Reliability and Validity of Patient-Reported, Rater-Based, and Hybrid Physical Activity Assessments in COPD: A Systematic Review

Shweta Gore , Tirupathi Chindam , Allon Goldberg , Min H. Huang , Michael Shoemaker & Jennifer Blackwood

View supplementary material

Published online: 15 Oct 2020.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

REVIEW

Check for updates

Reliability and Validity of Patient-Reported, Rater-Based, and Hybrid Physical Activity Assessments in COPD: A Systematic Review

Shweta Gore[a] (iD), Tirupathi Chindam[b], Allon Goldberg[c], Min H. Huang[c], Michael Shoemaker[d], and Jennifer Blackwood[c]

[a]School of Health and Rehabilitation Sciences, Department of Physical Therapy, MGH Institute of Health Professions, Boston, MA, USA; [b]Genesis Rehabilitation Services, Richmond, VA, USA; [c]Physical Therapy Department, University of Michigan-Flint, Flint, MI, USA; [d]Department of Physical Therapy, Grand Valley State University, Grand Rapids, MI, USA

ABSTRACT

Selecting valid and reliable PA assessments in chronic obstructive pulmonary disease (COPD) is crucial to ensure that the information obtained is accurate, valuable, and meaningful. The purpose of this systematic review was to compare the validity and reliability among PA assessments in COPD. An electronic database search of PubMed and CINAHL was completed in December 2019 using MeSH terms on physical activity, COPD, validation, and questionnaires. Transparency in reporting was assessed with the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) checklist while methodological quality was assessed with the modified Quality Appraisal tool for Reliability studies (QAREL) for reliability studies and the Quality Appraisal of Validity Studies (QAVALS) for validity studies. The search yielded fifteen different measures. The Stanford 7-day recall (PAR) demonstrated the strongest correlations with SenseWear Armband on energy expenditure ($r = 0.83$; $p < 0.001$) and moderate correlations for time spent in activity over 3 METs ($r = 0.54$, $p < 0.001$). The Multimedia Activity Recall (MARCA) also demonstrated moderate to good correlations with both SenseWear and Actigraph GT3X + accelerometers ($r = 0.66$–0.74). Assisted and computerized PRO measures (PAR and MARCA) and hybrid measures (C-PPAC and D-PPAC) demonstrate better psychometric properties as compared to other subjective measures and may be considered for quantification of PA in COPD. However, observations drawn from single validation studies limit strength of recommendations and further research is needed to replicate the findings.

Introduction

Patients with COPD often experience disabling symptoms such as dyspnea and fatigue, which result in decreased levels of physical activity (PA) as compared to individuals who do not have the disease. Despite recommendations to maintain regular PA by the Global Obstructive Lung Disease guidelines, PA levels in COPD continue to progressively decline starting right from the early stages of the disease [1, 2]. Decreased PA contributes to a downward spiral of immobility in COPD, which results in further deterioration of physical condition, decline in muscle strength, endurance, worsening of dyspnea and social isolation and depression [2, 3]. PA therefore is of prognostic significance in COPD with lower levels of activity related to a higher risk of hospital admissions and a shorter survival rate in COPD [1–3]. Considering the close relationship between PA and health in COPD, assessment and monitoring of PA is essential [2].

Several methods for assessment of PA are currently available that include objective, measures such as pedometers and accelerometers, subjective measures such as questionnaires, logs, or diaries, and a combination of objective and subjective (hybrid) measures. All these methods quantify activity duration or movements, from which estimates of energy expenditure can be made and distinctions between differing lifestyles can be inferred [4]. Subjective measures include patient reported outcome (PRO) measures and rater-based measures. PRO measures are defined as self-reported outcomes that come directly from patients without interpretation of patients' responses by clinician [5]. PRO measures include both questionnaires as well as diaries, logs and activity checklists and may be administered in different ways such as by the patient unassisted (self-administered), via an interviewer (assisted) or by entering responses via a computer-based program (computerized). Rater-based measures, on the other hand, include measures where patients' responses can be interpreted differently by the interviewer or rater in case of disparity between patients' responses and the PA classification used on the identified measure.

Subjective measures of PA often form an integral part of PA assessment in COPD since these are simple, cost-effective, and require minimal training. Subjective measures are frequently used to measure multiple dimensions of PA including not only the duration, frequency and intensity of activity but also reporting the type, location, domain, and context of the activity. Subjective PA measures provide estimates of time spent in activities of various levels of intensity and may be able to rank individuals according to intensity levels of reported activity [6]. Despite the simplicity and ease of use, the accuracy of these measures has been questioned due to the subjective nature of these tools which may lead to over- or under-estimation of PA [7].

While purely objective measures such as pedometers and accelerometers can quantify movement and provide objective estimates of energy expenditure, they lack the ability to quantify all the aspects of PA such as the type of activity and patients' experience of an activity [4]. To overcome some of these limitations, hybrid measures have been developed more recently that combine PRO measures with objective assessments such as pedometers and accelerometers to better reflect both the quantity and experience of PA [8].

Previous reviews on PA assessments have focused mainly on comparing objective and subjective measures in older adults without known comorbidities. However, in COPD, there is inadequate evidence regarding the reliability and validity of various PA measures including the PROs, rater based, and hybrid measures in COPD. The purpose of this review therefore, was to systematically compare the reliability and validity of the various PROs, rater based, and hybrid measures PA measures in COPD. This review will help clinicians in making well-informed decisions regarding the selection of appropriate subjective measures in patients with COPD.

Material and methods

The frame work of the review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement for reporting systematic reviews and meta-analyses. The study protocol was registered in the International prospective register of systematic reviews (PROSPERO) (registration no.CRD42016042588; available from http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016042588) which provided a guideline for development of this systematic review.

Data sources and searches

A systematic literature search was performed with in electronic databases, PubMed and CINAHL until December 8, 2019 using the following MeSH terms and keywords: ((((((((((((((((("physical activity") OR "activities of daily living") OR "motor activities") OR walking) OR exercise) AND "copd") OR "chronic obstructive pulmonary disease") OR "chronic obstructive lung disease") AND "surveys") OR "questionnaire") OR "interview") OR "log") OR "self-report")

OR "diary, health") AND "validation") AND "validation studies") AND validity)). To be more inclusive, no start date was used for the search. The search was conducted using filters for species (humans) and language (English). In addition to electronic database searches, searches were conducted from back references of articles and, also by contacting authors for access to English versions of articles if available.

Study selection

Validation studies published in English describing PRO (including self-administered, assisted, or computerized), rater-based, or hybrid PA measures in COPD were included in this review. Study designs included all observational study designs evaluating psychometric properties of PA measures in the COPD population. Studies were excluded from the review if the data of reliability or validity of the outcome measure was not evaluated. Studies were also excluded if these were systematic reviews, meta-analyses, narrative reviews, or letters to editors. Only measures directly related to PA such as those measuring the duration, intensity, frequency as well as the type and nature of PA were included. Measures that assessed related constructs of PA such as physical function, mobility, symptoms limiting activity and quality of life with no questions directed toward the amount, duration or type of PA were excluded from this review. Measures that assessed more than one construct of PA were included only if information on PA duration, intensity, frequency, or type was analyzed as a separate component and scoring of these components was separate from PA symptoms, difficulty, or experiences etc. No restrictions were applied on the publication date or type of publication (gray literature) to decrease chances of publication bias.

Titles and abstracts were independently reviewed by two reviewers who were board certified in geriatric physical therapy by the American Board of Physical Therapy Specialties. Full texts were gathered following a review of titles and abstracts. A rating form was developed and used by both reviewers to determine the eligibility of retrieved articles for inclusion in the review. The rating form included a list of inclusion and exclusion criteria. The reviewers independently completed the form and selected articles that met the criteria (Supplementary table 3). All disagreements were resolved by mutual consensus. In case of difficulty reaching a consensus, a third reviewer was contacted.

Data extraction

Following review of the full text articles, data related to sample demographics, details of outcome measures and their measurement properties were extracted. Specific information that was extracted included age, gender, BMI, lung function, type of outcome measure, method of administration, length of recall, scoring and interpretation, type of PA assessed, outcomes obtained, training, and cost. Data extraction forms created by the study team were used for extracting data. In case of missing information on pertinent items (e.g. cost, training requirements, and statistical analysis methods) in
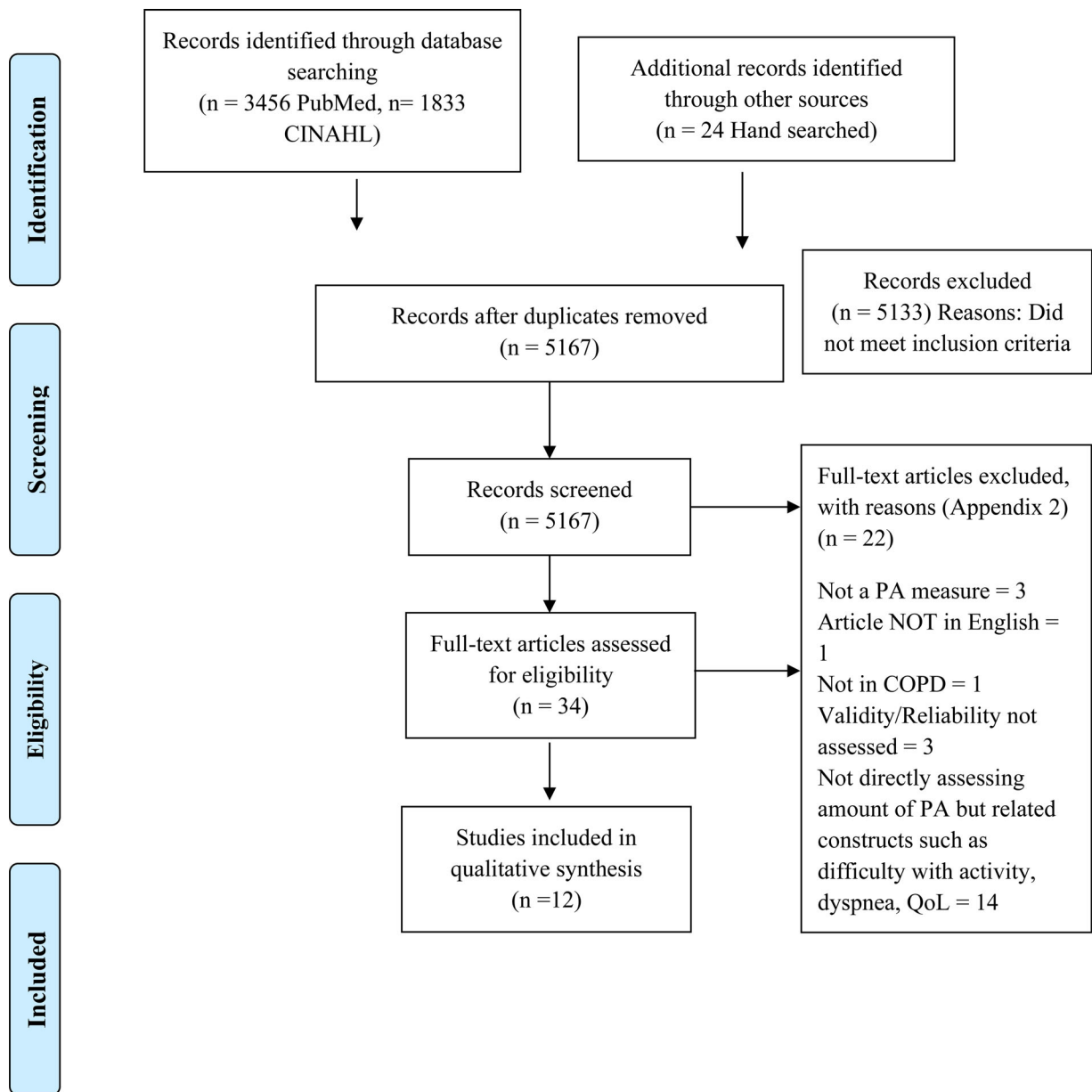
Figure 1. PRISMA Flow Diagram.

the studies selected, attempts were made to contact the authors *via* email. When no information was received from the individual authors or when the information was unavailable, these items were entered as 'not available' in the data extraction forms. Outcome measures were categorized as PRO measures if they were: 1) either self-administered, assisted (patient's responses were recorded by the interviewer), or computerized (responses were recorded by a computer) without any alteration or interpretation of the stated response, 2) rater-based measures when subject's responses were interpreted and scored differently by the interviewer, or 3) hybrid measures when a PRO was used in combination with an objective measure Measures were also included under PROs when prompts to aid recall (by the interviewer or computer) were provided but the patient responses were recorded as reported without alteration.

Different statistical measures used for testing validity, reliability and diagnostic accuracy were extracted. For

validity specifically, Pearson's and Spearman's correlation coefficients, variance, inter-quartile range, and observed differences in outcome variables assessed using t test and analysis of variance (ANOVA) were extracted. Correlation coefficients less than 0.25 were reported as having little or no relationship, 0.25 to 0.50 fair, correlation coefficients more than 0.50 and less than 0.75 as moderate-to-good, and those with coefficients above 0.75 were reported as having good-to-excellent relationship [9]. For diagnostic accuracy, area under the curve, sensitivity, specificity and cut off values were extracted and reported. Test-retest intervals and reliability coefficients including the intra-class correlation coefficients (ICC), coefficient of variation, limits of agreement and internal consistency were extracted for reliability. As a general guideline, ICC values below 0.50 were reported to have poor reliability, coefficients from 0.50 to 0.75 as moderate reliability and coefficients above 0.75 were considered to have good reliability [9].

The data from the identified articles was independently extracted and rated for methodological quality and transparency in reporting by the two reviewers.

Quality assessment

The methodological quality of reliability studies was assessed using the Quality Appraisal tool for Reliability studies (QAREL) [10]. QAREL is an 11-item checklist that assesses quality of reliability studies under seven domains including the subjects, assessors, blinding, and order effects of examination, time interval between repeated measures, statistical analyses, and appropriate test application [10, 11]. The items on the QAREL were rated as 0 or 1 with higher scores indicating better quality. The Quality Appraisal tool for Validity Studies (QAVALS) was used for methodological quality assessment of validity studies [12]. The QAVALS is a 24-item tool addressing different types of validity (content, concurrent, construct) and various aspects of methodological quality including design, participants, statistical analysis, and confounding [12]. Each item on the QAVALS is individually assessed on a categorical scale as a "yes", "no" or "other" response.

The transparency in reporting of studies was assessed using the Strengthening the Reporting of OBservational Studies in Epidemiology (STROBE) checklist. The STROBE addresses 22 fundamental aspects of reporting of observational cross-sectional studies [13]. Each aspect of individual studies was assigned a numerical value of 1 if explicitly described and present, and 0 if inadequately described or absent with a maximum possible score of 22. Higher scores on the STROBE reflected a better overall transparency in reporting of the study. The inter-rater reliability for quality appraisal scores among reviewers was tested using the weighted kappa coefficient and was found to be excellent (Kappa = 0.91, 95% CI = 0.79–1.03).

Results

Initial search yielded 5,289 studies including 3,456 studies on PubMed and 1,833 on CINAHL (Figure 1). Additionally, 24 records were identified *via* hand search. After removal of duplicates, a total of 5,167 titles and abstracts were retrieved. Following a review of the abstracts and removal of duplicates, 5133 studies were excluded as they did not meet the selection criteria. Thirty-four studies were obtained for full text review. Twenty-two studies were removed after review of full texts, leaving 12 studies that were ultimately included for this systematic review. The reasons for exclusion of articles at different stages of selection are described in Figure 1 and Supplementary table 4.

All the studies that were included in the final review were published between 2005 and 2015, with eight studies assessing the European population and two studies each assessing United States and Australian populations. The study participants were identified from community dwelling older adults *via* hospital or medical records (2 studies), outpatient clinics (8 studies) and previous pulmonary rehabilitation programs (2 studies). Participants included both males and females with a higher representation of males across the studies (605 males out of the total sample of 869 across studies reporting gender distribution). The mean age of the sample in the included studies ranged from 61 to 74. 4 years. The sample participants in the included studies demonstrated varying degrees of airway obstruction, with the mean percentage predicted forced expiratory volume in one second (FEV1) values ranging from 33 to 57%, and COPD severity ranging from mild to very severe. The sample size varied widely across the studies starting from 13 to as many as 236 (Table 1).

Of the 12 included studies, 7 studies assessed validity of PA measures [1, 14–19], 4 assessed both reliability and validity [4, 8, 20, 21], 1 study assessed reliability [22]. Four studies had additional information on diagnostic accuracy [1, 14, 19, 21]. Fifteen different PA measures were described in the studies of which only one was a rater-based measure (the Physical Activity Scale for the Elderly (PASE)). Among the PRO measures, 7 PA measures were self-administered, 2 were assisted (semi structured or structured interviews), 2 were computerized, and 1 was either self or assisted. Two types of hybrid measures (Daily and Clinical PRO Physical Activity in COPD measures: D-PPAC and C-PPAC) were described. Description on the type, scoring and other features of the measures can be found in Supplementary Table 1.

Differences in PA measures were observed in the type of administration, length of recall, number of items, outcomes assessed and scoring methods. Length of recall ranged from one day to a lifetime recall of activity. Outcomes assessed included activity duration, frequency, intensity, and PA type (both structured and unstructured). Scoring methods differed across the measures including continuous scores computed by summing item scores to obtain a total PA score, total energy expenditure using intensity codes, PA levels using item weights and categorical scores describing the level of PA (inactive to hard). Leisure-time activity was the most frequently included item on the PA measures (66.66% or 10 out of 15). Household/daily PA on the other hand, was the least frequently included item (33.33% or 5 out of 15 measures) (Supplementary Table 1).

The studies included in the review also differed in their transparency in reporting and methodological quality. The breakdown of scores on the QAREL, QAVALS and STROBE checklists is described in supplementary tables 5, 6, and 7. Very few studies included information on key elements of study design and methods. Only 3 studies included information on the representativeness of sample, study setting and period of recruitment, confounding variables, and confidence intervals. The selection criteria for participants was clearly identified in 6 studies, attrition of sample in 7, standardization of procedures in 8 and priori sample size calculations in 4 studies. Only one study described adjustments for multiple comparisons (Supplementary table 6). For studies assessing concurrent validity ($n = 11$), a sound rationale for the selection of reference standard and justification for the time interval between tests was provided only in 6 studies

Table 1. Characteristics of the study population.

Author, Year	Sample size	Gender (M/F)	BMI M (SD)	Age (years) M (SD)	FEV1 (%) M (SD)	FEV1/FVC M (SD)	COPD severity	Reliability and validity reported
Pitta [17], 2005	13	10/3	23.00 (5.00)	61.00 (8.00)	33.00 (10.00)	NA	Moderate - severe	Validity only
Moore [15], 2009	76	NS	28.20 (6.70)	71.10 (9.60)	46.30 (18.50)	41.70 (13.10)	Moderate - severe	Validity only
Moy [20], 2009	17	17/0	28.00 (4.00)	73.00 (8.00)	57.00 (22.00)	NA	Mild = 1; Moderate = 10; Severe = 2; Very severe = 4	Validity & reliability
Donaire-Gonzalez [14], 2011	172	161/11	47.00 (28.00)	70.00 (8.00)	52.00 (15.00)	54.00 (13.00)	Mild = 8; Moderate = 80; Severe = 57; Very severe = 14	Validity only
Gouzi [22], 2011	129	85/44	24.70 (4.40)	61.30 (9.60)	56.60 (22.80)	NA	Moderate - severe	Reliability only
Slinde [18], 2011	68	26/42	25.30 (5.50)/ 23.70 (4.40) ᵋ	66.00 (6.00)/ 64.00 (7.00) ᵋ	41.00 (16.00)/ 44.00 (15.00) ᵋ	NA	Severe	Validity only
DePew [21], 2012	67	38/29	28.90 (6.86)	71.40 (7.91)	46.70 (19.98)	NA	Moderate - severe	Validity only
Garfield [1], 2012	43	18/25	25.90 (5.60)	68.00 (9.00)	46.00 (22.00)	NA	Moderate – very severe	Validity only
Van Gestel [19], 2012	70	49/21	25.00 (7.70)	62.40 (7.40)	43.00 (22.00)	43.60 (15.00)	Mild = 23.9%; Moderate = 8.5%; Severe = 31; Very severe = 36.6	Validity only
Hunt [4], 2013	24	18/6	27.60 (4.30)	74.40 (7.90)	54.00 (13.00)	47.80 (12.90)	Mild = 5; Moderate = 7; Severe = 12; Very severe = 0	Validity & reliability
Nyssen [16], 2013	30	23/7	24.60 (4.70)	68.00 (10.00)	48.00 (14.90)	NA	Moderate – severe	Validity only
Gimeno-Santos [8], 2015	236	160/76	27.00 (5.50)	67.40 (8.40)	57.00 (20.50)	46.90 (13.20)	Mild = 34; Moderate = 108; Severe = 71; Very severe = 22	Validity & reliability

M (SD) = Mean (Standard Deviation), BMI = Body mass index, FEV1% = percentage of Forced expiratory volume in 1 s predicted value, FVC = forced vital capacity, ᵋ = males/females, NA = Not available, NS = not specified.

each (54.5% of the total sample). Information on the number of raters was provided in one study. However, despite the use of 2 raters, blinding of raters was not performed and inter-rater reliability of raters was not reported. Among the studies assessing the reliability of outcomes ($n = 5$), three studies used appropriate statistical measures and only one study varied the order of examination (Supplementary table 5).

Seven out of the total 12 studies scored in the top 25th percentile for transparency in reporting on the STROBE (Supplementary table 7). Very few studies reported key elements of study design such as bias and study size (25% and 42% of all studies). Other elements of study design were reported in 75–100% of the studies. When reporting results, 56% of the studies included details on participant recruitment, and only 25% included confidence intervals and ancillary analyses (sensitivity and sub-group analyses). Other important elements such as reporting of limitations and generalizability were missing in 25% and 33% of the included studies, respectively.

Validity outcomes

Validity of PA measures was assessed in 11 studies that were included in this review. Among the studies that assessed validity, eight studies addressed concurrent validity and 3 studies

addressed both concurrent and construct validity of the outcome measures. No study assessed the content validity (Supplementary Table 2). Concurrent validity was assessed on outcomes including energy expenditure, PA levels and time spent in PA. Construct validity was assessed against constructs such as six-minute walk distance and lung function. Four studies described diagnostic properties including sensitivity, specificity and/or area under the curve of five measures PASE, Stanford seven-day recall (PAR), Yale physical activity questionnaire (YPAS), modified Baecke, and Zutphen Physical Activity Questionnaire (ZPAC).

The self-administered PRO, modified Baecke questionnaire was the most frequently reported questionnaire in COPD with 3 studies [1, 16, 22] evaluating its validity, followed by the rater-based measure, PASE, and self-administered PRO, ZPAC (2 studies each) [1, 19, 21]. The remaining measures were described in single validation studies (Supplementary Table 1). Five studies used the SenseWear Pro Arm Band [1, 4, 14, 19, 21] (SAB) making it the most frequently used reference standard for testing concurrent validity of the subjective measures. Other reference standards used included the Yamax dig walker SW-700 [15, 16] (2 studies), ActiReg, Actigraph GT3X+, DynaPort Activity Monitor (DAM) and the Actihealth (Actiped) accelerometer (separate single studies) (Supplementary Table 2) [4, 17, 18, 20, 22].

Concurrent validity

Energy expenditure

The assisted PRO, PAR, demonstrated the strongest correlations with SenseWear Armband on the measured EE ($r = 0.83$; $p < 0.001$) followed by the self-administered PRO, YPAS ($r = 0.40$, $p < 0.001$) (Table 3) [1, 14]. The ZPAC demonstrated high variability across studies in the measured associations for EE ($r = 0.01–0.50$) with significant underestimation of EE as compared to the SAB (mean difference in EE $= 922$ KJ, $p < 0.001$) [1, 19].

Physical activity level

PA levels measured by the computerized PRO, Multimedia Activity Recall (MARCA), demonstrated moderate to good correlations with both SenseWear and Actigraph GT3X + accelerometers ($r = 0.66–0.74$) (Table 3) [4]. Modified Baecke ($r = 0.15$, $p = 0.35$) and PASE questionnaires ($r = 0.19$, $p = 0.23$), on the other hand, demonstrated weak correlations with reference standards in terms of PA levels [1].

Time spent in PA

The assisted PRO, PAR, was found to demonstrate moderate correlations with the SenseWear arm band for time spent in activity over 3 METs ($r = 0.54$, $p < 0.001$) [1]. The YPAS showed fair but significant correlations for time spent in PA ($r = 0.38–0.41$, $p < 0.001$) (Table 3) [14]. The self-administered PRO, ZPAC, on the other hand, demonstrated only weak correlations with the reference standard on time spent in activity ($r = 0.18$, $p = 0.25$) [1].

Steps per day

Of the four measures that were assessed for relationship with daily step counts, the modified Baecke questionnaire demonstrated the strongest agreement with pedometer steps with slight overestimation of counts (mean difference between pedometer and modified Baecke steps/day $= -0.129$, $p = 0.496$) [16]. The Follick diary showed weak correlations with pedometer counts as did the International Physical Activity Questionnaire short form (IPAQ-SF) [15, 16].

Construct validity

The construct validity of four measures (YPAS, physical activity checklist, Daily-PRO Physical Activity in COPD (D-PPAC) and Clinical- PRO Physical Activity in COPD (C-PPAC)) was reported in three studies against measures of functional capacity such as the six-minute walk distance (6MWD), lung function (FEV1) and dyspnea and disease severity measures (mMRC scale, BODE index etc.) (Supplementary Table 2). Of the four measures studied, the hybrid measure, C-PPAC, demonstrated the strongest correlations with 6MWD, both when combined with DynaPort activity monitor and with the Actigraph ($r = 0.62$, $p < 0.05$ and $r = 0.65$, $p < 0.05$ respectively followed closely by the D-

PPAC ($r = 0.55$, $p < 0.05$) [8]. The YPAS showed fair correlations with 6MWD ($r = 0.37- 0.40$, $p < 0.01$ (Table 3) [14, 20].

Diagnostic accuracy

The assisted PRO, PAR, also showed good diagnostic accuracy with ability to detect very inactive patients with a sensitivity of 0.73 and specificity of 0.76. The PASE (rater-based) and YPAS (assisted PRO) demonstrated high sensitivity values (0.85 and 0.75), but fair specificity (0.66 and 0.59), decreasing their overall accuracy in the ability to detect severe inactivity [1, 14, 21].

Reliability outcomes

Only five studies out of the total included studies examined reliability outcomes for the PA measures. The outcome measures studied for reliability included the PASE, activity checklist, Stanford Brief Assessment Scale (SBAS), MARCA, Quantification de l' Activité Physique (QUANTAP) and the C-PPAC and D-PPAC measures [4, 8, 15, 21, 22]. Test-retest reliability was reported for all measures, with test-retest intervals ranging from 4 h to 14 days. Both computerized PRO measures, MARCA and QUANTAP questionnaires, demonstrated excellent test-retest reliability values (ICC's $= 0.95–0.96$ and 0.92 respectively) [4, 22]. Among the hybrid measures, the C-PPAC showed better reliability over 1 week as compared to the D-PPAC (ICC's $= 0.74–0.88$ and 0.71–0.87 for the D-PPAC DAM and Actigraph respectively). The C-PPAC measures demonstrated excellent test-retest reliability with both the accelerometer combinations (C-PPAC DynaPort, ICC $= 0.92$ and C-PPAC Actigraph, ICC $= 0.90$ respectively) [8]. Details on reliability outcomes for these questionnaires can be found in Tables 2 and 3. Both PRO PPAC measures demonstrate good internal consistency between items.

Discussion

The purpose of this review was to systematically review the psychometric properties of various subjective PA assessments validated in adults with COPD. The results of this review showed that assisted and computerized PRO measures (PAR and MARCA) as well as the hybrid measures (C-PPAC and D-PPAC) demonstrated better validity than other PA assessments. Unsupervised logs such as the activity checklist and Follick's diary, on the other hand, failed to accurately assess PA dimensions [15, 17, 20].

Validity outcomes

Concurrent validity

The PAR, which is an assisted PRO measure, demonstrated strong correlations with reference standards on energy expenditure and moderate-to-good correlations on time spent in moderate PA [1]. As compared to other self-reported, unsupervised PRO measures, the close relationship

Table 2. Reliability of various subjective physical activity measures.

Author, Year; Questionnaire	Type of reliability	Outcomes measured for reliability	Test retest interval	Reliability measures
Moy [20], 2009; Activity checklist	Test retest	Number of activities	Over 14 days	*Intra subject coefficient of variation in number of daily activities*: Median CV = 0.28; IQR 0.22–0.32
Gouzi [22], 2011; Quantification de l' Activité Physique (QUANTAP)	Intra rater	Responses to questionnaire	1 week	*Intra-observer correlation coefficient (CI) between test 1 and test 2:* ICC = 0.92 (NA)
DePew [21], 2013; Physical Activity Scale for the Elderly (PASE)	Test retest	PASE scores	5–7 days	Pearson's r = 0.75 BA plot
DePew [21], 2013; Stanford Brief Assessment Scale (SBAS)	Test retest	SBAS scores	5–7 days	Kappa coefficient K = 0.41
Hunt [4], 2013; Multimedia Activity Recall in Children and Adults (MARCA)	Test retest	MVPA ≥ 3 METs (min/day), TDEE (MET.min)	4 h	*ICC, limits of agreement for TDEE:* ICC = 0.95- 0.96, −196 to + 164 MET.min *ICC, limits of agreement for MVPA ≥ 3 METs:* ICC = 0.93–0.95, −61 to + 66 min/day
Gimeno-Santos [8], 2015 PROactive Physical Activity in COPD (PPAC) tools: Clinical PAC (C-PPAC)	Test rest; Internal consistency	Amount of activity (PAQ + DAM) Amount of activity (PAQ + DAM)	1 week	*Intra class correlation (CI) coefficient for amount of PA using PAQ and DAM combined between:* Week 1 and week 2 = ICC 0.92 (NA) *Intra class correlation (CI) coefficient for amount of PA using PAQ and Actigraph combined between:* Week 1 and week 2 = ICC 0.90 (NA) *Internal consistency for D-PPAC factor 1 items using PAQ and DynaPort for examination 1 and 2:* Cronbach's α = 0.813 and 0.800 *Internal consistency for D-PPAC factor 1 items using PAQ and Actigraph for examination 1 and 2:* Cronbach's α = 0.803 and 0.781
Gimeno-Santos [8], 2015; PROactive Physical Activity in COPD (PPAC) tools: Daily PAC (D-PPAC)	Test rest; Internal consistency	Amount of activity (PAQ + DAM) Amount of activity (PAQ + DAM)	1 week	*Intra class correlation (CI) coefficient for amount of PA using PAQ and DAM combined between:* Day 1of week 1and day 1 of week 2 = ICC 0.74 (NA) Day 2 of week1 and day 2 of week 2 = ICC 0.84 (NA) Day 3 of week1 and day 3 of week 2 = ICC 0.80 (NA) Day 4 of week1 and day 4 of week 2 = ICC 0.74 (NA) Day 5 of week1 and day 5 of week 2 = ICC 0.75 (NA) Day 6 of week1 and day 6 of week 2 = ICC 0.88 (NA) Day 7 of week1 and day 7 of week 2 = ICC 0.86 (NA) *Intra class correlation (CI) coefficient for amount of PA using PAQ and Actigraph combined between:* Day 1of week 1and day 1 of week 2 = ICC 0.81 (NA) Day 2 of week1 and day 2 of week 2 = ICC 0.81 (NA) Day 3 of week1 and day 3 of week 2 = ICC 0.82 (NA) Day 4 of week1 and day 4 of week 2 = ICC 0.73 (NA) Day 5 of week1 and day 5 of week 2 = ICC 0.71 (NA) Day 6 of week1 and day 6 of week 2 = ICC 0.87 (NA) Day 7 of week1 and day 7 of week 2 = ICC 0.87 (NA) *Internal consistency for D-PPAC factor 1 items using PAQ and DynaPort for examination 1 and 2:* Cronbach's α = 0.862 and 0.845 *Internal consistency for D-PPAC factor 1 items using PAQ and Actigraph for examination 1 and 2:* Cronbach's α = 0.860 and 0.839

MVPA = moderate to vigorous physical activity, METs = metabolic equivalents, min/day = minutes per day, TDEE = total daily energy expenditure, MET.min = MET minutes, ICC = intra class correlation coefficient, CV = coefficient of variation, CI = confidence interval, IQR = interquartile range, NA = not available, BA plot = Bland Altman plot.

Table 3. Comparison of validity and reliability across PA measures.

Concurrent Validity		
Reference standards used for concurrent validity	PA measure	Correlation Coefficients (r)
Energy Expenditure with SAB	PAR	0.83
	YPAS	0.40
	ZPAC	0.01–0.50
PA Level (SAB and actigraph for MARCA and SAB for others)	MARCA	0.66–0.74
	Modified Baecke	0.15
	PASE	0.19
Time Spent in PA (min)	PAR	0.54
	YPAS	0.38–0.41
	ZPAC	0.18
Construct Validity		
Reference standards used for concurrent validity	PA measure	Correlation Coefficients
6MWD	C-PPAC	0.62
	D-PPAC	0.55
	YPAS	0.37–0.40
Reliability		
Test Retest Intervals	PA measure	ICC values
4 h	MARCA	0.95–0.96
1 week	QUANTAP	0.92
1 week	C-PPAC	0.90–0.92
Daily	D-PPAC	0.74–0.87

PA = physical activity, 6MWD = six-minute walk distance, ICC = intraclass correlation coefficient, SAB = SenseWear armband, Actigraph = actigraph GT3X, min = minutes.

of PAR with reference standards on time spent in activity can be explained by the nature of the interview. The PAR is a supervised interviewer-led questionnaire where the interviewer directs the recall process day by day using guided memory techniques to minimize errors in recall, which may explain the stronger correlations seen with this measure as compared to the unsupervised PROs [1]. However, one must note that these observations were drawn from the findings of a single study [1] and from a local hospital in a specific geographical location with no details on sampling, and other pertinent demographic factors that may affect PA (race, socio economic status, depression, and mobility parameters). Therefore, it is difficult to draw conclusions on the validity of the PAR for all COPD patients. Additionally, despite strong correlations on energy expenditure, the limits of agreement on the difference between SAB and PAR varied widely (−262 to 1062 Kcals/day) [1]. This, along with modest ICC values for time spent in moderate PA, further limit the strength of recommendations on the utility of this measure.

MARCA, a computerized PRO measure, was also shown to perform better than the self-administered measures in its relationship with reference standards [4]. The MARCA aids in recall by reconstructing entire days instead of asking open questions (e.g. How much time did you spend in walking, sleeping etc. during the day?") [4]. MARCA utilizes a segmented day format, where the entire day is divided into time segments of 5 min and users get a chance to systematically recall all activities in the context and order in which they were performed. The time segmentation and specific previous day recall combined with additional prompts to fill out activities for missed time segments, may decrease recall errors [23]. As with the PAR, the results on the MARCA were drawn from observations of a single study [4], making it difficult to draw definite conclusions. Although, the authors of the study assessing MARCA used both accelerometers (SAB and Actigraph GT3X+) and a pedometer as reference standards to address limitations of a single

reference standard [4], issues of methodological quality, including unclear selection criteria, limited representativeness of the sample, lack of information on the time frame of recruitment, failure to identify possible confounding variables and missing information on the confidence intervals limit the generalizability of findings (Supplementary tables 5, 6, and 7). Despite strong correlations of MARCA with both the accelerometers, the limits of agreement for the difference on time spent in moderate to vigorous PA varied greatly, both between Actigraph and MARCA (−104.31–153.81 min) and between SAB and MARCA (−118.09–144.82) [4].

Other self-administered PRO measures including the ZPAC, Baecke and activity diaries failed to demonstrate good concurrent validity in patients with COPD. The PASE, a rater-based measure, was found to have poor correlations with overall PA, despite of high sensitivity in identifying severe physical inactivity [1, 21]. One possible explanation for poor to modest correlations of the PASE with objectively-measured PA could be the inclusion of items requiring minimal physical exertion. Recall of activities that are moderate to high intensity is often easier than light PA and a component of recall bias using the PASE cannot be ruled out, especially in the absence of prompts or cues to facilitate recall [21].

Construct validity
Both the assisted PRO measure (YPAS) and hybrid measures (C-PPAC and D-PPAC) demonstrated statistically significant, fair-to-moderate correlations with related constructs of PA. The C-PPAC and D-PPAC were recently created based on the conceptual framework of PA in COPD that was drafted from patients' experience [24]. These tools combine objective and subjective components of activity assessment by utilizing two accelerometers, the Actigraph GT3x and the DynaPort Activity monitor in addition to the PRO measures [8]. Each of these measures examined two factors, amount of PA and difficulty with PA, which were analyzed

separately [8]. For the purpose of this review, only the objective components of both measures (amount of PA) were reported. Both measures were shown to demonstrate construct validity, with the clinical version (C-PPAC) showing better convergence with related constructs such as dyspnea and 6MWD as compared to the daily version (D-PPAC) [8]. Findings from the study also showed that the accelerometer-PRO combination yielded better associations than when either was used alone (accelerometer or PRO separately) [8]. This is a promising finding that warrants further exploration of different combinations of hybrid measures in future research.

The assisted PRO, YPAS, demonstrated fair correlations ($r = 0.38–0.40$) with related constructs such as the 6MWD [9]. The YPAS also is an assisted questionnaire where the interviewer uses structured questions to guide the user [25, 26]. It is noteworthy that significant correlations were obtained despite the inclusion of light PA items in the YPAS that are typically excluded from instruments [26]. The face to face nature of the questionnaire and nature of recall (recall from a previous specified week instead of general recall) [2, 27] may have contributed to this finding. However, findings obtained only from a single study limit the ability to draw conclusive statements on the applicability of YPAS at this time.

The distance walked in six minutes (6MWD) has been used as a related construct in previous research of construct validity [28–30]. However, this association between 6MWD and PA have mostly been determined via cross-sectional studies, with a few studies showing inconsistent relations in the COPD population [31–33]. A recent systematic review identified 6MWD as a weak but consistent determinant of PA in COPD. Several other determinants of PA have been identified in literature including dyspnea, quality of life, lung function, systemic inflammation, mortality and exacerbations etc [34]. Considering that the relationships of PA with 6MWD have mainly been identified via cross-sectional studies, use of additional determinants along with 6MWD for construct validity studies would be of interest.

Reliability outcomes

Both the computerized PRO measures (MARCA and QUANTAP) and the clinical hybrid measure (C-PPAC) demonstrated excellent test-retest reliability (ICC = 0.90–0.96). Between the two hybrid tools, the C-PPAC showed better test-retest reliability, possibly due to the test interval (one-week v/s daily), that may have affected recall (Table 2). The unsupervised PROs and rater-based measures, on the other hand demonstrated weak reliability outcomes [8]. However, it is noteworthy that Pearson's or Spearman's correlation coefficients were used to report reliability for the unsupervised PROs and rater-based measures. Correlation coefficients have been reported as less effective methods of determining reliability as these often ignore systematic errors and individual variations in performance [9, 35]. The difference in methods used for reliability analysis make it difficult to draw true comparisons among these measures.

Limitations

The present systematic review was not without limitations. The review was limited to studies published in the English language. Also, the review was limited to the search of 2 large databases. The authors tried to address these problems by being more inclusive in their search via inclusion of all gray literature, performance of manual searches in the reference list of articles and contacting authors for English versions of the studies, if available. Very few studies reported other measurement properties and therefore a decision was made to only include reliability and validity for this review.

The authors also advise caution to the readers when interpreting the findings of this review. The results reported here are from conclusions drawn from either single studies or a combination of very few (1 or 2) studies, with poor methodological quality and quality of reporting (Supplementary tables 5, 6, and 7). In the absence of more high-quality validation studies on PA measures in COPD, it is difficult to draw substantive conclusions. Additionally, the reference standards used in most studies was the SenseWear Pro armband which, has shown good criterion validity in controlled lab conditions, but has demonstrated questionable validity in daily living conditions, in those with slower walking speeds, or those using assistive devices [36, 37]. Since the measures were validated in daily living conditions, the use of SAB as a reference standard might not reflect a true representation of PA in these patients. Three studies also used pedometers as reference standards, which have been previously found to underestimate PA in COPD [4, 15, 16]. Mobility parameters such as the walking speed, use of assistive device, use of supplemental oxygen etc. and social factors such as marital status, social interaction, depression etc., that might have an effect on PA, were not well described in the studies, limiting our ability to draw definite conclusions [38].

The PRO tools for assessment of PA have been criticized in recent literature on their inability to capture the construct of PA [24, 39]. However, further exploration on the distinction between self-administered and assisted PROs for PA quantification is needed. For quantification of PA in terms of amount and duration of activity, the assisted and computerized PROs demonstrate better psychometrics than other measures. Further research to validate these measures in COPD should be performed.

Future implications

Future research is needed to examine the validity of assisted and computerized PRO measures using valid reference standards. Newer hybrid tools such as the C-PPAC and D-PPAC demonstrate promising good construct validity but need further research to establish validity in assessing various dimensions of PA against objective reference standards. There is a potential for further research to compare different accelerometer- PRO combinations and to examine their validity across different COPD severity to further validate these measures.

Conclusion

Assisted and computerized PRO measures (PAR and MARCA) and hybrid measures (C-PPAC and D-PPAC) demonstrate better psychometric properties as compared to other PRO or rater-based measures; and should be considered for PA assessment in COPD. Unsupervised self-administered PRO measures failed to demonstrate good validity and/or reliability and must be used with caution in this population. However, observations drawn from single validation studies and methodological quality concerns limit the strength of recommendations and further research is needed to fill this knowledge gap.

Disclosure statement

The authors report no conflicts of interest.

ORCID

Shweta Gore 🔟 http://orcid.org/0000-0002-5802-182X

References

1. Garfield BE, Canavan JL, Smith CJ, et al. Stanford seven-day physical activity recall questionnaire in COPD. Eur Respir J. 2012;40(2):356–362. DOI:10.1183/09031936.00113611
2. Pitta F, Troosters T, Probst VS, et al. Quantifying physical activity in daily life with questionnaires and motion sensors in COPD. Eur Respir J. 2006;27(5):1040–1055. DOI:10.1183/09031936.06.00064105
3. Bossenbroek L, de Greef MHG, Wempe JB, et al. Daily physical activity in patients with chronic obstructive pulmonary disease: A systematic review. Chronic Obstr Pulm Dis. 2011;8(4):306–319. DOI:10.3109/15412555.2011.578601
4. Hunt T, Williams MT, Olds TS. Reliability and validity of the multimedia activity recall in children and adults (MARCA) in people with chronic obstructive pulmonary disease. PLoS One. 2013;8(11):e81274. DOI:10.1371/journal.pone.0081274
5. Williams K, Frei A, Vetsch A, et al. Patient-reported physical activity questionnaires: a systematic review of content and format. Health Qual Life Outcomes. 2012;10(1):28. DOI:10.1186/1477-7525-10-28
6. Helmerhorst HJF, Brage S, Warren J, et al. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. Int J Behav Nutr Phys Act. 2012;9(1):103. DOI:10.1186/1479-5868-9-103
7. Ainsworth B, Cahalin L, Buman M, et al. The current state of physical activity assessment tools. Prog Cardiovasc Dis. 2015;57(4):387–395. DOI:10.1016/j.pcad.2014.10.005
8. Gimeno-Santos E, Raste Y, Demeyer H, et al. The PROactive instruments to measure physical activity in patients with chronic obstructive pulmonary disease. Eur Respir J. 2015;46(4):988–1000. DOI:10.1183/09031936.00183014
9. Portney L, Watkins M. Foundations of clinical research: applications to practice. Philadelphia: F.A. Davis Company; 2009.
10. Lucas N, Macaskill P, Irwig L, et al. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010;63(8):854–861. DOI:10.1016/j.jclinepi.2009.10.002
11. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). BMC Med Res Methodol. 2013;13(1):111. DOI:10.1186/1471-2288-13-111
12. Gore S, Goldberg A, Huang MH, et al. Development and validation of a quality appraisal tool for validity studies (QAVALS). Physiother Theory Pract. 2019;1–9. DOI:10.1080/09593985.2019.1636435
13. Vandenbroucke JP, Elm EV, Altman DG, STROBE Initiative, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007;18(6):805–835. DOI:10.1097/EDE.0b013e3181577511
14. Donaire-Gonzalez D, Gimeno-Santos E, Serra I, et al. Validation of the Yale physical activity survey in chronic obstructive pulmonary disease patients. Arch Bronconeumol. 2011;47(11):552–560. DOI:10.1016/j.arbr.2011.07.004
15. Moore R, Berlowitz D, Denehy L, et al. Comparison of pedometer and activity diary for measurement of physical activity in chronic obstructive pulmonary disease. J Cardiopulm Rehabil Prev. 2009;29(1):57–61. DOI:10.1097/HCR.0b013e318192786c
16. Nyssen SM, Santos JGD, Barusso MS, et al. Levels of physical activity and predictors of mortality in COPD. J Bras Pneumol. 2013;39(6):659–666. DOI:10.1590/S1806-37132013000600004
17. Pitta F, Troosters T, Spruit MA, et al. Activity monitoring for assessment of physical activities in daily life in patients with chronic obstructive pulmonary disease. Arch Phys Med Rehabil. 2005;86(10):1979–1985. DOI:10.1016/j.apmr.2005.04.016
18. Slinde F, Grönberg AM, Svantesson U, et al. Energy expenditure in chronic obstructive pulmonary disease-evaluation of simple measures. Eur J Clin Nutr. 2011;65(12):1309–1313. DOI:10.1038/ejcn.2011.117
19. van Gestel AJR, Clarenbach CF, Stöwhas AC, et al. Predicting daily physical activity in patients with chronic obstructive pulmonary disease. PLoS One. 2012;7(11):e48081. DOI:10.1371/journal.pone.0048081
20. Moy ML, Matthess K, Stolzmann K, et al. Free-living physical activity in COPD: assessment with accelerometer and activity checklist. J Rehabil Res Dev. 2009;46(2):277–286. DOI:10.1682/jrrd.2008.07.0083
21. DePew ZS, Garofoli AC, Novotny PJ, et al. Screening for severe physical inactivity in chronic obstructive pulmonary disease: The value of simple measures and the validation of two physical activity questionnaires. Chron Respir Dis. 2013;10(1):19–27. DOI:10.1177/1479972312464243
22. Gouzi F, Préfaut C, Abdellaoui A, et al. Evidence of an early physical activity reduction in chronic obstructive pulmonary disease patients. Arch Phys Med Rehabil. 2011;92(10):1611–1617. DOI:10.1016/j.apmr.2011.05.012
23. Ridley K, Olds TS, Hill A. The Multimedia Activity Recall for Children and Adolescents (MARCA): development and evaluation. Int J Behav Nutr Phys Act. 2006;3(1):10. DOI:10.1186/1479-5868-3-10
24. Dobbels F, de Jong C, Drost E, PROactive consortium, et al. The PROactive innovative conceptual framework on physical activity. Eur Respir J. 2014;44(5):1223–1233. DOI:10.1183/09031936.00004814
25. Dipietro L, Caspersen CJ, Ostfeld AM, et al. A survey for assessing physical activity among older adults. Med Sci Sports Exerc. 1993;25(5):628.
26. Young DR, Jee SH, Appel LJ. A comparison of the Yale Physical Activity Survey with other physical activity measures. Med Sci Sports Exerc. 2001;33(6):955–961. DOI:10.1097/00005768-200106000-00015
27. Bonnefoy M, Normand S, Pachiaudi C, et al. Simultaneous validation of ten physical activity questionnaires in older men: A doubly labeled water study. J Am Geriatr Soc. 2001;49(1):28–35. DOI:10.1046/j.1532-5415.2001.49006.x
28. Altenburg WA, Bossenbroek L, de Greef MHG, et al. Functional and psychological variables both affect daily physical activity in COPD: A structural equations model. Respir Med. 2013;107(11):1740–1747. DOI:10.1016/j.rmed.2013.06.002
29. Garcia-Rio F, Lores V, Mediano O, et al. Daily physical activity in patients with chronic obstructive pulmonary disease is mainly

associated with dynamic hyperinflation. Am J Respir Crit Care Med. 2009;180(6):506–512. DOI:10.1164/rccm.200812-1873OC

30. Pitta F, Troosters T, Spruit MA, et al. Characteristics of physical activities in daily life in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2005;171(9):972–977. DOI:10.1164/rccm.200407-855OC

31. Durheim MT, Smith PJ, Babyak MA, et al. Six-minute-walk distance and accelerometry predict outcomes in chronic obstructive pulmonary disease independent of Global Initiative for Chronic Obstructive Lung Disease 2011 Group. Ann Am Thorac Soc. 2015;12(3):349–356. DOI:10.1513/AnnalsATS.201408-365OC

32. Fastenau A, van Schayck OCP, Gosselink R, et al. Discrepancy between functional exercise capacity and daily physical activity: a cross-sectional study in patients with mild to moderate COPD. Prim Care Respir J. 2013;22(4):425–430. DOI:10.4104/pcrj.2013.00090

33. Sperandio EF, Arantes RL, da Silva RP, et al. Screening for physical inactivity among adults: the value of distance walked in the six-minute walk test. A cross-sectional diagnostic study. Sao Paulo Med J. 2016;134(1):56–62. DOI:10.1590/1516-3180.2015.00871609

34. Gimeno-Santos E, Frei A, Steurer-Stey C, PROactive consortium, et al. Determinants and outcomes of physical activity in patients with COPD: a systematic review. Thorax. 2014;69(8):731–739. DOI:10.1136/thoraxjnl-2013-204763

35. Vaz S, Falkmer T, Passmore AE, et al. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS One. 2013;8(9):e73990. DOI:10.1371/journal.pone.0073990

36. Cavalheri V, Donária L, Ferreira T, et al. Energy expenditure during daily activities as measured by two motion sensors in patients with COPD. Respir Med. 2011;105(6):922–929. DOI:10.1016/j.rmed.2011.01.004

37. Furlanetto KC, Bisca GW, Oldemberg N, et al. Step counting and energy expenditure estimation in patients with chronic obstructive pulmonary disease and healthy elderly: Accuracy of 2 motion sensors. Arch Phys Med Rehabil. 2010;91(2):261–267. DOI:10.1016/j.apmr.2009.10.024

38. Larson JL, Vos CM, Fernandez D. Interventions to increase physical activity in people with COPD: systematic review. Annu Rev Nurs Res. 2013;31(1):297–326. DOI:10.1891/0739-6686.31.297

39. Gimeno-Santos E, Frei A, Dobbels F, PROactive consortium, et al. Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review. Health Qual Life Outcomes. 2011;9(1):86. DOI:10.1186/1477-7525-9-86