrics, cancer, intensive care unit, heart and lung, neurology, and orthopedics. The assessments on test-day 1 (T1) were conducted by one rater. Another rater carried out the assessments on the second test-day (T2) and this rater was blinded from the results from T1. Finally, the patients were instructed not to talk about testing on T1.

Test procedures

Patients were instructed not to do any vigorous activities 3 hours prior to testing and to take their prescribed medication as usual. All assessments on T1 and T2 followed the same procedures (Figure 1) and were conducted under the same conditions including the same location and time frame from 10 am to 2 pm, Monday to Friday, with a 7–10 day interval from T1 to T2. The test protocol was chosen to simulate conditions in everyday clinical practice, where several performance tests and questionnaires are conducted within a narrow time frame. If a patient used a walking aid in daily life (eg, a rollator or a crutch), he/she was allowed to use that aid during the 6MWT at T1 and T2. Patients who used portable oxygen carried it themselves in a shoulder bag or on a rollator, and oxygen consumption (L/min) was noted in the case report form.

6MWT

On each test-day (T1 and T2), two 6MWTs were performed (Figure 1) in accordance with standardized guidelines:[10] the

Assessment and progression:

1. Subject history/introduction, while seated: resting blood pressure, resting heart rate, resting SpO_2, resting dyspnea. Standing: anthropometric measures (weight and height) (30 minutes).
2. Instruction and performing the 6MWT, end-heart rate, end-SpO_2, end-dyspnea (10 minutes).
3. Seated rest and measurements: heart rate, SpO_2, and dyspnea (5 minutes).
4. Instruction and performing the 30sec-STS (5 minutes).
5. Seated rest for 30 minutes, completion of four questionnaires.
6. Seated: resting blood pressure, resting heart rate, resting SpO_2, resting dyspnea (5 minutes).
7. Instruction and performing the 6MWT, end-heart rate, end-SpO_2, end-dyspnea (10 minutes).
8. Seated rest and measurements: heart rate, SpO_2, and dyspnea (5 minutes).
9. Instruction and performing the 30sec-STS (5 minutes).
10. Assessment session completed. Total time 145 minutes.

Figure 1 Assessment procedures at day 1 (T1) and day 2 (T2 reassessment).
Abbreviations: Dyspnea, perceived dyspnea; end-, measure taken immediately after test completion; 6MWT, six-minute walk test; 30sec-STS, 30-second sit-to-stand test; SpO_2, arterial oxygen saturation as measured by pulse oximetry.

walking course was 20 m due to walking space shortage at some hospitals and to ensure the same standardized walking course at all five locations.[10,22] Patients were instructed to walk as far as possible in 6 minutes; they received the recommended standardized encouragement; and a 30-minute seated rest was mandatory between the first and second 6MWT on both test-days (Figure 1). Heart rate (HR), arterial oxygen saturation measured by pulse oximetry (SpO_2), and perceived dyspnea (Borg cr-10) were assessed before and after each 6MWT trial. Oxygen supplementation was used if required and prescribed by a chest physician.

30sec-STS

On each test-day (T1 and T2), two 30sec-STS were performed in accordance with existing protocols (Figure 1).[15,37] The same chair with a seat height of 45–47 cm was used throughout all tests, and the patients were asked to stand up fully and sit down as many times as possible in 30 seconds with arms across the chest starting from the seated position. The number of full stands was recorded. Score zero was recorded if the patient was unable to rise from the chair without using the arms. A 30-minute pause between the first and second trial was mandatory in this study.

Other variables

Demographic and descriptive variables such as age, gender, height, weight, body mass index, marital status, education, smoking status, years with COPD, GOLD, A/B/C/D stratification,[38] Charlson morbidity index, body mass index, airflow obstruction, dyspnea, and exercise capacity (BODE index), and lung medications were registered on T1. A chest physician or a respiratory nurse performed the lung spirometry test at the Respiratory Department at the referral hospital prior to study referral. All hospitals used clinically approved spirometry equipment, but manufacture trademark varied between hospitals in the Capital Region. The spirometry procedure followed the guideline from the Danish Society of Respiratory Medicine.[39]

Statistical analysis

Descriptive data are presented as means with SD for continuous data and as medians with range for ordinal data and data not normally distributed. Data distribution was inspected by histogram and Q–Q plots and verified by Shapiro–Wilk test to determine an approximately normal distribution. Paired *t*-test and Wilcoxon signed rank test were used to compare systematic bias between two assessments conducted on the same day and unpaired *t*-test and Mann–Whitney *U*-test were used to compare differences between the two raters. Fisher's exact test was used for analyses of categorical data. Intraclass correlation coefficient (ICC) was calculated to describe the reliability. The $ICC_{1.1}$ model was used because the assessments were conducted at five centers, and all raters did not assess each patient.[6,40] The ICC for inter-rater reliability was calculated from the highest value/best performance registered on test-day 1 (T1) and test-day 2 (T2) as recommended in the standard recommendation procedure.[9,10] The $ICC_{1.1}$ is a fixed model addressing both systematic and random error. ICCs values between 0 and 0.5 were considered weak, ≥0.5–0.75 moderate, ≥0.75–0.9 good, and ≥0.9 to have excellent reliability. Agreement was calculated as standard error of measurement (SEM) using the equation SD*√1–ICC to establish the typical error in a single measurement of repeated measurements.[6,40] The corresponding smallest real difference (SRD95%) was calculated by the equation 1.96×√2× SEM to express the variation with 95% certainty for individual subjects which represents the smallest change to be detected beyond the measurement error.[40–42] SEM and SRD95% are presented in actual units, and expressed as a percentage of the mean of the two test sessions (grand mean), making comparisons between tests and other studies easier. Bland–Altman plots were used to visualize potential systematic bias around the zero line as well as heteroscedasticity. Identification of the mean difference with 95% CI and limits of agreement (95% LOA) were included in the plots.[6,43] The significance level was set as $P<0.05$ for all analyses. Data were analyzed using SPSS version 20.0 (IBM Corporation, Armonk, NY, USA).

Results

Participants vs non participants

Fifty of the 108 eligible patients agreed to participate in the intra- and inter-rater reproducibility study. Twenty-three declined participation due to the extra testing date while 35 patients could not be included because they performed the baseline assessment in an RCT <1 week before the scheduled randomization and intervention. The 58 patients who did not participate in the reproducibility study did not differ significantly from the included patients on baseline characteristics (Table 1).

Intra-rater reproducibility

The reliability and agreement of the tests are presented in Table 2. Intra-rater reliability for the 6-minute walk distance (6MWD) and 30sec-STS were 0.98 ($ICC_{1.1}$) and 0.94 ($ICC_{1.1}$), respectively. Agreement was 14.8 m (SEM) for the 6MWD

Table 1 Characteristics

Variables	Patients with COPD	Not included
Sex, men/women (n)	28/22	21/37
Age, years (SD)	66.6±9.0	69.4±9.1
Body mass index, kg/m^2, mean (SD)	25.4±5.6	25.8±5.6
FEV_1 % predicted, mean (SD)	32.3±9.0	35.1±9.4
FEV_1/FVC, mean (SD)	41.4±10.6	45.1±11.8
GOLD I/II/III/IV, %	0/0/54/46	0/0/67/33
A/B/C/D,[36] %	0/36/0/64	3/33/7/57
MRC dyspnea scale, median (range)	3.5 (3–5)	3.0 (2–5)
CAT symptoms, mean (SD)	20.8±6.1	18.7±7.6
BODE index points, median (range)	5 (3–9)	5 (3–8)
Charlson index 1/2/≥3, %	52/30/18	28/47/26
LTOT, n (%)	4 (8)	9 (16)
Walking aid, stick/walker, n (%)	1/8 (18)	4/17 (36)
Highest 6MWD (SD)	347 (102)	330 (103)
Highest 30sec-STS (SD)	10.8 (4.1)	9.8 (4.8)

Notes: Data are presented as mean ± SD, median (range), or percentage in non-normally distributed variables.
Abbreviations: A/B/C/D, risk stratification; BODE index, body mass index, airflow obstruction, dyspnea, and exercise capacity; CAT, COPD assessment test; LTOT, long-term oxygen therapy; MRC, Medical Research Council; 6MWD, six-minute walk distance; 30sec-STS, 30 second sit-to-stand test.

and 0.97 repetitions (SEM) for the 30sec-STS. There was no systematic bias between the first and second test trial in the 6MWD and 30sec-STS, while a significant increase ($P<0.05$) was found in end HR and self-perceived dyspnea from the first to second 6MWT trial. Oxygen saturation was similar in both 6MWT trials (Table 2). Three patients were unable to rise from the chair at both trials and got score zero. Bland–Altman plots with 95% LOA for both tests are shown in Figure 2A and B.

Inter-rater reproducibility

The inter-rater reliability and agreement for the 6MWD and 30sec-STS are presented in Table 3. Inter-rater reliability $ICC_{1.1}$ for the 6MWD and 30sec-STS were 0.96 and 0.92, respectively. Agreement was 20.5 m (SEM) for the 6MWD and 1.14 repetitions (SEM) for the 30sec-STS. There were significant improvements in the best test results on T2 compared with T1. The mean differences were 7.9 m (95% CI: 0.03; 15.8, $P=0.049$) for the 6MWD and 0.6 repetitions (95% CI: 0.2; 1.1, $P<0.01$) for the 30sec-STS. End HR increased by 5.3 beats per minute (95% CI: 2.2; 8.4, $P=0.001$), while saturation and perceived dyspnea remained unchanged (Table 3). Three patients were unable to rise from the chair on both T1 and T2. Bland–Altman plots with 95% LOA for the 6MWD and 30sec-STS are shown in Figure 2C and D.

Explorative results from the 6MWT

Results from T1 showed that 56% of the patients walked <350 m (the predefined threshold cut off from the BODE index)[44] in the first test trial (Table 4). Improvements from the first to the second trial occurred in 73% of those who walked ≥350 m in the first trial and in 43% of those who walked <350 m (Table 4). The proportion difference was statistically significant (OR: 3.6 [95% CI: 1.1; 11.8], Fisher's exact test, $P=0.047$). However, when comparing the best results on T1 and T2, the proportion difference was no longer present (OR: 1.4 [95% CI: 0.4; 4.2], Fisher's exact test, $P=0.772$).

Explorative results from the 30sec-STS

Results from T1 showed that 30% of the patients demonstrated a poor 30sec-STS performance based on a criterion referenced score, that is, ≤8 chair rises, which is associated with a risk for loss of functional mobility in all age groups[18,19,45] (Table 5). Improvements from the first to the second trial occurred in 27% of those who did ≤8 chair rises and in 23% who did >8 chair rises (Table 5). The proportion difference was not statistically significant (OR: 1.2 [95% CI: 0.3; 4.9], Fisher's exact test, $P=1.0$), and there was no difference when comparing the best 30sec-STS results on T1 and T2 (OR: 1.4 [95% CI: 0.4; 5.2], Fisher's exact test, $P=0.752$).

Table 2 Intra-rater reproducibility test-day 1 (T1)

Variables	Test one	Test two	Difference	$ICC_{1.1}$ (LL_{95})	SEM (SEM%)	SRD95% (SRD%)
6MWD	329.5±100.9	330.9±109.8	1.4 [–5.3; 8.1]	0.98 (0.96)	14.84 (5)	41.13 (13)
End SpO_2	90.6±7.9	90.6±7.9	0.0 [–1.1; 1.1]	0.87 (0.78)	2.82 (3)	7.82 (7)
End HR	104.0±19.6	106.5±18.9	2.5 [0.4; 4.7]*	0.91 (0.85)	5.75 (5)	15.94 (15)
End dyspnea	5.7±2.7	6.5±2.6	0.8 [0.5; 1.1]*	0.88 (0.81)	0.93 (15)	2.58 (42)
30sec-STS	9.9±3.9	9.7±4.1	–0.2 [–0.5; 0.3]	0.94 (0.90)	0.97 (10)	2.68 (27)

Notes: Test one and two and difference are presented as mean ± SD or delta difference [SE 95% CI]. Significant difference between tests is denoted as *$P<0.05$.
Abbreviations: Dyspnea, perceived dyspnea (Borg cr-10); HR, heart rate (beats per minute); $ICC_{1.1}$, intraclass correlation coefficient model 1.1; LL_{95}, lower 95% confidence limit; 6MWD, six-minute walk distance (meters); SEM, standard error of measurement; SEM%, standard error of measurement expressed as a percentage of the mean; 30sec-STS, 30 second sit-to-stand test (repetitions); SpO_2, arterial oxygen saturation as measured by pulse oximetry (%); SRD95%, smallest real difference at the 95% confidence level; SRD%, smallest real difference as a percentage of the mean.

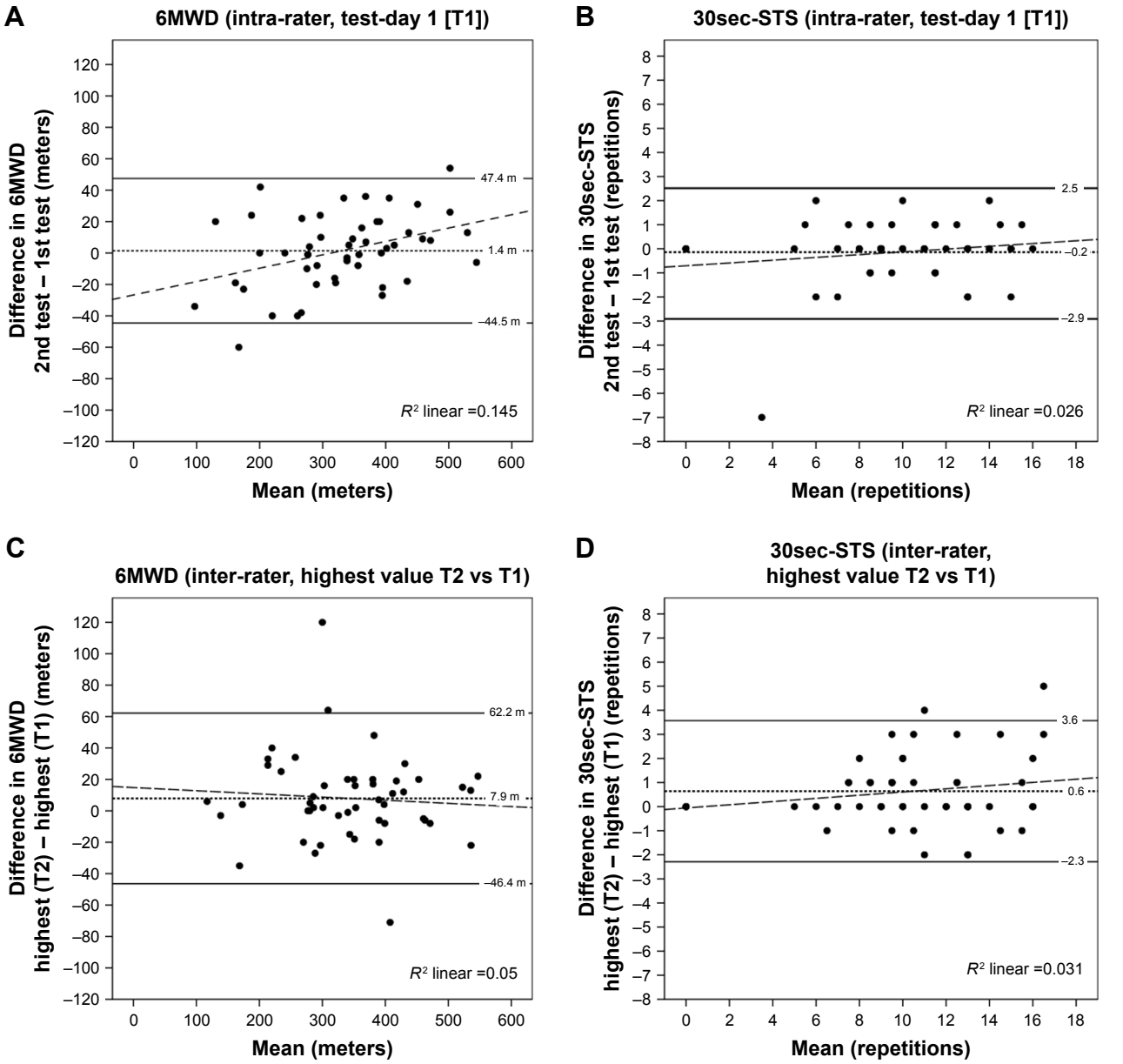

Figure 2 Bland–Altman plots of the 6MWD and 30sec-STS.

Notes: Mean difference between tests/or raters (dotted line) with limits of agreement 95% CI (black lines). The dashed line is the regression of change against the mean value. (**A**) 6MWD scores against 6MWD difference from a single rater at test-day 1 (T1). (**B**) 30sec-STS scores against 30sec-STS difference from a single rater at test-day 1 (T1). (**C**) 6MWD score difference obtained by two different raters on two separate test-days (T2 vs T1). (**D**) 30sec-STS score difference obtained by two different raters on two separate test-days (T2 vs T1).

Abbreviations: 30sec-STS, 30-second sit-to-stand test; 6MWD, 6-minute walk distance.

Table 3 Inter-rater reproducibility (highest value T1 vs highest value T2)

Variables	Rater T1	Rater T2	Difference	$ICC_{1.1}$ (LL_{95})	SEM (SEM%)	SRD95% (SRD%)
6MWD	339.3±103.7	347.2±101.7	7.9 [0.02; 15.8]*	0.96 (0.94)	20.46 (6)	56.70 (16)
End SpO_2	90.6±8.1	89.8±7.6	–0.8 [–0.3; 2.0]	0.86 (0.77)	2.94 (3)	8.15 (9)
End HR	105.4±19.1	110.7±17.7	5.3 [2.2; 8.4]*	0.79 (0.66)	8.48 (8)	23.50 (22)
End dyspnea	6.2±2.6	6.4±2.7	0.2 [–0.1; 0.6]	0.85 (0.75)	1.01 (16)	2.80 (45)
30sec-STS	10.2±3.9	10.8±4.1	0.6 [0.2; 1.1]*	0.92 (0.86)	1.14 (11)	3.15 (30)

Notes: Rater T1, T2, and difference are presented as mean ± SD or delta difference [SE 95% CI]. Significant difference between raters is denoted *$P<0.05$.

Abbreviations: Dyspnea, perceived dyspnea (Borg cr-10); HR, heart rate (beats per minute); $ICC_{1.1}$, intraclass correlation coefficient model 1.1; LL_{95}, lower 95% confidence limit; 6MWD, six-minute walk distance (meters); 30sec-STS, 30 second sit-to-stand test (repetitions); SEM, standard error of measurement; SEM%, standard error of measurement expressed as a percentage of the mean; SpO_2, arterial oxygen saturation as measured by pulse oximetry (%); SRD95%, smallest real difference at the 95% confidence level; SRD%, smallest real difference as a percentage of the mean.

Table 4 Explorative results for 6MWD

	Intra-rater test-day 1 (T1)			Inter-rater (highest value T1 vs highest value T2)		
	Improvers (n)	Non-improvers (n)	Total count (n)	Improvers (n)	Non-improvers (n)	Total count (n)
≥350 m	16	6	22[a]	14	10	24[c]
<350 m	12	16	28[b]	17	9	26[d]
Total count (n)	28	22	50	31	19	50

Notes: Number of improvers and non-improvers with a BODE index[38] cut off for the 6MWD. [a]18% (4/22) who walked ≥350 m improved their second 6MWD by ≥30 m and none showed a decrease by ≥30 m. [b]17% (2/28) who walked <350 m improved their second 6MWD by ≥30 m while 18% (5/28) decreased their walking distance by ≥30 m. [c]8% (2/24) who walked ≥350 m improved their 6MWD by ≥30 m from T1 to T2 while 4% (1/24) decreased their walking distance by ≥30 m. [d]19% (5/26) who walked <350 m improved their 6MWD by ≥30 m from T1 to T2 while 8% (2/26) decreased their walking distance with ≥30 m.
Abbreviations: 6MWD, six-minute walk distance; intra-rater, values obtained by the same rater on the same day; inter-rater, highest values obtained by two different raters on two separate test-days (T1 vs T2); BODE index, body mass index, airflow obstruction, dyspnea, and exercise capacity.

Discussion

To the best of our knowledge, this is the first study to investigate the intra- and inter-rater reproducibility of the 30sec-STS in patients with COPD. Also, this is only the second study to present inter-rater reproducibility of the 6MWT in patients with COPD. We found excellent intra- and inter-rater reliability for both tests and acceptable agreement indicating that both tests can be used for evaluative purpose even if measured by different raters.

6MWT

Based on our findings, the 6MWT appears to be highly reliable in patients with severe and very severe COPD. ICC for intra-rater reliability was excellent and consistent with previous findings from studies of patients with varying severity of COPD (ICC ranging from 0.88 to 0.98),[22,25–28] while ICC for inter-rater reliability was superior (lower limit [LL] 95% ICC: 0.94) to that reported by Labadessa et al[28] (LL95% ICC: 0.74). In contrast to previous studies, we did not find any systematic bias or clinical relevant learning effect when the test was conducted on the same day with 30 minutes of rest as recommended by the American Thoracic Society (ATS)/European Respiratory Society (ERS).[23,24,28,46] The results improved on average 7.9 m from T1 to T2. We cannot quantify to which extent the systematic bias from T1 to T2 was related to a learning effect or the effect of different raters. However, we believe that an average improvement of that magnitude is of minor clinical significance if the purpose is to measure a treatment effect. The findings in our study group of patients with severe and very severe COPD contradict previous findings of regular systematic bias in studies on heterogeneous groups of patients with varying severity of COPD (stage II–IV/moderate to very severe).[21,22,24–26,28,46] Without making the reservation that four of the previous studies used 30 minutes of rest between the two test trials and three studies used 24 hours and up to 7 days, their average intra-rater differences ranged from 16 to 37 m.[21,22,25,26,28,34,46] Previous studies on intra-rater reliability have shown learning improvements from the first to the second 6MWT in 70%–82% of the patients while we found improvements in 56% of the patients on T1 and in 62% of the patients from T1 to T2.[22,25–27] Due to the concern that the learning effect exceeds the minimal important difference (MID) of 30 m documented in previous studies it is currently recommended by the ERS/ATS to complete at least two 6MWT when assessing patients with COPD.[9,10] In our study, we found not only a proportion of 12% exceeding the MID of 30 m but also a counterpart of 10% decreasing by 30 m within the recommended minimum break of 30 minutes.

Although agreement parameters are preferable when measurement instruments are used for evaluation of changes over time,[6–8] only one reliability study has reported the SEM

Table 5 Explorative results for 30sec-STS

	Intra-rater test-day 1 (T1)			Inter-rater (highest value T1 vs highest value T2)		
	Improver (n)	Non-improver (n)	Total count (n)	Improver (n)	Non-improver (n)	Total count (n)
>8 repetitions	8	27	35	16	20	36
≤8 repetitions	4	11	15	5	9	14
Total count (n)	12	38	50	21	29	50

Note: Number of improvers and non-improvers with a criterion-based cut off[16,17,43] for the 30sec-STS.
Abbreviations: 30sec-STS, 30 second sit-to-stand test (repetitions); intra-rater, values obtained by the same rater on the same day; inter-rater, highest values obtained by two different raters on two separate test-days (T1 vs T2).

and the smallest real difference (SRD) of 6MWD in patients with COPD.[28] The time points for the measurements were similar to ours; however, our agreement parameters showed substantially less error variance (SEM-intra-rater: 15 m; SRD95%-intra-rater: 41 m; SEM-inter-rater: 21 m; SRD95%-inter-rater: 57 m) compared to the findings in the study by Labadessa et al (SEM-intra-rater: 31 m; SRD95%-intra-rater: 72 m; SEM-inter-rater: 37 m; SRD95%-inter-rater: 86 m).[28] Our intra-rater LOA from −45 to 47 m were approximately half the variance that has been stated in published studies with the lower limit ranging from −92 to −67 m and upper limits ranging from 103 to 120 m.[25,27,28]

Our findings regarding 6MWD do not confirm typical findings from previous studies. We can only speculate on the reasons for these findings. Our participants were patients with severe and very severe COPD (stage III–IV) who had limited reserve capacity. Thus, a potential learning effect may have been counteracted by fatigue or exhaustion when they performed two strenuous tests with 30 minutes pause between trials. In addition, the patients were not used to carrying out strenuous tests, and this may have influenced their motivation and performance. Ideally, the test and retest could have been performed on two separate days to allow for better restitution, but this would require two extra assessment visits. Since transportation and lack of energy were well-recognized barriers for participation, we anticipated that it would be difficult to persuade the patients to meet on two additional test-days. In support for this concern, 23 patients refused to participate in the study due to one extra assessment visit. Our results demonstrate that the magnitude of a potential learning effect is less dominant and prevalent in patients with severe and very severe COPD, especially in those with low initial walking capacity. The explorative analyses and Bland–Altman plot (Figure 2A) point toward a proportional difference indicating that those with low initial walking capacity, that is, a walking distance <350 m, improved less in the second trial compared to those who walked longer distances. Thus, it could be hypothesized that a threshold cut off of <350 m from the BODE index could be useful to discriminate between these two groups. However, this hypothesis needs confirmation in future research. The recommendation of two test trials with a break of minimum 30 minutes is based on results from studies among patients with heterogeneous disease severity, and it may not be as appropriate in a group of patients with severe and very severe COPD.

In our study, we also showed good reliability and an acceptable agreement regarding changes in oxygen saturation (end SpO_2), HR (end HR), and perceived dyspnea (end dyspnea) during the 6MWT. Our findings are in line with previous findings in groups with varying severity of COPD,[25,27,28] and emphasize that these measurements can be used for evaluative purposes, for example, monitoring SpO_2 and HR during 6MWT.[9,10]

30sec-STS

The main findings from the 30sec-STS were excellent reliability and acceptable agreement. The average improvement of 0.6 repetitions from T1 to T2 was small and most likely of minor clinical significance if the purpose is to measure a treatment effect. Our findings on reliability and agreement in patients with severe and very severe COPD were identical to the results reported in other patient groups with renal disease (ICC 0.93; SEM: 0.9; SRD: 2.6),[47] type 2 diabetes (ICC 0.92; SEM: 1.2; SRD: 3.3),[33] hip osteoarthritis (ICC 0.88; SEM: 1.5; SRD: 3.5),[31] acute medical illness (ICC 0.82; SEM: 1.32; SRD: 3.7),[48] hip replacement (ICC 0.88; SEM: 1.0; SRD: 2.8),[30] cognitive impairment (ICC 0.94; SEM: 0.9; SRD: 2.4),[49] and hospitalized stroke (ICC 0.87; SEM: 1.0; SRD: 3.0).[50]

Our results showed that the measurement error in 30sec-STS was small, indicating that the 30sec-STS would be sensitive to measuring relative small changes over time in patients with severe and very severe COPD. All our patients could complete the test, and in contrast to the 5STS the 30sec-STS did not show a floor effect. Two studies have shown a significant association between results in the 30sec-STS and leg extension (r=0.48)[15] and isokinetic quadriceps strength (r=0.78–0.81),[17] which is crucial for the ability to perform chair rise. In addition, one study has shown that the test is valid for assessing muscle performance of the lower limbs in patients with COPD pre and post a COPD rehabilitation program.[15] Thus, the 30sec-STS seems feasible for use in research and daily clinical practice. However, the MID and responsiveness of the test have yet to be investigated and established in patients with COPD.

Strength and limitations and future perspectives

Our study has meticulously followed the GRRAS, including reports on all relevant reproducibility domains for both intra- and inter-rater reproducibility, and a recommended sufficient sample of 50 patients. We used a rigorous standardized methodological assessment approach, that is, using the same walking course and avoiding influence of length and track layout on the walked distance, same chair height, tests were performed on the same time of the day, same rest

intervals, standardized instruction, and calibrated raters. However, there are some limitations. We cannot rule out that leg fatigue due to the many tests may have obliterated a possible learning effect. However, to limit the influence of fatigue we ensured that every patient felt rested and that oxygen saturation, HR, and perceived dyspnea were fully normalized to the resting level values before the retests were performed. Secondly, it was impossible to blind the raters to the results from the first test trial because they performed the second trial on the same test-day and this increases the risk of recall bias. The disclosed limitations to restrict the learning effect[21,27,30,33,47] and a possible recall bias are similar to those known from existing publications.[21–28] Regarding the inter-rater reproducibility, we were unable to estimate how large a proportion in the measurement difference from T1 and T2 relates to an actual learning effect, to the effect of different raters, or biological variation. Finally, variables such as anxiety, depression, and lung symptoms were not adjusted for.

Conclusion

In summary, the reproducibility of the 6MWT and 30sec-STS was excellent in patients with severe and very severe COPD. In contrast to previous studies, we found no learning effects when following the standardized guidelines for testing. Compared with the results on T1, there was a significant average improvement of 7.9 m in 6MWD on T2, which was performed 7–10 days apart and assessed by different raters. We consider this difference minor and without clinical relevance, while the SEM of 20.5 m was acceptable and below the established MID for the 6MWT. The reproducibility of cardiorespiratory variables (saturation, HR, and perceived dyspnea) was good and acceptable. Based on our findings, repeated 6MWT and 30sec-STS can be conducted by different raters in clinical practice, and one 6MWT and one 30sec-STS may be sufficient to assess patients with severe and very severe COPD. However, the responsiveness of 30sec-STS needs to be investigated in future studies.

Acknowledgments

The authors would like to thank the patients for taking part in this study and all the raters who assisted with the blinded data collection. The authors acknowledge the financial support from the Danish Lung Foundation (charitable funding), Telemedical Center Regional Capital Copenhagen (governmental funding), and TrygFonden foundation (charitable funding).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Mccarthy B, Casey D, Devane D, Murphy K, Murphy E, Lacasse Y. Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database Syst Rev*. 2015;2015(2):1–209.
2. Paneroni M, Simonelli C, Vitacca M, Ambrosino N. Aerobic Exercise Training in Very Severe Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis. *Am J Phys Med Rehabil*. 2017;96(8):541–548.
3. Puhan MA, Gimeno-Santos E, Cates CJ, Troosters T, Scharplatz M, Walters EH, Steurer J. Pulmonary rehabilitation following exacerbations of chronic obstructive pulmonary disease. *Cochrane Database Syst Rev*. 2016;12(12):CD005305.
4. Bui KL, Nyberg A, Maltais F, Saey D. Functional Tests in Chronic Obstructive Pulmonary Disease, Part 1: Clinical Relevance and Links to the International Classification of Functioning, Disability, and Health. *Ann Am Thorac Soc*. 2017;14(5):778–784.
5. Bui KL, Nyberg A, Maltais F, Saey D. Functional Tests in Chronic Obstructive Pulmonary Disease, Part 2: Measurement Properties. *Ann Am Thorac Soc*. 2017;14(5):785–794.
6. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59(10):1033–1039.
7. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48(6):661–671.
8. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010;10(1):22.
9. Singh SJ, Puhan MA, Andrianopoulos V, et al. An official systematic review of the European Respiratory Society/American Thoracic Society: measurement properties of field walking tests in chronic respiratory disease. *Eur Respir J*. 2014;44(6):1447–1478.
10. Holland AE, Spruit MA, Troosters T, et al. An official European Respiratory Society/American Thoracic Society technical standard: field walking tests in chronic respiratory disease. *Eur Respir J*. 2014;44(6):1428–1446.
11. Johnston KN, Potter AJ, Phillips A. Measurement Properties of Short Lower Extremity Functional Exercise Tests in People With Chronic Obstructive Pulmonary Disease: Systematic Review Background. An increasing variety of short functional exercise tests are reported in. *Phys Ther*. 2017;97(9):926–943.
12. Vaidya T, Chambellan A, de Bisschop C. Sit-to-stand tests for COPD: A literature review. *Respir Med*. 2017;128:70–77.
13. Jones SE, Kon SS, Canavan JL, et al. The five-repetition sit-to-stand test as a functional outcome measure in COPD. *Thorax*. 2013;68(11):1015–1020.
14. Crook S, Büsching G, Schultz K, et al. A multicentre validation of the 1-min sit-to-stand test in patients with COPD. *Eur Respir J*. 2017;49(3):1601871.
15. Zanini A, Aiello M, Cherubino F, et al. The one repetition maximum test and the sit-to-stand test in the assessment of a specific pulmonary rehabilitation program on peripheral muscle strength in COPD patients. *Int J Chron Obstruct Pulmon Dis*. 2015;10:2423–2430.
16. 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(2):M85–M94.
17. Butcher SJ, Pikaluk BJ, Chura RL, Walkner MJ, Farthing JP, Marciniuk DD. Associations between isokinetic muscle strength, high-level functional performance, and physiological parameters in patients with chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*. 2012;7:537–542.

18. National Board of Health. *Vaerktøjer Til Tidlig Opsporing 2/24. Copenhagen*. 2013. Available from: https://www.sst.dk/da/aeldre/aeldre-medicinske-patient/vaerktoejer-til-tidlig-opsporing. Accessed October 05, 2018.
19. Centers for Disease Control and Prevention. STEADI materials for healthcare providers. STEADI – older adult fall prevention. Available from: https://www.cdc.gov/steadi/materials.html. 2017. Accessed December 21, 2017.
20. Health Quality & Safety Commission New Zealand. Falls prevention toolkit for clinicians aids integrated approach. 2015. Available from: https://www.hqsc.govt.nz/our-programmes/reducing-harm-from-falls/publications-and-resources/publication/2232/. Accessed October 05, 2018.
21. Iriberri M, Gáldiz JB, Gorostiza A, Ansola P, Jaca C. Comparison of the distances covered during 3 and 6 min walking test. *Respir Med*. 2002;96(10):812–816.
22. Sciurba F, Criner GJ, Lee SM, et al. Six-Minute Walk Distance in Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med*. 2003;167(11):1522–1527.
23. Eiser N, Willsher D, Doré CJ, Reliability DCJ. Reliability, repeatability and sensitivity to change of externally and self-paced walking tests in COPD patients. *Respir Med*. 2003;97(4):407–414.
24. Spencer LM, Alison JA, McKeough ZJ. Six-Minute Walk Test as an Outcome Measure. *Am J Phys Med Rehabil*. 2008;87(3):224–228.
25. Hernandes NA, Wouters EF, Meijer K, Annegarn J, Pitta F, Spruit MA. Reproducibility of 6-minute walking test in patients with COPD. *Eur Respir J*. 2011;38(2):261–267.
26. Holland AE, Rasekaba T, Fiore JF, Burge AT, Lee AL. The 6-minute walk distance cannot be accurately assessed at home in people with COPD. *Disabil Rehabil*. 2015;37(12):1102–1106.
27. Osadnik CR, Borges RC, Mcdonald CF, Carvalho CR, Holland AE. Two 6-minute Walk Tests Are Required During Hospitalisation for Acute Exacerbation of COPD. *COPD*. 2016;13(3):288–292.
28. Labadessa IG, Arcuri JF, Sentanin AC, da Costa JN, Pessoa BV, Di Lorenzo VA. Should the 6-Minute Walk Test Be Compared When Conducted by 2 Different Assessors in Subjects With COPD? *Respir Care*. 2016;61(10):1323–1330.
29. Uszko-Lencer N, Mesquita R, Janssen E, et al. Reliability, construct validity and determinants of 6-minute walk test performance in patients with chronic heart failure. *Int J Cardiol*. 2017;240:285–290.
30. Mikkelsen LR, Mikkelsen S, Søballe K, Mechlenburg I, Petersen AK. A study of the inter-rater reliability of a test battery for use in patients after total hip replacement. *Clin Rehabil*. 2015;29(2):165–174.
31. Bieler T, Magnusson SP, Kjaer M, Beyer N. Intra-rater reliability and agreement of muscle strength, power and functional performance measures in patients with hip osteoarthritis. *J Rehabil Med*. 2014;46(10):997–1005.
32. Tveter AT, Dagfinrud H, Moseng T, Holm I. Measuring Health-Related Physical Fitness in Physiotherapy Practice: Reliability, Validity, and Feasibility of Clinical Field Tests and a Patient-Reported Measure. *J Orthop Sports Phys Ther*. 2014;44(3):206–216.
33. Alfonso-Rosa RM, del Pozo-Cruz B, del Pozo-Cruz J, Sañudo B, Rogers ME. Test-retest reliability and minimal detectable change scores for fitness assessment in older adults with type 2 diabetes. *Rehabil Nurs*. 2014;39(5):260–268.
34. Spencer L, Zafiropoulos B, Denniss W, Fowler D, Alison J, Celermajer D. Is there a learning effect when the 6-minute walk test is repeated in people with suspected pulmonary hypertension? *Chron Respir Dis*. Epub 2018 Jan 01.
35. Hansen H, Bieler T, Beyer N, Godtfredsen N, Kallemose T, Frølich A. COPD online-rehabilitation versus conventional COPD rehabilitation – rationale and design for a multicenter randomized controlled trial study protocol (CORe trial). *BMC Pulm Med*. 2017;17(1):140.
36. COSMIN. Cosmin checklist. Available from: https://www.cosmin.nl/tools/checklists-assessing-methodological-study-qualities/. Accessed October 05, 2018.
37. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport*. 1999;70(2):113–119.
38. Agusti A, Hurd S, Jones P, Fabbri LM, Martinez F, Vogelmeier C. Global Initiative for Chronic Obstructive Lung Disease. 2017. Available from: https://goldcopd.org/wp-content/uploads/2017/11/GOLD-2018-v6.0-FINAL-revised-20-Nov_WMS.pdf. Accessed October 05, 2018.
39. Danish Society of Respiratory Medicine. *Lungefunktionsstandard Spirometri Og Peakflow*. 2007. Available from: https://www.lungemedicin.dk/fagligt/klaringsrapporter/5-lfu-standard/file.html. Accessed December 22, 2017.
40. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231–240.
41. Hopkins WG. Measures of Reliability in Sports Medicine and Science. *Sports Med*. 2000;30(1):1–15.
42. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217–238.
43. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
44. Celli BR, Cote CG, Marin JM, et al. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med*. 2004;350(10):1005–1012.
45. Rikli RE, Jones CJ. Development and validation of criterion-referenced clinically relevant fitness standards for maintaining physical independence in later years. *Gerontologist*. 2013;53(2):255–267.
46. Jenkins S, Cecins NM. Six-minute walk test in pulmonary rehabilitation: do all patients need a practice test? *Respirology*. 2010;15(8):1192–1196.
47. Overend T, Anderson C, Sawant A, Perryman B, Locking-Cusolito H. Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiother Can*. 2010;62(2):122–128.
48. Bodilsen AC, Juul-Larsen HG, Petersen J, Beyer N, Andersen O, Bandholm T. Feasibility and inter-rater reliability of physical performance measures in acutely admitted older medical patients. *PLoS One*. 2015;10(2):e0118248.
49. Hesseberg K, Bentzen H, Bergland A. Reliability of the senior fitness test in community-dwelling older people with cognitive impairment. *Physiother Res Int*. 2015;20(1):37–44.
50. Johansen KL, Stistrup RD, Schjøtt CS, Madsen J, Vinther A. Absolute and relative reliability of the timed "Up & Go" test and "30second chair-stand" test in hospitalised patients with stroke. *PLoS One*. 2016;11(10):e0165663.

Supplementary material

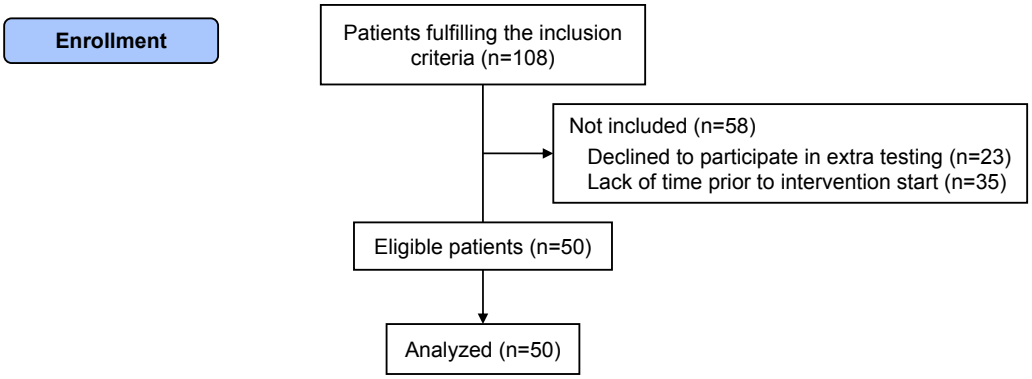

Figure S1 Recruitment flowchart.