Subjective Assessments of Physical Activity in Chronic Obstructive Pulmonary Disease

by

Shweta Gore PT, DPT, GCS, CLT

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Physical Therapy)
School of Health Professions and Studies
University of Michigan-Flint
2017

Doctoral Committee:

Associate Professor Jennifer Blackwood, Chair
Professor Allon Goldberg
Associate Professor Min H. Huang
Associate Professor Michael Shoemaker, Grand Valley State University

Shweta Gore

shwetag@umflint.edu

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the immense support of several people, who helped me in several ways throughout this process.

First of all, I would like to thank my doctoral committee chair, Dr. Jennifer Blackwood, for guiding and supporting me over the years, and for mentoring me to a career of research. You patiently provided feedback and answered my numerous emails even when you were not working. You kept me focused towards my goal and motivated me when I doubted myself. Because of you, I am a better clinician and a better researcher. Thank you.

Thank you to my committee members, Dr. Allon Goldberg, Dr. Min Huang and Dr. Michael Shoemaker for their immense support and valuable feedback, and for their readiness to help when I needed their guidance.

To my best friend, and husband, Devashish Tiwari, who supported me every step of the way, despite working on his own PhD coursework. I remember getting up at early hours in the morning, going to the library after work, and numerous weekends spent in writing, when he was there right by my side. I thank you Dev for patiently listening to my frustrations and always believing in me.

To my mom and dad, Sarita and Shirish Gore, who always supported my beliefs and encouraged me to pursue my goals in life. It was their strong will that made me move to the U.S. to pursue

my dreams. To my sister, Swati Gore, who was there to listen to me and to distract me with her stories that were always refreshing and therapeutic.

To Brady West and Josh Erickson from the Center of Statistical Consulting and Research (CSCAR), I thank you for your support and guidance with statistical analysis.

I thank Dr. Bara Alsalaheen, for being a constant source of inspiration and support throughout my time at University of Michigan-Flint.

Finally, I thank and all my friends and colleagues for their love, support and constant encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS		ii
LIST OF TABLES		vii
LIST OF FIGURES		viii
LIST OF APPENDICES		ix
PUBLICATION STATUS		x
BROAD ABSTRACT		xi
CHAPTER		
I	**Background**	1
	Classification, domains and dimensions of activity	3
	Methods of assessment of PA	5
	Need for systematic review on PA assessment in COPD	11
	Validity of Measures for Population PA Surveillance	15
	Approach and Methodology	19
	Summary	24
	References	25

II	**Development Of A Quality Appraisal Tool For Validity Studies (QAVALS)**	30
	Abstract	31
	Introduction	33
	Methods	36
	Analysis	42
	Results	43
	Discussion	46
	Limitations	49
	Conclusion	50
	References	51
III	**Subjective Physical Activity Assessments in adults with COPD: A Systematic Review**	75
	Abstract	76
	Introduction	79
	Methods	82
	Quality Assessment	85
	Results	87
	Discussion	93
	Limitations	97
	Conclusion	99
	References	100
IV	**Validity of the Global Physical Activity Questionnaire in older adults with COPD**	133
	Abstract	134
	Introduction	136

Methods 139
Analysis 143
Results 145
Discussion 148
Limitations 151
Conclusion 153
References 154

V Overview 165

Summary of Research Design and Results 166
Discussion of Results 169
Limitations 174
Recommendations for Future Research 175
Conclusion and Clinical Implications 177
References 178

LIST OF TABLES

II.1	Quality Assessment of Validity Studies (QAVALS)	54
II.2	Inter rater reliability of individual items on the QAVALS	55
III.1	Characteristics of the study population	107
III.2	Characteristics of subjective physical activity measures	108
III.3	Validity and diagnostic properties of subjective physical activity measures	115
III.4	Reliability of various subjective physical activity measures	124
IV.1	Comparisons of demographic and clinical characteristics between COPD and non COPD participants	160
IV.2	Logistic regression to test the ability of GPAQv2 PA in identifying the presence or absence of COPD	161
IV.3	Multiple regression examining associations between PA scores and lung function – FEV1 in older adults with COPD using GPAQv2	162
IV.4	Logistic regression examining associations between PA scores and shortness of breath in older adults with COPD using GPAQv2	163
IV.5	Logistic regression examining association between GPAQv2 PA and household income in older adults with COPD	164

LIST OF FIGURES

PRISMA Flow Diagram 105

Participant flow chart 158

LIST OF APPENDICES

II.A	Results of first round of review on the QAVALS	56
II.B	Results of the second round of review on the QAVALS	61
II.C	Explanation of items and instructions for scoring QAVALS	65
III.A	Screening criteria for inclusion of articles in the review	126
III.B	Reasons for exclusion of full text articles after review	127
III.C	Methodological quality assessment of reliability studies (QAREL)	129
III.D	Methodological quality assessment of validity studies	130
III.E	Assessment of reporting quality of the studies using STROBE	132
V.A	Approval Letter from the Institutional Review Board, University of Michigan-Flint	181
V.B	PEERRS certification	182

PUBLICATION STATUS

II	Development of a Quality Appraisal Tool for Validity Studies (QAVALS)	Not yet submitted
III	Subjective Physical Activity Assessments in adults with COPD: A Systematic Review	Not yet submitted
IV	Validity of the Global Physical Activity Questionnaire in older adults with COPD	Not yet submitted

BROAD ABSTRACT

The aim of this three-paper dissertation is to examine the measurement properties of available subjective physical activity assessments in chronic obstructive pulmonary disease (COPD) in order to aid clinicians and researchers in making sound clinical judgements by systematically reviewing available evidence and then examining the validity of a widely used physical activity assessment tool in COPD. This dissertation also provides researchers with a new valid and reliable tool that can be used for quality assessment of validation studies and that can be used to improve the rigor of systematic reviews in observational studies.

Within this dissertation, there are three independent studies. The first study describes the development and initial validation of a tool for quality assessment of validity studies. The second study is built on Study 1, where it utilizes the newly developed tool from study 1 to systematically review the reliability and validity of various subjective physical activity assessments in adults with COPD. The final study then investigates the construct validity of a widely used national physical activity surveillance tool - the Global Physical Activity Questionnaire, in older adults with COPD.

Background information for each of the three studies is provided in Chapter I of this dissertation. Chapters II-IV include three individual studies with each containing methods,

results and discussion sections for three studies. The findings of the three studies are presented in an integrated discussion in the final chapter of the dissertation.

The following section provides background information for the three studies to help frame the purpose of the three studies within this dissertation

CHAPTER I

BACKGROUND

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the United States and worldwide.[1,2] By 2020, COPD has been predicted to become the fifth largest contributor to disability in the United States.[3,4] Physical inactivity is common in individuals with COPD, even in the early stages of the disease.[5,6] Individuals with COPD have 40 – 60% lower physical activity (PA) levels as compared to those without the disease.[5-8] Disabling symptoms of COPD including dyspnea and fatigue have been frequently attributed to decreased PA in COPD.[9] Several other disease-specific factors have reported associations with low PA in COPD, such as measures of impaired lung function (forced expiratory volume in the first second (FEV1), diffusing capacity and expiratory muscle strength), decreased functional capacity (peak aerobic capacity, 6-minute walk distance and quadriceps strength), accelerated inflammation (fibrinogen, C-reactive protein, tumor necrosis factor) and disease severity (degree of disease severity, frequency of COPD exacerbations and health status).[9-11] Physical inactivity contributes to a downward spiral of immobility in COPD.[9] Individuals with COPD often report dyspnea and fatigue induced by daily PA.[12] In response to these symptoms, they often decrease their level of PA, which in itself can contribute to further worsening of muscle strength, endurance, cardiopulmonary conditioning, and worsening of dyspnea.[10,13,14] This contributes to a vicious cycle where low PA in COPD results as a consequence of the disease, but also individually contributes to the further progression and worsening of the disease.[8] This downward spiral of symptom induced activity limitation and immobility can lead to social isolation and depression, thereby, negatively affecting overall quality of life.[9,12]

Physical inactivity, independent of lung function abnormality, is associated with poor outcomes in COPD, including an increased mortality risk.[6,9] PA was found to be a stronger predictor of survival in COPD than any other factors including lung function, 6-minute walk distance, body mass index, fat-free mass index, dyspnea, health status, depression symptoms, and multiple systemic biomarkers.[14]

Low levels of PA have also been identified to predict hospitalization or re-hospitalization and have been associated with a faster decline in lung function.[13,15,16] Increasing daily PA, on the other hand, has demonstrated the potential to break the vicious downward spiral of mobility in COPD.[8] Higher levels of PA have shown a protective effect in people with COPD, including slower lung function decline, fewer exacerbations, decreased risk of hospitalization and decreased risk of mortality.[7,14-16] Based on the complex relationships between PA, dyspnea and deconditioning, positive gains can be obtained in disease outcomes by interrupting any one of these interlinked mechanisms.[8] Since lack of PA in individuals with COPD is associated with poor outcomes, increasing the overall PA level is a desired goal for management.[4,6]

Accurate assessment of activity plays an important role in understanding the severity of the disease and is also important in the prognosis and prevention of the downward spiral of immobility.[8,9] Given the growing interest in this area, the American College of Sports Medicine and Kaiser Permanente in a joint consensus meeting in 2015 discussed the development and implementation of a physical activity vital sign (PAVS) to be recorded at every medical visit. They further called for all current and future health care providers to use PA assessment in their routine practice with every patient.[17]

Physical Activity

PA was first described in 1985[18] as any bodily movement produced by skeletal muscles resulting in energy expenditure.[18] PA has been listed as one of the top ten health indicators in the healthy people objectives for the American population.[19] A ten percent relative reduction in the prevalence of inactivity has also been documented as one of the top four global targets to decrease the burden of non-communicable disease.[20] The 2008 Federal guidelines for PA recommended that all adults should engage in some form of activity and that substantial gains in health can be observed by performing150 minutes per week of moderate-intensity or 75 minutes per week of vigorous-intensity aerobic PA, or an equivalent combination of both, accumulating in bouts of ≥10 minutes. Additional gains in health outcomes can be observed by increasing PA to 300 minutes per week at moderate intensity or 150 minutes per week at vigorous intensity or an equivalent combination of both along with performance of muscle-strengthening activities of moderate to high intensity performed ≥2 days per week.[19,21,22]

Classification, dimensions and domains of PA

Classification of PA. PA can be classified into two broad categories, namely, structured PA and incidental PA. Structured PA is otherwise referred to as 'exercise', which is defined as a planned and purposeful activity performed with an intention to promote health and fitness.[19] Incidental PA, on the other hand, is unplanned. Incidental physical activities are usually associated with daily living tasks that are performed at home, during transport or at work.[19]

Dimensions of PA. PA is described under four major dimensions including mode, frequency, duration and finally the intensity of performing an activity.[19] Mode refers to the type of PA performed (e.g. walking, running, swimming, cycling etc.) or different forms of activity

such as aerobic or anaerobic, strengthening, agility, flexibility or balance activity. Frequency refers to the number of sessions an activity is performed in a day or per week. For an activity to be qualified to be counted in the number of sessions, the activity should be performed for at least ≥10 minutes. Duration of PA refers to the time measured in minutes or hours that is spent in an activity during a specified time frame such as a day, week, year or month. Intensity of PA refers to the rate of energy expenditure. This is a measure of the amount of energy spent during an activity.[19] It can be objectively quantified with physiological measures (e.g., oxygen consumption, heart rate, respiratory exchange ratio), subjectively assessed by perceptual characteristics (e.g., rating of perceived exertion, walk-and-talk test), or quantified by body movement (e.g., stepping rate, 3-dimensional body accelerations)

Domains of PA. PA is also described under four common domains. These are the 1) occupational domain which includes work related tasks that may involve manual labor, lifting, carrying or walking tasks; 2) domestic domain that includes house work, self-care, shopping, child-care, yard work and incidental activities; 3) transport domain including tasks related to the purpose of traveling or going somewhere such as walking, bicycling, climbing stairs, and public transportation; or 4) leisure time activities that include recreational activities such as sports, hobbies, exercise, or volunteer work.[19]

Since PA pertains to body movements that result in energy expenditure, it is often measured in terms of the amount of energy spent during an activity. Various units of PA measurement are used.

Units of measurement of PA

The amount of energy expended by a person above the resting level during PA may range from low to high and is determined by the amount of muscle mass contracting, as well as the

intensity, duration and frequency of the muscle contractions.[18,19] The quantification of PA may be performed in relation to either movement or energy expenditure.[23] Total energy expenditure is comprised of three major components, including resting energy expenditure, thermic effect of food and physical activity energy expenditure. PA energy expenditure (PAEE) is defined as the measure of energy cost of performing an activity[24] and is the most variable, constituting about 15 – 30% of the total energy spent daily.[19,25]

Commonly used expressions of the amount of energy expenditure include kilocalories, metabolic equivalent units (MET) or as a measure of time spent in a specific PA (measured in minutes or hours).[19] Kilocalorie is a measure of heat and when expressed as a rate describes the amount of energy expended per unit of time (kcal per unit time).[18] Energy expenditure is also expressed in terms of oxygen consumption, where approximately 5 Kcal of energy are spent for consumption of one liter of oxygen.[25] Finally, since energy expenditure is directly related to the amount of body mass moved, energy expenditure is also sometimes expressed relative to the body mass (kcal per kilogram body mass per minute).[19]

Total PA level can be estimated using the resting MET as a baseline. One MET refers to the amount of energy expended at rest or during quiet sitting and is 3.5 milliliters of oxygen per kilogram body weight per minute, and METs are commonly used to express the amount of energy expended during PA. [19]

Methods of assessment of PA

Various methods have been used to quantify PA in COPD both in clinical as well as research settings. These include objective measures such as direct observation, complex laboratory methods including doubly labelled water and indirect calorimetry, pedometers and accelerometers, and subjective measures including PA questionnaires, logs and diaries. All of

these methods quantify activity duration or movements, from which estimates of energy expenditure can be made and distinctions between differing activity levels can be inferred.[26]

Objective Measures of PA

Direct Observation. Direct observation is carried out by trained observers who watch or video record the activities of an individual and then quantify these activities by manually going through the frames and scoring the movement intensity and type of activity. Because of the time consuming, intrusive and demanding nature of this method, it is often not the method of choice for assessment of PA in the adult population.[8]

Doubly labelled water. Doubly labelled water (DLW) and indirect calorimetry are accepted as the criterion standards for measurement of PAEE.[23] The DLW method measures total energy expenditure by assessing the carbon dioxide production over time. This method uses the difference between the elimination rates of stable and non-radioactive oxygen and deuterium isotopes, which, in turn can help determine the carbon dioxide production over a specified period of time.

Indirect Calorimetry. Indirect calorimetry assesses energy expenditure by measuring the amount of oxygen consumed and carbon dioxide exhaled. This method is an indirect assessment of the amount of heat generated by the body via measurement of the amount of substrate use and byproducts production.[19,27] The total average daily energy expenditure in kcal is usually calculated by having the individual breathe through an open circuit system where the exhaled oxygen and carbon di oxide are analyzed.[27]

Pedometers. Pedometers provide an estimation of the number of steps taken by assessing vertical movements at the hip by detecting the amount of force from a heel strike during gait. Pedometers have improved in their accuracy with the utilization of microelectromechanical

systems for step detection. However the validity of pedometers is compromised at slower walking speeds and when used in individuals with gait impairments.[28] Despite the low cost, applicability of pedometers is limited owing to their inability to measure activity intensity in terms of energy expenditure, non-ambulatory activities and posture.[28,29]

Accelerometers. These are small wearable monitors that measure body accelerations in one or more planes (uni-, bi-, tri- or multi-axial). These measured accelerations are converted to lower resolution epoch units or counts and further summarized to an activity output.[29] Accelerometers overcome the limitations of pedometers by providing comprehensive and relatively more precise information on the frequency, duration, and intensity of activity in terms of energy expenditure. Accelerometers also provide information on non-ambulatory activities and body positions.[23]

Despite the advantages of objective measures in assessing PA, accelerometer-based measures have considerable limitations.[29] Although both indirect calorimetry and DLW methods provide accurate measurement of the quantification of the energy spent on PA, the use of these measures in clinical settings is limited due to the exorbitant cost, cumbersome instrumentation and the inability to capture daily activities.[8,23] While pedometers and accelerometers can quantify movement and provide objective estimates of energy expenditure, these wearable activity monitors lack the ability to quantify all the aspects of PA such as the type of activity and patients' experience of an activity.[30] Additionally, interpretation of data from the monitors can be challenging due to the large volume of information obtained, the need for cleaning and summarizing the data and inconsistencies with calibration and validation.[17,31] Inconsistencies in the literature also exist pertaining to the scoring of data and analysis of the large quantity of information produced.[17,31] Although they are a more practical option compared to direct

observation, DLW, and indirect calorimetry, wearable activity monitors have also seen limited utilization in large scale studies and in routine clinical practice owing to their relatively higher cost, subject burden, and the need to be worn over several days to record meaningful information and patient compliance issues when compared to subjective measures of PA.[26,32] Subjective measures of PA may therefore form an integral part of assessment in the COPD population at this time.

Subjective Measures of PA

Subjective measures are often the most feasible methods of PA assessment in COPD. These are frequently used to assess multiple dimensions of PA by reporting type, location, domain and context of the activity. Subjective measures also 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.[33] Their low cost and convenience of use along with their ability to describe PA in different forms make these a practical choice for PA assessment in surveillance systems, in large observational studies as well as in clinical settings.[34] Subjective PA measures include patient-reported outcome (PRO) measures, rater-based and hybrid measures. These measures mentioned above are available as activity logs or diaries as well as structured and semi-structured recall questionnaires.

PRO tools are defined as self-reported outcome tools that come directly from the patient without interpretation of patient's response by the clinician.[35] PRO tools can be administered in several different ways, either by the individual himself (self-administered), via an interviewer (assisted) or via a phone or computer based system (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' response and the PA classification used on the identified measure. Rater-based measures can be administered either via self or in the form of an interview. Hybrid tools have recently been developed in an attempt to better capture the conceptual framework of PA and include a combination of a short patient- reported outcome and two different types of objective assessments in the form of accelerometers. [1,36]

PA logs and diaries. PA logs and diaries are PRO measures that can either be self-administered or computerized.[19] PA logs are checklists that individuals complete at a specified time of the day to list the various activities that were completed during that time by the individual.[29,37] PA diaries include more detailed hour by hour information about different activities and behaviors. The information is expansive and may include details on the different activity domains, type of PA, body positions assumed when performing activities, along with information on the frequency and duration.[19,29] Both diaries and logs can be scored using the "Compendium of Physical Activities" that helps in assigning MET values to specific activities.[38,39]

Physical Activity Questionnaires. Physical activity questionnaires help in the identification of PA dimensions and domains using either a self-administered or assisted (interview) format.[19] All three forms of subjective assessments (PRO, rater based and hybrid measures), can be delivered as physical activity questionnaires. Questionnaires can vary from being very short with only a few items to long, detailed questionnaires based on recall of activities from the one month to a lifetime of recall. PA questionnaires fall under three main categories: global, short-term recall and quantitative history questionnaires.[19] Global questionnaires are very short questionnaires that comprise of one to four items and provide a quick overview of an individual's PA status. Global questionnaires are self-administered PRO

and have been used in clinical and surveillance settings as well as epidemiological studies for their ease of administration, short and concise format and ability to derive a PA score.[19,29] Short-term recall questionnaires are either self-administered or assisted (interview) questionnaires that require individuals to respond to 7 to 20 questions based on the duration, intensity and frequency of specific activities performed over a week or a month.[19,29] Scores on these questionnaires are calculated by multiplying the frequency, intensity and duration of specific PA. Quantitative history recall questionnaires are very long, up to 60 items or more involving recall of various dimensions of different activities performed in the last year or even in a lifetime. The quantitative history recall questionnaires are usually assisted PROs that are administered by an interviewer.[29]

Out of over 130 subjective PA measures currently available for use in different types of population sub-groups, only 15 measures have examined reliability and/or validity in the COPD population. Previous systematic reviews on PA assessments have focused on comparisons between both objective and subjective measures of PA in older adults and COPD[8,9,34] and comparisons between various PRO measures in older adults, not specific to COPD group.[35] However, since the mode and type of delivery of the tool (interviews versus self-report) may affect the overall reliability and validity of the tool,[5] it is important to explore the differences between the types of subjective measures of PA in COPD.

Need for systematic review on subjective PA assessment in COPD

Systematic reviews form the highest level of evidence that can aid in clinical decision making by summarizing large bodies of evidence and explaining the differences among the studies on the same clinical question.[40, 41] Systematic reviews help the reader to draw conclusions on the available evidence by critical analyzing the quality of the studies and providing us with a synthesis of all studies.[40,42] When conducting systematic reviews, a high level of rigor is required in terms of how these assess and rate the available evidence, in order for the information from these reviews to be relevant to the clinician. At this time, a review has not been completed examining the reliability and validity of subjective PA measures in COPD including the PRO, rater based and hybrid measures making it difficult for clinicians to make informed decisions regarding selection of appropriate tools for PA assessment. This lack of evidence identified the need for a systematic review to assess the reliability and validity of various subjective PA assessments in adults with COPD.

Tools for Assessing Methodological Rigor in Systematic Review Research

Several strategies are utilized when conducting systematic reviews that make the investigation structured and systematic, including a comprehensive search in several databases, use of clear and reproducible search terms, explicit selection criteria and the use of quality rating scales for assessment of the quality of the included studies.[43] By adding the level of rigor in the synthesis of systematic reviews, quality rating scales provide the clinicians with tools to keep abreast with the recent literature by explaining the differences among studies on a specific clinical question.[40] One of the most important aspects of conducting a systematic review is the assessment of risk of bias.[44] If the results of individual studies are biased and if these are

synthesized without any consideration of quality, then the results of the review will also be biased and can influence the conclusions.[45] Assessment of quality of individual studies in terms of risk of bias, applicability and to a certain extent, the quality of reporting, is therefore essential to ensure that the quality of evidence synthesized from the systematic reviews are free from bias.[45] A formal assessment of the quality of primary studies included in a review allows investigation of the effect of different biases and sources of variation on study results.

Researchers rely on valid and reliable tools for quality appraisal of the available literature when conducting systematic reviews. Owing to the recognition of randomized controlled trials as the study design of choice for assessing the effectiveness of an intervention, systematic reviews have largely focused on randomized trials[44] and over 25 checklists are currently available to judge the methodological quality of these designs.[41] However, randomized controlled trials are not able to answer some important questions of interest, such as assessment of rare outcomes, long term effects of exposure, or development of assessments, establishing norms for certain measures and assessing measurement properties of tests.[44]

Quality Appraisal Tools for Non-randomized Designs

Non-randomized studies include case control studies, epidemiological cohort studies and cross-sectional studies. Since non-randomized designs are more susceptible to bias as compared to randomized designs, quality appraisal of these studies is of critical importance.[46] At this time, however, limited tools are available to assess the quality of non-randomized designs. Previous quality rating tools for non-randomized designs have been specific to reliability studies (Quality Appraisal tool for Reliability studies – QAREL)[47,48], case control and cohort studies (The Risk of Bias Assessment tool for Non-randomized Studies-RoBANs and Risk of Bias Assessment tool

for Non-randomized Intervention Studies - RoBINs)[44,49] and diagnostic accuracy studies (Quality Assessment of Diagnostic Accuracy Studies -QUADAS and QUADAS-2).[50,51]

Validation studies are types of non-randomized designs that determine if a specific assessment measures what it intends to measure and that assessment scores are statistically related to performance data.[52] Validity is a broad construct and includes different types such as face, content, criterion, construct validity and diagnostic accuracy. Studies examining different types of validity have their own unique design and conduct. Diagnostic accuracy studies assess the ability of a test to detect the presence or absence of a condition by comparing the results of an index test with a reference standard.[50] Diagnostic accuracy studies provide information on the sensitivity, specificity, positive and negative predictive values, likelihood ratios and receiver operating curves of the index test.[50] Content validity studies assess the degree to which the elements within a measure are relevant and representative of the construct of interest and require a subjective process of development of a tool and qualitative and quantitative approaches to validation.[53] Criterion validity studies examine the ability of a test to predict results obtained on an external criterion and involve comparison of an index measure with a gold standard or a reference standard.[52] Criterion-related validity is often categorized into two distinct components: concurrent and predictive validity. Concurrent validity studies involve concurrent measurements of the test to be validated and the criterion measure, in order for both the tests to reflect the same incident of behavior.[52] Studies of construct validity assess the degree to which a test measures the construct it is intended to measure.[54] Known-groups validity is a type of construct validity that determines the ability of a test to discriminate between individuals who are known to have the trait and those that do not.[52] Convergent and discriminant validity studies evaluate the ability

of a test in terms of how its measures relate to other tests of the same construct and tests of different construct.[52]

The QUADAS and QUADAS-2 tools assess the quality of diagnostic accuracy studies. However, considering the different facets of validity, a tool that can assess the quality of all types of validity studies is needed. The consensus-based standards for the selection of health measurement instruments (COSMIN) checklist is another quality appraisal tool of studies on measurement properties.[55] However, feasibility of use of this tool is limited by the length of the tool (12 boxes with 119 items), complexity of administration, item redundancy and inconsistencies in terminology.[56,57] Preliminary work on the development of such a tool has been previously reported. Rennie et al., designed a checklist to address quality of validation studies specific to PA outcome measures.[58] This checklist had six items and was not comprehensive to cover all elements of study quality. Hagstromer et al. used items from this checklist to develop a more comprehensive checklist for PA validation studies.[59] However, since this checklist was never published, information on the administration of the checklist, scoring and interpretation of some of the items remained unclear. Also, since both these studies were specifically designed for PA-related research, these checklists could not be used for validity studies of other outcome measures. At this time, no tool exists to rate the methodological quality of different types of validity studies which identified the need to develop a checklist to assess the methodological quality of validation studies. The lack of research in this area identified the need to first develop a quality appraisal tool and then to systematically review the reliability and validity of subjective PA assessments in COPD using this tool. These individual studies are described in Chapters II and III of this dissertation.

Validity of Measures for PA Surveillance in COPD

PA surveillance

PA has been identified as a major independent, modifiable risk factor for chronic disease including COPD, making population-based survey of PA an important part of global health initiatives.[60,61] Considering the impact of low PA levels towards increased burden of disease, premature death and associated health care costs in COPD, initiatives to monitor population levels of PA using standardized measures have been considered a national health priority and an important component of public health practice.[61,62]

Owing to their relatively low cost, higher clinical utility and low burden on participants, subjective measures are more accessible and widely used in large epidemiological studies and surveillance systems.[63] Surveillance of PA in large population groups is most often undertaken using self-reported questionnaires.[60] Of the various subjective PA measures available for use, the Global PA Questionnaire version 2 (GPAQv2) is one of the most commonly used questionnaires in national surveillance systems and has been recommended by the World Health Organization (WHO) to be used for the surveillance of PA across countries.[64] Although the GPAQv2 has been previously used to assess and report PA in COPD, the validity of this measure in older adults with COPD has not yet been examined.

Global PA Questionnaire (GPAQ). The GPAQ was developed by the WHO in 2002 as part of the WHO STEPwise approach to chronic disease risk-factor surveillance (STEPS).[60] The STEPS approach provides a framework to strengthen and sustain the capacity of countries to perform an ongoing surveillance of chronic disease risk factors.[65] The GPAQ was designed to overcome the limitations of the only other available national PA surveillance measure, the

International Physical Activity Questionnaire (IPAQ).[60,66-68] Thirty one items in the long form of the IPAQ were considered too complex and lengthy to be used in a surveillance tool.[60] The short form was also found to have major limitations in accurately estimating physical activity.[60,68]

Therefore, the original GPAQ version 1 was developed with an intention to compare regional and global differences in PA levels and to inform decisions about PA policy.[60] The original version was comprised of 19 items and underwent revisions based on quantitative analysis of research on GPAQ as well as qualitative feedback from interviewers and interviewees to develop the final, now commonly used GPAQ version 2 (GPAQv2). The GPAQv2 was shorter than the original version after removal of some redundant screening questions. The GPAQv2 is an interviewer assisted questionnaire with 16 questions that are designed to provide an estimate of the PA in three broad domains, including work (6 items), transport (3 items) and leisure time (6 items). The GPAQv2 also has an additional item that looks at time spent in sedentary behavior (1 item).[66,67] The work and leisure domains assess duration and frequency of PA in different intensities such as moderate and vigorous PA, whereas the transport domain assesses the duration and frequency of walking and cycling with no distinction between the activities based on the intensity.[60,66,67]

Scoring of GPAQv2. The GPAQv2 measures PA in three different ways, one of which is total PA, which is a continuous variable measured in MET minutes per week.[60,66,69] The GPAQ guidelines indicate that compared to sitting quietly, a person's caloric consumption is four times as high when being moderately active, and eight times as high when being vigorously active. Therefore, when calculating a person's overall energy expenditure using GPAQ data, activities are classified into intensity levels of vigorous (8 METs), moderate (4 METs) and inactivity (1 MET).[66] The duration and frequency of different intensity activities in each domain (work,

transportation and recreation) is recorded to get an estimate of weekly PA in minutes per week. The amount of activity in minutes per week in then multiplied by the intensity levels of activity (8, 4 or 1) to get the PA score in each intensity category. A summary score for total PA in MET-minutes per week can then be obtained by combining the activity scores for moderate and vigorous intensity activities in each domain.[66] In addition to total PA, the GPAQv2 also allows for assessment of sedentary time measured in minutes per day. Finally, the GPAQv2 also produces a categorical outcome that classifies individuals as inactive or sufficiently active based on the total PA levels.[66] Based on the GPAQ guidelines, individuals are classified as sufficiently active if they demonstrate at least 1) 150 minutes of moderate intensity PA in a week or 2) 75 minutes of vigorous intensity PA in a week or 3) a combination of moderate and vigorous-intensity PA achieving a minimum of 600 MET-minutes per week.[66]

Research on measurement properties

Although this measure has been validated in the general community dwelling, younger (< 60 years) population, [62,67,69-71] most research on this tool has been performed on smaller samples without any reported comorbidities. Previous studies examining the validity and reliability of the GPAQv2 have shown inconsistent findings. Herman et al., found low to moderate validity (r = 0.28 - 0.48) of the GPAQv2 when tested on a group of healthy American adults (mean age 43.1±11.4).[69] Cleland et al., in their study on Irish adults (mean age 44±14), found that although the GPAQv2 showed moderate correlations (r = 0.48) with objective standards of moderate to vigorous-intensity PA, correlations with objective standards for sedentary behavior or inactivity were poor (r = 0.19).[67] Riviere et al., developed the French version of the GPAQv2 where they found moderate to good concurrent validity with the IPAQ (r = 0.41 – 0.86) but poor criterion validity when tested with an objective measure of PA (r = 0.22 – 0.42).[72]

At this time, there is a lack of population based studies examining the validity of the GPAQv2 in older adults with COPD. Considering that the GPAQv2 was designed and has been used as a surveillance tool, both locally and internationally, it is important that the validity of this tool be further explored on a larger population based sample. This lack of evidence identified the need to examine the construct validity of the GPAQv2 in older adults with COPD. Details of this study are available in Chapter IV of this dissertation.

Approach and Methodology

Although the three papers in this dissertation are inter-related and emerge from a common theme, the approach to collection of data and the design of each paper is unique to each individual paper. The individual methodology for each of the three papers is described below.

Study One

This study aimed to design and validate a quality appraisal tool specific to validity studies. The focus of the analysis was to develop a reliable and valid tool that can be used as a method for quality assessment in systematic reviews of non-randomized validity studies.

Procedures

Following identification of possible items for inclusion on the tool, content experts were identified and invited to review the preliminary checklist with 34 items. Content experts rated each item on the tool as ‘essential’, ‘essential, but not necessary’, and ‘not essential’. Content validity ratio was calculated for each item as the criteria for inclusion or exclusion of the item. The tool went through two rounds of review before the final quality appraisal tool for validity studies (QAVALS) was developed. Inter- rater and test-retest reliability of the QAVALS was then examined.

Data Analysis

The content validity of the QAVALS was established using the content validity ratio for individual items and the content validity index for the entire checklist. Weighted kappa coefficients were used to assess the inter-rater and test-retest reliability. SPSS 24.0 and (SPSS Inc., Armonk, NY) was used for the statistical analyses. This study is found in Chapter II.

Study Two

This study aimed to systematically review and compare various available validated subjective measures of PA assessment in COPD. The focus of the analysis was to describe the most reliable and valid subjective PA assessment tools that could be used by clinicians for assessment of activity in the COPD population.

Procedures

The available literature of subjective PA assessments was screened and reviewed by two independent reviewers in the electronic databases of PubMed and CINAHL. Manual searches of back references of relevant articles were also performed to supplement the electronic search. Relevant data on participant demographics, specifics of the PA outcome measures and the measurement properties of the outcomes were extracted.

Data Analysis

The studies included in the review were assessed for their quality of reporting as well as their methodological quality. The quality appraisal tool – QAVALS, developed in Study 1 was used to appraise the quality of the included validity studies. The included studies on reliability were assessed using the quality appraisal tool for reliability studies (QAREL). This study is found in Chapter III.

Study Three

This study examined the construct validity of the Global Physical Activity Questionnaire (GPAQ v2) in a large population based cohort of older adults with COPD. The focus of the analysis was to answer the research question: Is the GPAQv2 a valid tool to assess PA in community dwelling older adults with COPD?

Sample Description

The sample for this study was identified and selected from the publically available secondary data of the National Health and Nutrition Examination Survey (NHANES) collected yearly between the years 2007 and 2012. The NHANES is a cross-sectional, multistage, stratified, clustered probability sample of civilian, non-institutionalized, U.S. populations conducted by the National Center for Health Statistics. The NHANES utilizes standardized interviews and physical examinations in order to collect demographic and clinical information from the participants.[73] All individuals age 65 years and older with COPD were identified from the NHANES. Individuals with COPD were identified as those who responded 'yes' to having either emphysema or chronic bronchitis or both. In addition to patient reported diagnosis, diagnosis of COPD was also established using spirometry data where the ratio of post-bronchodilator forced expiratory volume in the 1st second and forced vital capacity (FEV1/FVC) less than 0.70 was considered as evidence of COPD.[74]

Procedures

Demographic and clinical information of the participants was extracted from the NHANES dataset. The GPAQv2 was used to assess the total weekly PA performed at various

intensity levels of moderate and vigorous intensities and inactivity (MET minutes per week). The percentage of participants who were inactive or sufficiently active was also extracted based on the WHO cutoff of PA less than 600 MET- minutes in a week.[66]

Analysis

Demographic data and clinical characteristics were expressed as means and standard deviations for continuous variables and as percentages for categorical variables. In order to establish the known-groups validity, differences in PA (GPAQv2 scores) between COPD and non-COPD groups were first performed followed by logistic regression models with presence or absence of COPD as the dependent variable and PA (GPAQv2 scores) as independent variables while controlling for potential confounding variables. In the COPD group, multiple regression models were used to examine the independent relationships between GPAQv2 outcomes and PA related constructs of FEV1 and shortness of breath for convergent validity and unrelated construct of household income for discriminant validity, while controlling for known clinical and demographic covariates. SPSS 24.0 and (SPSS Inc., Armonk, NY) was used for the statistical analyses. This study is found in Chapter IV.

Summary

PA has emerged as a non-invasive, cost effective tool to reduce the disease burden, health care utilization and cost of care in individuals with COPD. Use of valid and reliable tools in the COPD population can help in accurately detecting activity and sedentary behavior, both of which have important implications to health outcomes in COPD. Research on measurement properties of PA measures has been undertaken, but the area of subjective PA assessments in COPD is still under-explored. Within this dissertation, this gap in literature was addressed by conducting an expansive review of current evidence on the available subjective measurements of physical activity in COPD and then further examining the validity of a national PA surveillance tool in a population sub-group of older adults with COPD. The quality of evidence produced by the systematic review conducted in this dissertation was further strengthened by using a quality appraisal tool for validity studies (QAVALS), which was developed and validated as part of this dissertation, for the methodological quality assessment of validity studies.

References

1. Dobbels F, de Jong C, st E, et al. The PROactive innovative conceptual framework on physical activity. *European Respiratory Journal.* 2014;44(5):1223-1233.
2. Statistics NCfH. Health, United States 2015 with Special Feature on Racial and Ethnic Health Disparities. *Hyattsville, MD: US Dept Health and Human Services.* 2015.
3. Murray CJL, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet.* 1997;349(9064):1498-1504.
4. Park SK, Richardson CR, Holleman RG, Larson JL. Physical activity in people with COPD, using the National Health and Nutrition Evaluation Survey dataset (2003–2006). *Heart & Lung: The Journal of Acute and Critical Care.* 2013;42(4):235-240.
5. 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.
6. Spruit M, Singh S, Garvey C, et al. An official American thoracic society/European respiratory society statement: Key concepts and advances in pulmonary rehabilitation. *American Journal of Respiratory and Critical Care Medicine.* 2013;188(8):e13-e64.
7. Larson JL, Vos CM, Fernandez D. Interventions to increase physical activity in people with COPD: systematic review. *Annual review of nursing research.* 2013;31:297.
8. Pitta F, Troosters T, Probst VS, Spruit MA, Decramer M, Gosselink R. Quantifying physical activity in daily life with questionnaires and motion sensors in COPD. *European Respiratory Journal.* 2006;27(5):1040.
9. Bossenbroek L, De Greef MHG, Wempe JB, Krijnen WP, Ten Hacken NHT. Daily physical activity in patients with chronic obstructive pulmonary disease: A systematic review. *COPD: Journal of Chronic Obstructive Pulmonary Disease.* 2011;8(4):306-319.
10. Pitta FT, Thierry;Spruit, Martijn A;Probst, Vanessa S;et al. Characteristics of Physical Activities in Daily Life in Chronic Obstructive Pulmonary Disease.pdf. *American Journal of Respiratory and Critical Care Medicine.* 2005;171(9):972 - 977.
11. Waschki B, Spruit MA, Watz H, et al. Physical activity monitoring in COPD: compliance and associations with clinical characteristics in a multicenter study. *Respir Med.* 2012;106(4):522-530.
12. Rabe KF, Hurd S, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary. *American Journal of Respiratory and Critical Care Medicine.* 2007;176(6):532-555.
13. Seidel D, Cheung A, Suh ES, Raste Y, Atakhorrami M, Spruit MA. Physical inactivity and risk of hospitalisation for chronic obstructive pulmonary disease. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease.* 2012;16(8):1015.
14. Waschki B, Kirsten A, Holz O, et al. Physical activity is the strongest predictor of all-cause mortality in patients with COPD: a prospective cohort study. *Chest.* 2011;140(2):331.
15. Garcia-Aymerich J, Farrero E, Félez MA, et al. Risk factors of readmission to hospital for a COPD exacerbation: a prospective study. *Thorax.* 2003;58(2):100-105.
16. Garcia-Rio F, Rojo B, Casitas R, et al. Prognostic value of the objective measurement of daily physical activity in patients with COPD. *Chest.* 2012;142(2):338.

17. Sallis RE, Matuszak JM, Baggish AL, et al. Call to Action on Making Physical Activity Assessment and Prescription a Medical Standard of Care. *Current sports medicine reports* 2016;15(3):207.
18. Caspersen CJ, Powell KE, Christenson GM. Physical Activity, Exercise, and Physical Fitness: Definitions and Distinctions for Health-Related Research. *Public Health Reports (1974-).* 1985;100(2):126-131.
19. Strath SJ, Kaminsky LA, Ainsworth BE, et al. Guide to the Assessment of Physical Activity: Clinical and Research Applications: A Scientific Statement From the American Heart Association. *Circulation.* 2013;128(20):2259-2279.
20. Smith JSC, Collins A, Ferrari R, et al. Our time: a call to save preventable death from cardiovascular disease (heart disease and stroke). *Journal of the American College of Cardiology.* 2012;60(22):2343-2348.
21. Physical Activity Guidelines Report *US Department of Health and Human Services.* 2008.
22. Physical Activity Guidelines Advisory Committee report, 2008. To the Secretary of Health and Human Services. Part A: executive summary. *Nutrition reviews* 2009;67(2):114.
23. Dhillon SS, Sima CA, Kirkham AR, Syed N, Camp PG. Physical Activity Measurement Accuracy in Individuals With Chronic Lung Disease: A Systematic Review With Meta-Analysis of Method Comparison Studies. *Archives of physical medicine and rehabilitation.* 2015;96(11):2079-2088.
24. Schutz Y, Weinsier RL, Hunter GR. Assessment of free-living physical activity in humans: an overview of currently available and proposed new measures. *Obesity research.* 2001;9(6):368-379.
25. McArdle WD, Katch, F. L., and Katch, V. L. Essentials of exercise physiology (4th ed.). In: Katch VL, McArdle, W. D., Katch, F. L., ed. *Energy expenditure during rest and physical activity*. 4 ed. Baltimore, USA: Lippincott Williams & Wilkins; 2011:237 - 262.
26. 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.
27. Haugen HA, Chan L-N, Li F. Indirect calorimetry: a practical guide for clinicians. *Nutrition in clinical practice : official publication of the American Society for Parenteral and Enteral Nutrition.* 2007;22(4):377-388.
28. Harris TJ, Owen CG, Victor CR, Adams R, Ekelund U, Cook DG. A comparison of questionnaire, accelerometer, and pedometer: measures in older people. *Medicine and science in sports and exercise.* 2009;41(7):1392.
29. Ainsworth B, Cahalin L, Buman M, Ross R. The Current State of Physical Activity Assessment Tools. *Progress in Cardiovascular Diseases.* 2015;57(4):387-395.
30. Cavalheri V, Donaria 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.
31. Troiano RP. A Timely Meeting: Objective Measurement of Physical Activity. *Medicine & Science in Sports & Exercise.* 2005;37(11 Suppl):S487-S489.
32. Donaire-Gonzalez D, Gimeno-Santos E, Serra I, et al. Validation of the Yale Physical Activity Survey in chronic obstructive pulmonary disease patients. *Archivos de Bronconeumología ((English Edition)).* 2011;47(11):552.

33. Taylor-Piliae RE, Norton LC, Haskell WL, et al. Validation of a new brief physical activity survey among men and women aged 60-69 years. *American journal of epidemiology.* 2006;164(6):598-606.
34. Helmerhorst HJF, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *The international journal of behavioral nutrition and physical activity.* 2012;9(1):103-103.
35. Williams K, Frei A, Vetsch A, Dobbels F, Puhan MA, Rüdell K. Patient-reported physical activity questionnaires: a systematic review of content and format. *Health and quality of life outcomes.* 2012;10(1):28-28.
36. Gimeno-Santos E, Raste Y, Demeyer H, et al. The PROactive instruments to measure physical activity in patients with chronic obstructive pulmonary disease. *The European respiratory journal* 2015;46(4):988.
37. Moy ML, Matthess K, Stolzmann K, Reilly J, Garshick E. Free-living physical activity in COPD: assessment with accelerometer and activity checklist. *Journal of rehabilitation research and development.* 2009;46(2):277.
38. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. *Medicine and science in sports and exercise.* 2011;43(8):1575.
39. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Medicine and science in sports and exercise.* 2000;32(9 Suppl):S498.
40. Cook DJ, Mulrow CD, Haynes RB. Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions. *Annals of Internal Medicine.* 1997;126(5):376-380.
41. Downs SH, Black N. The Feasibility of Creating a Checklist for the Assessment of the Methodological Quality Both of Randomised and Non-Randomised Studies of Health Care Interventions. *Journal of Epidemiology and Community Health (1979-).* 1998;52(6):377-384.
42. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic Reviews on the Basis of Methodological Criteria. *Physiotherapy.* 1997;83(6):284-289.
43. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the potsdam consultation on meta-analysis. *Journal of Clinical Epidemiology.* 1995;48(1):167-171.
44. Kim SY, Park JE, Lee YJ, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of clinical epidemiology.* 2013;66(4):408-414.
45. (CRD) CfRaD. CRD's guidance for undertaking reviews in healthcare. 2009.
46. Jarde A, Losilla J-M, Vives J, F. Rodrigo M. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology.* 2013;13(2):138-146.
47. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC medical research methodology.* 2013;13(1):111.
48. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology.* 2010;63(8):854-861.

49. Sterne JAC, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ : British Medical Journal (Online)* 2016;355.
50. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC medical research methodology.* 2003;3(1):25-25.
51. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* 2011;155(8):529.
52. Portney LaWS. Foundations of clinical research; applications to practice, 2d ed. Vol 24. Portland: Ringgold Inc; 2000.
53. Haynes SN, Richard DCS, Kubany ES. Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment.* 1995;7(3):238-247.
54. DeVon HA, Block ME, Moyle-Wright P, et al. A Psychometric Toolbox for Testing Validity and Reliability. *Journal of Nursing Scholarship.* 2007;39(2):155-164.
55. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research.* 2010;19(4):539-549.
56. 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 Medical Research Methodology.* 2010;10(1):22-22.
57. Winser SJ, Smith CM, Hale LA, Claydon LS, Whitney SL, Mehta P. COSMIN for quality rating systematic reviews on psychometric properties. *Physical Therapy Reviews.* 2015;20(2):132-134.
58. Rennie KL, Wareham NJ. The validation of physical activity instruments for measuring energy expenditure: problems and pitfalls. *Public Health Nutrition.* 1998;1(4):265-271.
59. Hagstromer M. A checklist for evaluating the validit and suitability of existing physical activity and sedentary behavior instruments. *Measurement of active an sedentary behavors: closing the gaps in self-report methods.* 2010.
60. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). *Journal of Public Health.* 2006;14(2):66-70.
61. Fulton JE, Carlson SA, Ainsworth BE, et al. Strategic Priorities for Physical Activity Surveillance in the United States. *Medicine & Science in Sports & Exercise.* 2016;48(10):2057-2069.
62. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. *Journal of physical activity & health.* 2009;6(6):790.
63. Anne HYC, Sheryl HXN, Koh D, Müller-Riemenschneider F. Reliability and Validity of the Self- and Interviewer-Administered Versions of the Global Physical Activity Questionnaire (GPAQ). *PLoS One* 2015;10(9).
64. World Health Organization. Global Strategy on diet, physical activity and health. 2004.
65. Armstrong T, Bonita R. Capacity building for an integrated noncommunicable disease risk factor surveillance system in developing countries. *Ethnicity & disease.* 2003;13(2 Suppl 2):S13.

66. World Health Organization. Global physical activity questionnaire (GPAQ) analysis guide. *World Health Organization.*Geneva, Switzerland.
67. Cleland CL, Hunter RF, Kee F, Cupples ME, Sallis JF, Tully MA. Validity of the global physical activity questionnaire (GPAQ) in assessing levels and change in moderate-vigorous physical activity and sedentary behaviour. *BMC public health.* 2014;14(1):1255.
68. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. *The international journal of behavioral nutrition and physical activity.* 2011;8(1):115-115.
69. Herrmann SD, Heumann KJ, Der Ananian CA, Ainsworth BE. Validity and reliability of the Global Physical Activity Questionnaire. *Measurement in Physical Education & Exercise Science.* 2013;17(3):221.
70. Au TB, Blizzard L, Schmidt M, Pham LH, Magnussen C, Dwyer T. Reliability and validity of the global physical activity questionnaire in Vietnam. *Journal of physical activity & health.* 2010;7(3):410.
71. Hoos T, Espinoza N, Marshall S, Arredondo EM. Validity of the Global Physical Activity Questionnaire (GPAQ) in Adult Latinas. *Journal of physical activity & health* 2012;9(5):698.
72. Rivière F, Widad FZ, Speyer E, Erpelding M-L, Escalon H, Vuillemin A. Reliability and validity of the French version of the global physical activity questionnaire. *Journal of Sport and Health Science.* 2016.
73. National Health and Nutrition Examination Survey; About the National Health and Nutrition Examination Survey. *Centers for Disease Control and Prevention.*
74. Vestbo J, Hurd SS, Agustí AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine.* 2013;187(4):347.

CHAPTER II

Development of a Quality Appraisal Tool for Validity Studies (QAVALS)

ABSTRACT

Background. Presence of bias in studies that are included in a systematic review can be detrimental to the overall quality of evidence. Quality appraisal tools help to provide a high level of rigor in evaluating included studies, thereby improving the quality of information synthesized from these reviews. Although there are over 25 different checklists currently available for quality appraisal of intervention studies, limited tools are available to assess the quality of non-randomized designs. Currently available tools for quality appraisal demonstrate limited ability to evaluate different aspects of validity of outcome measures. The purpose of this study was to develop a quality appraisal tool for validity studies (QAVALS) and to examine its reliability and validity.

Methods. Following identification of key concepts of the design and objectives of the tool, an initial version of the tool was developed with a list of 34 possible items. Content validity was established by inviting content experts to rate each item on the tool as either 'essential', 'useful but not essential', and 'not necessary'. The content validity ratio (CVR) was calculated and items were deleted or modified based on the CVR and feedback from content experts. The modified version of the tool was sent to content experts a second time for additional review. The content validity index of the tool was calculated to establish content validity of the final checklist. Inter-rater and test-retest reliability were assessed by two external reviewers using weighted kappa coefficients.

Results. Eight out of 10 content experts completed the initial and follow-up reviews on the checklist. Based on the CVR, all items that did not reach the critical value of 0.75 were reviewed. Items that were below a CVR of 0.50 or those where 3 or more experts disagreed were

eliminated from the tool. Following the results of the initial review, 5 items on the tool were deleted, 5 items were reframed or modified using the feedback received from the experts, and 4 items were combined in to 2 items. The preliminary draft of the QAVALS was modified to 27 items following the first review. The revised tool was then sent to the experts a second time. Following the second round, an additional item was deleted and four items were combined into two items. The final tool had 24 items. The content validity index of the final tool was 0.90.

Discussion. QAVALS is the first quality appraisal tool specifically designed to address different types of validity. Items on the tool are rated using one of three categories (yes, no, or other). Summary scores were not used as they have previously been noted to increase the likelihood of producing different quality scores depending on the method of weighting used.

Conclusion. The QAVALS is a 24-item, quality appraisal tool that was designed and validated to be used in systematic reviews of validity studies. The QAVALS demonstrates strong content validity, good overall inter-rater reliability, and excellent test-retest reliability. However, the reliability of individual items was low. Further research is warranted to examine reliability using larger number of studies and raters with different experience levels.

Introduction

Assessment of risk of bias and quality appraisal of studies is one of the most important aspects of conducting a systematic review. Bias in included studies can be detrimental to the overall quality of evidence produced by systematic reviews.[1] Quality appraisal tools help provide a high level of rigor when evaluating the included studies and add more relevance to the information synthesized from these reviews. Use of valid and reliable tools for quality appraisal of the available literature helps in improving the chances of achieving valid results from a systematic review.[2]

Quality of included studies in a systematic review can vary greatly in terms of internal validity, external validity and to some extent, the quality of reporting, making quality appraisal of these individual studies an integral component of systematic reviews.[3] Quality appraisal is even more important for non-randomized designs as these are more susceptible to bias as compared to randomized controlled trials.[2] Quality appraisal tools for intervention studies have been well researched, with over 25 different checklists currently available for assessment of randomized designs.[4] However, limited tools are available at this time to assess the quality of non-randomized designs.[5]

Validity studies are types of non-randomized designs that assess the "extent to which an instrument measures what it is intended to measure".[6] Validity studies examine the ability to make inferences from measurements and are of several types including face, content, construct, criterion validity, and diagnostic accuracy.[6] Studies of specific types of validity have their own unique designs that follow different methods and analyses. Diagnostic accuracy studies "evaluate the ability of one or more medical tests to correctly classify participants as having a target

condition".[7] These studies provide information on the sensitivity, specificity, positive and negative predictive values, likelihood ratios and receiver operating curves of the index test.[3] Studies of face and content validity assess the degree to which the items within a measure are relevant to, and representative of the construct of interest. These studies require a subjective process of development of a tool and the qualitative and quantitative approaches to validation.[8,9] Criterion validity involves comparison of an index measure with a reference standard. Studies of construct validity asses the ability of a test to measure an abstract concept or construct.[6,10] Known-groups validity is a type of construct validity, where the outcomes of the test are compared between groups. Known-groups validity provides evidence of construct validity when a test can differentiate between individuals with a known trait and those without.[6,11]

Currently available validated quality appraisal tools for studies of measurement properties include the quality appraisal of reliability studies (QAREL) and the quality assessment of diagnostic accuracy studies (QADAS and QADAS-2).[3,5,11] The QAREL, as the name suggests is specific to studies of reliability.[5] The QADAS and QADAS-2, although widely used as quality appraisal tools, fail to represent and assess all the aspects of validity.[3,11] The consensus-based standards for the selection of health measurement instruments (COSMIN) checklist is another tool that was developed for quality appraisal of studies on measurement properties.[12] However, limitations in the tool, including the length of the tool (12 boxes with 119 items), complexity of administration, redundancy of some items and inconsistencies in terminology used make it difficult for routine use of this tool.[13,14] Preliminary efforts to design quality assessment tools for validity studies have been reported in literature. Rennie et al.[15] designed a checklist to address quality of validation studies specific to physical activity outcome measures. This checklist of only six items was not sufficiently comprehensive to cover all elements of study quality.[16]

Hagstromer and Bowles[16] used items from Rennie et al.'s checklist to develop a more comprehensive checklist for PA validation studies. However, since this checklist was never published, questions remained unanswered on the administration of the checklist, and scoring and interpretation of the items. Additionally, the lack of systematic development and evaluation of both these tools along with the specificity of these tools to only physical activity outcomes, limit their applicability as comprehensive tools for quality appraisal of validation studies.[17]

Currently, no tool exists that can be specifically used to evaluate quality of validity studies. Therefore, the purpose of this study was to design a valid and reliable quality appraisal tool for validity studies (QAVALS). Creating a tool for quality appraisal of validity studies would aid in improving the synthesis of information in a systematic review of outcome measures' validity.

Methods

The development of the QAVALS followed methods utilized for the design of existing quality appraisal tools which included a four stage process involving preliminary conceptual decisions, item generation, assessment of content validity, and finally the assessment of reliability.[3,5]

Development

Preliminary conceptual decisions. The preliminary conceptual decisions were based on the design used for the development of the QUADAS and QAREL tools.[3,5] For this study, quality was defined as the extent to which the design, methods, and reporting of the study were in line with the objectives of the study and the extent to which the results of the study were applicable to the target population.[2,3,5] The intent of this study was to design a quality appraisal tool for validity studies and based on consensus among the research team, it was agreed that the tool should be able to: 1. Be used in systematic reviews of validity studies, 2. Be a generic tool that could be used to assess quality of any validity study, 3. Be simple to use and easy to understand, 4. Allow for a reliable assessment of quality by different raters, and 5. Allow for an individual assessment of each item rather than a summary score.

Previous quality appraisal tools have used numeric summary scores for ranking quality of included studies.[5,18] However, the use of summary scores has been questioned in the literature due to problems associated with weighting of individual items.[19] Summed quality scores have been shown to differ when items from the same quality appraisal tools were weighted using different methods. Since each item on a quality appraisal tool can individually impact the overall quality of the study, use of summary scores for quality assessment in systematic reviews have

been discouraged.[5,19,20] Based on these observations, it was decided that each item on the QAVALS would be considered separately instead of using an overall quality score.

Item generation. A team of four researchers with experience in research design and methodology, statistics and systematic reviews of measurement properties was formed to provide feedback throughout the process. The principal investigator mediated the feedback between the researchers via online or face to face meetings. An initial list of 34 possible items for inclusion on the QAVALS was drafted by the primary investigator using the conceptual principles following a review of existing quality appraisal tools for both randomized and non-randomized designs. The rating criteria developed by McNeely et al.[21], De Vet et al.[22], and Down and Black[4], and checklists developed by Rennie et al.[15] and Hagstromer and Bowles[16] were reviewed. Additionally, quality appraisal tools including the QUADAS 1 and 2[3,11], QAREL[5], STARD[7], Risk of Bias Assessment tool for Non-randomized designs (RoBANs and RoBINs)[1,23], New Castle Ottawa scale, the NIH Quality assessment tool for observational cohort and cross sectional studies,[24] and quality reporting tools including the STROBE and STARD were reviewed.[7,25]

For each item on the checklist, a detailed list of instructions were developed to aid the rater in the interpretation and scoring of the item and to standardize the rating process.[5] Each item on the tool was designed to be rated on one of the three possible options based on previous rating systems[21,22,24]: 'Yes' = meets the criterion, 'No' = does not meet the criterion, and 'Other' = cannot be determined (CD)/not applicable (NA)/not reported (NR). An item could be rated as CD if the answer to the question could not be determined from the study, as NA if the question was not applicable to the design of the study, and as NR if the information was not reported.

Assessment of content validity

A panel of content experts in the area of research design and methodology, statistics, and systematic reviews of measurement properties, were invited to establish the content validity of the checklist. The number of experts chosen was based on previous recommendations.[26,27] A panel of 5 – 10 experts has been documented to be preferred and more than 10 experts have been rendered unnecessary for the purpose of content validation.[26,27]

Content validation process round 1. Ten experts were invited to be a part of the content evaluation panel. Each of them was provided with the initial version of the QAVALS with 34 items. The content validity ratio (CVR) was used for the purpose of validation. The CVR is internationally recognized and one of the most widely used methods for establishing content validity.[26,28-30] The CVR provides a statistical measure that helps in the rejection or retention of individual items and has been used extensively in previous research.[26-28] Content experts were asked to independently rate each item on the checklist as essential, useful but not essential, or not necessary according to the criteria originally developed by Lawshe.[26,28] A period of three weeks was provided to the experts to complete their ratings, and reminder emails were sent at the end of each week. Individual responses from all the content experts were collected and the number of responses marked as "essential" was identified for each item. For each item, the content experts were also asked to provide reasons for their responses. The CVR for each item was then calculated based on Lawshe's formula as:

$$CVR = \frac{ne - N/2}{N/2}$$

where ne is the number of experts identifying an item as “essential” and N is the total number of experts.[28]

The CVR values range between -1 (perfect disagreement) and + 1 (perfect agreement). CVR values above zero indicate that over half of the panelists agree on an item as “essential”. Lawshe and Schipper’s table of critical values was then used to assess the critical values of CVR ($CVR_{critical}$) in order to eliminate any chance agreements between experts.[28] $CVR_{critical}$ is the lowest level of CVR such that for a given level of significance (alpha or type 1 error probability), the level of agreement exceeds that of chance for a given item.[29] Items were retained on the tool in their original form only if the CVRs of the items were above the critical value listed in the table.[28] All items below the critical CVR value were reviewed and modified. According to Lawshe and Schipper’s table, different values for $CVR_{critical}$ are used for different number of content experts.[28,29] Eight experts were used in this study and so, a $CVR_{critical}$ of 0.75, identified from the critical value table, was used to retain items on the tool.

Content validation process round 2. Based on the responses and feedback from the content experts and following calculation of the CVR, items were either modified or deleted. After modifications on the initial checklist were completed, the revised version of QAVALS was sent out to each of the experts a second time for independent review. Finally, the content validity index (CVI) of the entire checklist was also calculated. The CVI is calculated as the mean of the overall CVRs for all items included in the final instrument.[26] The CVI thus provides a numeric values to the content validity of the total scale. A CVI value greater than 0.80 was considered as evidence of good content validity.[26]

Reliability

Following assessment of content validity, two graduate students (not involved in the initial development of the QAVALS) with experience in research methods and critical appraisal of studies were invited as raters for reliability testing. Raters were first asked to independently rate one study that was not included in the reliability testing as a trial assessment. The raters were provided with instructions on the interpretation of items on the checklist prior to its use. Following rating of this trial study, the raters met with the primary investigator to discuss the criteria for interpretation of each item. After this meeting, each of the two raters was asked to independently rate 10 validity studies. In order to avoid any form of bias, the raters did not discuss these 10 studies during the trial meeting. The raters were blinded from each other's ratings and were not permitted to discuss their ratings during this process.

Since QAVALS was designed as an appraisal tool for systematic reviews that typically evaluate a group of articles of similar content, it was decided to select all ten studies from validity research on one specific research area for testing reliability. A comprehensive search of PubMed and CINAHL was conducted to locate potential papers on the validity of physical activity monitors. A total of 5392 records were screened to include 25 articles that met the inclusion criteria. Only articles that reported validity of accelerometers were included for this purpose. Of the 25 articles identified, 10 were randomly selected for reliability testing of the QAVALS. [4] The raters were provided instructions on the interpretation of items on the checklist prior to its use.

For test-retest reliability, the raters were asked to rate the same set of 10 studies a second time after a period of 2 weeks. The test-retest interval was set at 2 weeks based on previous literature.[31] The raters were advised to destroy their initial ratings after the forms were returned

to the primary investigator following the first review in order to avoid bias. Data from completed forms were then used for analysis of reliability.

Analysis

The content validity for the QAVALS was calculated using the CVR and CVI. Inter-rater and test-retest reliability of the checklist were assessed using weighted kappa coefficients. Inter-rater reliability was examined first by assessing the overall agreement between raters on all items (total number of yes, no or other responses that were common between both raters) as well as agreements between raters on individual items. For reliability, kappa coefficient values of 0.8 or more were interpreted as excellent agreement, 0.6 – 0.8 as good or substantial level of agreement, 0.4-0.6 as moderate, 0.2 – 0.4 as fair agreement, and values below 0.2 as poor agreement.[6,32] Reliability of 0.4 or above was considered as acceptable.[18]

The level of significance was set at 0.05. All analyses were performed in SPSS version 24.0 (SPSS Inc., Armonk, NY).

Results

Validity outcomes

Eight out of the ten experts who initially agreed to take part in the process returned completed checklists. One reviewer dropped out due to lack of time and one reviewer did not return the first review in time.

Based on the $CVR_{critical}$, items were included only if they met the desired critical value of 0.75 for a panel of 8 experts.[28] All items that did not reach the critical value of 0.75 were reviewed. Items that were below a CVR of 0.50 or items where 3 or more experts disagreed on were eliminated from the checklist. Items that had a CVR of less than 0.75 but more than 0.5 were modified based on the feedback from the experts.[4] Following the results of the initial review, 5 items on the checklist were deleted, 5 items were reframed or modified using the feedback received from the experts, and 4 items were combined into 2 items (Appendix II.A). Two items on confounding addressing questions on similar concepts were combined into one item. Similarly, two related items on homogeneity were combined (Appendix II.A). The preliminary draft of the QAVALS was modified to have 27 items following the first review. The items that were removed were:

1) Purpose of the study: Panel members felt that since the QAVALs was intended for quality appraisal of validity studies, asking if the purpose of the study was assessing validity was a redundant item. 2) Original source citations: Panel members provided suggestions to combine this item with other items or exclude it. This item as a stand-alone item did not contribute to the quality of validity studies. 3) Unplanned analysis: Panel members suggested that as this item was not addressing the main effects of the study. It was an item that may be

useful to know, but not essential to determine the quality. 4) Generalizability: Panel members thought that individual factors already discussed in the checklist were sufficient to determine the generalizability. A separate item was useful but not necessary. 5) Clinical application: Panel members thought that although this was important, clinical applicability may not always hold true for a validity study as there are studies that validate lab-based instruments.

The revised checklist with 27 items was then sent to the experts a second time. All eight experts reviewed and returned the checklists with feedback. Following the second round, one item on tests of normality was deleted from the checklist as it had a CVR of 0.25. Based on the feedback received, four additional items were combined to form two items, resulting in the final checklist with 24 items. All other items with values above $CVR_{critical}$ were retained. The details on CVR ratings and items modified are listed in Appendix II.B.

The CVI of the items retained on the checklist was calculated and was found to be 0.90, which indicated good content validity.

Reliability

Test-retest reliability for both raters was found to be excellent (k = 0.84, 95% CI = 0.76 – 0.90 for one rater, and k = 0.80, 95% CI = 0.76 – 0.90 for the other rater). The inter-rater reliability of the overall tool was found to be good (k = 0.70, 95% CI = 0.61 – 0.79). When inter-rater reliability of individual items was calculated, it was found to be low for some items and high for some items. Moderate to excellent agreement was observed for 7 items (0.41 – 0.87), fair agreement for 2 items (0.21 – 0.34), and poor to no agreement for 2 items (-0.11 -0.09). Weighted kappa coefficients for 13 out of the 24 items could not be assessed due to both raters

having the same responses to all items resulting from a lack of variability between studies on the items. Table II.2 reports the inter-rater reliability of individual items.

The QAVALS checklist

QAVALS checklist comprises of 24 questions addressing various aspects quality. Each item on the checklist can be rated as “yes”, “no”, or “other”. The checklist is presented in Table II.1. A detailed description of each item along with criteria for rating is described in Appendix II.C.

Discussion

The QAVALS checklist was developed using an evidence-based systematic approach to assess the quality of validity studies. The final tool has 24 items where each item can be rated as 'yes', 'no', or 'other'. For this study, it was intentionally decided not to use summary scores for the tool. Although an overall quantitative score on quality makes the decision making process easier, different methods of item weighting can influence the scores and result in erroneous results.[19,20] Since each item on a quality appraisal tool has its own importance in determining the quality of the study, it is very difficult to find an objective method to weigh individual items on a scale. The criteria for weighting of items are usually subjective and arbitrary, thereby increasing the likelihood of producing different quality scores when different weighting methods are used. Using summary scores in a systematic review can lead to different conclusions, thereby affecting the overall quality of a systematic review.[19]

For this checklist, it was also decided not to use an overall subjective quality grading (good, fair, poor or low, moderate, high risk of bias) to determine the quality of the studies. Variable responses were received when content experts were asked on their opinion regarding inclusion of an overall subjective grading. Although the majority of experts thought that the use of a subjective quality grading would be useful to the reader and more informative, there was mutual consensus that inclusion of this rating system would include the possibility of ambiguity in the rating. Another problem with inclusion of overall quality grading was the development of criteria to establish overall quality. A straightforward method would be to count the number of 'yes' responses and establish a cut off for 'yes' responses beyond which the study would qualify as a good quality study. However, use of this method defies the very concept of not using summary scores and would automatically weight each item evenly. The other problem with the

use of this method was that it would only consider the 'yes' responses to establish whether a study is of good quality. Since the checklist has several items that could be rated as 'not applicable', a greater number of 'not applicable' items may inadvertently bring the quality of the study down irrespective of strong design. Another way to approach this problem would be to have the rater use his discretion based on his rating of individual items to grade quality. However, since these ratings would then be highly subjective and based on the rater's perspective of quality, the tool would become highly specific to the use of experienced raters or would require a high level of training before use of this tool.

Different cut-offs for CVI values have been reported in literature as criteria for establishing good content validity. Values above 0.70 have been considered as acceptable.[33] Davis et al. in 1992 recommended that a CVI cut off of 0.80 should be considered to establish content validity.[34] The CVI for QAVALS was found to be 0.90 in this study which is considered as evidence of strong content validity.

Overall, the QAVALS tool was found to have excellent test-retest and good inter-rater reliability. Although, low reliability was seen in the inter-rater reliability of few individual items, a majority of items demonstrated moderate to good agreement, with the exception of 4 items that showed fair to poor agreement between raters (Table II.2). Based on these results, the QAVALS is considered a sufficiently reliable tool to assess quality of studies of validity.

The two items that demonstrated poor agreement between raters were items describing the outcomes to be validated and the procedures for testing validity. Possible explanation for the low reliability of these items may be that studies included for the review were specific to physical activity monitors, a content area not familiar to the raters. Future research to examine

reliability of QAVALS using studies on different outcome measures should be performed. Weighted kappa coefficients could not be calculated for 13 items on this tool where both raters gave the same responses across all studies. This was because of a lack of variability between the studies on these response categories.[35] Since weighted kappa coefficients compare the variability between pairs of items to the total variability across studies, low variance between studies may result in large error variance in relation to the study variance and hence, this measurement cannot be performed.[35,36] For example in item number 1: 'was the study design reported?', since study design was reported in all included studies. Hence, there was no variability across the studies on this item, thereby resulting in both raters giving the same response for all 10 studies on this item.

The studies included for reliability testing in this study were identified from a previous systematic review performed. These studies were selected in order to limit the studies to one particular area of interest, making it consistent with the usual systematic review quality appraisal process. Lucas et al.[18] indicated that using studies of similar topic is a preferred method for reliability testing of quality appraisal studies. On the other hand, Hollingworth et al.[37] criticized this method, stating that limiting the studies to only one area of diagnostic technology when testing reliability, may result in low reliability among raters. However, despite the use of studies of similar interest, low reliability was not observed in our study. Since the raters used for reliability testing were graduate students, their experience with quality appraisal of studies was limited. More experienced researchers may demonstrate better reliability. However, despite the limited experience of the raters in this study, it was found that the overall reliability of the tool was good validating our preliminary conceptual decision that the QAVALS may be used as a reliable assessment by raters of different experience.

Limitations

Although this checklist was developed via a systematic evaluation process, limitations exist. First, an overall quality grading for the studies evaluated using QAVALS was not available. The Cochrane collaboration has endorsed the use of a rating system that identifies studies on a 'low' to 'high risk of bias'.[38] An overall subjective rating of quality as 'good, fair, or poor' has also been used in previous literature.[24] However, due to problems with determining the criteria and ambiguity in the subjective rating, it was decided to not use an overall quality grade. Further exploration of ways to incorporate a better overall subjective rating system is warranted.

Second, this checklist was limited in its ability to distinguish between reporting quality and methodological quality. The quality of a study strongly depends not only on the design of methods, but also on the reporting of methods and results. Several items on reporting were included as part of quality assessment in this checklist. Although the final version of the checklist was formed after removing several items from the original pool of items in the two rounds of review, the checklist still had 24 items and was considerably lengthy. Future studies to develop a shorter version of this checklist may be helpful.

Finally, it was found that the raters had difficulty understanding the distinction between 'unclear' and 'cannot be determined' responses. It was noted that several items could be rated as either of the two responses and more clarity on these responses would aid in better rating. Since these responses were part of a single rating category of 'other', the difference in responses within this category did not affect the overall reliability of the tool. However, a shorter version of this tool in future may be developed with use of a single response rather than a compilation of responses in one category.

Conclusion

This study presents a new valid and reliable tool for quality appraisal of validity studies and can be used in the risk of bias assessment of the included studies in systematic reviews of validation studies. Further research on testing its reliability using different raters and studies of different outcome measures is needed.

References

1. Kim SY, Park JE, Lee YJ, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of clinical epidemiology.* 2013;66(4):408-414.
2. Jarde A, Losilla J-M, Vives J, F. Rodrigo M. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology.* 2013;13(2):138-146.
3. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC medical research methodology.* 2003;3(1):25-25.
4. Downs SH, Black N. The Feasibility of Creating a Checklist for the Assessment of the Methodological Quality Both of Randomised and Non-Randomised Studies of Health Care Interventions. *Journal of Epidemiology and Community Health (1979-).* 1998;52(6):377-384.
5. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology.* 2010;63(8):854-861.
6. Portney L, Watkins M. *Foundations of Clinical Research : Applications to Practice.* Vol 3: Pearson Health Science; 2009.
7. Bossuyt PMM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical chemistry.* 2003;49(1):7-18.
8. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health.* 2007;30(4):459-467.
9. Haynes SN, Richard DCS, Kubany ES. Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment.* 1995;7(3):238-247.
10. DeVon HA, Block ME, Moyle-Wright P, et al. A Psychometric Toolbox for Testing Validity and Reliability. *Journal of Nursing Scholarship.* 2007;39(2):155-164.
11. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* 2011;155(8):529.
12. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research.* 2010;19(4):539-549.
13. 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 Medical Research Methodology.* 2010;10(1):22-22.
14. Winser SJ, Smith CM, Hale LA, Claydon LS, Whitney SL, Mehta P. COSMIN for quality rating systematic reviews on psychometric properties. *Physical Therapy Reviews.* 2015;20(2):132-134.
15. Rennie KL, Wareham NJ. The validation of physical activity instruments for measuring energy expenditure: problems and pitfalls. *Public Health Nutrition.* 1998;1(4):265-271.

16. Hagstromer M. A checklist for evaluating the validit and suitability of existing physical activity and sedentary behavior instruments. *Measurement of active an sedentary behavors: closing the gaps in self-report methods.* 2010.
17. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *Journal of Clinical Epidemiology.* 2005;58(1):1-12.
18. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC medical research methodology.* 2013;13(1):111.
19. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC medical research methodology.* 2005;5(1):19-19.
20. Jüni P, Witschi A, Bloch R, Egger M. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. *JAMA.* 1999;282(11):1054-1060.
21. McNeely ML, Olivo SA, Magee DJ. A Systematic Review of the Effectiveness of Physical Therapy Interventions for Temporomandibular Disorders. *Physical Therapy.* 2006;86(5):710.
22. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic Reviews on the Basis of Methodological Criteria. *Physiotherapy.* 1997;83(6):284-289.
23. Sterne JAC, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ : British Medical Journal (Online)* 2016;355.
24. <NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies.pdf>.
25. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Annals of internal medicine* 2007;147(8):W163.
26. Gilbert GE, Prion S. Making Sense of Methods and Measurement: Lawshe's Content Validity Index. *Clinical Simulation in Nursing.* 2016;12(12):530-531.
27. Lynn MR. Determination and Quantification Of Content Validity. *Nursing Research.* 1986;35(6):382-386.
28. Lawshe CH. A Quantitative Approach to Content Validity. *Personnel Psychology.* 1975;28(4):563.
29. Scally AJ, Ayre C. Critical values for lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development.* 2014;47(1):79.
30. Wilson FR, Pan W, Schumsky DA. Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development.* 2012;45(3):197.
31. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology.* 2003;56(8):730-735.
32. Portney LaWS. Foundations of clinical research; applications to practice, 2d ed. Vol 24. Portland: Ringgold Inc; 2000.
33. Tilden VP, Nelson CA, May BA. Use of Qualitative Methods to Enhance Content Validity. *Nursing Research.* 1990;39(3):172-175.

34. Davis LL. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research.* 1992;5(4):194-197.
35. Mokkink LB, Terwee CB, Gibbons E, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC medical research methodology.* 2010;10(1):82-82.
36. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Statistics in Medicine.* 2002;21(14):2109-2129.
37. Hollingworth W, Medina LS, Lenkinski RE, et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Academic radiology.* 2006;13(7):803.
38. Reitsma JB RA, Whiting P, Vlassov VV, Leeflang MM, Deeks JJ. Assessing methodological quality. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 100 The Cochrane Collaboration.* 2009.
39. Rolenz E, Reneker JC. Validity of the 8-Foot Up and Go, Timed Up and Go, and Activities-Specific Balance Confidence Scale in older adults with and without cognitive impairment. *Journal of rehabilitation research and development.* 2016;53(4):511-518.
40. Attia A. Why should researchers report the confidence interval in modern research? *Evidence-Based Medicine Corner.* 2005;10(1):78 - 81.
41. Téllez A, García CH, Corral-Verdugo V. Effect size, confidence intervals and statistical power in psychological research. *Psychology in Russia.* 2015;8(3):27-47.

Table II.1: Quality Assessment of Validity Studies (QAVALS)

Item	Item criteria	Yes	No	Other (CD, NR, NA*)
1	Was the study design reported?			
2	Did the study provide an accurate description of the type of validity tested?			
3	Was the study setting and time frame of participant recruitment clearly described?			
4	Were the criteria for participant selection clearly described?			
5	Were the participants in the study representative of the sample population from which they were recruited?			
6	Did the study clearly describe the outcome measures to be validated?			
7	Did the study provide a clear description of the procedures for testing validity?			
8	Was the testing procedure standardized for all participants?			
9	Was a priori sample size calculation performed to ensure that the study had sufficient power?			
10	Did the study describe and justify any attrition that may have occurred?			
11	Were statistical analyses used to test validity appropriate for the study?			
12	When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?			
13	Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?			
14	Were primary findings of the study clearly described?			
15	Were validity coefficients reported for primary outcomes?			
16	For primary outcomes, did the study report standard deviations or confidence intervals for normally distributed data? If non-normally distributed data, did the study report inter-quartile ranges for the main outcomes?			
Face and Content Validity:				
17	Was the process of selecting expert panel and their qualifications described?			
Criterion validity:				
18	Did the study provide a rationale for the selection of the reference standard?			
19	When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?			
20	When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?			
21	Was the time interval used between administration of reference standard and the test measure appropriate?			
Construct Validity (Known Groups):				
22	Were subjects in different groups homogenous at baseline? If they weren't homogenous at baseline, were differences between groups accounted for during the analysis?			
Construct Validity (Convergent):				
23	Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?			
Construct Validity (Discriminant):				
24	Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?			

*CD = cannot be determined; NA = not applicable; NR = not reported

Table II.2: Inter rater reliability of individual items on the QAVALS

	QAVALS Item	Weighted Kappa	*p* (95% CI)
1	Was the study design reported?	-	
2	Did the study provide an accurate description of the type of validity tested?	0.54	0.53 (0.44 – 1.04)
3	Was the study setting and time frame of participant recruitment clearly outlined and described?	0.44	0.48 (-0.2 – 1.08)
4	Were the criteria for participant selection clearly described?	0.87	0.001 (0.62 – 1.12)
5	Were the participants in the study representative of the sample population from which they were recruited?	-	
6	Did the study clearly describe the outcome measures to be validated?	-0.11	0.72 (-0.26 – 0.42)
7	Did the study provide a clear description of the procedures for testing validity?	0.09	0.87 (-0.47 – 0.65)
8	Was the testing procedure standardized for all participants?	0.34	0.08 (-0.03 – 0.71)
9	Was a priori sample size calculation performed to ensure that the study had sufficient power?	0.60	0.03 (0.14 – 1.05)
10	Did the study describe and justify any attrition that may have occurred?	-	
11	Were the statistical analyses used to test validity appropriate for the study?	-	
12	When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?	0.41	0.10 (-0.18 – 1.00)
13	Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?	-	
14	Were the primary findings of the study clearly described?	-	
15	Were validity coefficients reported for primary outcomes?	-	
16	For primary outcomes, did the study report the standard deviation or confidence intervals for normally distributed data? Or, if non-normally distributed data, did the study report the inter-quartile range for the main outcomes?	-	
17	Was the process of selecting expert panel and their qualifications described?	-	
18	Did the study provide a rationale for the selection of the reference standard?	0.48	0.04 (-0.04 – 1.01)
19	When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?	-	
20	When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?	-	
21	Was the time interval used between administration of reference standard and the test measure appropriate?	0.61	0.03 (-0.04 – 1.27)
22	Were subjects in different groups homogenous at baseline or if they weren't homogenous at baseline, were differences between groups accounted for during the analysis?	-	
23	Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?	0.21	0.49 (-0.43 – 0.85)
24	Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?	-	

CI = confidence interval; '-'= the *k* statistic could not be calculated because both raters had same responses

Appendix II.A: Results of first round of review on the QAVALS

No	ITEM	*R1	*R2	*R3	*R4	*R5	*R6	*R7	*R8	Ne	CVR	KEEP DELETE MODIFY	MODIFIED ITEM
1	Was the study design reported?	1	1	1	1	1	1	1	1	8	1.0	Keep	
2	Was the purpose of the study to establish validity of a tool?	1	1	1	2	1	2	1	2	5	0.25	Delete	
3	Did the study provide an accurate description of the type of validity tested?	1	1	1	1	1	1	2	1	7	0.75	Keep	
4	Was the source and characteristics of the study sample clearly outlined and described?	1	1	2	1	1	2	1	1	6	0.5	Modify	Was the study setting, location and time periods of participant recruitment clearly outlined and described?
5	Were the criteria for participant selection clearly described?	1	1	1	1	1	1	1	1	8	1.0	Keep	
6	Were the participants in the study representative of the population from which they were recruited?	1	1	2	2	1	1	1	1	6	0.5	Modify	Were participants in the study representative of the sample population from which they were recruited?"
7	Did the study clearly describe the outcome measures to be validated?	1	1	1	1	1	1	1	1	8	1.0	Keep	
8	Were original source citations of the outcome measures to be validated provided?	2	1	1	2	1	2	2	1	4	0	Delete	
9	Did the study provide a clear description of the procedures	1	1	1	1	1	1	1	1	8	1.0	Keep	

	for testing validity?												
10	Was the testing procedure standardized and same for all participants?	1	1	1	1	1	1	1	1	8	1.0	Keep	
11	Was the sample size used adequate to ensure that the study had sufficient power?	1	1	0	1	1	1	1	1	7	0.75	Keep	
12	Did the study describe and justify any attrition that may have occurred?	1	1	1	0	1	1	1	1	7	0.75	Keep	
13	Were the data tested for normality?	2	1	1	2	1	1	1	1	6	0.5	Modify	Were tests of normality performed on primary outcome and variables?
14	Were the statistical analyses used to test validity appropriate for the study?	1	1	1	1	1	1	1	1	8	1.0	Keep	
15	When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?	1	1	1	1	1	1	2	1	7	0.75	Keep	
16	Did the study provide clarity on unplanned analyses?	2	2	2	0	1	1	1	1	4	0	Delete	
17	Were potential confounding variables identified and described in the study?	1	1	1	1	1	1	1	1	8	1.0	Keep	
18	Did the study adjust for potential confounding	1	1	1	2	1	2	1	1	6	0.5	Modify	Delete 18 Combined 17 and 18: Did the study identify potential

	variables?												confounding variables and if so, were measures taken to adjust for these confounders? OR Did the study screen and adjust for potential confounders?
19	Were the main findings of the study clearly described?	1	1	1	1	1	1	2	1	7	0.75	Keep	Add 'primary' findings
20	Were validity coefficients reported for primary outcomes?	1	1	1	1	1	1	1	1	8	1.0	Keep	
21	For the main outcomes, did the study report the standard deviation or confidence intervals for normally distributed data? Or, if non-normally distributed data, did the study report the inter-quartile range for the main outcomes?	1	1	1	1	1	1	1	1	8	1.0	Keep	For primary outcomes, did the study report standard deviations or confidence intervals for normally distributed and/or inter-quartile ranges for non-normally distributed data
22	Could the results be generalized from the sample to the target population of interest?	2	1	2	1	1	1	0	2	4	0	Delete	
23	Were the results of the study applicable in a clinical setting?	1	1	0	2	1	2	2	2	3	-0.25	Delete	
24	Was evidence of expertise provided for the panel of experts used for establishing validity?	1	1	0	1	1	1	1	1	7	0.75	Keep - reframe	Was the process of selecting expert panel and their qualifications described?
25	Did the study provide a	1	1	1	2	1	1	1	1	7	0.75	Keep	

	rationale for the selection of the reference standard?												
26	Was the reference standard used valid and reliable?	1	1	1	2	1	1	1	1	7	0.75	Keep	
27	When reference standards were assessed by more than one rater, was the inter-rater reliability between raters established and reported?	1	1	1	1	2	1	2	1	6	0.5	Modify	When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?
28	Was the time interval used between administration of reference standard and the test measure appropriate?	1	1	1	1	1	2	1	1	7	0.75	Keep	Change the word appropriate
29	Were the raters blinded to the outcomes of the tests?	1	1	2	1	1	1	0	1	6	0.5	Modify	Move this up to number 26
30	Were groups tested for homogeneity at baseline?	2	1	1	1	1	1	1	1	7	0.75	Keep	
31	Were baseline differences between groups accounted for during the analysis?	1	1	1	1	1	1	1	1	8	1.0	Keep	
32	Were subjects in different groups recruited from the same population?	2	1	1	1	1	1	0	1	6	0.5	Modify	Move this to item 31 and more 31 down to 32
33	Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?	1	1	1	1	1	1	1	1	8	1.0	Keep	

34	Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?	1	1	1	1	1	1	1	1	8	1.0	Keep	

*R1, R2…R8 = Reviewer 1 through Reviewer 8; Ne = Number of experts rating the item as essential; CVR = content validity ratio calculated as (Ne – N/2)/N/2, where N = total number of experts;

Appendix II.B: Results of the second round of review on the QAVALS

NO	ITEM	R1	R2	R3	R4	R5	R6	R7	R8	Ne	CVR	KEEP DELETE MODIFY	MODIFIED ITEM
1	Was the study design reported?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
2	Did the study provide an accurate description of the type of validity tested?	1	1	2	1	1	1	2	1	6	0.5	MODIFY	
3	Was the study setting, location and time periods of participant recruitment clearly outlined and described?	1	1	1	1	1	1	1	1	8	1.0	KEEP	Reframe as: Was the study setting and time frame of participant recruitment clearly outlined and described?
4	Were the criteria for participant selection clearly described?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
5	Were the participants in the study representative of the sample population from which they were recruited?	1	1	2	1	1	1	1	1	7	0.75	KEEP	
6	Did the study clearly describe the outcome measures to be validated?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
7	Did the study provide a clear description of the procedures for testing validity?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
8	Was the testing procedure standardized for all participants?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
9	Was priori sample size calculation performed to ensure that the study had sufficient power?	1	1	0	1	1	1	1	1	7	0.75	KEEP	
10	Did the study describe and justify any attrition that may have occurred?	1	1	1	2	1	1	1	1	7	0.75	KEEP	
11	Were tests of normality	1	1	1	2	1	2	1	2	5	0.25	DELETE	Delete this question

	performed on primary outcome variables?												
12	Were the statistical analyses used to test validity appropriate for the study?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
13	When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
14	Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?	1	1	1	2	1	1	1	1	7	0.75	KEEP	
15	Were the primary findings of the study clearly described?	1	1	1	1	1	1	2	1	7	0.75	KEEP	
16	Were validity coefficients reported for primary outcomes?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
17	For primary outcomes, did the study report the standard deviation or confidence intervals for normally distributed data? Or, if non-normally distributed data, did the study report the inter-quartile range for the main outcomes?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
18	Was the process of selecting expert panel and their qualifications described?	1	1	2	1	1	1	1	1	7	0.75	KEEP	
19	Did the study provide a rationale for the	1	1	1	2	1	1	1	1	7	0.75	KEEP	Combine 19 and 20 as:

	selection of the reference standard?												Did the study provide a rationale for the selection of the reference standard?
20	Was the reference standard used valid and reliable?	1	1	1	1	1	1	0	1	7	0.75	KEEP	Redundant question : Delete this
21	When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?	1	1	2	1	1	1	1	1	7	0.75	KEEP	
22	When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
23	Was the time interval used between administration of reference standard and the test measure appropriate?	1	1	1	1	1	1	1	1	8	1.0	KEEP	
24	Were subjects in different groups homogenous at baseline and were they recruited from the same population?	1	1	1	1	1	1	1	1	8	1.0	KEEP	Combine 24 -25 as: Were subjects in different groups homogenous at baseline? If not, were the baseline differences between groups accounted for during the analysis?
25	Were baseline differences between groups accounted for during the analysis?	1	1	1	1	1	2	1	2	6	0.5	MODIFY	25 may not be needed – delete and combine with item 24 above
26	Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?	1	1	1	1	1	1	1	1	8	1.0	KEEP	

27	Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?	1	1	1	1	1	1	1	1	8	1.0	KEEP	

R1, R2..R8 = Rater 1 through Rater 8; Ne = Number of experts rating the item as essential; CVR = content validity ratio calculated as (Ne – N/2)/N/2, where N = total number of experts;

Appendix II.C: Explanation of items and instructions for scoring QAVALS

Items 1 and 2

Was the study design reported? Did the study provide an accurate description of the type of validity tested?

Description of the item. Description of details of the elements of design of the study, including a clear description of the type of validity tested, is important for the reader to understand the basic structure of the study. Since variables are not manipulated in a validity study, it is important to highlight these differences in design from other studies by identifying these as 'observational', 'cross sectional', or 'validation' designs. Also, since there are different types of validity that can be tested in different ways, it is important to describe the type of validity that the study was assessing by providing details on how the validity was intended to be established and using what reference standards. For example, when assessing the concurrent validity of a study, it must be clarified that the index test was assessed against an established reference standard.

Rating of the item. Item 1 should be rated as 'yes' when the study identifies the design of the study and clearly mentions the design either as an observational, cross sectional, or methodological design, and rated as 'no' when these elements are missing. Item 2 should be rated as 'yes' when the study either clearly lists the type of validity tested (concurrent, predictive, convergent or discriminant, known groups , content or face validity), or provides a detailed description of these types of validity leading the reader to understand the type tested early in the background or methods of the study. When details on the type of validity are missing, the item should be rated as 'no'. Since both these items are relevant to all studies of validity, these items should always be included for quality assessment and only be rated as 'yes' or 'no' responses.

Item 3

Was the study setting and time frame of participant recruitment clearly outlined and described?

Description of the item. This item is important to help the reader understand the generalizability of the results. Understanding of the time period and settings from where the participants were recruited can help to interpret the findings and help future studies to replicate the findings. Participants can be recruited from different settings including outpatient setting, acute care, rehabilitation settings, community centers, or national databases.

Rating of the item. Rate this item as 'yes' if there is information on the sites of recruitment as well as the time period of recruitment. An example of this would be "community-dwelling women over the age of 65 years with moderate to severe COPD who visited the Biomed respiratory outpatient clinic between July 2011 and Dec 2011 were recruited for the study". If

information on these items is missing, then the item should be rated as 'no'. When insufficient information is given such as only the time frame but no study setting, then the item should be rated as 'other - unclear'.

Item 4

Were the criteria for participant selection clearly described?

Description of the item. This item refers to the inclusion and exclusion criteria for selection of participants in the study. In order to eliminate bias, it is important the selection criteria for recruitment are developed prior to the start of the study and the same selection criteria are used for all participants.

Rating of the item. To rate this item as 'yes', the rater should assess whether the selection criteria (inclusion and exclusion) are listed with sufficient details to understand the study population. If the selection criteria are not provided, then this item should be rated as 'no'. If there is insufficient information available, then this item should be rated as 'other – unclear or cannot be determined'. Since this item is applicable to all validity studies, this item should not be rated as 'not applicable'.

Item 5

Were the participants in the study representative of the sample population from which they were recruited?

Description of the item. Validity of measures can vary considerably based on differences in the sample demographics and clinical characteristics. For example, for assessing known groups validity of a balance measure in diabetes, if the cases are recruited from patients in a hospital setting who have multiple comorbidities, such as depression and heart failure, and the controls are identified from community-dwelling healthy adults, then the differences in the balance measure between groups may result from the differences in sample characteristics between groups rather than due to the ability of the test to exclusively discriminate between diabetics and non-diabetics.

Rating of the item. To rate this item as 'yes', the rater should determine how the sample was identified. It is important to ask whether the sample was a convenience sample or whether they were randomly sampled. If random sampling was not performed, did the entire identified source population get a chance to participate in the study? Patients are representative if they include either the entire source population, an unselected sample, or a random sample.[24] For example, if the source population was older adults with COPD identified from a source population of a pulmonary rehabilitation setting, were all COPD patients attending the program screened for recruitment to provide an equal chance to all patients to be recruited? Along with this, the selection criteria for inclusion in the study should be comprehensive and listed clearly. If the sample was preselected or

convenience sampling was used, then this item should be rated as ‘no’. If insufficient information on the item is given making it difficult to determine how the sample was selected, this item should be marked as ‘other- unclear’.

Item 6

Did the study clearly describe the outcome measures to be validated?

Description of the item. This item refers to the design of the study. Clear description of the index measure provides the reader with a better understanding of the outcome and can help in selection of an appropriate reference standard. For example, for assessing validity of an accelerometer, it is important that a clear description is provided on what the accelerometer measures, and what outcomes it produces in order to select the appropriate reference standard against which it can be tested.

Rating of the item. For a ’yes’ response, the study should identify the outcome measures to be validated and provide a detailed description of the outcomes. The authors should identify which outcome measure or measures they intend to validate in the beginning of the study, provide a detailed description of how these outcome measures can assess the construct of interest, and describe the components of the outcome measure and how it is performed. This item should be rated as ‘no’ when the outcome measures are not described. When some information on the outcome measure is given but not in sufficient details, this item should be rated as ‘other - unclear’.

Items 7 and 8

Did the study provide a clear description of the procedures for testing validity? Was the testing procedure standardized for all participants?

Description of the items. A sufficient explanation on how the index measure and the reference standard was performed and how these were executed on the sample is crucial to ensure that the test can be replicated When sufficient description of the procedures of testing is provided, differences in the performance of the index measure and reference standard may be traced back to the differences in the administration of the tests.[3] Additionally, standardizing the protocol means that all participants would undergo the test in the same way. This is important to eliminate bias in the design that may occur from improper instruction as well as instrument bias.

Rating of the item. Item 7 should be rated as ‘yes’ if the study provides a step by step description of the procedures for testing validity to allow replication of the test. To rate this item, the rater(s) should consider asking whether they a trained researcher would be able to perform all the procedures from the information provided, if they had to conduct this study again. If procedures are not described in details to allow for replication, this tem should be rated as ‘no’. Item 8 should be rated as ‘yes’ if

the study uses a standardized testing protocol in terms of the same testing procedures used for all participants. For example, the time of the day of testing, testing sequence and time intervals between tests should be consistent for all participants. If testing procedures were not similar for all participants, then this item should be rated as 'no'. When there is incomplete information on the testing procedures, the item should be rated as 'other- unclear'. When the rater believes that some standardization was used but it is difficult for the rater to determine this from the text, the item should be rated as 'other-cannot be determined'.

Item 9

Was a priori sample size calculation performed to ensure that the study had sufficient power?

Description of the item. This item is important as it refers to the generalizability of the study. The study must have sufficient power in order for the results to be generalized to the population of interest. When conducting a new study, a priori calculation of sample size is helpful in determining the number of subjects needed to be recruited to achieve sufficient statistical power.

Rating of the item. For a 'yes' response, the study should provide a justification for the sample size chosen by using a-priori sample size and power estimation for detection of a true association, if one existed. The item should also be rated as 'yes' if estimates of effect sizes are provided instead of sample size calculations. However, if neither is mentioned, then this item is rated as 'no'. An example of this may be: "It was estimated that 45 subjects would be needed to detect a correlation of 0.5 and 85% power at alpha of 0.05."

Item 10

Did the study describe and justify any attrition that may have occurred?

Description of the item. This item refers to the withdrawal of participants from the study. When participants enrolled in a study withdraw prior to completion of the study, the results of these participants on their performance on tests will be missing. Identifying participants who completed the study and distinguishing them from those who did not is important. Bias may be introduced in a study if participants who withdraw have characteristics different from those that completed the study.

Rating of the item. For a 'yes' response, the study should provide both the number as well as the specific reasons for any or all participants that dropped out of the study before the results of the study were known. If there were no drop outs, this should be rated as a 'yes'. If the study provides clarity on how many participants were initially recruited and how many withdrew, either as a flow diagram or as descriptions in a paragraph, this item should be rated as 'yes'. If no information on the

drop outs is available, the item should be rated as ‘no’. The item should also be rated as ‘no’ when it is clear from the study that there were some withdrawals but no justification or reasons for the withdrawals is noted.

Items 11and 12

Were the statistical analyses used to test validity appropriate for the study? When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?

Description of the items. In general, all normally distributed parametric data should be tested using parametric tests. Appropriate validity coefficients should be used to answer appropriate research questions. Continuous data should be analyzed using Pearson’s correlation coefficient for establishing concurrent and construct validity. Concurrent validity can also be tested by assessing the difference between the outcomes of the index test and reference standards using analysis of variance (ANOVA) or limits of agreement with Bland Altman plots. Non-normally distributed data can be tested with Spearman’s correlation coefficients. Linear regression models could also be used to assess the strength of relationships between the outcomes measured by the reference standard and the index test, where multiple confounders are identified. Between-group differences for known-groups validity can be established using independent t tests or ANOVA for normally distributed data, and Mann Whitney U test or Friedman’s ANOVA for non-normally distributed data. When multiple comparisons are performed in a study, the chances of type 1 error are inflated. Statistical tests should be performed to adjust for these comparisons. In general, statistical adjustments such as the Bonferroni, Scheffe’s, Tukey’s or Newman-Keul’s are used to control for type 1 error.

Rating of the items. For item 11, a ‘yes’ response should be used when appropriate statistical measures should be used for analyses of the data. When statistical tests are inappropriately used and cannot answer the research question, the item should be rated as ‘no’. Item 12 should be rated as ‘yes’ if the study controlled for the likelihood of a type 1 error in case of multiple comparisons using statistical methods. When there is evidence of multiple comparisons, but no statistical adjustments have been performed, the item should be marked as “no”. When multiple comparisons are not performed, this item should be rated as ‘other- not applicable’.

Item 13

Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?

Description of the item. Identification and adjustment for potential confounding variables is important because the results of the study may be biased if controlling for confounders is not performed. All the key factors that may be associated with the outcome of interest that are not directly related to the research question should be controlled to eliminate bias. For example, findings of a study assessing the relationship between gait speed and balance in community-dwelling older adults may be

confounded by participants' smoking status, body mass index (BMI), and presence of neuropathy. Therefore, these potential confounders must be controlled for in the analyses. Regression methods are often used to account for the influence of confounding variables that are not of interest. Other methods including stratification, matching and multivariable analysis have been used to control for confounding.[25]

Rating of the item. For a 'yes' response, the study should describe potential confounding variables and methods to control for confounding variables. The item should be rated as 'no' when no measures are taken to control for confounding despite the presence of confounding factors. When no potential confounding factors are expected in a study, this item should be rated as 'other- not applicable'. When potential confounding variables are identified but it is unclear what measures were taken to control for confounders, the item should be rated as 'other- unclear'.

Items 14 and 15

Were the primary findings of the study clearly described? Were validity coefficients reported for primary outcomes?

Description of items. Item 14 refers to the main results of the study. Adequate description of the main findings provides clarity to the reader of both the positive and negative outcomes of the study. The results should provide a factual summary of what was found with no interpretation of findings from the authors. This is important for generalizing the findings to the population of interest and for establishing the need for further research. Item 15 refers specifically to validity coefficients. For the main outcomes, the results of validity testing should be described. An example of this would be: "The correlation of the total time in seconds for the 8-foot up and go (8UG) and the TUG demonstrated excellent concurrent validity for those with MCI ($r = 0.92$, $p < 0.001$) and those without MCI ($r = 0.85$, $p < 0.001$)."[39]

Rating of the items. These items should be rated as 'yes' if the study provides an adequate description of the main findings of the study, both quantitatively in a table, graph, and written text. Additionally, the study should report validity coefficients for the main outcomes.

Item 16

For primary outcomes, did the study report the standard deviation or confidence intervals for normally distributed data? Or, if non-normally distributed data, did the study report the inter-quartile range for the main outcomes?

Description of the item. "A confidence interval gives an estimated range of values which is likely to include an unknown population parameter."[40] Confidence intervals, in addition to providing information on the strength of association between variables, also provide the direction and range that helps the reader to determine the probability of the results being due

to chance.[41] Category boundaries also report the range of data within the boundaries, which helps the reader to make informed decisions.

Rating of the item. The item should be rated as 'yes' if the study reports standard deviations or confidence intervals for normally distributed data. For non-normally distributed data, the inter-quartile range or range should be reported. If testing for normality was not reported, the rater is asked to assume that the data was normally distributed when answering this question.[4] As an example, age being a continuous variable should be reported as means and standard deviation. Alternatively, if a non-normal distribution was reported, age should be reported in terms of median and range. Categorical variables such as gender should be reported in percentages. The items should be rated as 'no' when no category boundaries are mentioned for the outcomes.

Item 17

Was the process of selecting expert panel and their qualifications described?

Description of the item. The process of content validation depends on the agreement between content experts on the item's relevance.[9] To ensure that the content validation process is systematic and free from bias, identification of an expert panel that is experienced in testing the domain or content area is important.

Rating of the item. The item should be rated as 'yes' if the study provides a rationale for the selection of the content expert panel. The study should describe the qualifications and experience of the panel, providing evidence their expertise in the area of study.

Item 18

Did the study provide a rationale for the selection of the reference standard?

Description of the item. The reference standard is an important determinant of validity of a test as this serves as the standard against which the index test results are compared. The reference standard therefore must demonstrate high sensitivity and specificity, well established psychometrics, and must be a well-researched outcome for the construct being tested. The reference standard used should be described in enough details to provide the rater a good idea of why this was chosen.

Rating of the item. For a 'yes' response, a sound reasoning based on previous literature should be provided for the selection of the reference standard. If a relatively new outcome measure, which has not been validated thoroughly for its psychometric properties, is used as the reference standard, this item should be rated as "no". If there is insufficient information on the reasons for selection, it should be rated as 'other- unclear or cannot be determined'.

Item 19

When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?

Description of the item. Blinding is important as knowledge of the test results may increase chances of introducing bias in the study. As an example, when testing the validity of a balance outcome measure, the researcher's own bias of getting a significant correlation with the reference standard, may be a potential threat to the overall results. Also, when more than one rater performs a measure, it is important that the two raters are unaware of each other's findings as this may bias the results. The study should standardize the test procedures and blind the raters to the aims of the study (test or measure being validated) so that raters are unaware of the test results.

Rating of the item. For a 'yes' response, raters should be blinded to the results of the tests. When blinding is not performed, the item should be rated as 'no'. When insufficient information is available to accurately make a decision, the item should be rated as 'other- unclear or cannot be determined'.

Item 20

When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?

Description of the item. When more than one rater performs an index test on the participants, the performance of the raters can affect the results, thereby making it important to ensure that the performance between the raters is consistent. Common measures of testing reliability include intra-class correlation coefficients, percent agreement, Kappa statistic, and the agreement value.

Rating of the item. The item should be rated as 'yes' when the inter-rater reliability between the raters is tested and described in cases where more than one rater is used in the study. In case of instrumented measures like posturography or accelerometers, establishing reliability may not be applicable, then this item should be marked as 'other- not applicable' and an explanation should be provided from the selection.

Item 21

Was the time interval used between administration of reference standard and the test measure appropriate?

Description of the item. The time interval between administration of reference standard and the test measure is crucial for validity studies to eliminate any form of bias that may arise from changes in a participant's condition during the time interval between tests. Preferably, both the test results should be collected at the same testing session. However, in case of a delay, a

justification should be provided for the delay to ensure caution was taken to decrease chances of bias in the findings. As an example, when assessing the validity of a balance outcome measure, if the index test is performed at an interval of two weeks following administration of the reference test in a frail older adult, then this may introduce bias. This is due to the participant's frailty condition, which may rapidly change over time, making two weeks too long of a time interval. Since it is possible that the participant may show a decline in balance in two weeks, the time interval chosen should be shorter.

Rating of the item, The item should be rated as 'yes' when the index test and the reference standard are either tested at the same time or appropriate justification is provided for selection of the time interval used. When it is believed that the time interval chosen between the two tests is too long or too short to record the same results, this item should be rated as 'no'. The item should be rated 'other- unclear' if there is insufficient information to come to a conclusion.

Item 22

Were subjects in different groups homogenous at baseline or if they weren't homogenous at baseline, were differences between groups accounted for during the analysis?

Description of the item. Baseline differences between groups in demographic and clinical factors that are not the interest of the research question may act as confounding variables when testing known-groups validity and must be addressed. Baseline comparisons should be performed to identify any differences between groups. Any significant between-group differences should be adjusted during statistical analysis to control for these differences.

Rating of the item. The item should be rated as 'yes' if the baseline demographic and clinical characteristics between known groups of interest are homogenous, or statistical adjustments are performed to control for any differences, if found. The item should be rated as 'no' when baseline comparisons are not performed to assess for differences or when baseline comparisons are performed but no adjustments are performed to control for these differences. If insufficient information is available, then the item should be rated as 'other- unclear or cannot be determined'.

Item 23.

Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?

Description of the item. The study should provide a sound reasoning based on moderate to high correlation coefficients in previous literature for the selection of the measures against which the outcome was tested. As an example, for validating a physical activity outcome, aerobic capacity, lung function and BMI outcomes fall under the same construct (based on previous literature) and may be used to test convergent validity.

Rating of the item. For a ‘yes’ response, the outcome measures used for the purpose of convergent validity should represent a similar construct or characteristic as the outcome measure to be validated. When no justification is provided for the use of a construct as a reference standard, this item should be rated as ‘no’. In case of insufficient information limiting the reader to make an informed decision, the item should be rated as ‘other- unclear or cannot be determined’.

Item 24.

Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?

Description of the item. An outcome measure is shown to have discriminant validity if it shows a weak correlation with the outcome measure(s) against which it is tested. As an example, a physical activity outcome could be shown to have discriminant validity if it shows no correlation with outcomes of unrelated constructs such as balance.

Rating of the item. For a ‘yes’ response, the outcome measures used for the purpose of discriminant validity should represent a totally different construct or characteristic from the outcome measure to be validated. If an item that demonstrates a similar construct as the index measure is used as a reference standard, or if no justification is provided on the selection of the reference standard, this item should be rated as ‘no’. In case of insufficient information limiting the reader to make an informed decision, the item should be rated as ‘other- unclear or cannot be determined’.

CHAPTER III

Subjective Physical Activity Assessments in adults with COPD: A Systematic Review

ABSTRACT

Background. Reduced physical activity (PA) is associated with an increase in severity of COPD and is a poor prognostic indicator in the disease course. Several methods for assessment of PA are currently available including both objective and subjective measures. While objective measures can provide objective estimates of energy expenditure (EE), they lack the ability to quantify all the domains of PA such as the type of activity and patients' experiences of activities. These measures have also seen limited utilization owing to their high cost, subject burden and patient compliance. Subjective measures are therefore an integral part of PA assessment in COPD. At least 130 different types of subjective assessments are currently available to clinicians, providing a wide range of options from which to choose from. Selecting valid and reliable assessments in COPD is crucial to ensure that the information obtained is accurate, valuable, and meaningful.

Purpose. The purpose of this study was to systematically review and report the reliability and validity of various subjective PA measures used in studies of people with COPD.

Data Sources. An electronic database search of Medline and CINAHL was performed with no start date up to April 2017 using MeSH terms and keywords related to PA, COPD, questionnaire and validation.

Study Selection. Observational study designs published in English, evaluating the measurement properties of subjective PA measures in the COPD population.

Data Extraction. Data related to sample demographics, details about the outcome measure assessed, and the psychometric properties including validity, reliability and diagnostic properties were extracted. The Quality Appraisal tool for Reliability studies (QAREL) was used

for assessment of the methodological quality of reliability studies. Methodological quality of validity studies was assessed using the Quality Appraisal tool for Validity Studies (QAVALS).

Results. The search yielded 5164 studies of which 12 studies were included in the final review. Fifteen different measures were described of which 7 were self-administered, 2 were assisted (semi structured or structured interviews), 2 were computerized, 1 was either self or assisted, 1 was rater based and 2 were hybrid measures including a combination of patient report and objective activity monitoring. The Stanford 7-day recall (PAR) demonstrated the strongest correlations with SenseWear Armband on EE ($r = 0.83$; $p < 0.001$) 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 ($r = 0.66 – 0.74$).

Limitations. The present study was limited to only studies published in the English language which may have limited the scope of the search. The studies included in this review had small sample sizes and were of poor methodological quality, limiting the generalizability of the findings.

Conclusions. Assisted and computerized measures (PAR and MARCA) demonstrate better psychometric properties as compared to other subjective measures; and may be considered for quantification of PA. However, observations drawn from single validation studies limit the strength of recommendations and further research is needed in this area to replicate the findings. Newer hybrid tools such as the C-PPAC and D-PPAC demonstrate good construct validity, but need further research to establish accuracy against objective reference standards. Further research is warranted to compare different accelerometer-PRO combinations and to examine their validity across varying COPD severity to further validate these measures.

Key Words: physical activity questionnaire, COPD, validity, reliability

Introduction

Physical activity (PA) was first defined in 1985 as "any bodily movement produced by a contraction of skeletal muscles resulting in energy expenditure".[18] More recently, the American Heart Association elaborated this definition further to include both, structured and incidental PA. Structured PA refers to planned and purposeful activity performed to improve fitness, whereas incidental PA indicates activities that are associated with activities of daily living.[19] PA has been associated with higher levels of health-related fitness and a lower risk of development and progression of disabling medical conditions and various chronic diseases including chronic obstructive pulmonary disease (COPD).[21] The federal guidelines for PA recommend that all individuals should engage in 150 minutes per week of moderate-intensity or 75 minutes per week of vigorous-intensity PA, or a combination of both for substantial health benefits. Additionally, further gains in health can be obtained by increasing PA to 300 minutes per week of moderate-intensity or 150 minutes per week of vigorous-intensity PA, or a combination of both.[19,21] For COPD specifically, the Global Obstructive Lung Disease guidelines recommend regular PA for all patients.[71]

Patients with COPD often experience disabling symptoms such as dyspnea and fatigue, which result in decreased levels of PA as compared to individuals who do not have the disease. Decreased levels of PA are noted even in the early stages of the disease and decline further as the condition progresses.[5,8] 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.[8,9] 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.[5,8,9] Considering the close relationship between PA and health in COPD, assessment and monitoring of PA is necessary.[8]

Several methods for assessment of PA are currently available that include objective measures such as pedometers and accelerometers and subjective measures such as questionnaires, logs or diaries. All these methods quantify activity duration or movement, from which estimates of energy expenditure can be made and distinctions between differing lifestyles can be inferred.[26] While 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' experiences of a given activity.[26] Activity monitors have also seen limited utilization in large scale studies and in routine clinical practice owing to their high cost, subject burden, patient compliance issues, and the need to be worn over several days to record meaningful information.[32,33]

Subjective measures of PA therefore, form an integral part of PA assessment in COPD and 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.[34]

Subjective measures include patient reported outcomes (PRO), rater-based and hybrid measures. PRO measures are defined as self-reported outcomes that come directly from patients without interpretation of patients' responses by clinician.[35] 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 himself (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. More recently, hybrid measures have been developed that combine PRO measures with objective assessments to better reflect PA.[36]

Previous reviews of PA assessments have focused more on comparing objective and subjective measures in older adults. However, there is inadequate evidence on the reliability and validity of various subjective 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 subjective PA measures used in studies of adults with COPD. This review will help clinicians in making well-informed decisions regarding the selection of appropriate subjective measures in patients with COPD.

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 using electronic databases, PubMed and CINAHL until December 1, 2016 and later updated on April 23, 2017 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)). In order 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 subjective measures of assessment of PA in COPD were included in this review. Study designs included all observational study designs evaluating psychometric properties of subjective PA measures in the COPD population. Studies were excluded from the review if reliability or validity data 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 towards 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. No restrictions were applied on the publication date or type of publication (grey literature) in order to decrease chances of publication bias.

Data extraction

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 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 (Appendix III.A). All disagreements

were resolved by mutual consensus. In case of difficulty reaching a consensus, a third reviewer was contacted. The data from the identified articles was independently extracted and rated for quality of reporting and methodological quality by the two reviewers. The inter-rater reliability among reviewers was tested using the weighted kappa coefficient and was found to be excellent (Kappa = 0.91, 95% CI = 0.79 – 1.03).

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. Information on measurement properties, including validity, reliability and diagnostic accuracy were also extracted. Standard data extraction forms created by the study team were used. In the case of missing information on pertinent items (e.g. cost, training requirements, and statistical analysis methods) in the studies selected, attempts were made to contact the authors via email. When no information was received from the authors or when the information was unavailable, these items were entered as 'not available' in the data extraction forms. Outcome measures were categorized as self-reported or PRO measures if they were either self-administered or if the patient's responses were recorded by the interviewer (assisted) or the computer (computerized) without any alteration or interpretation of the stated response. 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. Where subject's responses were interpreted and scored differently by the interviewer, the measure was categorized as rater-based.

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, Bland and Altman plots, and/or 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.[72] 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 form 0.50 to 0.75 as moderate reliability and coefficients above 0.75 were considered to have good reliability.[72]

Quality Assessment

The methodological quality of reliability studies was assessed using the Quality Appraisal tool for Reliability studies (QAREL).[48] 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.[47,48] 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. 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. Each

item on the QAVALS is individually assessed on a categorical scale as a “yes”, “no” or “other” response.

The reporting quality of studies was performed 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.[83] 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 reporting quality of the study.

Results

Initial search yielded 5,156 studies including 3323 studies on PubMed and 1833 on CINAHL (Fig. 1). Additionally, 23 records were identified via hand search. After removal of duplicates, a total of 5166 titles and abstracts were retrieved. Following a review of the abstracts, 5,133 studies were excluded as they did not meet the selection criteria. Thirty-three studies were obtained for full text review. Twenty-one 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 Fig. 1 and Appendix III.B.

Insert figure 1 about here

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 from 13 to 236 (Table III.1).

Of the 12 included studies, 7 studies assessed validity of PA measures,[5,32,97-101] 4 assessed both reliability and validity[26,36,37,102] and 1 study assessed only reliability.[103] Fifteen different subjective measures were described in the studies of which only one was a rater based measure (the Physical Activity Scale for the Elderly (PASE). Seven PA measures were self-administered, 2 were assisted (semi structured or structured interviews), 2 were computerized, 1 was either self or assisted, 1 was rater based and 2 were hybrid measures including patient report and objective activity monitoring (PRO Physical Activity in COPD measures, PPAC) (Table III.2). Description on the type, scoring and other features of the measures can be found in Table III.2.

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) (Table III.2).

The studies included in the review also differed in their methodological quality. The breakdown of scores on the QAREL, QAVALS and STROBE checklists is described in appendices III.C, III.D and III.E. 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.

(Appendix III.D) The selection criteria for participants was identified in 6 studies, attrition in 7, standardization of procedures in 8 and priori sample size calculations in 4 studies. Only one study described adjustments for multiple comparisons.[102] 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 (54.5% of the total sample). Information on the number of raters was provided in one study.[32] However, despite the use of 2 raters, blinding of raters was not performed and inter-rater reliability of raters was not reported.[32] 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. (Appendix III.C)

Seven out of 12 studies scored in the top 25th percentile for reporting quality on the STROBE (Appendix III.E). 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 subjective 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. (Table III.3) 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 (Physical Activity Scale for the Elderly (PASE), Stanford seven-day recall (PAR), Yale physical activity questionnaire (YPAS), modified Baecke and Zutphen Physical Activity Questionnaire (ZPAC).

The modified Baecke questionnaire was the most frequently reported questionnaire in COPD with 3 studies evaluating its validity, followed by the PASE and the ZPAC (2 studies each).[5,98,101-103] The remaining measures were described in single validation studies (Table III.2). Five studies used the SenseWear Pro Arm Band (SAB) [5,26,32,101,102] 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[97,98] (2 studies), ActiReg, Actigraph GT3X+, Dynaport Activity Monitor (DAM) and the Actihealth (Actiped) accelerometer (separate single studies).[26,37,99,100,103] (Table III.3)

Concurrent Validity

Energy Expenditure. The PAR demonstrated the strongest correlations with SenseWear Armband on the measured EE ($r = 0.83$; $p < 0.001$) followed by the YPAS ($r = 0.40$, $p < 0.001$).[5,32] 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, $p < 0.001$).[5,101]

Physical activity level. PA levels measured by the Multimedia Activity Recall (MARCA) demonstrated moderate to good correlations with both SenseWear and Actigraph GT3X+ accelerometers ($r = 0.66 – 0.74$).[26] 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.[5]

Time spent in PA. The PAR was found to demonstrate moderate correlations with the SenseWear arm band for time spent in activity over 3 METs ($r = 0.54$, $p < 0.001$).[5] The YPAS showed fair but significant correlations for time spent in PA ($r = 0.38 – 0.41$, $p < 0.001$).[32] However, the self-administered PRO measure, ZPAC, demonstrated only weak correlations with the reference standard on time spent in activity ($r = 0.18$, $p = 0.25$).[5]

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, $p = 0.496$).[98] The Follick diary showed weak correlations with pedometer counts as did the International Physical Activity Questionnaire short form (IPAQ-SF).[97,98]

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.) (Table III.3). Of the four measures studied, the 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 ($r = 0.55$, $p < 0.05$).[36] The YPAS showed fair correlations with 6MWD ($r = 0.37$- 0.40, $p < 0.01$.[32,37]

Diagnostic accuracy. The 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 and YPAS 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.[5,32,102]

Reliability outcomes

Only five 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.[26,36,97,102,103] Test-retest reliability was reported for all measures, with test-retest intervals ranging from 4 hours to 14 days. Both MARCA and QUANTAP questionnaires demonstrated excellent test-retest reliability values (ICC's = 0.95 – 0.96 and 0.92 respectively).[26,103] Among PRO-PAC 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).[36] Details on reliability outcomes for these questionnaires can be found in Table III.4. 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, demonstrated better validity than other subjective PA assessments. Unsupervised logs such as the activity checklist and Follick's diary, on the other hand, failed to accurately assess PA dimensions.[37,97,99]

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.[5] As compared to other self-reported, unsupervised measures, the close relationship 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.[5] However, these observations were drawn from the findings of a single study[5] 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 adults with COPD. 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).[5] 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 measure was also shown to perform better than the self-administered measures in its relationship with reference standards.[26] 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?").[26] MARCA utilizes a segmented day format, where the entire day is divided into time segments of 5 minutes 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.[104] As with the PAR, the results on the MARCA were drawn from observations of a single study,[26] 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,[26] 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.(Appendices III.C, III.D and III.E) 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 minutes) and between SAB and MARCA (-118.09 – 144.82).[26]

Other self-reported measures including the ZPAC, Baecke and activity dairies failed to demonstrated good concurrent validity in patients with COPD. The PASE, a rater-based questionnaire was found to have poor correlations with overall PA, despite high sensitivity in

identifying severe physical inactivity.[5,102] 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.[102]

Construct Validity. Both the assisted (YPAS) and hybrid (C-PPAC and D-PPAC) measures 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.[105] 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 .[36] Each of these measures examined two factors, amount of PA and difficulty with PA, which were analyzed separately.[36] 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).[36] Findings from the study also showed that the accelerometer-PRO combination yielded better associations than when either was used alone (accelerometer or PRO separately).[36] This is a promising finding that warrants further exploration of different combinations of hybrid measures in future research.

The YPAS demonstrated fair correlations (r = 0.38 – 0.40) with related constructs such as the 6MWD.[32] The YPAS also is an assisted questionnaire where the interviewer uses structured questions to guide the user.[106,107] It is noteworthy that significant correlations were obtained despite the inclusion of light PA items in the YPAS that are typically excluded from

instruments.[106] The face to face nature of the questionnaire and nature of recall (recall from a previous specified week instead of general recall) [8,108] 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.[10,109,110] 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.[111-113] A recent systematic review identified 6MWD as a weak but consistent determinant of PA in COPD.[114] Several other determinants of PA have been identified in literature including dyspnea, quality of life, lung function, systemic inflammation, mortality and exacerbations.[114] 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

The computerized PRO measures as well as the clinical hybrid measure, both 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 III.4). The unsupervised PROs and rater based measures, on the other hand demonstrated weak reliability outcomes.[36] However, 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.[52,115] The difference

in methods used for reliability analysis makes it difficult to draw true comparisons among these measures.

The PRO tools for assessment of PA have been criticized in recent literature on their inability to capture the construct of PA.[1,116] 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.

Limitations

The present systematic review was not without limitations. The review was limited to studies published in the English language and within studies found by searching of 2 large databases. These problems were addressed by being more inclusive in the search via inclusion of all grey literature, performance of manual searches in the reference list of articles and contacting authors for English versions of the studies, if available.

Caution should be used 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 (Appendices III.C, III.D and III.E). In the absence of more high quality validation studies on subjective measures of PA 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.[30,117] 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.[26,97,98] Mobility parameters (e.g. walking speed, use of assistive device, use of supplemental oxygen) and social factors (marital status, social interaction, depression) might have an effect on PA, were not well described in the studies, limiting the ability to draw definite conclusions.[7]

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

Subjective PA assessments remain an active area of research in the COPD population. Assisted and computerized PROs (PAR and MARCA) demonstrate better psychometric properties as compared to other subjective measures; and may be considered for quantification of PA. 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.

References

1. Caspersen CJ, Powell KE, Christenson GM. Physical Activity, Exercise, and Physical Fitness: Definitions and Distinctions for Health-Related Research. *Public Health Reports (1974-).* 1985;100(2):126-131.
2. Strath SJ, Kaminsky LA, Ainsworth BE, et al. Guide to the Assessment of Physical Activity: Clinical and Research Applications: A Scientific Statement From the American Heart Association. *Circulation.* 2013;128(20):2259-2279.
3. Physical Activity Guidelines Report *US Department of Health and Human Services.* 2008.
4. Vestbo J, Hurd SS, Agustí AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine.* 2013;187(4):347.
5. 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.
6. Pitta F, Troosters T, Probst VS, Spruit MA, Decramer M, Gosselink R. Quantifying physical activity in daily life with questionnaires and motion sensors in COPD. *European Respiratory Journal.* 2006;27(5):1040.
7. Bossenbroek L, De Greef MHG, Wempe JB, Krijnen WP, Ten Hacken NHT. Daily physical activity in patients with chronic obstructive pulmonary disease: A systematic review. *COPD: Journal of Chronic Obstructive Pulmonary Disease.* 2011;8(4):306-319.
8. 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.
9. Donaire-Gonzalez D, Gimeno-Santos E, Serra I, et al. Validation of the Yale Physical Activity Survey in chronic obstructive pulmonary disease patients. *Archivos de Bronconeumología ((English Edition)).* 2011;47(11):552.
10. Taylor-Piliae RE, Norton LC, Haskell WL, et al. Validation of a new brief physical activity survey among men and women aged 60-69 years. *American journal of epidemiology.* 2006;164(6):598-606.
11. Helmerhorst HJF, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *The international journal of behavioral nutrition and physical activity.* 2012;9(1):103-103.
12. Williams K, Frei A, Vetsch A, Dobbels F, Puhan MA, Rüdell K. Patient-reported physical activity questionnaires: a systematic review of content and format. *Health and quality of life outcomes.* 2012;10(1):28-28.
13. Gimeno-Santos E, Raste Y, Demeyer H, et al. The PROactive instruments to measure physical activity in patients with chronic obstructive pulmonary disease. *The European respiratory journal* 2015;46(4):988.
14. Portney L, Watkins M. *Foundations of Clinical Research : Applications to Practice.* Vol 3: Pearson Health Science; 2009.
15. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology.* 2010;63(8):854-861.
16. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC medical research methodology.* 2013;13(1):111.

17. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Annals of internal medicine* 2007;147(8):W163.
18. Moore R, Berlowitz D, Denehy L, Jackson B, McDonald CF. Comparison of Pedometer and Activity Diary for Measurement of Physical Activity in Chronic Obstructive Pulmonary Disease. *Journal of Cardiopulmonary Rehabilitation and Prevention.* 2009;29(1):57-61.
19. Nyssen SM, Santos JGd, Barusso MS, Oliveira Junior ADd, Lorenzo VAPD, Jamami M. Levels of physical activity and predictors of mortality in COPD. *Jornal Brasileiro de Pneumologia.* 2013;39(6):659-666.
20. Pitta F, Troosters T, Spruit MA, Decramer M, Gosselink R. 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.
21. Slinde F, Gronberg AM, Svantesson U, Hulthen L, Larsson S. Energy expenditure in chronic obstructive pulmonary disease-evaluation of simple measures. *Eur J Clin Nutr.* 2011;65(12):1309-1313.
22. 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.
23. DePew ZS, Garofoli AC, Novotny PJ, Benzo RP. 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.
24. Moy ML, Matthess K, Stolzmann K, Reilly J, Garshick E. Free-living physical activity in COPD: assessment with accelerometer and activity checklist. *Journal of rehabilitation research and development.* 2009;46(2):277.
25. Gouzi F, Préfaut C, Abdellaoui A, et al. Evidence of an Early Physical Activity Reduction in Chronic Obstructive Pulmonary Disease Patients. *Archives of Physical Medicine and Rehabilitation.* 2011;92(10):1611-1617.e1612.
26. Ridley K, Olds TS, Hill A. The Multimedia Activity Recall for Children and Adolescents (MARCA): development and evaluation. *The international journal of behavioral nutrition and physical activity.* 2006;3(1):10-10.
27. st E, van der Molen T, Kulich K, et al. The PROactive innovative conceptual framework on physical activity. *European Respiratory Journal.* 2014;44(5):1223-1233.
28. Young DR, Jee SH, Appel LJ. A comparison of the Yale Physical Activity Survey with other physical activity measures. *Medicine and science in sports and exercise.* 2001;33(6):955-961.
29. Dipietro L, Caspersen CJ, Ostfeld AM, Nadel ER. A survey for assessing physical activity among older adults. *Medicine and science in sports and exercise.* 1993;25(5):628-642.
30. Bonnefoy M, Normand S, Pachiaudi C, Lacour JR, Laville M, Kostka T. Simultaneous Validation of Ten Physical Activity Questionnaires in Older Men: A Doubly Labeled Water Study. *Journal of the American Geriatrics Society.* 2001;49(1):28-35.
31. Altenburg WA, Bossenbroek L, de Greef MHG, Kerstjens HAM, ten Hacken NHT, Wempe JB. Functional and psychological variables both affect daily physical activity in COPD: A structural equations model. *Respiratory Medicine.* 2013;107(11):1740-1747.
32. 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. *American Journal of Respiratory and Critical Care Medicine.* 2009;180(6):506.
33. Pitta FT, Thierry;Spruit, Martijn A;Probst, Vanessa S;et al. Characteristics of Physical Activities in Daily Life in Chronic Obstructive Pulmonary Disease.pdf. *American Journal of Respiratory and Critical Care Medicine.* 2005;171(9):972 - 977.
34. 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. *Annals of the American Thoracic Society.* 2015;12(3):349-356.
35. Fastenau A, van Schayck OCP, Gosselink R, Aretz KCPM, Muris JWM. Discrepancy between functional exercise capacity and daily physical activity: a cross-sectional study in patients with mild to moderate COPD. *Primary care respiratory journal : journal of the General Practice Airways Group* 2013;22(4):425.
36. 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. *São Paulo medical journal = Revista paulista de medicina.* 2016;134(1):56-62.
37. Gimeno-Santos E, Frei A, Steurer-Stey C, et al. Determinants and outcomes of physical activity in patients with COPD: a systematic review. *Thorax.* 2014;69(8):731-739.
38. Portney LaWS. Foundations of clinical research; applications to practice, 2d ed. Vol 24. Portland: Ringgold Inc; 2000.
39. 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.
40. Dobbels F, de Jong C, st E, et al. The PROactive innovative conceptual framework on physical activity. *European Respiratory Journal.* 2014;44(5):1223-1233.
41. Gimeno-Santos E, Frei A, Dobbels F, et al. Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review. *Health and quality of life outcomes.* 2011;9(1):86-86.
42. Cavalheri V, Donaria 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.
43. 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.
44. Larson JL, Vos CM, Fernandez D. Interventions to increase physical activity in people with COPD: systematic review. *Annual review of nursing research.* 2013;31:297.
45. Marques A, Jácome C, Gonçalves A, et al. Validation of the Comprehensive ICF Core Set for obstructive pulmonary diseases from the patient's perspective. *International journal of rehabilitation research Internationale Zeitschrift für Rehabilitationsforschung Revue internationale de recherches de réadaptation.* 2014;37(2):152.
46. Bennett AV, Amtmann D, Diehr P, Patrick DL. Comparison of 7-day recall and daily diary reports of COPD symptoms and impacts. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2012;15(3):466.
47. Garcia-Aymerich J, Félez MA, Escarrabill J, et al. Physical activity and its determinants in severe chronic obstructive pulmonary disease. *Medicine and science in sports and exercise.* 2004;36(10):1667-1673.

48. Morimoto M, Takai K, Nakajima K, Kagawa K. Development of the COPD Activity rating scale. *Nursing & Health Sciences.* 2003;5(1):23-30.
49. Marín Royo M, Pellicer Císcar C, González Villaescusa C, et al. Physical activity and its relationship with the state of health of stable COPD patients. *Archivos de Bronconeumología ((English Edition)).* 2011;47(7):335.
50. Vilaró J, Gimeno E, Sánchez Férez N, et al. Daily living activity in chronic obstructive pulmonary disease: validation of the Spanish version and comparative analysis of 2 questionnaires. *Medicina clínica.* 2007;129(9):326.
51. Bisca GW, Proença M, Salomão A, Hernandes NA, Pitta F. Minimal Detectable Change of the London Chest Activity of Daily Living Scale in Patients With COPD. *Journal of Cardiopulmonary Rehabilitation and Prevention.* 2014;34(3):213-216.
52. Carpes MF, Mayer AF, Simon KM, Jardim JR, Garrod R. The Brazilian Portuguese version of the London Chest Activity of Daily Living scale for use in patients with chronic obstructive pulmonary disease. *Jornal brasileiro de pneumologia : publicaçẫo oficial da Sociedade Brasileira de Pneumologia e Tisilogia* 2008;34(3):143.
53. Guo AM, Han JN, Kline Leidy N, Wu ZL, Wang P, Lin YX. Validation of the Chinese version of the Functional Performance Inventory Short Form in patients with chronic obstructive pulmonary disease. *Journal of Clinical Nursing.* 2011;20(11-12):1613-1622.
54. Irwin DE, Atwood JCA, Hays RD, et al. Correlation of PROMIS scales and clinical measures among chronic obstructive pulmonary disease patients with and without exacerbations. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.* 2015;24(4):999-1009.
55. Klijn P, Legemaat M, Beelen A, et al. Validity, Reliability, and Responsiveness of the Dutch Version of the London Chest Activity of Daily Living Scale in Patients With Severe COPD. *Medicine.* 2015;94(49):e2191.
56. Kovelis D, Zabatiero J, Oldemberg N, et al. Responsiveness of three instruments to assess self-reported functional status in patients with COPD. *COPD* 2011;8(5):334.
57. Larson JL, Kapella MC, Wirtz S, Covey MK, Berry J. Reliability and validity of the functional performance inventory in patients with moderate to severe chronic obstructive pulmonary disease. *Journal of nursing measurement.* 1998;6(1):55.
58. Leidy NK, Hamilton A, Becker K. Assessing patient report of function: content validity of the Functional Performance Inventory-Short Form (FPI-SF) in patients with chronic obstructive pulmonary disease (COPD). *International journal of chronic obstructive pulmonary disease.* 2012;7:543-554.
59. Partridge MR, Miravitlles M, Ståhl E, Karlsson N, Svensson K, Welte T. Development and validation of the Capacity of Daily Living during the Morning questionnaire and the Global Chest Symptoms Questionnaire in COPD. *The European respiratory journal.* 2010;36(1):96-104.
60. So CT, Man DWK. Development and Validation of an Activities of Daily Living Inventory for the Rehabilitation of Patients with Chronic Obstructive Pulmonary Disease. *OTJR: Occupation, Participation and Health.* 2008;28(4):149-159.
61. Weldam SWM, Lammers J-WJ, de Bruin-Veelers MCC, Schuurmans MJ. The Dutch Functional Performance Inventory: Validity and Reliability in Patients With Chronic Obstructive Lung Disease. *Nursing Research.* 2015;64(1):44-52.

62. Yohannes AM, Greenwood YA, Connolly MJ. Reliability of the Manchester Respiratory Activities of Daily Living Questionnaire as a postal questionnaire. *Age and ageing.* 2002;31(5):355-358.
63. Yoza Y, Ariyoshi K, Honda S, Taniguchi H, Senjyu H. Development of an activity of daily living scale for patients with COPD: the Activity of Daily Living Dyspnoea scale. *Respirology (Carlton, Vic).* 2009;14(3):429-435.
64. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. *Medicine and science in sports and exercise.* 2003;35(8):1381-1395.

Figure 1: PRISMA Flow Diagram

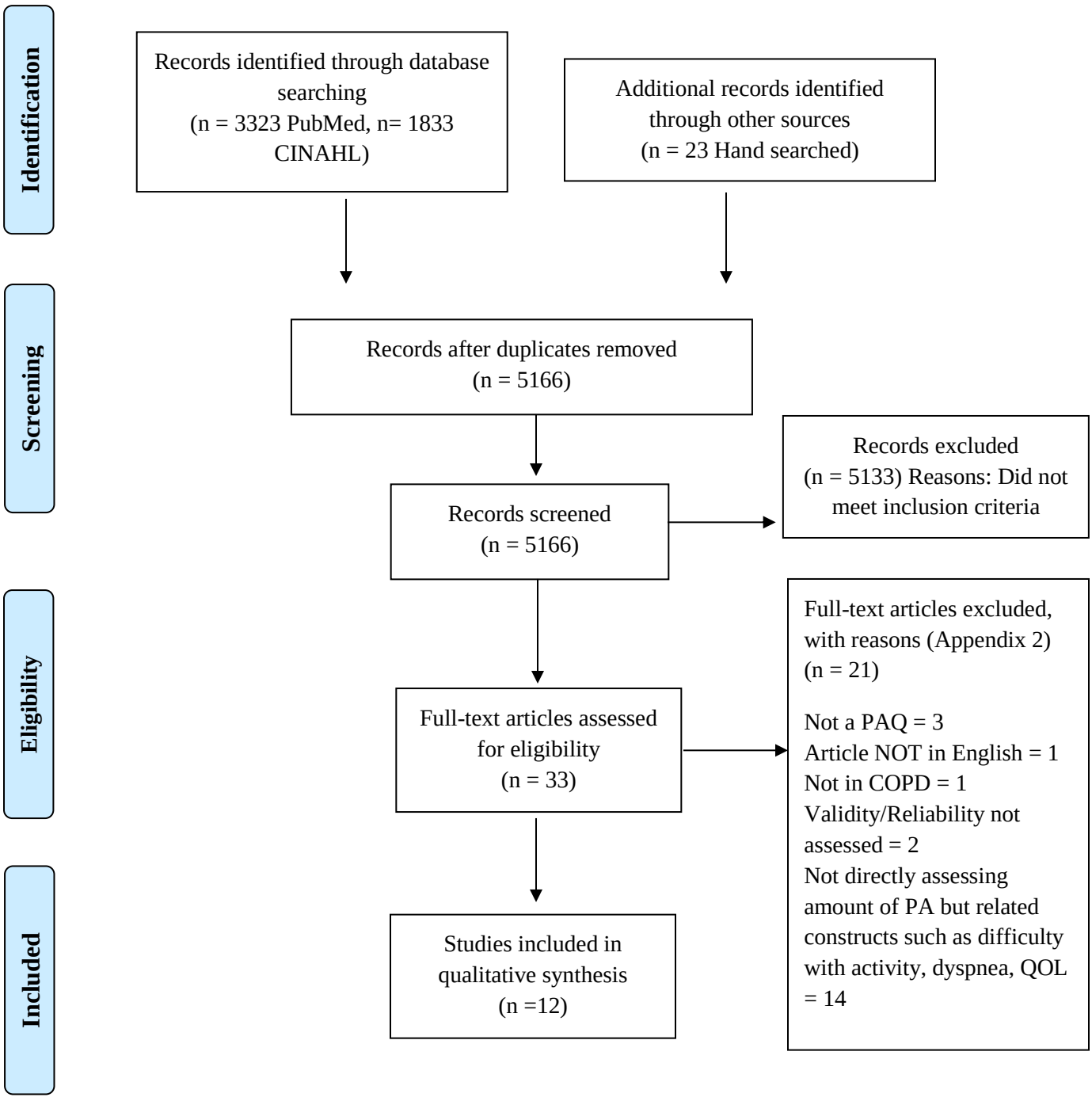
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 3323 PubMed, n= 1833 CINAHL)
Additional records identified through other sources (n = 23 Hand searched)
Records after duplicates removed (n = 5166)
Records screened (n = 5166)
Records excluded (n = 5133) Reasons: Did not meet inclusion criteria
Full-text articles assessed for eligibility (n = 33)
Full-text articles excluded, with reasons (Appendix 2) (n = 21)
Not a PAQ = 3
Article NOT in English = 1
Not in COPD = 1
Validity/Reliability not assessed = 2
Not directly assessing amount of PA but related constructs such as difficulty with activity, dyspnea, QOL = 14
Studies included in qualitative synthesis (n =12)

Table III.1. Characteristics of the study population

Author, Year	Sample size	Gender (M/F)	BMI *M* (SD)	Age (yrs.) *M* (SD)	FEV1 (%) *M* (SD)	FEV1/FVC *M* (SD)	COPD severity
Pitta, 2005	13	10/3	23.00 (5.00)	61.00 (8.00)	33.00 (10.00)	NA	Moderate - severe
Moore, 2009	76	NS	28.20 (6.70)	71.10 (9.60)	46.30 (18.50)	41.70 (13.10)	Moderate - severe
Moy, 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
Donaire-Gonzalez, 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
Gouzi, 2011	129	85/44	24.70 (4.40)	61.30 (9.60)	56.60 (22.80)	NA	Moderate - severe
Slinde, 2011	68	26/42	25.30 (5.50)/ 23.70 (4.40)[ε]	66.00 (6.00)/ 64.00 (7.00) [ε]	41.00 (16.00)/ 44.00 (15.00) [ε]	NA	Severe
DePew, 2012	67	38/29	28.90 (6.86)	71.40 (7.91)	46.70 (19.98)	NA	Moderate - severe
Garfield, 2012	43	18/25	25.90 (5.60)	68.00 (9.00)	46.00 (22.00)	NA	Moderate – very severe
Van Gestel, 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%
Hunt, 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
Nyssen, 2013	30	23/7	24.60 (4.70)	68.00 (10.00)	48.00 (14.90)	NA	Moderate – severe
Gimeno-Santos, 2014	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

M (SD) = Mean (Standard Deviation), BMI = Body mass index, yrs. = years, FEV1 = Forced expiratory volume in 1 second, FVC = forced vital capacity, [ε] = males/females, NA = Not available, V. severe = very severe

Table III.2. Characteristics of subjective physical activity measures

Name	Author (original validation/ COPD validation)	Method of administration	Interpretation	No. of items; Length of recall	Type of PA	Outcomes assessed	Scoring	Classification of PA	Training/ Cost
Physical Activity Scale for the Elderly (PASE)	Washburn, 1993/ DePew, 2012	Interviewer based: Self-report or interview via phone or in person	In case of conflict between patient's interpretation of activity as light or moderate, The PASE category and NOT the participant's choice is selected	12; 1wk	Leisure time activities Household work Occupational activities	Activity Frequency: never, seldom (1-2 days/wk), sometimes (3-4 days/wk) and often (5-7 days/wk). Activity duration: less than 1 hr., 1-2, 2-4 or more than 4 hrs. per day Total hrs. per week for work other than sitting. House work, lawn/yard, gardening, home repair recorded as yes/no	Computed by multiplying time spent in PA per wk or participation (yes/no) by item weights; Sum of the accrued points within each of the three components; Continuous scoring from 0 – 361; Higher scores reflect increased PA	NA	No/ $125.00
Stanford Brief Assessment Scale (SBAS)	Taylor-Piliae 2006, 2010/ DePew, 2012	Self-report (<5 mins)	Respondents select the one pattern each for their on-the-job activity, and Leisure-time activity.	2; Past year	Leisure time activities Occupational activities	Usual amount and intensity of PA throughout the day	Categorical scoring; Scored on 1 – 5 Likert scale from inactive to very hard intensity activity; The SBAS activity category is determined using a color-coded scoring table. The intersection of leisure time and occupational category responses on the color-coded	Inactive, light, moderate intensity (3.0–4.9 METs), hard (5.0–6.9 METs), and very hard (>7.0 METs) activity	No/Free

							scoring table, determines the respondent's current activity pattern		
Yale Physical Activity Survey (YPAS)	DiPietro, 1992/ Donaire-Gonzalez, 2011	Interviewer-administered (20 minutes)	Respondents' answers are documented and interpreted as reported	2 sects: Sect 1 – time spent in 5 categories; Sect 2 categorical questions on frequency; Typical week in the past month	Work, yard work, care taking, exercise, and recreational activities (25 activities total in the 5 categories)	Frequency, duration and intensity of PA performed (a) during a typical week in the past month and (b) in the past month	1. Total time summary index in hrs./wk 2. Energy expenditure summary index = time spent in each activity multiplied by an intensity code ($kcal \cdot min^{-1}$) and then summed across all activities ($kcal \cdot wk^{-1}$). 3. Activity dimensions = multiplying frequency score by duration score for each of 5 activities and then multiplying by a weighting factor scored from 0 - 137	NA	No/Free
Stanford Seven-Day Physical Activity Recall (PAR)/	Sallis, 1984/ Garfield, 2012	Interviewer-administered: Semi structured (30 - 45 mins + 15 mins for review and recoding)	The interviewer guides the participant through the recall process, day by day, to determine duration and intensity of the PA by appropriate prompts	Total 8 questions; 4 questions specific to PA; Past 7 days	Leisure and occupational	Duration (hrs.) of sleep , moderate, hard and very hard PA for weekend and week days	Only PA of moderate intensity and greater are counted; Total EE and Time spent.	Moderate≥ 3 METs, hard (mean 6 METs) and very hard PA (mean 10 METs)	No/Free

			when necessary; Respondents' answers are documented and interpreted as reported						
Zutphen Physical Activity questionnaire (ZPAC)	Casperson, 1991/ Slinde, 2003, Harrsion, 2009, van Gestel 2012, Garfield 2012	Self-report	Respondents' answers are documented and interpreted as reported	17; Past wk, past month or usual activity with no specified time period	Leisure time activities including walking, cycling, gardening, odd jobs, sporting activities, and hobbies	Frequency, duration and EE of PA in Kcals/kg/day	1. Summary MET score expressed in Kcal/Kg/day calculated by multiplying hrs. /day for an activity by an intensity code (Kcal/Kg/hr.), using published MET values (e.g., walking = 3.5 and swimming = 5.0). Total EE for PA is then calculated by summation 2. Total minutes spent in PA	Light < 2 kcal/kg/hr., moderate activity 2-4 kcal/kg/hr. and heavy ≥ 4 kcal/kg/hr.	No/Free
Modified Baecke	Pols MA, 1995; Hertogh EM, 2008/ Gouzi, 2011, Garfield 2012, Nyssen, 2013	Self-report	If answer to a question was not filled, the total index was assumed to be the mean score of remaining answers; If respondent did not fill out any of the work questions, the leisure time index was	19 items under 3 sects.; Past year	Occupational, Sports activity, Non-sports leisure activity	Frequency of PA, hrs./wk and months/yr. of sporting activity, hrs./day of sleep	Items rated on 5 point Likert scale ranging from never to always or very often; PA index = sum of work, sport and leisure index Individual index calculated by multiplying frequency, duration and intensity code.	Sores < 9.4 = low PA	No/Free

			computed twice to get the final PA index						
Multimedia Activity Recall in Children and Adults (MARCA)	Ridley, 2006; Gomersall, 2010/ Hunt, 2013;	Computer delivered self-administered or computer assisted personal/phone interview	Time diary approach where respondents sequentially recall their previous day in time slices of 5 min or more.	3 modules: 1-day activity recall, a compendium of energy costs, and an analytic module; 10 domains with > 500 activities; 24 hrs.	Inactivity, sport/recreation, occupation, self-care, home activities and other	Duration of PA, intensity and EE	Daily activity profiles created using individual physical activity level (PAL) (the time weighted average of MET values over the day); time spent within a given MET range (e.g. time spent in PA ≥3 METs); and time spent lying down, sitting, standing or in locomotion	Sedentary (1–1.9 METs); Light (2–2.9 METs); Moderate (3–5.9 METs); and Vigorous (≥ 6 METs) EE zones	Yes/ free for clinicians on request
Follick's diary	Follick, 1984/ Pitta, 2005; Moore, 2009	Self-report	Respondents' answers are documented and interpreted as reported	Variable; 0.5 hr. blocks over 24 hrs.	Time spent in activities and positions such as sitting, standing, walking, lying and sleeping	Duration of PA	Sum of duration in individual blocks to provide a total duration of PA in 24 hrs. for individual positions/activities	NA	No/Free
Physical activity checklist	Moy, 2009	Self-Report	Respondents' answers are documented and interpreted as reported	17; One day	Performance of activities such as walking (in a mall, grocery store, walking dog), travelling (in car, bus or train; visiting relatives; medical appointment) laundry, exercise, meal	Number of activities	Answers marked as Yes/No with total number of 'Yes' answers summed to reflect total number of activities in a day	NA	No/Free

					preparation, climbing stairs and yard work				
Quantification de l' Activité Physique (QUANTAP)	Vuillemin, 1998; Vuillemin 2000/ Gouzi, 2011	Structured computer assisted interview/ 30 mins	On-screen computer assistance is provided during the interview to aid recall and data collection, and to minimize inaccuracy	4 dimensions; Lifetime (birth to present)	Sports at school, leisure sports, occupation, daily activities	PA type, duration n of PA in hrs, no. of years/months per year/sessions per month in PA, frequency in hrs. per year, intensity in METs, PA indicators in MET-mins per year or METs per year	PA indicator in each dimension (METs-min per year) = multiplying duration of each PA by its frequency; the sum of time for each year is averaged over the period and expressed in hours per year; Hours per year is then multiplied by intensity (METs) and divided by 60.	NA	Yes/ NS
International Physical Activity Questionnaire Short Form (IPAQ-SF)	Craig, 2003/ Nyssen, 2013	Telephone interview/ self-report	Clarifications and probes provided in a standardized manner by the interviewer	9; Last 7 days/ usual week	Time spent per week in different PA including walking, vigorous and moderate intensity activity and in sedentary activity	Time spent in PA, EE in METs	Total time spent in each activity = sum of time per week in each PA; Total weekly PA in MET-minutes per week = duration x frequency per week (minutes per week) x MET EE estimate assigned to each category of activity summed across activity domains	Sedentary, irregularly active, active or very active	No/Free
PROactive Physical Activity in COPD (PPAC)	Gimeno-Santos, 2015	Hybrid tools including self-report plus activity monitoring	Respondents' answers are documented and interprete	9 items divided into 2 main categorical: Amount of PA	Duration of PA in one day; Number of activities in one day	Time spent in PA in a day; Total PA score	Items scored on a 0 to 4 scale and the total score = sum of the scores of the items	Lower scores indicate poor PA	No/Free

tools: Daily PAC (DPAC)			d as reported on the PAQ portion	(2items from PAQ and 2 from accelerometer) and difficulty during PA (5 items; Daily					
PROactive Physical Activity in COPD (PPAC) tools: Clinical PAC (CPAC	Gimeno-Santos, 2015	Hybrid tools including self-report plus activity monitoring	Respondents' answers are documented and interpreted as reported on the PAQ portion	14 items divided into 2 main categories: Amount of PA (2items from PAQ and 2 from accelerometer) and difficulty during PA (10 items); Over 7 days	Duration of weekly PA; Number of activities in the last 7 days	Time spent in PA in a day; Total PA score	Items scored on a 0 to 4 scale and the total score = sum of the scores of the items	Lower scores indicate poor PA	No/Free
Lindholm, Lundgren, and Saltin Questionnaire (Q1)	Saltin and Grimby, 1968/ Slinde, 2011	Self-report	Respondents' answers are documented and interpreted as reported	2; Usual	Occupational /housekeeping work and leisure time activity	Intensity and description of PA from one of the 5 alternatives provided; Category of PA decided on the patient's rating from 0 to 4 including 'not working, no housekeeping' to 'heavy manual work' for the occupational domain and 'sedentary leisure time' to 'regular hard exercise'; EE	Total score = sum of score from the first and second questions with possible results ranging from 1 (not working, no housekeeping and sedentary leisure time) to 8 (heavy manual work and regular hard exercise); TEE = predicted RMR (using WHO predicted values) multiplied by the mean PAL for different activity categories derived from ActiReg	NA	No/Free

Sonn et al Questionnaire (Q2)	Sonn, 1993/ Slinde, 2011	Self-report	Respondents' answers are documented and interpreted as reported	1; Usual	NA	Description and intensity of PA rated from the 6 choices provided ranging from "hardly any PA" to "Hard exercise regularly and several times per week, physical effort is large such as running and skiing"; EE	Category score out of 6 (respondent selects a value from 1 to 6) to describe the PA level; TEE = predicted RMR (using WHO predicted values) multiplied by the mean PAL for different activity categories derived from ActiReg	NA	No/Free

Wk = week, hrs. = hours, yr. = year, METs = Metabolic equivalents, NA = Not available, sects. = sections, PAL = Physical activity level, EE = energy expenditure, kcal·min-^{1} = Kilo calories per minute, kcal·wk-^{1} = Kilo calories per week, PA = physical activity, Kcals/kg/day = kilo calories per kilogram per day.

Table III.3. Validity and diagnostic properties of subjective physical activity measures

Author, Year; Questionnaire	Outcomes measured for validity	Type of validity assessed; Reference Standard	Validity coefficients	Cut off scores/AUC	Sensitivity/specificity (CI)
DePew, 2012; Physical Activity Scale for the Elderly (PASE)	PAL measured as total daily EE divided by resting EE and PASE scores	Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	Regression model with PASE as predictor on PAL: R^2 = 0.38, p = <0.001	For detection of severe physical inactivity: PASE score < 111/ NA	For detection of severe physical inactivity: 0.85/0.62 (NA); p <0.001
Garfield 2012; Physical Activity Scale for the Elderly (PASE)	TEE, time spent in PA≥ 3 METs	Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Spearman's correlation r; p between***: PASE score and PAL derived by SAB = 0.19, p = 0.23	For detecting active patients with COPD achieving at least ≥ 30 min per day PA of ≥ 3 metabolic equivalents: NA/ 0.63 For detecting active patients with COPD with increased survival rates (physical activity level (PAL) ≥ 1.55: NA/0.64 For detecting very inactive patients with COPD (PAL < 1.4): NA/0.60	
Garfield, 2012; Stanford Seven-Day Physical Activity Recall (PAR)	TEE, time spent in PA≥ 3 METs	Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Spearman's correlation r; p between***: TEE by PAR and SAB = 0.83, p < 0.001 Time ≥ 3 METs derived by PAR and SAB = 0.54, p < 0.001 Time ≥ 3 METs derived by PAR and SAB derived PAL = 0.46, p = 0.002 ***ICC for time ≥ 3 METs between PAR and SAB:*** ICC = 0.40 ***Limits of agreement for time ≥ 3 METs***	For detecting active patients with COPD achieving at least ≥ 30 min per day PA of ≥ 3 metabolic equivalents: NA/0.83 For detecting active patients with COPD with increased survival rates (physical activity level (PAL) ≥ 1.55 PAR time: NA/0.77 For detecting very inactive patients with COPD (PAL <	For detecting active patients with COPD achieving at least ≥ 30 min per day PA of ≥ 3 metabolic equivalents: 0.79/0.80 (NA); p = NA For detecting active patients with COPD with increased survival rates (physical activity level (PAL) ≥ 1.55 PAR time: 0.85/0.63(NA); p = NA For detecting very inactive patients with COPD (PAL < 1.4): 0.73/0.76

			between PAR and SAB: -112.9 to 114.7	1.4): NA/0.70	
Donaire – Gonzalez, 2011; Yale physical activity questionnaire (YPAS)	steps/ day time of activity ≥ 1.4MET in hrs./day, EE ≥ 1.4 MET, kcal/Day and 6MWD	Concurrent and Construct; Sense Wear Pro 2 Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Spearman's correlation r; p between***: YPAS MET and SAB steps/ day; time spent in PA ≥ 1.4 METs in hrs./day; EE ≥ 1.4 METs in kcal/Day and 6MWD = 0.38; 0.41; 0.40 and 0.37; $p < 0.001$ YPAS ***Spearman's correlation r; p between***: YPAS summary index of PA score (0 -137) and SAB steps/ day ; time spent in PA ≥ 1.4 METs in hrs./day; EE ≥ 1.4 METs in kcal/Day and 6MWD = 0.52; 0.38; 0.43 and 0.40; $p < 0.001$ ***ICC (CI) for Time spent in PA, intensity (METs) and EE between YPAS and SAB:*** ICC 0.397, 0.360 and 0.339 (NA)	For detection of sedentary patients (<30 minutes PA per day): 51/0.71 (95% CI: 0.63–0.79)	For detection of sedentary patients (<30 minutes PA per day): 0.75/0.59 (NA); p = NA
Hunt, 2013; MARCA	PAL (METs), MVPA (min)	Concurrent; New Lifestyles 1000 pedometers, New Lifestyles, Inc., Lee's Summit, Missouri, USA (NL-1000), Actigraph GT3X+ accelerometers, Actigraph, Pensacola, Florida USA and Sensewear Pro3H Armbands, Body Media, Pittsburgh USA (SAB).	***Spearman's correlation r; p; ICC, LOA between***: MARCA PAL and Actigraph GT3X counts = 0.74; p = NA; NA; ICC = NA, NA MARCA PAL and SAB PAL = 0.66; p = NA; NA; ICC = 0.33, -0.17 to +0.51	NA	NA

			MARCA MVPA and Actigraph MVPA = 0.68; p = NA; ICC = 0.30, -29 to +142 MARCA MVPA and SAB MVPA = 0.47; p = NA; ICC = 0.42, -102 to +132		
Garfield, 2012; Modified Baecke	TEE, time spent in PA≥ 3 METs	Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Spearman's correlation r; p between***: Modified Baecke score and PAL derived by SAB = 0.15, p = 0.35	For detecting active patients with COPD achieving at least ≥ 30 min per day PA of ≥ 3 metabolic equivalents: NA/0.64 For detecting active patients with COPD with increased survival rates (PAL) ≥ 1.55: NA/0.64 For detecting very inactive patients with COPD (PAL < 1.4): NA/0.55	
Nyssen, 2013; Modified Baecke	Steps per day, Modified Baecke total score	Concurrent; Yamax Digi-Walker SW-700 pedometer, Yamax, Tokyo, Japan	***Spearman's correlation r; p between***: Pedometer steps/day and modified Baecke total score = -0.129; p = 0.496	NA	NA
Nyssen, 2013; International IPAQ -SF	Steps per day, IPAQ-SF MET-min/wk of TPA	Concurrent; Yamax Digi-Walker SW-700 pedometer, Yamax, Tokyo, Japan	***Spearman's correlation r; p between***: Pedometer steps/day and IPAQ-SF MET-min/wk = 0.399; p = 0.029	NS	NS
Garfield, 2012; Zutphen Physical Activity Questionnaire (ZPAC)	TEE, time spent in PA≥ 2 METs	Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Spearman's correlation r; p between***: TEE by ZPAC and SAB = 0.01, p = 0.94 Time spent in ≥ 2	For detecting active patients with COPD achieving at least ≥ 30 min per day PA of ≥ 2 metabolic equivalents:	

			METs derived by ZPAC and SAB = 0.18, p = 0.25 ***ICC for time ≥ 2 METs between ZPAC and SAB:*** ICC = 0.37	NA/0.57 For detecting active patients with COPD with increased survival rates (physical activity level (PAL) ≥ 1.55: NA/0.66 For detecting very inactive patients with COPD (PAL < 1.4): NA/0.49	
Van Gestel, 2012; Zutphen Physical Activity Questionnaire (ZPAC)		Concurrent; Sense Wear Pro Arm Band (SAB), BodyMedia Inc., Pittsburgh, PA, USA	***Pearson's correlation r, p between***: TEE by ZPAC and steps/day by SAB = 0.50 (CI 0.30 – 0.66), p < 0.001 TEE by ZPAC and FEV1 = 0.31 (CI 0.28 – 0.44), p < 0.001 ***Mean difference (CI), p in TEE (Kcal/day) between:*** SAB and ZPAQ = 922 (703 – 1141), p < 0.001 ***Regression model with TEE by ZPAC as predictor on PA (steps/day***): R^2 = 0.479, p < 0.001	For detecting very inactive patients with COPD (PAL < 1.4: NA/0.43	NA
Moore, 2009; Follick Diary	Stand /walk time Sleep time Sit plus stand/walk time	Concurrent; Yamax Digiwalker SW – 700 pedometer, Yamax Corporation, Tokyo, Japan	***Pearson's correlation r; p between***: Diary stand/walk time and pedometer count = 0.374, p = 0.001 Diary sit plus stand/walk time and pedometer count = 0.236, p	NA	NA

			= 0.04 Diary sleep time and pedometer count = 0.001, *p* = 0.99 Diary outings time and pedometer count = 0.359, *p* = 0.001		
Pitta, 2005; Follick Diary	Time spent in walking, cycling, standing, sitting and lying over 12 hr. period	Concurrent; DynaPort Activity Monitor (DAM)	***Mean (SD) of time spent in activity recorded via DAM and diary using paired t test:*** Walk time = 45 (20) & 66 (47)* Cycling time = 4 (9) & 3 (9) Standing time = 191 (85) & 146 (71)* Sitting time = 390 (161) & 375 (134) Lying time = 88 (141) & 83 (83)	NA	NA
Moy, 2009; Physical activity checklist	Number of checklist activities, steps per day, FEV1% and BODE index	Concurrent and Construct; Actiped (Actihealth), FitSense Technology, Inc., Southborough, MA	***Unadjusted coefficients (95% CI) between:*** Number of checklist activities and steps per day on Actiped = 174 (7 – 341), *p* = 0.04 Number of checklist activities and FEV1% predicted = 0.023 (0.005 – 0.041), *p* = 0.01 Number of checklist activities and BODE index = -0.34 (-0.58 - -0.10), *p* = 0.008	NA	NA
Gimeno-Santos, 2014; D-PPAC	Total amount of activity (PAQ + DAM); Total amount of activity	Construct and Known-groups; CRQ dyspnea, CRQ fatigue, CCQ total, 6 MWD, mMRC and CAT	***Spearman's correlation r; p between***: PA amount with DAM and CRQ dyspnea = 0.36;	NA	NA

	(PAQ +Actigraph)		p <0.05 PA amount with DAM and CRQ fatigue = 0.18; p <0.05 PA amount with DAM and CCQ total = -0.20; p <0.05 PA amount with DAM and mMRC = -0.42; p <0.05 PA amount with DAM and 6MWD = 0.55; p <0.05 PA amount with DAM and CAT dyspnea = -0.23; p <0.05 ***Spearman's correlation r; p between***: PA amount with actigraph and CRQ dyspnea = 0.34; p <0.05 PA amount with actigraph and CRQ fatigue = 0.17; p <0.05 PA amount with actigraph and CCQ total = -0.19; p <0.05 PA amount with actigraph and mMRC = -0.41; p <0.05 PA amount with actigraph and 6MWD = 0.55; p <0.05 PA amount with actigraph and CAT dyspnea = -0.20; p <0.05		
Gimeno-Santos, 2014; C-PPAC	Total amount of activity (PAQ + DAM); Total amount of activity (PAQ +Actigraph)	Construct and Known-groups; CRQ dyspnea, CRQ fatigue, CCQ total, 6 MWD, mMRC and CAT	***Spearman's correlation r; p between***: PA amount with DAM and CRQ dyspnea = 0.36; p <0.05 PA amount with DAM and CRQ fatigue = 0.37; p <0.05 PA amount with DAM and CCQ	NA	NA

			total = -0.37; *p* <0.05 PA amount with DAM and mMRC = 0.53; *p* <0.05 PA amount with DAM and 6MWD = 0.62; *p* <0.05 PA amount with DAM and CAT dyspnea = -0.38; *p* <0.05 ***Spearman's correlation r; p between***: PA amount with actigraph and CRQ dyspnea = 0.35; *p* <0.05 PA amount with actigraph and CRQ fatigue = 0.36; *p* <0.05 PA amount with actigraph and CCQ total = -0.34; *p* <0.05 PA amount with actigraph and mMRC = 0.55; *p* <0.05 PA amount with actigraph and 6MWD = 0.65; *p* <0.05 PA amount with actigraph and CAT dyspnea = -0.37; *p* <0.05		
Slinde, 2011; Saltin Q1 and Sonn Q2	TEE	Concurrent; ActiReg, PreMed AS, Oslo, Norway	***Mean (SD); p values for calculated EE in Kj:*** Actireg = 8317 (2255) and Q1 = 8272 (1412); *p* = 0.76 Actireg = 8317 (2255) and Q2 = 8324 (1484); *p* = 0.87 ***Bland Altman limits of agreement between:*** Mean TEE from ActiReg and Q1 = > ± 2 S.D	NA	NA

			from the difference of means Mean TEE from ActiReg and Q2 = > ± 2 S.D from difference of means		

CI = confidence interval, R^2 = variance, NA = not available, TEE = total energy expenditure, MET = metabolic equivalent, PA = physical activity, PAL = physical activity level derived by total energy expenditure per minute divided by resting energy expenditure per minute, NT = note tested, ICC = intra class correlation coefficient, 6MWD = six minute walk distance, AUC = area under curve, MVPA = moderate to vigorous physical activity, min = minutes, LOA = limits of agreement, TPA = Total Physical activity, FEV1 = forced expiratory volume in 1 second , kcals/day = kilo calories per day, SD = standard deviation, BODE = body mass index, airway obstruction, dyspnea, exercise capacity index, CRQ = chronic respiratory questionnaire, CCQ = clinical---- mMRC = modified medical research council rating

Table III.4. Reliability of various subjective physical activity measures

Author, Year; Questionnaire	Type of reliability	Outcomes measured for reliability	Test retest interval	Reliability measures
DePew[102], 2013; Physical Activity Scale for the Elderly (PASE)	Test retest	PASE scores	5 – 7 days	Pearson's r = 0.75 BA plot
DePew[102], 2013; Stanford Brief Assessment Scale (SBAS)	Test retest	SBAS scores	5 – 7 days	Kappa coefficient K = 0.41
Hunt[26], 2013; Multimedia Activity Recall in Children and Adults (MARCA)	Test retest	MVPA ≥ 3 METs (min/day), TDEE (MET.min)	4 hours	***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
Moy[37], 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[103], 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)
Gimeno-Santos[36], 2014; 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
Gimeno-Santos[36], 2014 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

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

Appendix III.A: Screening criteria for inclusion of articles in the review

RELEVANCE CRITERIA	YES	NO	Criteria for inclusion
1. Does this study describe a subjective PA assessment?			Y
2. Does this study evaluate the reliability and/or validity of the subjective PA assessment?			Y
3. Do the participants in the study have COPD?			Y
4. Does the PA assessment described in the study assess one or more of the following? a. Duration, b. Frequency c. Type of PA d. Estimated energy expenditure			Y
4. Does the PA assessment described in the study assess: a. Symptoms associated with PA b. Functional performance or limitations c. Experiences with PA (shortness of breath, pain with PA) d. Quality of Life			N
5. Does the study include the following type/s? a. Systematic review b. Meta-analyses c. Narrative review d. Letter to the editor			N
REVIEWER DECISION:			
1. Include in final review for quality appraisal			
2. Additional references to be screened If yes, mark on the reference list of the study			
IN CASE OF DISCREPANCY			
Reason for discrepancy:			
1. oversight			
2. Differences in the interpretation of criteria			
3. Differences in the interpretation of study			
FINAL DECISION:INCLUDE IN REVIEW			

Appendix III.B: Reasons for exclusion of full text articles after review

No.	Author, year	Title	Reason for exclusion
1	Marques A et al[118], 2014	Validation of the Comprehensive ICF Core Set for obstructive pulmonary diseases from the patient's perspective	Not a PAQ; Assesses the health experience of patients living with specific health conditions
2	Benett et al[119], 2012	Comparison of 7-Day Recall and Daily Diary Reports of COPD Symptoms and Impacts	Dyspnea questionnaires; Amount of PA not separately analyzed
3	Dobbels et al[1], 2014	The PROactive innovative conceptual framework	Not a PAQ
4	Garcia-Ameyrich et al[120], 2004	Physical Activity and Its Determinants in Severe Chronic Obstructive Pulmonary Disease	Psychometric properties of PAQ not tested in the study
5	Morimoto et al[121], 2003	Development of the COPD Activities Rating Scale	Assesses difficulty with performing activities; Amount of PA not separately analyzed
6	Royo et al[122], 2011	Physical Activity and its Relationship With the State of Health of Stable COPD Patients	Psychometric properties of PAQ not tested in the study
7	Vilaro et al[123], J	Daily living activity in chronic obstructive pulmonary disease: validation of the Spanish version and comparative analysis of 2 questionnaires	Article not in English –translation n English not available
8	Bisca et al[124], 2014	Minimal Detectable Change of the London Chest Activity of Daily Living Scale in Patients With COPD	Difficulty with ADL; Amount of PA not separately analyzed
9	Carpes et al[125], 2008	The Brazilian Portuguese version of the London Chest Activity of Daily Living scale for use in patients with chronic obstructive pulmonary disease	Difficulty with ADL; Amount of PA not separately analyzed
10	Guo et al[126], 2011	Validation of the Chinese version of Functional Performance Inventory-Short Form (FPI-SF) in patients with chronic obstructive pulmonary disease (COPD	Amount of PA not separately analyzed
11	Irwin et al[127], 2015	Correlation of PROMIS scales and clinical measures among chronic obstructive pulmonary disease patients with and without exacerbations	Not a PAQ
12	Klijn et al[128], 2015	Validity, Reliability, and Responsiveness of the Dutch Version of the London Chest Activity of Daily Living Scale in Patients With Severe COPD	Symptoms with ADL; Amount of PA not separately analyzed
13	Kovelis et al[129], 2011	Responsiveness of three instruments to assess self-reported functional status in patients with COPD	Functional mobility assessed. Amount of PA not separately analyzed
14	Larson JL et al[130], 1998	Reliability and validity of the functional performance inventory in patients with moderate to severe chronic obstructive pulmonary disease	Functional mobility assessed. Amount of PA not separately analyzed
15	Leidy et al[131], 2012	Assessing patient report of function: content validity of the Functional Performance Inventory-Short Form (FPI-SF) in patients with chronic obstructive pulmonary disease (COPD)	Functional mobility assessed. Amount of PA not separately analyzed
16	Patridge et al[132], 2010	Development and validation of the Capacity of Daily Living during the Morning questionnaire and the Global Chest Symptoms Questionnaire in COPD	Symptoms with ADL; Amount of PA not separately analyzed
17	Tao So et al[133], 2008	Development and validation of activities of daily living inventory	ADL scale; Amount of PA not separately analyzed
18	Weldam et al[134],	The Dutch Functional Performance Inventory in COPD:	Amount of PA not separately

	2015	validity and reliability	analyzed
19	Yohannes et al[135], 2002	Reliability of the Manchester respiratory activities of daily living questionnaire as a postal questionnaire	ADL scale; Amount of PA not separately analyzed
20	Yoza et al[136], 2009	Development of an activity of daily living scale for patients with COPD: The Activity of Daily Living Dyspnoea scale	ADL scale; Amount of PA not separately analyzed
21	Craig et al[137], 2003	International physical activity questionnaire: 12-country reliability and validity	Not in COPD patients

Appendix III.C: Methodological quality assessment of reliability studies (QAREL)

Quality Appraisal of Diagnostic Reliability Checklist (QAREL)						
Item	**Description**	**DePew, 2012**	**Hunt, 2013**	**Moy, 2009**	**Gouzi, 2011**	**Gimeno-Santos, 2014**
1	Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied	YES	NO	YES	YES	YES
2	Was the test perfumed by raters who were representative of those to whom the authors intended the results to be applied?	YES	YES	YES	YES	YES
3	Were raters blinded to the findings of other raters during the study	NA	NA	NA	NA	NA
4	Were raters blinded to their own prior findings of the test under evaluation	NA	NA	NA	NA	NA
5	Were raters blinded to the results of the reference standard for the target disorder being evaluated	NA	NA	NA	NA	NA
6	Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design	NA	NA	NA	NA	NA
7	Were raters blinded to additional cues that were not part of the test	UNCLEAR	UNCLEAR	UNCLEAR	UNCLEAR	UNCLEAR
8	Was the order of examination varied	NO	NO	NO	NO	YES
9	Was the time interval between repeated measurements compatible with the stability of the variable being measured	YES	YES	YES	YES	YES
10	Was the test applied correctly and interpreted appropriately	YES	YES	YES	YES	YES
11	Were appropriate statistical measures of agreement used	NO	YES	YES	NO	YES

Appendix III.D: Methodological quality assessment of validity studies

		Quality Appraisal tool of Validity Studies – QAVALS											
No	**Description**	**Study Number***											**Total**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
1	Study design	1	1	0	1	0	1	0	1	0	1	0	6
2	Type of validity	0	1	0	1	0	0	0	0	1	0	0	1
3	Study setting and time frame	0	0	0	0	0	0	1	1	0	0	1	3
4	Participant selection	0	1	0	1	0	0	0	1	1	1	1	6
5	Representative of the sample population	1	1	0	1	0	0	0	0	0	0	1	4
6	Describe outcomes to be validated	1	1	1	1	1	1	1	1	1	1	1	11
7	Procedures for testing	0	1	1	1	1	1	1	1	1	1	0	9
8	Standardization of procedures	0	1	0	1	1	1	1	1	1	1	0	8
9	Priori sample size	0	1	1	1	1	0	0	0	0	0	0	4
10	Attrition	0	1	1	0	1	1	0	1	1	0	1	7
11	Statistical analyses	1	1	1	1	1	1	1	1	1	1	1	11
12	Multiple comparison adjustment	NA	0	NA	NA	NA	0	0	0	0	0	0	0
13	Confounding variables	0	0	0	0	0	0	1	1	0	0	1	3
14	Primary findings of the study	1	1	1	1	1	1	1	1	1	1	1	11
15	Validity coefficients reported	1	1	1	1	1	1	1	1	1	1	1	11
16	Confidence intervals/ range	0	0	0	0	0	0	1	0	1	0	1	3
FACE AND CONTENT VALIDITY													
17	Expert Panel	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
CRITERION VALIDITY													
18	Rationale for reference standard	0	1	1	NA	1	0	1	0	1	1	0	6
19	Blinding of raters	NR	0	NR	NA	NR	NR	NA	NR	NA	NR	NR	
20	Inter rater reliability between raters	NR	0	NR	NA	NR	NR	NA	NR	NA	NR	NR	
21	Time interval between ref standard and index test	CD	1	CD	NA	NR	1	1	1	1	1	0	6

CONSTRUCT VALIDITY (KNOWN GROUPS)													
22	Homogeneity between groups	NA	NA	NA	0	NA	NA	NA	0	NA	NA	NA	
CONSTRUCT VALIDITY (CONVERGENT)													
23	Measures represent similar construct	NA	NA	NA	1	NA	NA	NA	1	NA	NA	1	
CONSTRUCT VALIDITY (DISCRIMINANT)													
24	Measures represent different construct	NA	NA	NA	1	NA	NA	NA	1	NA	NA	NA	

*Study 1 = De Pew, 2012; study 2 = Donaire Gonzalez, 2011; study 3 = Garfield 2012; study 4 = Gimeno-Santos 2014; study 5 = Hunt2013; study 6 = Moore, 2009; study 7 = Moy, 2009; study 8 = Nyssen, 2013; study 9 = Pitta, 2005; Slinde, 2011; Van Gestel, 2012

Appendix III.E: Assessment of reporting quality of the studies using STROBE

S. No.	Item	Study Number*												STROBE Total
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
1	Title and abstract	1	1	1	1	1	1	0	0	0	1	1	1	9
	Introduction													
2	Background/ rationale	1	1	1	1	1	1	1	1	1	1	1	1	12
3	Objectives	1	1	1	1	1	1	1	1	1	1	1	1	12
	Study Design													
4	Study design	1	1	1	1	1	1	1	1	1	1	1	1	12
5	Setting	1	1	1	1	1	1	0	1	1	1	1	1	11
6	Participant	0	1	1	1	1	0	1	1	1	1	1	1	10
7	Variables	1	1	0	1	1	1	1	1	1	1	1	1	11
8	Data sources	1	1	0	0	1	1	0	1	1	1	1	1	9
9	Bias	0	0	1	0	1	0	0	0	0	1	0	0	3
10	Study Size	1	1	1	1	0	1	0	0	0	0	0	0	5
11	Quantitative variables	1	1	1	1	1	1	1	1	1	1	1	1	12
12	Statistical methods	1	1	1	1	1	1	1	1	1	1	1	1	12
	Results:													
13	Participant	0	1	1	1	1	1	1	0	1	0	0	1	8
14	Descriptive data	1	1	1	1	1	1	0	1	1	1	1	1	11
15	Outcome data	1	1	1	1	1	1	1	1	1	1	0	1	11
16	Main results	0	1	0	0	0	0	0	1	0	0	0	1	3
17	Other analyses	1	0	0	1	1	0	0	0	0	0	0	0	3
	Discussion													
18	Key Results	1	1	1	1	1	1	0	1	1	1	1	1	11
19	Limitations	0	1	1	1	1	1	0	0	1	1	1	1	9
20	Interpretation	1	1	0	1	1	1	1	1	1	1	1	1	11
21	Generalizability	0	0	1	1	0	1	1	1	0	1	1	1	8
	Other Information													
22	Funding	1	1	1	1	1	1	1	1	1	1	1	1	12
Total		**16**	**19**	**17**	**19**	**19**	**19**	**12**	**16**	**16**	**18**	**16**	**19**	

*Study 1 = De Pew, 2012; study 2 = Donaire Gonzalez, 2011; study 3 = Garfield 2012; study 4 = Gimeno-Santos 2014; study 5 = Hunt2013; study 6 = Moore, 2009; study 7 = Moy, 2009; study 8 = Nyssen, 2013; study 9 = Pitta, 2005; Slinde, 2011; Van Gestel, 2012

CHAPTER IV

Validity of the Global Physical Activity Questionnaire in older adults with COPD

ABSTRACT

Background: Physical activity (PA) is a major independent, modifiable risk factor that has a protective effect on health outcomes in chronic obstructive pulmonary disease (COPD). Population-based survey of PA has been an important part of global health initiatives. Subjective measures of PA are widely used in population surveillance, owing to their relatively low cost, higher clinical utility and low burden on participants. The Global Physical Activity Questionnaire version 2 (GPAQv2) is one of the most commonly used PA questionnaires which has been recommended by the World Health Organization for PA surveillance globally across countries. At this time, there is a lack of population-based studies to assess the validity of this tool in older adults with COPD.

Purpose: The purpose of this study was to examine the construct validity of the GPAQv2 using population-based data from a cohort of older adults with COPD.

Methods: All individuals age 65 and older who were interviewed in the National Health and Nutrition Examination Survey (NHANES) between the years 2007 and 2012 were included for this study. Individuals with COPD were identified from this group based on self-report of diagnosis and from spirometry values. The GPAQv2 was used to assess the total weekly PA expressed in metabolic equivalent (MET) minutes per week. Known-groups validity was assessed by testing the ability of GPAQv2 to identify the presence or absence of COPD among all older adults (N = 4329) included in the sample. Associations between related constructs of PA in COPD (lung function and shortness of breath) and GPAQ scores were examined to assess convergent validity in a group of older adults with COPD (N = 636). For discriminant validity,

associations between GPAQ scores and an unrelated construct of total household income were examined.

Results: The GPAQv2-derived total PA, sedentary time and activity levels were unable to explain the variance in all the main outcomes indicating a lack of known-groups and convergent validity. GPAQv2 demonstrated poor correlation with household income indicating the presence of discriminant validity.

Limitations: This study was limited by missing data and heterogeneity in the COPD population.

Conclusions: The GPAQv2 did not demonstrate construct validity in the sample of older adults with COPD. Future research is needed to assess the validity of GPAQv2 in older adults with COPD by age and disease severity. Future research to establish validity of GPAQv2 against objective reference standards such as accelerometers is also warranted.

Introduction

Chronic obstructive pulmonary disease (COPD) has been identified as the third leading cause of death in the United States.[1-3] COPD has been predicted to become the fifth largest cause of disability by the year 2020.[4] The disability found in individuals with COPD is largely associated with their sedentary lifestyle as they are notably less active as compared to those without the disease.[2,3,5] As compared to healthy older adults, individuals with COPD have 40 – 60% lower activity levels.[5-7] Inactivity in COPD is found even in the early stages of the disease and declines as the disease severity worsens, with individuals on long term oxygen therapy demonstrating a further decline in activity levels.[7,8] The disabling symptoms of COPD including dyspnea, peripheral and respiratory muscle dysfunction, along with age-related declines in endurance, result in activity limitations and a greater proportion of sedentary time.[9-11] Lower levels of activity, in turn result in further muscle weakness and deconditioning.[6] Physical inactivity seen in COPD has been shown to be a poor prognostic indicator in the disease course and is associated with increased risk of hospitalization and death.[12,13] Inactivity in these individuals can therefore significantly affect the quality of life by limiting their ability to perform daily tasks.[9]

Physical activity (PA) has been identified as a major independent, modifiable risk factor that has a protective effect on important health outcomes including functional status, lung function, risk of acute exacerbations, hospitalization and even death in individuals with COPD.[12-15] Lower levels of activity in COPD have been confirmed with the use of different PA assessments including accelerometers, pedometers and self-report measures.[5-7,16,17] Owing to these reasons, improving PA levels is an important goal in the management of COPD.[8,18] In order to accomplish this goal of improving overall PA, assessment and monitoring of PA and

understanding of the patterns of PA among individuals with varying severity of COPD is imperative. Considering the impact of low PA levels towards increased burden of chronic disease, premature death and associated health care costs, initiatives to monitor population levels of PA using standardized measures have been identified as a national health priority and an important component of public health practice.[19,20]

Various methods have been used to quantify PA levels in clinical and research settings. These methods range from simple self-reported questionnaires or activity logs to more complex accelerometers and laboratory methods such as indirect calorimetric methods.[9,11,21-23] Due to their relatively low cost, higher clinical utility, and low burden on participants, subjective measures are more accessible for use in large epidemiological studies and population based surveillance systems.[24] Subjective PA assessments are therefore widely used in the surveillance of PA in large population groups, and form an important part of global health initiatives.[18]

Of the various subjective PA measures available, the Global Physical Activity Questionnaire version 2 (GPAQ v2) is one of the most commonly used questionnaires in population based surveillance systems. The GPAQv2 has been recommended by the World Health Organization (WHO) for the national surveillance of PA across countries.[25] The GPAQ was developed by the WHO in 2002 as part of the WHO STEPwise approach to surveillance (STEPS).[26,27] The GPAQv2 was designed to compare regional and global differences in PA levels, to inform decisions about PA policy and to address the limitations of the only other available national surveillance measure, the International Physical Activity Questionnaire (IPAQ).[18,28,29] The long form of IPAQ was considered too lengthy and complex to be used as a surveillance measure.[18] The short form of IPAQ was found to largely overestimate PA, and did not allow for differentiation of data from different PA domains.[18,30] The GPAQv2 is an interviewer assisted

questionnaire with 16 questions that are designed to provide an estimate of PA in three broad domains, including work (6 items), transport (3 items) and leisure time (6 items). The GPAQv2 also assesses time spent in sedentary behavior (1 item).[28,29]

Because the GPAQv2 was designed and has been used as a surveillance tool of PA, nationally and internationally, it is important that the validity of this tool be explored. Despite its widespread use, there is limited published data demonstrating the reliability and validity of the GPAQv2.[31] Although previous studies have examined and confirmed validity of this measure, most research on assessment of measurement properties of the GPAQv2 have been performed on smaller, relatively younger (30 - 58 years), and healthy samples with no reported comorbidities.[19,29,31-33] At this time, there are no population based studies to assess the validity of this tool in older adults with COPD. Therefore, the purpose of this study was to examine the construct validity of the GPAQv2 in a large population based cohort of older adults with COPD.

Methods

The study was approved by the Institutional Review Board at the University of Michigan-Flint.

Study design

This study utilized publically available secondary data from the National Health and Nutrition Examination Survey (NHANES). The NHANES is a cross-sectional, multistage, stratified, clustered probability sample of civilian, non-institutionalized, U.S. populations conducted by the National Center for Health Statistics with oversampling of ages over 60 years, low income groups, African- Americans and Mexican- Americans.[34] The NHANES utilizes standardized interviews and physical examinations that are conducted in participants' homes and mobile medical centers via trained professionals in order to collect demographic and clinical information from participants.[34]

Inclusion and exclusion criteria

All individuals, age 65 and older who participated in the NHANES survey from year 2007 to 2012 were included for this study. Individuals with COPD were identified as those who responded 'yes' to having either emphysema or chronic bronchitis or both, on the medical conditions questionnaire of the NHANES and/or those who demonstrated a spirometry post-bronchodilator forced expiratory volume in the 1st second and forced vital capacity (FEV1/FVC) ratio of less than 0.70.[35]

Sample and setting

Among the 30,442 non-institutionalized community dwelling participants that completed NHANES interviews between years 2007 and 2012, 4329 were 65 years or older. Based on self-report of diagnosis on the medical questionnaire and/or from spirometry testing, 636 participants were identified as having COPD. The remaining older adult participants (n = 3693) were identified as non-COPD.[35]

Insert figure 1 here

Instruments

GPAQv2. The GPAQv2 is an interviewer based questionnaire is comprised of 16 questions assessing PA in three domains including work (n = 6), transport (n = 3) and leisure (n = 6) as well as the time spent in sedentary behavior (n = 1) in a typical week. The GPAQv2 takes about 5 minutes to administer and classifies PA according to intensity levels including moderate, vigorous and inactivity. The interviewer asks participants to determine intensity of PA based on the increase in heart rate or breathing (small or large increase) caused by PA. The work and leisure domains measure duration and frequency of PA at different intensities (moderate and vigorous), whereas the transport domain assesses duration and frequency of walking and cycling with no distinction between the activities based on intensity.[18,28,29] For each domain, the GPAQv2 assesses frequency (days per week) and the time spent in PA (minutes per week). The GPAQv2 guidelines indicate that compared to sitting quietly, a person's caloric consumption is four times as high when being moderately active, and eight times as high when being vigorously active. Therefore, when calculating a person's overall energy expenditure using GPAQv2 data, 1 metabolic equivalent (MET) is assigned for inactivity, 4 METs for time spent in moderate activities, and 8 METs for the time spent in vigorous activities.[28]

Scoring of GPAQv2 is performed according to the GPAQv2 analysis guide. Total PA score is expressed in MET-minutes per week. First, the total PA for different intensities across each individual domain (moderate-intensity work, moderate-intensity recreation, vigorous-intensity work and vigorous-intensity recreation) over a week is calculated. For example, moderate-intensity PA for work is calculated by multiplying time spent in moderate-intensity work, frequency of moderate-intensity work and 4METs. Similarly, vigorous-intensity PA for work is calculated by multiplying time spent in vigorous-intensity work, frequency of vigorous-intensity work and 8METs). The PA of individual domains of different intensities is then added to obtain the final total PA. Based on the total PA over a week, the GPAQ also allows for classification of individuals as sufficiently active or inactive depending on whether or not they meet the WHO established criteria of any combination of walking and moderate or vigorous intensity PA achieving a minimum of at least 600 MET minutes per week.[28]

Spirometry. Spirometry was performed to assess lung function parameters including FEV1 and FVC, using the Ohio 822/827 dry-rolling seal volume spirometers and following administration procedures recommended by the American Thoracic Society (ATS).[36] FEV1 was measured as the maximum volume of air expired in the first second and FVC was measured as the maximum volume of air forcefully exhaled after a maximal inspiration.[37]

Procedures

Two year survey cycles of NHANES from 2007 to 2012 (2007-2008, 2009-2010 and 2011-2012) were combined to obtain the sample. Following identification of the sample, demographic information including age, gender, body mass index (BMI), race and household income was extracted. Total household income was categorized as less than 25,000, 25,000 to

55,000 and more than 55,000. Potential confounding variables including self-report of current smoking status (‘yes’ or ‘no’ response), shortness of breath (‘yes’ or ‘no’ response), other cardiovascular comorbidities including history of heart failure, coronary heart disease, and heart attack (‘yes’ or ‘no’ response), and self-rated health (excellent, very good, good, fair or poor) were extracted. Additionally, information on specific lung function values, (FEV1 and FVC) was also extracted from the dataset. Information was considered as missing when the data was either missing, or where the individuals either refused to answer questions or did not know the answers. The GPAQv2 was used to assess the total weekly PA expressed in MET minutes per week and the time spent in sedentary behavior. Additionally, a categorical scoring of percentage of participants classified as sufficiently active or inactive based on the WHO criteria was also calculated.[28]

Construct validity demonstrates the ability of an instrument to measure an abstract concept and incorporates known-groups, convergent and discriminant validity.[38] Known-groups validity was examined by testing the ability of GPAQv2 to identify the presence or absence of COPD among all older adults included in the sample. Lung function, in terms of FEV1, and shortness of breath have been significantly associated with PA in COPD.[39] Therefore, convergent validity was assessed by examining associations between PA (GPAQv2 scores) and related constructs of PA in COPD (lung function and shortness of breath). Discriminant validity was performed by examining associations between PA (GPAQv2 scores) and household income as socioeconomic status has not been reported to have significant relationship with PA in COPD.[39,38]

Analysis

NHANES utilizes a complex survey design with oversampling of certain populations, where sample weights are assigned to individuals to create a sample that is representative of the American population. In order to account for this oversampling and stratification and to ensure an unbiased nationally representative sample, new population-based sampling weights were created prior to conducting further analyses.[28,40] New sample weight for the combined dataset of 6 years (2007 – 2012) were constructed using the formula provided in the NHANES analysis tutorial: 1/3 x examination weight for 2 year survey cycles (2007-2008, 2009-2010 and 2011-2012). Analyses were performed using STATA version 15.0 (StataCorp LLC, College Station, Texas) and SPSS 24.0 and (SPSS Inc., Armonk, NY).

Demographic and clinical data were expressed as means and standard deviations for continuous variables and frequencies and percentages for categorical variables. Baseline comparisons of demographic and clinical variables between COPD and non COPD participants were performed using the independent t-test for continuous variables and chi square test for categorical variables. To examine known-groups validity, differences in the GPAQv2 PA scores were first compared between COPD and non-COPD groups. To account for potential confounding variables, a logistic regression model was then constructed with COPD (yes/no) as the dependent variable and total PA, sedentary time and the percentage of individuals passing the WHO cut-off for sufficient activity as independent variables. Other covariates including age, BMI, gender, smoking status, self-rated health, and history of heart failure, coronary heart disease or heart attack were controlled for in the model. For convergent validity, associations between FEV1 and shortness of breath and GPAQv2 derived PA were examined using separate regression models. Multiple regression was performed to assess the relationship of FEV1 as the

dependent variable with GPAQv2 derived total PA, sedentary time and percentage of individuals who were sufficiently active as independent variables, while controlling for covariates including age, BMI, gender, smoking status, self-rated health and comorbidities. All independent variables were entered together in the model using the forced entry method as there was no evidence of multicollinearity. The variance inflation factor (VIF) was used to test multicollinearity, with an average VIF greater than 6 indicating evidence of multicollinearity.[41] The average VIF for the linear regression model was 1.28. Finally, in order to examine the relationship between shortness of breath and PA measured in the GPAQv2, a logistic regression model with presence or absence of shortness of breath as the dependent variable was performed. For discriminant validity, a logistic regression model with household income categorized into less than $25,000 or more as the dependent variable was performed. The two-tailed significance level was set at $p < 0.05$.

In order to determine the effect of missing data on the overall model, multiple imputation method with 5 imputations was performed. Multiple imputation is a statistical technique for analyzing incomplete datasets, where the missing entries of incomplete datasets are filled, not once, but several times (generally, 3 – 5 times) to create several plausible complete datasets. All these created datasets are then individually analyzed and the individual results are combined to obtain the final result.[42,43] Relative variance increase (RVI) was noted to assess the impact of missing values on overall variance of estimates. RVI indicates an increase in the variance of estimate because of loss of information about a parameter due to missing data or non-response as compared to the variance of the estimate when information was complete. The effect of missing data on the variance of estimate decreases as the RVI approaches zero.[44]

Results

Of the total sample (NHANES data from 2007 – 2012), 4329 participants were 65 and older and 14.69% (n = 636) had COPD. The mean age of the total sample was 73.79 (5.30) years, of which48.83% were males and 57.08% were non-Hispanic whites.(Table IV.1) Statistically significant differences in the baseline demographic and clinical variables were noted between the COPD and non COPD groups. The COPD group was younger with a higher percentage of males, current smokers, greater percentage of reported cardiovascular comorbidities, shortness of breath and lower self-rated health as compared to those without COPD. No differences were noted between groups in BMI, household income, PA and other non-cardiovascular comorbidities. (Table IV.1) Over 74% of the entire sample (COPD group: 75.60% and non-COPD group: 74.36%) met the WHO guidelines of being physically active with a total weekly PA over 600 MET-minutes per week. Information was missing in key areas including total PA scores (53.8% of total sample, 54.2% of COPD and 53.7% of non-COPD) and in smoking status (48.8% of total sample, 27.5% of COPD and 52.4% of non-COPD).

No differences in the results were found following multiple imputations with an average RVI using 5 imputations being 0.0026 indicating minimal impact of missing values on the overall variance of estimates. Since imputing for missing data did not change the statistical significance of any of the outcomes, and since this method did not provide individual odds ratios for predictors, the findings of logistic regression run without imputation were reported here to provide clarity in the interpretation.

Known-groups validity

No significant differences were noted between groups on either of the PA outcomes derived via GPAQv2 indicating a lack of known-groups validity. However, since significant baseline differences were noted between groups in smoking status, self-rated health, cardiovascular comorbidities and shortness of breath, a logistic regression controlling for confounding variables was performed to confirm the results. The model demonstrated an overall good fit ($\chi^2 = 25.69$, $p < 0.001$). Although GPAQv2 derived total PA demonstrated a trend towards an inverse relationship with COPD diagnosis where the odds of having COPD decreased by 3.99^{-6} times with a 1 unit increase in total PA, this relationship was not statistically significant (95% CI = -.0000266 to .0000186). Similarly, the categorical classification of being physically active or time spent in sedentary behavior showed no statistical significance in their ability to identify the presence or absence of COPD. Poor self-rated health and smoking emerged as the strongest predictors of COPD ($p < 0.05$). The odds ratios for all predictors with their 95% confidence intervals can be found in Table IV.2.

Convergent and discriminant validity

The overall model explained 3.6% of the variance in FEV1 and 1.90% of the variance in shortness of breath ($p < 0.001$). However, total PA, as measured with GPAQv2 was not significantly associated with either shortness of breath or FEV1, indicating a lack of convergent validity (Tables IV. 3 and IV.4). Age, gender, smoking status and having a history of heart failure were found to be significant predictors of FEV1. Table IV.3 and IV.4 provide details on the coefficients with confidence intervals for individual predictors.

The GPAQv2 demonstrated discriminant validity evidenced by weak non-significant association between total PA and annual household income (OR = .0000228, 95% CI =-.0000604 - .000106; p = 0.591). (Table IV.5)

Discussion

This study was the first to examine the construct validity of the GPAQv2 in a population-based cohort of older adults with COPD. Construct validity incorporates known-groups, convergent and discriminant validity.[38] The results of this study showed that, although the GPAQv2 demonstrated discriminant validity, it did not demonstrate known-groups validity when controlling for other demographic and clinical factors. Additionally, the GPAQv2 failed to demonstrate convergent validity with related constructs of PA.

Several explanations for these null findings are possible. First, this study utilized lung function and shortness of breath as reference standards to assess construct validity. Although previous research has identified both FEV1[39,45,46] and shortness of breath[39,45-48] as being significantly associated with PA in COPD, there are other constructs that have also been associated with PA in this population, such as quality of life, self-efficacy, exercise capacity, lung hyperinflation, exacerbations and mortality.[6,12,39,46,49-52] These variables were not collected in the NHANES and hence were not used to examine validity. The possibility of GPAQv2 relating better to functional measures of exercise capacity and quality of life rather than structural constructs such as the ones used in this study exists, and therefore further research is needed to assess the validity of GPAQv2 against these related constructs.

The total PA in GPAQv2 is calculated from the equation provided in the GPAQ analysis guide using a summed score of frequency, time spent in PA and assigned MET values based on PA intensity. Individual PA of work and recreation performed at different intensities (moderate and vigorous) is first calculated by multiplying the frequency and duration of PA at work or recreation at a specific intensity, and the MET units for that intensity (8 METs for vigorous and 4 METs for moderate intensity PA).[28] These individual PA scores for moderate and vigorous work

and recreation are then summed to get the total PA. Since the sample consisted of older adults, most of the participants did not engage in work-related tasks. Only 353 of the total 4329 individuals responded to performing any vigorous work and only 1047 of the total sample engaged in moderate-intensity work. In the COPD group, this was even lower with only 71 engaged in vigorous-intensity and 181 in moderate-intensity work related activities. Although, the GPAQv2 asks questions about transportation in terms of time spent in walking and cycling, it does not incorporate walking activities in the calculation of total PA. For older individuals with limited employment and work-related activities, consideration of walking tasks in the calculation of total PA may have better reflected PA of this group. More so, the WHO criteria incorporates a 'combination of walking and moderate or vigorous intensity PA achieving a minimum of at least 600 MET minutes' for classification of individuals as being sufficiently active or inactive. However, omission of transport (walking) domain in the calculation of total PA may have introduced a possibility of incomplete representation of the total PA as measured with GPAQv2.

According to the Global Obstructive Lung Disease (GOLD) guidelines, FEV1 values along with respiratory symptoms are typically used to establish the severity of disease in COPD.[53] FEV1 values less than or equal to 30% of the predicted normal indicate very severe disease, ≥30 – ≤50 % severe, ≥50 – ≤80% moderate and ≥ 80% indicate mild disease.[53] The disease severity of this group was difficult to determine as lung function values were only available for 57.7% of the COPD sample. Of the remaining COPD sample that had complete information on FEV1, heterogeneity was observed with participants belonging to different racial groups, gender and age-groups, and since predicted normal values for FEV1 and FVC vary across age, gender and race, it was difficult to establish the severity for this mixed group. The inability of GPAQv2 to discriminate between COPD and non-COPD may be due to having a

sample with a mixed disease severity group. Further research on examining validity of GPAQv2 in different COPD severities would be helpful in confirming this hypothesis.

The COPD group in this sample may not be representative of the general COPD populationas75.7% of the sample met the WHO criteria for being sufficiently active. This is in contrast to previous literature indicating that individuals at different stages of disease severity in COPD have been known to demonstrate varying levels of activity limitation. However, the total PA scores were only available for 45.75% of the COPD sample. PA levels higher than expected for a group with COPD, along with inadequacies of data, may have resulted in the inability of GPAQv2 to explain the variance in COPD.

The representation of PA by the GPAQv2 may also have been affected by the method of administration of the scale. The GPAQv2 scale uses heart rate and breathing as descriptors of intensity of activity (e.g. in order to ask whether or not participants engaged in vigorous-intensity activities, the interviewer asks participants if they engaged in activities that cause small or large increases in their breathing or heart rate). Although this may be a useful subjective gauge to assess activity intensity, it is highly dependent on the participant's understanding and perception of heart rate and breathing responses, which may lead to misclassification of PA. Use of a numeric scale to help the participants understand the intensity of activity may be more accurate.

Another limitation of the GPAQv2 was the lack of specification of the week or day for activity recall. Instead of using a specific week (previous week or day) to recall activities, the interviewer while administering GPAQv2 asks questions to participants using a typical week or day of their lives. Although this is a better way to get a general idea of their overall PA levels, it may make recall more difficult due to the non-specificity of the question, which may also lead to under or over-estimation of findings. Other subjective PA measures have used several

approaches to minimize recall errors by providing guided prompts to the participant, having participants recall activities in a particular segment of a specific day (segmented day format), or asking them questions about a specific week or day.[21,22] The lack of specificity of the day and absence of prompts by the interviewer may lead to poor recall, especially in older adults. Further research to compare the measurement properties of GPAQv2 in younger and older population is warranted to study the impact of recall errors.

This study is not without limitations. Data for this study was utilized from publically available NHANES survey cycles between 2007 and 2012. Although use of NHANES data provided an opportunity to access information on a large population, minimizing the issues of power in this study, over 40% data in the COPD group was missing which limits the generalizability of the findings. Completer-analyses is often justified only when the number of incomplete cases is small and when the missing data are independent of the outcome or group assignment.[38] Considering the large number of incomplete cases in this study, 5 different imputations were performed during analysis to adjust for missing data. Although, multiple-imputation is a general approach utilized to allow for uncertainty of missing data by creating several different plausible complete datasets, this method is not without limitations as it is a complex process based on intensive computation, and use of approximations.[42 54] The findings of our study did not vary with the use of multiple imputations from when the data was analyzed with only complete values. Nevertheless, this approach is in no way a replacement for the actual data.

Finally, the cohort selected for this study demonstrated various comorbidities that may have affected the findings. We controlled for cardiovascular comorbidities in the regression model as cardiovascular diseases have previously been identified as determinants of PA in

COPD.[39] However, this cohort also had other comorbidities including arthritis, stroke, gout and thyroid disease which may have had an influence on the overall PA levels.

Conclusion

The GPAQv2 did not demonstrate construct validity in older adults in this sample population of older adults with COPD. Future research is needed to assess the validity of GPAQv2 in different COPD severity groups and different age-groups. Further research to establish validity of GPAQv2 against objective reference standards in COPD should be performed.

References

1. Burt L, Corbridge S. COPD exacerbations. *The American journal of nursing.* 2013;113(2):34-43.
2. Perriot B, Argod J, Pepin J-L, Noury N. Characterization of Physical Activity in COPD Patients: Validation of a Robust Algorithm for Actigraphic Measurements in Living Situations. *IEEE Journal of Biomedical and Health Informatics.* 2014;18(4):1225-1231.
3. Park SK, Richardson CR, Holleman RG, Larson JL. Physical activity in people with COPD, using the National Health and Nutrition Evaluation Survey dataset (2003–2006). *Heart & Lung: The Journal of Acute and Critical Care.* 2013;42(4):235-240.
4. Murray CJL, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet.* 1997;349(9064):1498-1504.
5. Larson JL, Vos CM, Fernandez D. Interventions to increase physical activity in people with COPD: systematic review. *Annual review of nursing research.* 2013;31:297.
6. Pitta FT, Thierry;Spruit, Martijn A;Probst, Vanessa S;et al. Characteristics of Physical Activities in Daily Life in Chronic Obstructive Pulmonary Disease.pdf. *American Journal of Respiratory and Critical Care Medicine.* 2005;171(9):972 - 977.
7. Sandland CJ, Singh SJ, Curcio A, Jones PM, Morgan MDL. A profile of daily activity in chronic obstructive pulmonary disease. *Journal of cardiopulmonary rehabilitation.* 2005;25(3):181-183.
8. Spruit M, Singh S, Garvey C, et al. An official American thoracic society/European respiratory society statement: Key concepts and advances in pulmonary rehabilitation. *American Journal of Respiratory and Critical Care Medicine.* 2013;188(8):e13-e64.
9. Bonnie G. Steele P, RN; Lyn Holt, MS; Basia Belza, PhD, RN;, Scott Ferris MSL, MD; and David M. Buchner, MD, MPH. Quantitating Physical Activity in COPD Using Tri axial Accelerometer. *CHEST.* 2000;117:1359 -1367.
10. Make B, Garvey C, Benzo R, et al. An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation. *American Journal of Respiratory and Critical Care Medicine.* 2013;188(8):e13-64.
11. Mikael Andersson CJ, and Margareta Emtner. Accuracy of three activity monitors in patients with chronic obstructive pulmonary disease: a comparison with video recordings.pdf. *COPD.* 2014;11(5):560 - 567.
12. Benzo RP, Chang C-CH, Farrell MH, et al. Physical Activity, Health Status and Risk of Hospitalization in Patients with Severe Chronic Obstructive Pulmonary Disease. *Respiration.* 2010;80(1):10-18.
13. Garcia-Aymerich J, Lange P, Benet M, Schnohr P, Antó JM. Regular physical activity reduces hospital admission and mortality in chronic obstructive pulmonary disease: a population based cohort study. *Thorax.* 2006;61(9):772-778.
14. Garcia-Aymerich J, Farrero E, Félez MA, et al. Risk factors of readmission to hospital for a COPD exacerbation: a prospective study. *Thorax.* 2003;58(2):100-105.
15. Garcia-Rio F, Rojo B, Casitas R, et al. Prognostic value of the objective measurement of daily physical activity in patients with COPD. *Chest.* 2012;142(2):338.
16. Hirayama F, Lee AH, Binns CW, Leong CC, Hiramatsu T. Physical Activity of Patients With Chronic Obstructive Pulmonary Disease: Implications for Pulmonary Rehabilitation. *Journal of Cardiopulmonary Rehabilitation and Prevention.* 2008;28(5):330-334.

17. Tudor-Locke C, Washington TL, Hart TL. Expected values for steps/day in special populations. *Prev Med.* 2009;49(1):3-11.
18. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). *Journal of Public Health.* 2006;14(2):66-70.
19. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. *Journal of physical activity & health.* 2009;6(6):790.
20. Fulton JE, Carlson SA, Ainsworth BE, et al. Strategic Priorities for Physical Activity Surveillance in the United States. *Medicine & Science in Sports & Exercise.* 2016;48(10):2057-2069.
21. 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.
22. 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.
23. Annegarn J, Spruit MA, Uszko-Lencer NH, et al. Objective physical activity assessment in patients with chronic organ failure: a validation study of a new single-unit activity monitor. *Arch Phys Med Rehabil.* 2011;92(11):1852-1857 e1851.
24. Anne HYC, Sheryl HXN, Koh D, Müller-Riemenschneider F. Reliability and Validity of the Self- and Interviewer-Administered Versions of the Global Physical Activity Questionnaire (GPAQ). *PLoS One* 2015;10(9).
25. World Health Organization. Global Strategy on diet, physical activity and health. 2004.
26. Armstrong T, Bonita R. Capacity building for an integrated noncommunicable disease risk factor surveillance system in developing countries. *Ethnicity & disease.* 2003;13(2 Suppl 2):S13.
27. World Health Organization. WHO STEPS Surveillance Manual: The WHO STEPwise approach to chronic disease risk factor surveillance. *World Health Organization, Geneva.* 2005.
28. World Health Organization. Global physical activity questionnaire (GPAQ) analysis guide. *World Health Organization.*Geneva, Switzerland.
29. Cleland CL, Hunter RF, Kee F, Cupples ME, Sallis JF, Tully MA. Validity of the global physical activity questionnaire (GPAQ) in assessing levels and change in moderate-vigorous physical activity and sedentary behaviour. *BMC public health.* 2014;14(1):1255.
30. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review. *The international journal of behavioral nutrition and physical activity.* 2011;8(1):115-115.
31. Herrmann SD, Heumann KJ, Der Ananian CA, Ainsworth BE. Validity and reliability of the Global Physical Activity Questionnaire. *Measurement in Physical Education & Exercise Science.* 2013;17(3):221.
32. Au TB, Blizzard L, Schmidt M, Pham LH, Magnussen C, Dwyer T. Reliability and validity of the global physical activity questionnaire in Vietnam. *Journal of physical activity & health.* 2010;7(3):410.
33. Hoos T, Espinoza N, Marshall S, Arredondo EM. Validity of the Global Physical Activity Questionnaire (GPAQ) in Adult Latinas. *Journal of physical activity & health* 2012;9(5):698.

34. National Health and Nutrition Examination Survey; About the National Health and Nutrition Examination Survey. *Centers for Disease Control and Prevention.*
35. Vestbo J, Hurd SS, Agustí AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine.* 2013;187(4):347.
36. Redlich CA, Tarlo SM, Hankinson JL, et al. Official American Thoracic Society technical standards: spirometry in the occupational setting. *American journal of respiratory and critical care medicine.* 2014;189(8):983.
37. Watchie J, ebrary I. *Cardiovascular and pulmonary physical therapy: a clinical manual.* 2nd ed. St. Louis, Mo: Saunders/Elsevier; 2010.
38. Portney L, Watkins M. *Foundations of Clinical Research : Applications to Practice.* Vol 3: Pearson Health Science; 2009.
39. Gimeno-Santos E, Frei A, Steurer-Stey C, et al. Determinants and outcomes of physical activity in patients with COPD: a systematic review. *Thorax.* 2014;69(8):731-739.
40. Heeringa G WB, Berglund PA. Applied Survey Data Analysis. 2010.
41. Loprinzi PD, Walker JF, Lee H. Association between physical activity and inflammatory markers among U.S. adults with chronic obstructive pulmonary disease. *American journal of health promotion : AJHP.* 2014;29(2):81.
42. Jonathan ACS, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ: British Medical Journal.* 2009;339(7713):157-160.
43. Lee KJ, Simpson JA. Introduction to multiple imputation for dealing with missing data: Multiple imputation for missing data. *Respirology.* 2014;19(2):162-167.
44. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software.* 2011;45(4).
45. Garcia-Aymerich J, Félez MA, Escarrabill J, et al. Physical activity and its determinants in severe chronic obstructive pulmonary disease. *Medicine and science in sports and exercise.* 2004;36(10):1667-1673.
46. 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. *American Journal of Respiratory and Critical Care Medicine.* 2009;180(6):506.
47. Moy ML, Matthess K, Stolzmann K, Reilly J, Garshick E. Free-living physical activity in COPD: assessment with accelerometer and activity checklist. *Journal of rehabilitation research and development.* 2009;46(2):277.
48. Katajisto M, Kupiainen H, Rantanen P, et al. Physical inactivity in COPD and increased patient perception of dyspnea. *International journal of chronic obstructive pulmonary disease.* 2012;7:743-755.
49. Altenburg WA, Bossenbroek L, de Greef MHG, Kerstjens HAM, ten Hacken NHT, Wempe JB. Functional and psychological variables both affect daily physical activity in COPD: A structural equations model. *Respiratory Medicine.* 2013;107(11):1740-1747.
50. Berry MJ, Adair NE, Rejeski WJ. Use of Peak Oxygen Consumption in Predicting Physical Function and Quality of Life in COPD Patients. *Chest.* 2006;129(6):1516-1522.
51. Hartman JE, Boezen HM, de Greef MH, ten Hacken NH. Physical and Psychosocial Factors Associated With Physical Activity in Patients With Chronic Obstructive

Pulmonary Disease. *Archives of Physical Medicine and Rehabilitation.* 2013;94(12):2396-2402.
52. Pitta F, Troosters T, Probst VS, Lucas S, Decramer M, Gosselink R. Potential consequences for stable chronic obstructive pulmonary disease patients who do not get the recommended minimum daily amount of physical activity. *Jornal brasileiro de pneumologia : publicaçẫo oficial da Sociedade Brasileira de Pneumologia* 2006;32(4):301.
53. Rabe KF, Hurd S, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary. *American Journal of Respiratory and Critical Care Medicine.* 2007;176(6):532-555.
54. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ: British Medical Journal.* 2007;335(7611):136-141.

Figure 1: Participant flow chart

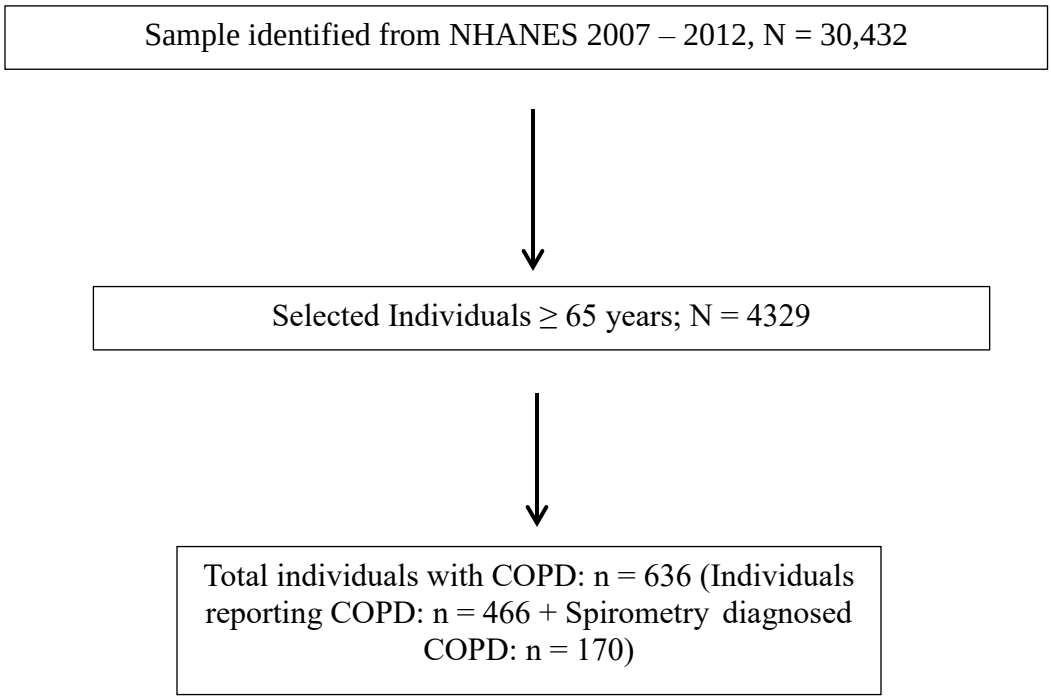
Sample identified from NHANES 2007 – 2012, N = 30,432
Selected Individuals ≥ 65 years; N = 4329
Total individuals with COPD: n = 636 (Individuals reporting COPD: n = 466 + Spirometry diagnosed COPD: n = 170)

Table IV.1: Comparisons of demographic and clinical characteristics between COPD and non COPD participants[#]

	Total sample (N = 4329)	**COPD (N = 636)**	**Non-COPD (N = 3693)**	***p* value**
	Mean (SD) or n (%)	Mean (SD) or n (%)	Mean (SD) or n (%)	
Age (years)	73.78 (5.30)	73.06 (5.07)	73.90(5.33)	< 0.001*
Gender (male)	2114 (48.83)	344 (54.09)	1770 (47.93)	0.004*
Annual Household Income ($) < 25,000 25,000 – 55,000 > 55,000	1655 (38.20) 1340 (31.00) 920 (21.30)	270 (42.50) 193 (30.30) 120 (18.90)	1655 (38.20) 1340 (31.00) 920 (21.30)	0.064
Race Non-Hispanic White Others[#]	 2471 (57.08) 1858 (42.92)	 427 (67.14) 209 (32.86)	 2471 (57.10) 1858 (42.90)	 < 0.001*
BMI (Kg/m^{-2})	28.55 (5.89)	28.85* (6.36)	28.53* (5.80)	0.231
Currently smoking cigarettes (Y)	422 (19.04)	124 (26.90)	298 (16.98)	< 0.001*
Self-rated Health: Good or better Fair or worse	 2625 (69.89) 1131 (30.11)	 346 (61.02) 221 (38.98)	 2279 (71.46) 910 (28.54)	 < 0.001*
Shortness of Breath (Y)	1581 (36.61)	1178 (74.9)	394 (25.1)	< 0.001*
Baseline FEV1 (ml)	2259.96 (681.01)	2043.99 (670.51)	2301.81 (675.26)	< 0.001*
Baseline FVC (ml)	3126.42 (934.21)	3205.26 (993.75)	3111.15 (927.73)	0.094
Comorbidities HF CHD Heart Attack Arthritis Stroke Thyroid disease	 371(8.65) 468 (10.93) 461 (10.69) 2279 (52.80) 443 (10.27) 731 (16.93)	 90 (14.33) 85 (13.69) 105 (16.51) 395 (62.52) 72 (11.36) 120 (18.96)	 281 (7.67) 383 (10.46) 356 (9.68) 1882 (51.13) 371 (10.09) 611 (16.58)	 < 0.001* 0.018* < 0.001* < 0.001* 0.320 0.152
GPAQv2 Total PA in MET minutes/ week	3068.79 (9064.29)	3005.61 (4889.98)	3079.56(9597.11)	0.858
Active Individuals (PA > 600 MET minutes / week)	1490 (74.54)	220 (75.60)	1708 (74.36)	0.716
Total Sedentary Time (minutes)	413.855 (830.917)	417.41 (688.38)	413.24 (853.11)	0.844

[#] Analysis using independent t tests for continuous and chi square test for categorical variables; N = sample size, SD = standard deviation, n = frequency, % = percentage, BMI = body mass index, ($Kg\ m^{-2}$) = kilograms per meter squared, MET = metabolic equivalents, Y/N = yes/no, ml = milli-liters, PA = physical activity, * = statistically significant difference between COPD and non COPD groups, [#] = Mexican American, Hispanic, non-Hispanic Black and other races, HF = heart failure, CHD = coronary heart disease

Table IV.2: Logistic regression to test the ability of GPAQv2 PA in identifying the presence or absence of COPD[#]

	Odds Ratio (95% CI)	***p* value**
GPAQv2 Total PA	$-3.99e^{-06}$ (-.000 - .000)	.730
Total Sedentary Time	-.000 (-.000 - .000)	.660
Active Individuals (PA > 600 MET minutes / week)	.184 (-.207 - .576)	.356
Age	.006 (-.026 - .039)	.690
Gender	-.096 (-.450 - .258)	.595
BMI	-.003 (-.034 - .026)	.805
Self-rated Health	-.519 (-.901 - -.136)	.008*
Current Smoker	.595 (.150 - 1.041)	.009*
History of Heart Failure	.615 (.024 - 1.207)	.041*
History of Coronary Heart Disease	-.180 (-.786 - .425)	.560
History of Heart Attack	.320 (-.253 - .894)	.274

[#]GPAQv2 derived PA (independent variable) in explaining the presence or of COPD (dependent variable); N = 925 out of 4392 with missing values = 3404; Overall model: Pseudo $R^2 = 0.0276$, $\chi^2 = 25.69$, p = 0.0072, where pseudo R^2 = pseudo variance, χ^2 = chi-square test, 95% CI = 95% confidence interval, BMI = body mass index, PA = physical activity, NA = not available, * $p < 0.05$;

Table IV.3: Multiple regression examining associations between PA scores and lung function – FEV1 in older adults with COPD using GPAQv2[#]

	Unstandardized beta (95% CI)	***p* value**
GPAQv2 Total PA	-.006 (-.022 - .009)	.421
Total Sedentary Time	.334 (-.237 - .907)	.238
Active Individuals (PA > 600 MET minutes /week)	-21.873 (-240.932 - 197.185)	.841
Age	-30.465 (-47.107 - -13.823)	<0.001*
Gender	447.266 (249.885 - 644.647)	<0.001*
BMI	-5.89 (-20.406 - 8.615)	.417
Self-rated Health	223.769 (19.928 - 427.610)	.032*
Current Smoker	-313.936 (-587.097 - -40.776)	.027*
History of Heart Failure	-503.583 (-954.217 - -52.948)	.031*
History of Coronary Heart Disease	103.457 (-225.312 - 432.228)	.526
History of Heart Attack	93.8362 (-189.7452 - 377.4177)	.514

[#]GPAQv2 derived PA (independent variable) and lung function – FEV1 (dependent variable); N = 287; Overall model: R^2 = 0.360, df = 11, F = 5.68, p <0.001; where N = sample size, R^2 = variance, df = degrees of freedom, F = F test for ANOVA, *p* = statistical significance, BMI = body mass index, ml = milli-liters, PA = physical activity, * $p < 0.05$

Table IV.4: Logistic regression examining associations between PA scores and shortness of breath in older adults with COPD using GPAQv2[#]

	Odds ratio (95% CI)	***p* value**
GPAQv2 Total PA	-.000 (-.000 - .000)	.392
Total Sedentary Time	.001 (-.000 - .003)	.243
Active Individuals (PA > 600 MET minutes/week)	-.333 (-1.283 - .616)	.491
Age	.077 (-.000 - .154)	.052
Gender	-.507 (-1.271 -.257)	.193
BMI	.127 (.054 - .206)	.001*
Self-rated Health	-1.448 (-2.356 - -.541)	.002*
Current Smoker	1.132 (.124 - 2.141)	.028*
History of Heart Failure	.798 (-.920 - 2.516)	.363
History of Coronary Heart Disease	.322 (-1.026 - 1.670)	.640
History of Heart Attack	.722 (-.665 - 2.111)	.308

[#]GPAQv2 derived PA scores (independent variable) and shortness of breath (Y/N) (dependent variable); N = 187; Overall model: Pseudo $R^2 = 0.1909$, $\chi^2 = 47.40$, <0.001, where pseudo R^2 = variance, χ^2 = chi-square test, 95% CI = 95% confidence interval, BMI = body mass index, SOB = shortness of breath, PA = physical activity, NA = not available, * $p < 0.05$

Table IV.5: Logistic regression examining association between GPAQv2 PA and household income in older adults with COPD[#]

	Odds ratio (95% CI)	***p* value**
GPAQv2 Total PA	.000 (-.000 - .000)	.591
Total Sedentary Time	.001 (-.000 - .003)	.272
Active Individuals (PA > 600 MET minutes /week)	-.049 (-.894 - .794)	.908
Age	-.057 (-.136 - .020)	.148
Gender	.678 (-.038 - 1.395)	.064
BMI	-.045 (-.105 - .014)	.135
Self-rated Health	.986(.225 - 1.746)	.011*
Current Smoker	-.841 (-1.762 - .079)	.073
History of Heart Failure	.416 (-.831 - 1.665)	.513
History of Coronary Heart Disease	.967 (-.336 - 2.271)	.146
History of Heart Attack	-.772 (-1.880782 - .3366452)	.172

[#]GPAQv2 derived PA (independent variable) and household income (categorical dependent variable); N = 177; Overall model: Pseudo $R^2 = 0.0984$, $\chi^2 = 23.10$, p = 0.017, where pseudo R^2 = variance, χ^2 = chi-square test, 95% CI = 95% confidence interval, BMI = body mass index, PA = physical activity, NA = not available, * $p < 0.05$

CHAPTER V

Overview

The aim of this dissertation was to examine subjective assessments of PA in COPD by systematically reviewing available evidence, and then examining the validity of a widely used PA assessment tool in COPD. An additional aim was to develop and validate a quality appraisal tool that could be used for methodological quality assessment of validity studies.

This dissertation utilized the three paper method with three individual studies. The focus of this dissertation was to systematically review the reliability and validity of various subjective PA assessments in adults with COPD. However, in the absence of a risk of bias assessment tool for validity studies, the quality appraisal tool for validity studies (QAVALS) was developed in Study One and used along with another quality appraisal tool for a systematic review of PA assessments in COPD in Study Two. In Study Three, the construct validity of a widely used national surveillance tool, the Global Physical Activity Questionnaire version 2 (GPAQv2), was examined in a large population-based sample of older adults with COPD.

This final chapter summarizes the major findings of the three studies in this dissertation and their implications for future research. It also discusses the limitations of the research and how those limitations may have affected the results.

Summary of Research Design and Results

The following section describes a summary of the methods and results of the three individual studies in this dissertation.

Study One

The purpose of Study One was to design a reliable and valid quality appraisal tool specific to validity studies that could be used for risk of bias assessment in systematic reviews. For this study, a preliminary checklist with 34 possible items for inclusion on the tool was created. Content experts reviewed each of these items to determine if items were 'essential', 'useful but not essential' or 'not necessary' for quality assessment of validity studies. Based on the content experts' ratings of each item, a content validity ratio (CVR) was calculated to determine whether or not the item should be included on the tool. Following two rounds of review, the final quality appraisal tool for validity studies - QAVALS was developed. The content validity index was used to examine the content validity of the developed tool and weighted Kappa coefficients were used to examine inter-rater and test-retest reliability. The QAVALS emerged as a 24 item, evidence-based quality appraisal tool for different types of validity, with evidence of strong content validity (content validity index 0.90), good overall inter-rater (k = 0.70, 95% CI = 0.61 – 0.79) and excellent test-retest reliability (k = 0.80 - 0.84; 95% CI = 0.76 – 0.90). However, the individual item reliability for items on the tool was highly variable in this study warranting further investigation.

Study Two

The purpose of Study Two was to complete a systematic review of the reliability and validity of various available subjective measures of PA in COPD. The QAVALS was used for methodological quality assessment of validity studies. For reliability studies, the Quality Appraisal of RELiability studies (QAREL) tool was used.[1] Fifteen different subjective PA measures that were examined for reliability and/or validity were identified from this review, of which 7 were self-administered, 2 were assisted (semi-structured or structured interviews), 2 were computerized, 1 was either self-administered or assisted, 1 was rater based and 2 were hybrid measures. The Stanford 7-day recall[2] (PAR) and the Multimedia Activity Recall[3] (MARCA) demonstrated good concurrent validity with the strongest correlations with the reference standards, indicating that assisted and computerized patient reported outcome (PRO) measures demonstrated better validity than other subjective PA assessments. Hybrid PA measures demonstrated good construct validity ($r = 0.55 – 0.65$, $p < 0.05$).[4] Additionally, computerized and clinical hybrid measures showed excellent test-retest reliability (ICC = 0.90 – 0.96).[3-5] Observations drawn from single validation studies, methodological inconsistencies and use of improper reference standards limited the ability to draw conclusions and make recommendations. Further studies using valid reference standards would be worth examining for the assisted, computerized and hybrid tools.

Study Three

The purpose of Study Three was to examine the construct validity (known-groups, convergent and discriminant validity) of the GPAQv2 in a large population based cohort of older adults with COPD. For this study, publically available secondary data of the National Health and Nutrition Examination Survey (NHANES) collected between the years 2007 and 2012 was utilized to identify the sample of older adults (COPD, n= 636 and non-COPD, n = 3693). Regression analysis was used to examine the known-groups, convergent and discriminant validity of the GPAQv2 while controlling for potential confounders. The results of the study indicated that PA outcomes derived from GPAQv2 (total PA, sedentary time and activity status) were neither able to identify the presence or absence of COPD nor were unable to demonstrate significant associations with related constructs of PA, indicating a lack of known groups and convergent validity. GPAQv2 was able to identify the unrelated construct of household income from PA outcomes, indicating evidence of discriminant validity. Inadequacies of the data, including missing data and insufficient representation of the COPD population in the sample influenced the findings of this study. Future research is needed to examine the validity of GPAQv2 in different COPD severity groups and different age-groups.

Discussion of Results

This dissertation reviewed available subjective PA measures and sought to explore the validity and reliability of these measures in adults with COPD. Results of the three studies add to the existing body of knowledge in this area.

Study One was developed to bridge the gap in literature on reliable and valid quality appraisal tools for methodological quality assessment of validity studies. This study formed an integral part of the systematic review performed in Study Two. The absence of a reliable and valid quality appraisal tool specific to addressing different types of validity designs, led to the identification of the objectives and key concepts of the designs for the QAVALS. Additionally, related literature on the development and validation of other quality appraisal tools was reviewed to help create this tool.[1,6-8]

The QAVALS emerged as a 24-item tool that went through 2 rounds of methodological review and content validation. The CVR, a widely recognized statistical measure for establishing content validity, was utilized to help retain or eliminate items on the QAVALS.[9,10] According to Lawshe and Shipper, different cut-off values for CVRs are used for different number of content experts.[10] Based on the ratings of 8 experts used in this study, items were retained on the tool in their original form if an individual item's CVR was above 0.75, which was the cut-off identified from the critical value table.[10] The final content validity index of the QAVALS was 0.90 indicating evidence of strong content validity.[9]

Although the QAVALS demonstrated good to excellent inter-rater and test-retest reliability of the overall tool, the reliability of individual items on the tool was variable. This was not a surprising finding, considering the small number of highly specific studies included for reliability testing in this study. Since the raters were not subject matter experts and may have

been unfamiliar with the outcome measures examined in the studies that they rated, it may have affected the reliability.

The development of QAVALS helped in the risk of bias assessment of included validity studies in Study Two. QAVALS assisted in identification of inconsistencies in study quality of the included validity studies in the systematic review. The results of the systematic review demonstrated that assisted and computerized PRO measures as well as hybrid measures demonstrated better validity as compared to other subjective measures.[2,3] These findings conflicted with those previously reported which that have criticized all types of PRO measures for their inability to capture the construct of PA.[11,12] However, this review indicated that this was true only for the unsupervised and rater-based PRO measures that did not demonstrate good measurement properties in people with COPD. Specific types of PRO measures, especially the assisted and computerized measures, demonstrated moderate to good ($r = 0.54 – 0.83$, $p < 0.001$) correlations with the reference standards in detecting energy expenditure, PA levels and time spent in PA in COPD. These measures also demonstrated moderate ($r = 0.37 – 0.40$, $p < 0.001$) associations with related constructs of PA, indicating evidence of construct validity. Both these type of measures utilize structured prompts that aid in better recall and, in part, may be responsible for their close associations with reference standards. In addition, both computerized and hybrid measures demonstrated excellent test-retest reliability ($r = 0.90 – 0.96$, $p < 0.001$), which could be attributed to the highly structured format of these tools.[3-5]

Study Three examined the validity of a widely used assisted PRO measure, the GPAQv2, in older adults with COPD. The GPAQv2 is recommended by the WHO as a tool for national and international PA surveillance.[13] Based on the results of the systematic review that indicated better performance of assisted PRO tools in COPD, it was hypothesized that the GPAQv2 would

demonstrate good construct validity. However, unlike other assisted tools in the systematic review that demonstrated moderate correlations with the six minute walk distance (6MWD)[14], the findings of this study indicated poor construct validity of the GPAQv2 in older adults with COPD. The GPAQv2 was not able to predict the presence of COPD (poor known-groups validity) nor was it significantly associated with related constructs of PA, indicating poor convergent validity. One possible explanation for the difference in performance of GPAQv2 from other assisted and computerized PRO measures was the reference standard used to assess construct validity. In contrast to studies that utilized exercise capacity (6MWD) as a related construct in Study Two, Study Three utilized shortness of breath and lung function to examine construct validity. It is possible that assisted measures such as the GPAQv2 may relate better to functional measures of exercise capacity rather than structural constructs such as the ones used in Study Three. Since all three constructs (lung function, shortness of breath and exercise capacity) have been identified in previous research to have a relationship with PA in COPD, it would be worthwhile to assess the validity of GPAQv2 against the construct of exercise capacity.[15]

Another reason for the difference in the performance of GPAQv2 was the way it measured total PA. The GPAQv2 utilizes a formula to calculate total PA that is different from how it is calculated in other assisted measures such as the PAR. The PAR calculates energy expenditure for light, medium and heavy activities (irrespective of the activity domain) by assigning a range of metabolic equivalent (MET) values based on the activity intensity.[16] This measure distinguishes between sedentary and light PA by assigning a different MET value to light PA.[16] The MARCA also calculates the PA level as the weighted mean MET per day score, by counting multiples of resting metabolic rate in a day.[17] The GPAQv2, on the other hand, does not distinguish between sedentary and light PA and assigns 1 MET to all activities that fall below

moderate intensity PA. By not distinguishing between light and sedentary PA, the possibility of under estimation of PA by the GPAQv2 cannot be ruled out.

The other difference in the way GPAQv2 measures PA is that it assigns MET values based on the activity domain. The GPAQv2 focuses mainly on the job-related and recreation domains in their calculation of the total PA and ignores the transport domain, that includes walking and cycling PA.[18] This is in contrast with other assisted tools discussed in Study Two that use energy expenditure of all moderate to vigorous intensity activities regardless of the PA domain.[2] Using domain-specific calculations and a lack of including the transport domain in the total PA could introduce a possibility of improper representation of total PA. This difference may hold even more importance in the older population with chronic conditions, such as the one studied in this research, as the possibility of the older population not engaging in job-related activities is higher. Measuring total PA using the walking domain could provide a better overall representation of the total PA in this sample of older adults with COPD. Alternatively, it would be of value to compare PA of each domain individually instead of a total PA value.

Unlike other assisted tools (PAR/YPAS), the interviewer administering the GPAQv2 did not use guided memory techniques to help with recall of questions.[18] The omission of structured prompts to aid in recall may also explain the difference in the performance of GPAQv2 from other assisted tools. Additionally, unlike other assisted PRO measures, the GPAQv2 questions did not pertain to PA performed on a specific day or week. Instead, the interviewer asked participants to give a general idea of their activity levels under different domains in a typical week.[18] In contrast, other assisted tools such as the PAR require participants to fill out activities performed either on the previous day or previous week, making the recall more specific. Not limiting the questions to a specific week or day of the week, decreased the specificity of the date

of reference for the questions asked, thereby making the questions more abstract, and therefore, difficult to recall.

Finally, demographic differences in the NHANES sample may have affected the findings of Study Three. Majority of the COPD group in the sample demonstrated a sufficiently active lifestyle, with no significant differences in the activity levels between the COPD and non COPD groups. This is in contrast to the abundance of previous literature indicating lower PA levels in individuals with COPD.[2,19-25] Activity levels higher than what are expected of in a group with COPD in this sample may have resulted in the inability of GPAQv2 to identify COPD from the sample of older adults.

Limitations

This dissertation examined subjective PA assessments in COPD. Several limitations were observed in this dissertation. Since a quality appraisal tool that could specifically look at methodological quality of validity studies was not identified, a quality appraisal tool had to be developed prior to conducting the systematic review. Although this tool demonstrated acceptable measurement properties, this tool was limited as the reliability of the QAVALS was examined using raters with similar backgrounds and studies specific to one outcome measure.

Another limitation of this dissertation was the use of a secondary dataset (NHANES) that did not allow for selection or manipulation of the variables that were already collected. Several determinants of PA in COPD have been identified in literature, including exercise capacity[15,26-28], quality of life[15,28,29], exacerbations[15,30], hyperinflation[15,28], self-efficacy[15,26,31], systemic inflammation and gas exchange.[15] This dissertation was limited in using only two determinants (lung function and shortness of breath) to examine the construct validity due to the lack of other variables in the dataset.

Study Two demonstrated that not only assisted tools but hybrid measures also demonstrated promising findings in terms of measurement properties for PA assessment in COPD. However, these could not be examined or compared with the GPAQv2 due to limitations of the dataset. Finally, this dissertation was also limited due to missing data in the sample.

Recommendations for Future Research

The limitations discussed above provide directions for potential future research. The QAVALS tool developed in Study One fills an important gap in literature regarding quality assessment of validity studies. Future studies are needed to develop a shorter version of this tool with a single response category, rather than a compilation of responses in one category, to improve clarity and to decrease the length of time for administration. Also future reliability studies using a larger number of studies examining different outcome measures and different researchers of varying levels of experience should be considered.

Study Two revealed that assisted and computerized PRO measures demonstrate better measurement properties and may be used for assessment of PA in COPD. Future research is needed to examine the validity of these measures further using valid reference standards. Additional research is also needed on the construct validity of these measures using other potential determinants of PA in COPD. Newer hybrid measures also demonstrated good reliability and construct validity warranting further research to examine their validity against objective reference standards. Future studies should also examine the validity of hybrid measures using different accelerometers in the accelerometer-PRO combinations.

The findings of Study Three indicated that the GPAQv2, which is a widely used assisted PRO- measure, may not be a valid tool for PA assessment in older adults with COPD. Since the possibility that older adults with COPD may not be engaged in job related activities is high, future studies should examine the validity of GPAQv2 by assessing each domain (work, leisure and transport) individually instead of just using the total PA score that ignores the transport domain. Since the construct validity was assessed using only lung function and dyspnea, future

studies should be performed to examine the validity of GPAQv2 against other related constructs of exercise capacity and quality of life. Finally, studies to examine the validity of GPAQv2 against valid objective reference standards including the Actigraph GT3X and Stepwatch 3 accelerometers are warranted.

Conclusion and Clinical Implications

The results of these three studies provide evidence in the area of use and quality appraisal of subjective PA assessments in COPD.

First, this dissertation provides researchers with a new tool to assess methodological quality of validity studies. Development of the QAVALS facilitated the risk of bias assessment in the systematic review of subjective PA measures. In contrast to previous quality appraisal tools for validity that were specific to one outcome measure of interest, the QAVALS was designed to assess quality of validity studies of different types of outcome measures.[32,33] Additionally, this tool provides the clinician and researcher with a better method for quality assessment as it does not use summary scores thus promoting the individual importance of each item on the tool instead of equal weighting of all the items.[34]

Second, this dissertation highlights the lack of valid and reliable subjective PA measures to assess PA in COPD and the importance of conducting future research in this area. One of the important implications from this research was that contrary to the current criticism of PRO measures in assessing PA, some PRO measures, especially the assisted and computerized PRO measures, demonstrate better measurement properties than unsupervised PROs and rater based measures in COPD. Assisted and computerized PRO measures may therefore be used in the assessment of PA in COPD and need to be further examined for measurement properties.

Finally, despite having established validity in younger adults without reported comorbidities, the GPAQv2 did not demonstrate known-groups or convergent validity in older adults with COPD.[36-40] Although this tool has been recommended by the WHO to be used nationally and internationally for population based surveillance of PA, the GPAQv2 may not be an appropriate tool to measure PA in older adults with COPD until further studies are conducted.

References

1. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology.* 2010;63(8):854-861.
2. 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.
3. 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.
4. Gimeno-Santos E, Raste Y, Demeyer H, et al. The PROactive instruments to measure physical activity in patients with chronic obstructive pulmonary disease. *The European respiratory journal* 2015;46(4):988.
5. Gouzi F, Préfaut C, Abdellaoui A, et al. Evidence of an Early Physical Activity Reduction in Chronic Obstructive Pulmonary Disease Patients. *Archives of Physical Medicine and Rehabilitation.* 2011;92(10):1611-1617.e1612.
6. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC medical research methodology.* 2013;13(1):111.
7. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC medical research methodology.* 2003;3(1):25-25.
8. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* 2011;155(8):529.
9. Gilbert GE, Prion S. Making Sense of Methods and Measurement: Lawshe's Content Validity Index. *Clinical Simulation in Nursing.* 2016;12(12):530-531.
10. Lawshe CH. A Quantitative Approach to Content Validity. *Personnel Psychology.* 1975;28(4):563.
11. Dobbels F, de Jong C, st E, et al. The PROactive innovative conceptual framework on physical activity. *European Respiratory Journal.* 2014;44(5):1223-1233.
12. Gimeno-Santos E, Frei A, Dobbels F, et al. Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review. *Health and quality of life outcomes.* 2011;9(1):86-86.
13. World Health Organization. Global Strategy on diet, physical activity and health. 2004.
14. Donaire-Gonzalez D, Gimeno-Santos E, Serra I, et al. Validation of the Yale Physical Activity Survey in chronic obstructive pulmonary disease patients. *Archivos de Bronconeumología ((English Edition)).* 2011;47(11):552.
15. Gimeno-Santos E, Frei A, Steurer-Stey C, et al. Determinants and outcomes of physical activity in patients with COPD: a systematic review. *Thorax.* 2014;69(8):731-739.
16. Sallis JF, Haskell WL, Wood PD, et al. Physical activity assessment methodology in the Five-City Project. *American journal of epidemiology.* 1985;121(1):91.
17. Ridley K, Olds TS, Hill A. The Multimedia Activity Recall for Children and Adolescents (MARCA): development and evaluation. *The international journal of behavioral nutrition and physical activity.* 2006;3(1):10-10.
18. World Health Organization. Global physical activity questionnaire (GPAQ) analysis guide. *World Health Organization.*Geneva, Switzerland.

19. Bossenbroek L, De Greef MHG, Wempe JB, Krijnen WP, Ten Hacken NHT. Daily physical activity in patients with chronic obstructive pulmonary disease: A systematic review. *COPD: Journal of Chronic Obstructive Pulmonary Disease.* 2011;8(4):306-319.
20. Larson JL, Vos CM, Fernandez D. Interventions to increase physical activity in people with COPD: systematic review. *Annual review of nursing research.* 2013;31:297.
21. Pitta F, Troosters T, Probst VS, Spruit MA, Decramer M, Gosselink R. Quantifying physical activity in daily life with questionnaires and motion sensors in COPD. *European Respiratory Journal.* 2006;27(5):1040.
22. Pitta FT, Thierry;Spruit, Martijn A;Probst, Vanessa S;et al. Characteristics of Physical Activities in Daily Life in Chronic Obstructive Pulmonary Disease.pdf. *American Journal of Respiratory and Critical Care Medicine.* 2005;171(9):972 - 977.
23. Seidel D, Cheung A, Suh ES, Raste Y, Atakhorrami M, Spruit MA. Physical inactivity and risk of hospitalisation for chronic obstructive pulmonary disease. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease.* 2012;16(8):1015.
24. Spruit M, Singh S, Garvey C, et al. An official American thoracic society/European respiratory society statement: Key concepts and advances in pulmonary rehabilitation. *American Journal of Respiratory and Critical Care Medicine.* 2013;188(8):e13-e64.
25. Waschki B, Kirsten A, Holz O, et al. Physical activity is the strongest predictor of all-cause mortality in patients with COPD: a prospective cohort study. *Chest.* 2011;140(2):331.
26. Altenburg WA, Bossenbroek L, de Greef MHG, Kerstjens HAM, ten Hacken NHT, Wempe JB. Functional and psychological variables both affect daily physical activity in COPD: A structural equations model. *Respiratory Medicine.* 2013;107(11):1740-1747.
27. Berry MJ, Adair NE, Rejeski WJ. Use of Peak Oxygen Consumption in Predicting Physical Function and Quality of Life in COPD Patients. *Chest.* 2006;129(6):1516-1522.
28. 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. *American Journal of Respiratory and Critical Care Medicine.* 2009;180(6):506.
29. Pitta F, Troosters T, Probst VS, Lucas S, Decramer M, Gosselink R. Potential consequences for stable chronic obstructive pulmonary disease patients who do not get the recommended minimum daily amount of physical activity. *Jornal brasileiro de pneumologia : publicaçấo oficial da Sociedade Brasileira de Pneumologia* 2006;32(4):301.
30. Benzo RP, Chang C-CH, Farrell MH, et al. Physical Activity, Health Status and Risk of Hospitalization in Patients with Severe Chronic Obstructive Pulmonary Disease. *Respiration.* 2010;80(1):10-18.
31. Hartman JE, Boezen HM, de Greef MH, ten Hacken NH. Physical and Psychosocial Factors Associated With Physical Activity in Patients With Chronic Obstructive Pulmonary Disease. *Archives of Physical Medicine and Rehabilitation.* 2013;94(12):2396-2402.
32. Hagstromer M. A checklist for evaluating the validit and suitability of existing physical activity and sedentary behavior instruments. *Measurement of active an sedentary behavors: closing the gaps in self-report methods.* 2010.

33. Rennie KL, Wareham NJ. The validation of physical activity instruments for measuring energy expenditure: problems and pitfalls. *Public Health Nutrition.* 1998;1(4):265-271.
34. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC medical research methodology.* 2005;5(1):19-19.
35. Garcia-Aymerich J, Félez MA, Escarrabill J, et al. Physical activity and its determinants in severe chronic obstructive pulmonary disease. *Medicine and science in sports and exercise.* 2004;36(10):1667-1673.
36. Au TB, Blizzard L, Schmidt M, Pham LH, Magnussen C, Dwyer T. Reliability and validity of the global physical activity questionnaire in Vietnam. *Journal of physical activity & health.* 2010;7(3):410.
37. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. *Journal of physical activity & health.* 2009;6(6):790.
38. Cleland CL, Hunter RF, Kee F, Cupples ME, Sallis JF, Tully MA. Validity of the global physical activity questionnaire (GPAQ) in assessing levels and change in moderate-vigorous physical activity and sedentary behaviour. *BMC public health.* 2014;14(1):1255.
39. Herrmann SD, Heumann KJ, Der Ananian CA, Ainsworth BE. Validity and reliability of the Global Physical Activity Questionnaire. *Measurement in Physical Education & Exercise Science.* 2013;17(3):221.
40. Hoos T, Espinoza N, Marshall S, Arredondo EM. Validity of the Global Physical Activity Questionnaire (GPAQ) in Adult Latinas. *Journal of physical activity & health* 2012;9(5):698.

Institutional Review Board Letter of Determination

Subject: Notice of Exemption for [HUM00128563]

SUBMISSION INFORMATION:

Title: Validity of the Global Physical Activity Questionnaire in older adults with COPD

Full Study Title (if applicable): Validity of the Global Physical Activity Questionnaire in older adults with chronic obstructive pulmonary disease

Study eResearch ID: HUM00128563

Date of this Notification from IRB: 4/21/2017

Date of IRB Exempt Determination: 4/21/2017

UM Federalwide Assurance: FWA00004969 (For the current FWA expiration date, please visit the UM HRPP Webpage)

OHRP IRB Registration Number(s): IRB00000248

IRB EXEMPTION STATUS

The IRB Flint has reviewed the study referenced above and determined that, as currently described, it is exempt from ongoing IRB review, per the following federal exemption category:

EXEMPTION #4 of the 45 CFR 46.101.(b)

Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Note that the study is considered exempt as long as any changes to the use of human subjects (including their data) remain within the scope of the exemption category above. Any proposed changes that may exceed the scope of this category, or the approval conditions of any other non-IRB reviewing committees, must be submitted as an amendment through eResearch.

Although an exemption determination eliminates the need for ongoing IRB review and approval, you still have an obligation to understand and abide by generally accepted principles of responsible and ethical conduct of research. Examples of these principles can be found in the Belmont Report as well as in guidance from professional societies and scientific organizations.

SUBMITTING AMENDMENTS VIA eRESEARCH

You can access the online forms for amendments in the eResearch workspace for this exempt study, referenced above.

ACCESSING EXEMPT STUDIES IN eRESEARCH

Click the "Exempt and Not Regulated" tab in your eResearch home workspace to access this exempt study.

Marianne McGrath

Chair, IRB Flint