PREFERENCE-BASED MEASURES IN COPD

MEASUREMENT PROPERTIES OF GENERIC PREFERENCE-BASED MEASURES IN INDIVIDUALS WITH CHRONIC OBSTRUCTIVE PULMONARY DISEASE

By AVA MEHDIPOUR, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2020) Hamilton, Ontario (Rehabilitation Science)

TITLE: The performance of generic preference-based measures in individuals with chronic obstructive pulmonary disease

AUTHOR: Ava Mehdipour, B.Sc. (McMaster University)

SUPERVISOR: Dr. Ayse Kuspinar

COMMITTEE: Dr. Marla Beauchamp & Dr. Joshua Wald

NUMBER OF PAGES: xi, 112

Lay Abstract

Chronic obstructive pulmonary disease is a disabling lung disease that affects many Canadians, and policymakers require tools to help them decide how to best use limited healthcare resources to help patients. Such tools are called generic preference-based measures and they help tell us how effective a treatment is based on quality of life and cost. However, before these tools can be used to make healthcare decisions, they have to be valid in the target population. Therefore, we conducted a review of studies evaluating the reliability and validity of these measures in people with chronic obstructive pulmonary disease. We also checked whether these tools accurately reflected the areas of life important to this population. Our findings showed that generic preference-based measures were not sensitive to the quality of life of patients with chronic obstructive pulmonary disease, and that there is a need for the development of condition-specific tools.

Abstract

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide. Luckily, many interventions are available for patients with COPD to improve their symptoms and exercise tolerance, and reduce exacerbation events. Generic preference-based measures are measures of health-related quality of life that can be used for cost-utility analysis. However, before these measures can be used to make healthcare decisions, their psychometric properties (i.e., reliability, validity, responsiveness) have to be assessed. The aim of this thesis was to evaluate the psychometric properties of generic preference-based measures in people with COPD. First, a systematic review was conducted to evaluate the existing evidence on the psychometric properties of these measures in people with COPD. Then, a content validation study was conducted to examine whether these measures accurately reflect the areas of life important to people with COPD. Findings from these two studies showed that generic preference-based measures were not sensitive or fully reflective of patients' health concerns. Findings highlighted the need for properly designed studies (e.g., using correct methodology) when evaluating the psychometric properties of generic preference-based measures in COPD. In addition, our results suggest the need for development of a COPD-specific preference-based measure to improve the sensitivity of cost-utility analyses in this population. This in turn would enable the health-related quality of life of individuals with COPD to be accurately captured when making healthcare decisions.

Acknowledgements

I would like thank my supervisor Dr. Ayse Kuspinar, who has guided and helped me tremendously throughout my degree. She was always available and happy to help. Her passion for measurement research and her vast knowledge in the field is admirable. Her care and patience created a safe and comfortable learning environment. I was very lucky to have her guidance throughout my master's. I would also like to thank my committee members, Dr. Marla Beauchamp and Dr. Joshua Wald, for their academic mentorship and support. Thank you both for always being so welcoming and supportive. Your insights and expertise have enhanced my knowledge and improved the quality of my work. I am extremely grateful to have worked with such caring and considerate supervisors.

I am grateful to have worked in such a supportive and collaborative environment. Thank you to the faculty, staff and graduate students at the School of Rehabilitation Science for guiding me through my academics.

Thank you to all the research participants, who volunteered their time to further the field of quality of life and measurement. I would also like to thank the Firestone Institute for Respiratory Health and the West Park Healthcare Centre for allowing my team and I to collect data.

I am also beyond grateful to have gained amazing, compassionate and genuine friends throughout this journey. Thank you for being understanding through the rough times and uplifting through the good.

I would finally like to thank my parents for encouraging me to always persevere and never give up. I am forever indebted.

Table of Contents

List of Tables and Figures

Chapter 2

Chapter 3

List of Abbreviations and Symbols

15D: 15-Dimensional
AQoL-6D: Assessment of Quality of Life 6-Dimensions
AQoL-8D: Assessment of Quality of Life 8-Dimensions
AUC: Area under the Curve
COPD: Chronic Obstructive Pulmonary Disease
COSMIN: Consensus-based Standards for the selection of health Measurement Instruments
CSPBM: Condition-Specific Preference-Based Measure
EQ-5D: EuroQol 5-Dimensions
EQ-5D-3L: EuroQol 5-Dimensions 3-Levels
EQ-5D-5L: EuroQol 5-Dimensions 5-Levels
ES: Effect Size
FEV1: Forced Expiratory Volume in 1 second
FVC: Forced Vital Capacity
GOLD: Global Initiative for Chronic Obstructive Lung Disease
GPBM: Generic Preference-Based Measure
GRADE: Grading of Recommendations Assessment, Development, and Evaluation;
GRS: Global Rating Scale
HRQoL: Health-Related Quality of Life
HUI: Health Utilities Index
HUI 2: Health Utilities Index Mark 2
HUI 3: Health Utilities Index Mark 3
ICC: Intra-class Correlation Coefficient
ICF: International Classification of Functioning, Disability and Health
MID: Minimal Important Difference
PGI: Patient-Generated Index
QALYs: Quality-Adjusted Life Years
QWB: Quality of Wellbeing
QWB-SA: Quality of Well-Being Self-Administered
ROC: Receiver Operating Characteristic
SG: Standard Gamble
SF-36: Short Form Health Survey
SF-6D: Six-Dimensional Short Form Survey
SRM: Standardized Response Mean
TTO: Time Trade-Off
VAS: Visual Analogue Scale

r= Pearson’s correlation coefficient
rho= Spearman’s correlation coefficient

Declaration of Academic Achievement

I, Ava Mehdipour, am the first author for all the thesis components/chapters. Chapter 1 and 4 have been primarily completed by myself with feedback and guidance from Dr. Ayse Kuspinar, and review from Dr. Marla Beauchamp and Dr. Joshua Wald.

For Chapter 2, I contributed to the study's conceptualization, design and interpretations, with collaboration from all co-authors (Dr. Ayse Kuspinar, Dr. Marla Beauchamp, Dr. Joshua Wald and Nicole Peters). Screening of articles and quality assessments were performed by myself and Nicole Peters, under the supervision of Dr. Ayse Kuspinar. Contents of this chapter were written by myself with preliminary edits from Dr. Ayse Kuspinar, and review from all co-authors.

For Chapter 3, I contributed to the study's conceptualization, design and interpretations, with collaboration from all co-authors (Dr. Ayse Kuspinar, Sachi O'Hoski, Dr. Marla Beauchamp and Dr. Joshua Wald). Data collection and analysis was performed by myself and Sachi O'Hoski, under the supervision of Dr. Ayse Kuspinar. Contents of this chapter were written by myself with preliminary edits from Dr. Ayse Kuspinar, and review from all co-authors.

CHAPTER 1

Introduction and Literature Review

1.0 <u>Summary of Problem</u>

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide (Mannino & Buist, 2007). Outcome measures are available to help policymakers determine how to effectively allocate healthcare resources among individuals with COPD (Brazier, 2007). These measures are called generic preference-based measures (GPBMs). They are health-related quality of life (HRQoL) measures that provide a single preference-based score of HRQoL, which can be multiplied by the number of years an intervention is expected to extend life to generate quality-adjusted life years (QALYs) (Brazier, 2007). Policy decisions are made using cost-utility ratios; weighing the QALYs and costs of interventions (Laupacis et al., 1992). However, before GPBMs can be used to make such decisions, they must be reliable, valid and responsive in COPD (De Vet et al., 2011). Therefore, the overall aim of this thesis was to evaluate the psychometric properties of GPBMs in individuals with COPD.

2.0 <u>Chronic Obstructive Pulmonary Disease</u>

COPD is a life-threating and disabling respiratory condition that affects millions of people around the world (World Health Organization, 2017). COPD is defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) (2020) as “a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases”. COPD symptoms include shortness of breath (i.e., dyspnea), cough, sputum production, wheezing, chest-tightness and fatigue (Global Initiative for Chronic Obstructive Lung Disease, 2020). These symptoms can make daily and physical activities difficult, and be detrimental to one's mental health and overall quality of life (Miravitlles & Ribera, 2017; Zamzam et al., 2012).

COPD is a leading cause of morbidity and mortality in the world and is the fifth leading cause of death in Canada (Global Initiative for Chronic Obstructive Lung Disease, 2020; Statistics Canada, 2019). Over 10% of Canadians, over the age of 35, are living with COPD (*Report from the Canadian Chronic Disease Surveillance System: Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada*, 2018). Direct healthcare costs associated with COPD in Ontario (e.g., hospitalization, emergency visits, healthcare professional costs, medications, rehabilitation programs) were estimated to be approximately $3.3 billion in 2011, and are expected to cost the province $172.3 billion by 2041 (Smetanin et al., 2011).

COPD is a very complex and variable disease; its cause, physiological impact and manifestation vary greatly from patient to patient (Agusti et al., 2010; Global Initiative for Chronic Obstructive Lung Disease, 2020). It can develop and progress from various risk factors, including both genetic and environmental factors (Mannino & Buist, 2007). Smoking is the leading risk factor for COPD (Global Initiative for Chronic Obstructive Lung Disease, 2020; Mannino & Buist, 2007). Long-term exposure to inhaled particulate (e.g., smoking) leads to inflammation of the airways, mucous production and alveolar destruction (emphysema) (Global Initiative for Chronic Obstructive Lung Disease, 2020; MacNee, 2006).

A diagnosis of COPD can only be established with the administration of a spirometry test (Global Initiative for Chronic Obstructive Lung Disease, 2020). Spirometry is a breathing test that assesses the volume of air that an individual can forcibly exhale. It examines one's forced vital capacity (FVC); the volume of air exhaled following a deep inhalation, and forced expiratory volume in 1 second (FEV1); the amount of air expelled in the first second of exhalation (Ranu et al., 2011). An FEV1/FVC ratio less than 0.70 is required for a diagnosis of airflow obstruction (Global Initiative for Chronic Obstructive Lung Disease, 2020). Unlike in asthma, in COPD this airflow obstruction is fixed; meaning that it never returns to normal despite treatment (Welte & Groneberg, 2006). The severity of airflow obstruction is determined by comparing FEV1 values to reference values for the individual, which are dependent on age, sex, height and race (Global Initiative for Chronic Obstructive Lung Disease, 2020). Spirometry is the most reliable and objective measure available to assess airflow limitations, however, it should not be used alone as it has poor specificity (Çolak et al., 2019). It is suggested that an accurate diagnosis of COPD should be made with other factors (i.e., exacerbation history and symptoms) taken into consideration (Global Initiative for Chronic Obstructive Lung Disease, 2020).

Both pharmacological and non-pharmacological interventions are available for patients with COPD to improve their symptoms (e.g., dyspnea), health status and exercise tolerance, and reduce infections and exacerbation events (Global Initiative for Chronic Obstructive Lung Disease, 2020). Two common pharmacological therapies include bronchodilators and anti-inflammatory agents. Bronchodilators are designed to relax airway muscles, and anti-inflammatory agents (e.g. inhaled steroids) are designed to reduce airway inflammation and exacerbations (Global Initiative for Chronic Obstructive Lung Disease, 2020). Pulmonary rehabilitation is a non-pharmacological intervention that has been proven to improve shortness of breath, quality of life and exercise tolerance (McCarthy et al., 2015). Other interventions for COPD include oxygen therapy (i.e., the delivery of oxygen to the patient’s body for a prolonged period of time), lung volume reduction surgery and lung transplantation (Global Initiative for Chronic Obstructive Lung Disease, 2020).

3.0 Health-Related Quality of Life

The International Society for Quality of Life defines HRQoL as "an individual's perception of how an illness and its treatment affect the physical, mental and social aspects of his or her life" (Mayo, 2015). HRQoL is a multidimensional construct that can be measured through an individual's perception of their own health status (De Vet et al., 2011; Karimi & Brazier, 2016).

3.1 Measures of Health-Related Quality of Life

There are three different types of HRQoL measures, each serving a different purpose: individualized HRQoL measures, health profiles and preference-based measures. Both health profiles and preference-based measures can be generic or condition-specific. Generic measures of HRQoL allow for health status comparisons across different diseases, demographics and groups, whereas, condition-specific measures are designed for a specific population or health condition and allow for comparisons within a disease (Patrick & Deyo, 1989).

3.1.1 Individualized Measures

Individualized measures are designed to capture the areas of life respondents consider most important and/or enable respondents to weigh their importance (Fayers et al., 2005). A well-known individualized measure of HRQoL is the Patient-Generated Index (PGI) (Martin et al., 2007). The PGI is patient nominated and weighted (Fayers et al., 2005); allowing participants to nominate areas impacted by their health condition, rate them on their severity and allocate points according to their desire for improvement (Ruta et al., 1994). An advantage to these measures is that they allow patients' perspectives and adaptations to be captured (Fayers et al., 2005). However, due to the personalized nature of these measures, it can be difficult to use them to make comparisons between different groups and determine meaningful cut-off scores (Tang et al., 2014).

3.1.2 Health Profiles

Health profiles assess HRQoL by providing multiple outcome scores; a score for each domain of health (Fayers et al., 2005). A common generic health profile is the Short Form Health Survey (SF-36), a 36-item questionnaire that assesses 8 domains (physical functioning, role limitations due to physical problems, role limitation due to emotional problems, bodily pain, general health perceptions, vitality, social functioning and mental health). Each domain is scored on a scale from 0 to 100, with higher scores representing better health (Fayers et al., 2005). Most health profiles do not provide information on the relative importance attached to each domain, as a result, the domains cannot be combined into an overall score. For example, an intervention can have a positive effect on physical health but a negative effect on mental health. Unless the relative importance of each domain is known, it is difficult to establish whether the intervention resulted in a net improvement or decline in HRQoL (Kuspinar & Mayo, 2013).

3.1.3 Preference-Based Measures

Preference-based measures are HRQoL measures used for economic evaluation purposes. They are designed to provide a single preference-based score of HRQoL, with anchors at 0 (death) and 1 (perfect health). This single value of HRQoL can be multiplied by the number of life years that is expected to be gained by an intervention to generate QALYs (Brazier et al., 2007). QALYs reflect the number of years gained in perfect health from an intervention, which is useful when comparing different interventions via cost-utility ratios (Brazier et al., 2007).

Cost-utility ratios are calculated by dividing the additional cost by the QALY(s) gained from the new intervention, with respect to the current intervention (Brazier et al., 2007). Policymakers and researchers can use QALYs to decide which intervention(s) to implement in healthcare (Laupacis et al., 1992). There are two types of preference-based measures; direct and indirect (Fayers et al., 2005).

3.1.3.1 Direct Preference-Based Measures

Direct preference-based measures allow respondents to directly value health states (Fayers et al., 2005). Common examples of direct preference-based measures include standard gamble (SG) and time trade-off (TTO) (Brazier et al., 2007). When using the SG technique, respondents are provided with two alternatives: 1) outcomes of an impaired health state and 2) the treatment with a given probability of returning to full health (Brazier et al., 2007). The respondent is given this choice with different probabilities of returning to full health with the treatment, and the point of indifference is used to calculate the health utility value (Brazier et al., 2007). Similarly, the TTO technique provides respondents with two choices: 1) impaired health state for a fixed period of time and 2) perfect health (with treatment) for a shorter period of time (Brazier et al., 2007). The time period for perfect health varies until the point of indifference (Brazier et al., 2007), which is used to calculate the health utility value. Unfortunately, these methods may introduce biases (e.g., risk aversion bias for SG and time preference bias for TTO) (Brazier et al., 2007) and involve burdensome administration processes (Fayers et al., 2005).

3.1.3.2 Indirect Preference-Based Measures

Conversely, indirect preference-based measures are typically administered in the form of a short questionnaire, making them less burdensome and easier to administer (Fayers et al., 2005). Indirect preference-based measures, also known as GPBMs, are commonly used for economic evaluation purposes because of their ease of use and their generic nature (Brazier et al., 2007). They are labelled 'generic' because they are intended for cost-utility analyses across different diseases (Brazier et al., 2007). They are developed using the general population's preferences for health states, usually by employing a direct valuation method (e.g. SG), and are designed to assess HRQoL across different populations and interventions (Brazier et al., 2007).

There are 7 well-recognized and documented GPBMs (Brazier et al., 2017), each with a unique descriptive system (i.e., content and dimensions) and valuation method (i.e., technique used for deriving weights for health states). An overview of each GPBM is provided below.

3.1.3.2.1 The EuroQol Five-Dimensions Questionnaire (EQ-5D)

The EuroQol Five-Dimensions questionnaire (EQ-5D) is the most widely used GPBM (Brauer et al., 2006; Brazier et al., 2017). It was developed in 1990 by a group of European researchers (EuroQol Group, 1990). Their intent was to develop a general measure of HRQoL that would be efficient in clinical trial settings (i.e., quick and cognitively simple), could be administered alongside other quality of life measures and be used for health state comparisons across nations (EuroQol Group, 1990). Over the years, its use for cost-utility analyses of healthcare interventions became increasingly popular (Brooks & De Charro, 1996). The EQ-5D's descriptive system was developed by examination of existing health status measures' contents (e.g., the Sickness Impact Profile, the Nottingham Health Profile, the Rosser Index and the Quality of Well-Being (QWB) scale) (EuroQol Group, 1990). It consists of 5 dimensions/items: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, with 3 levels (no problems, some problems and extreme problems) each, defining 243 health states (*EQ-5D-3L User Guide*, 2018). Fifteen years later, the EuroQol Group added 2 levels to each dimension to increase the measure's sensitivity and reduce previously reported ceiling effects (*EQ-5D-5L User Guide*, 2015). This revised version was named the EQ-5D-5L with the original becoming the EQ-5D-3L. EQ-5D-5L response levels consist of: no problems, slight problems, moderate problems, severe problems and extreme problems, and define 3125 health states. The EQ-5D-3L and EQ-5D-5L have been valued in many countries around the world, using visual analogue scale (VAS) or TTO methods (*EQ-5D-3L | Valuation*, 2020). The Canadian value set for the EQ-5D-3L and EQ-5D-5L were developed using TTO methods (Bansback et al., 2012; Xie et al., 2016). Health state utilities range from -0.34 (worst possible health state; 33333) to 1.00 (best possible health state; 11111) (Bansback et al., 2012) for the EQ-5D-3L and from -0.148 (55555) to 0.949 (11111) for the EQ-5D-5L (Xie et al., 2016).

3.1.3.2.2 <u>The Six-Dimensional Short Form Survey (SF-6D)</u>

The Six-Dimensional Short Form Survey (SF-6D) was developed from the well-known generic health profile; the SF-36, by Brazier and his colleagues in 1998 and finalized in 2002 (Brazier et al., 2002). The SF-6D was developed to produce single preference-based index scores of HRQoL that could be used for cost-utility analyses (Brazier et al., 2002). The SF-6D includes 6 dimensions/items: physical functioning, role limitation, social functioning, pain, mental health and vitality, with 4-6 response levels (e.g., limiting none to all the time) each, defining 18,000 health states (Brazier et al., 2002). The UK value set was developed using the SG technique

(Brazier et al., 2002). The UK value set ranges from 0.301 (worst possible health state; 645655) to 1.00 (best possible health states; 111111) (Brazier et al., 2017). Recently, a new algorithm has been developed from the UK data set using a non-parametric Bayesian approach, and this has been proven to have better predictive ability of health states (Kharroubi et al., 2007). It ranges from 0.203 (worst possible health state; 645655) to 1.00 (best possible health states; 111111) (Kharroubi et al., 2007).

3.1.3.2.3 The Quality of Well-Being (QWB) Scale

The QWB scale is the oldest GPBM, with its development beginning in 1970 (Fanshel & Bush, 1970). The QWB scale was specifically developed to measure QALYs for economic evaluation (Seiber et al., 2008). The QWB scale is interviewer-administered and involves formal training to properly probe respondents (Read et al., 1987). It consists of 3 dimensions: mobility, physical activity and social activity, which generates 46 functional levels and 27 symptom and problem complexes. When combined, these produce 945 health states (Brazier et al., 2007). In 1998, Andresen et al. (1998) developed a self-administered version of the scale (QWB-SA) to widen its use. The QWB-SA consists of 58 symptom complexes (chronic, acute physical and mental health symptoms) and items related to mobility, physical activity and social activity (Seiber et al., 2008). Weights for the QWB scale were estimated from a sample of adults from San Diego, US, using the VAS technique (Seiber et al., 2008). The worst possible health state is valued at 0.08 and the best possible health state at 1.00 (Brazier et al., 2017).

3.1.3.2.4 Health Utilities Index (HUI)

The Health Utilities Index (HUI) are a family of measures including the Health Utilities Index Mark 1 (HUI1), Mark 2 (HUI2) and Mark 3 (HUI3). HUI1 was developed in 1982 and used to measure neonatal intensive care outcomes for low birth-weight infants (Boyle et al., 1983). The HUI1 evolved into the HUI2 and eventually into the HUI3. The HUI2 was developed many years later for its application in childhood cancer (Torrance et al., 1996). The HUI2 consists of 7 dimensions: sensation, mobility, emotion, cognition, self-care, pain and fertility, with 3-5 levels each, defining 24,000 states (Torrance et al., 1996). The HUI3 was designed to be applicable to a general population with dimensions chosen to be structurally independent (Feeny et al., 2002; Horsman et al., 2003). The HUI3 consists of 8 dimensions: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain, with 5-6 levels each, defining 927,000 health states (Feeny et al., 2002). Both measures (HUI2 and HUI3) were valuated using VAS and SG techniques (Feeny et al., 2002; Torrance et al., 1996). Weights for HUI2 were estimated using a sample of parents of childhood cancer and school-aged children in Hamilton, Ontario, Canada (Torrance et al., 1996). Weights for HUI3 were estimated using adults in the same location (Feeny et al., 2002). The HUI2 ranges from -0.03 (worst possible health state) to 1.00 (best possible health state), and the HUI3 ranges from -0.36 to 1.00 (Horsman et al., 2003).

3.1.3.2.5 The Fifteen-Dimensional (15D)

The Fifteen-Dimensional (15D) consists of 15 dimensions/items: mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental functions, discomfort and symptoms, depression, distress, vitality and sexual activity, with 5 response levels each, defining billions of health states (Sintonen, 2001). Valuations were obtained from Finnish population samples using modified VAS techniques (ratio scale) (Sintonen, 2001). The worst possible health state is valued at 0.11 and the best possible health state at 1.00 (Brazier et al., 2017).

3.1.3.2.6 The Assessment of Quality of Life Eight-Dimensions (AQoL-8D)

The Assessment of Quality of Life Eight-Dimensions (AQoL-8D) questionnaire evolved from the 6 dimensions (AQoL-6D) to increase sensitivity in the mental health domain (Richardson et al., 2014). It consists of the following 8 dimensions: independent living, happiness, mental health, coping, relationships, self-worth, pain and senses, and 35 items each with 4-6 response levels, defining $2.37 \text{x} 10^{23}$ health states (Brazier et al., 2017; Richardson et al., 2014). The AQoL-8D was valuated in a sample of Australians and mental health patients (Richardson et al., 2014). VAS and TTO techniques were utilized to obtain population values for health states (Richardson et al., 2014). Health state values for this measure range from -0.04 to 1.00 (worst to best health state) (Brazier et al., 2017).

4.0 Psychometric Properties

Psychometric properties of a measure include reliability, validity and responsiveness (De Vet et al., 2011), which need to be evaluated before a measure is used in research and practice.

4.1 Reliability

According to the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN), reliability is defined as “the proportion of the total variance in the measurements which is because of ‘true’ differences among patients”; meaning if the patient is stable in terms of the outcome, then results between different measurements should be consistent (Mokkink et al., 2010). There are three types of reliability in accordance with this definition: test-retest, inter-rater and intra-rater (Koo & Li, 2016). Test-retest reliability is how consistent a measure is over time. Inter-rater reliability is how consistent a measure is between two raters on the same occasion. Intra-rater reliability is how consistent a measure is with one rater between two occasions. Since GPBMs tend to be self-report (the rater is the respondent), test-retest reliability would be the most appropriate type of reliability to examine. For continuous measurement scales, which include GPBMs since a single 0-1 index score of HRQoL is obtained, intra-class correlation coefficients (ICCs) should be utilized to assess reliability (De Vet et al., 2011).

4.2 *Validity*

As defined by COSMIN, validity is "the degree to which an instrument truly measures the construct(s) it purports to measure" (Mokkink et al., 2010). There are three types of validity: content validity, construct validity and criterion validity (De Vet et al., 2011).

4.2.1 *Content Validity*

Content validity is defined as "the degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured" (Mokkink et al., 2010). Content validity evaluates the relevance and comprehensiveness of the content (i.e., items) to the construct (De Vet et al., 2011). Evaluation of this type of validity is performed by asking a panel of experts, which in the case of patient-reported measures would be patients themselves, to provide insight on aspects important to the construct (De Vet et al., 2011). Subsequently, responses can then be quantified using a framework (e.g. the International Classification of Functioning, Disability and Health) (De Vet et al., 2011).

4.2.2 Criterion Validity

Criterion validity is defined as "the degree to which the scores of a measurement instrument are an adequate reflection of a gold standard" (Mokkink et al., 2010). Measures that are well-accepted by experts or are the longer version of the measure under study are considered to be gold standards (De Vet et al., 2011). There are two types of criterion validity: concurrent and predictive (De Vet et al., 2011). Concurrent validity examines the association between the instrument's and gold standard's score at the same time (i.e., concurrently) (De Vet et al., 2011). Predictive validity examines whether the instrument's score predicts the gold standard's score or the expected event/outcome (e.g., falls) in the future (De Vet et al., 2011). The statistical parameters used to calculate criterion validity for continuous measures (i.e., GPBMs) are: receiver operating characteristic (ROC) curves, Pearson's correlation coefficient (r), Spearman's correlation coefficient (rho), Bland-Altman Plots or ICC (De Vet et al., 2011). ROC curves are computed if the gold-standard is dichotomous, Spearman's r is calculated if the gold standard is ordinal or continuous, and Pearson's r, Bland-Altman or ICC are calculated if the gold standard is continuous (De Vet et al., 2011).

4.2.3 Construct Validity

Construct validity is assessed by establishing a hypothesis regarding the relationship between the instrument's scores and other measures or variables (De Vet et al., 2011). There are two types of construct validity utilized in psychometric evaluation of GPBMs: convergent validity and known-groups validity.

Convergent validity examines the relationship between the score on the measurement instrument under study and the score on an instrument measuring a similar construct. The statistical parameters used to calculate convergent validity are similar to criterion validity (ROC curves if comparator is dichotomous, Spearman's rho if comparator is ordinal or dichotomous, Pearson's r, Bland-Altman or ICC if comparator is continuous) (De Vet et al., 2011).

Known-groups validity is the ability of an instrument to discriminate between subgroups that are known to be different, for example, if a measure can discriminate between people with mild versus severe disabilities (De Vet et al., 2011). Statistical parameters used to calculate known-groups validity include mean differences (e.g., t-tests), ROC curves or effect sizes.

4.3 <u>Responsiveness</u>

Responsiveness is a form of validity; it is criterion and construct validity within a longitudinal context. COSMIN defines responsiveness as "the ability of an instrument to detect change over time in the construct to be measured" (Mokkink et al., 2010). Two main methods for assessing responsiveness are the criterion and construct approach. The criterion approach evaluates the relationship between change scores on the measurement instrument and the gold standard. Gold standards for patient-reported outcome measures may either be the longer version of a questionnaire or a global rating scale (GRS) (De Vet et al., 2011). The construct approach examines the relationship between the change in scores on the measurement instrument and the change in scores on another instrument, or the change in scores on the measurement instrument between different subgroup (e.g., different disease severities) (De Vet et al., 2011). Hypotheses regarding expected statistical outcomes should be made a priori (De Vet et al., 2011). When assessing responsiveness, similar statistical methods to validity should be employed (De Vet et al., 2011).

5.0 Rationale and Objectives of Thesis

COPD is a leading cause of death and disability in the world (World Health Organization, 2017). The disease not only impairs patients' well-being, but it also causes a significant burden on provincial healthcare costs. Due to these costs, policymakers and researchers rely on HRQoL measures to assess the cost-utility of different interventions for COPD; in order to efficiently allocate scarce healthcare resources. Cost-utility analysis is the most widely used method for economic evaluation as incremental costs of an intervention are compared to its incremental health improvement, expressed in QALYs (Brazier et al., 2007). The Canadian Agency for Drugs and Technologies in Health (2017) recommends the use of GPBMs to obtain the 'Q' in QALYs. GPBMs can be widely used across different populations and are easy to administer alongside other measures in clinical trials (Brazier et al., 2017). However, before these measures can be used to evaluate interventions for COPD, their psychometric properties need to be evaluated to ensure that they are reliable (i.e., provide the same outcomes in stable conditions), valid (i.e., accurately capture HRQoL) and responsive (i.e., accurately capture change in HRQoL over time). Therefore, the overall goal of this thesis was to evaluate the psychometric properties of GPBMs in individuals with COPD. The specific aims were:

1) To conduct a systematic review to examine the psychometric properties of GPBMs in individuals with COPD (Chapter 2) and;
2) To evaluate the content validity of GPBMs in individuals with COPD (Chapter 3).

Findings from these two studies will provide a comprehensive overview on the current performance of these measures in individuals with COPD and help inform the suitability of these measures for use in cost-utility analyses.

References

Agusti, A., Calverley, P. M. A., Celli, B., Coxson, H. O., Edwards, L. D., Lomas, D. A., MacNee, W., Miller, B. E., Rennard, S., Silverman, E. K., Tal-Singer, R., Wouters, E., Yates, J. C., Vestbo, J., & investigators, the E. of C. L. to I. P. S. E. (ECLIPSE). (2010). Characterisation of COPD heterogeneity in the ECLIPSE cohort. *Respiratory Research*, *11*(1), 122. https://doi.org/10.1186/1465-9921-11-122

Andresen, E. M., Rothenberg, B. M., & Kaplan, R. M. (1998). Performance of a Self-Administered Mailed Version of the Quality of Well-Being (QWB-SA) Questionnaire among Older Adults. *Medical Care*, *36*(9), 1349–1360.

Bansback, N., Tsuchiya, A., Brazier, J., & Anis, A. (2012). Canadian valuation of EQ-5D health states: Preliminary value set and considerations for future valuation studies. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0031115

Boyle, M. H., Torrance, G. W., Sinclair, J. C., & Horwood, S. P. (1983). Economic Evaluation of Neonatal Intensive Care of Very-Low-Birth-Weight Infants. *New England Journal of Medicine*. https://doi.org/10.1056/NEJM198306023082206

Brauer, C. A., Rosen, A. B., Greenberg, D., & Neumann, P. J. (2006). Trends in the measurement of health utilities in published cost-utility analyses. *Value in Health*. https://doi.org/10.1111/j.1524-4733.2006.00116.x

Brazier, J, Ara, R., Rowen, D., & Chevrou-Severac, H. (2017). *A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models*.

Brazier, J, Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford University Press.

Brazier, J, Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, *21*, 271–292.

Brooks, R., & De Charro, F. (1996). EuroQol: The current state of play. *Health Policy*. https://doi.org/10.1016/0168-8510(96)00822-6

Canadian Agency for Drugs and Technologies in Health. (2017). Guidelines for the Economic Evaluation of Health Technologies: Canada 4th Edition. *CADTH Methods and Guidelines*.

Çolak, Y., Nordestgaard, B. G., Vestbo, J., Lange, P., & Afzal, S. (2019). Prognostic significance of chronic respiratory symptoms in individuals with normal spirometry. *European Respiratory Journal*, *54*(3), 1900734.

De Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*.

EQ-5D-3L | Valuation. (2020). EuroQol Research Foundation. https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/valuation/

EQ-5D-3L User Guide. (2018). EuroQol Research Foundation. https://euroqol.org/publications/user-guides/

EQ-5D-5L User Guide. (2015). EuroQol Research Foundation. https://euroqol.org/publications/user-guides

EuroQol Group. (1990). EuroQol—A new facility for the measurement of health-related quality of life. *Health Policy*. https://doi.org/10.1016/0168-8510(90)90421-9

Fanshel, S., & Bush, J. W. (1970). A Health-Status Index and its Application to Health-Services Outcomes. *Operations Research*. https://doi.org/10.1287/opre.18.6.1021

Fayers, P. M., Hays, R., & Hays, R. D. (2005). *Assessing Quality of Life in Clinical Trials: Methods and Practice*. Oxford University Press. https://books.google.ca/books?id=wWRLN5U9-3YC

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., Denton, M., & Boyle, M. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care*. https://doi.org/10.1097/00005650-200202000-00006

Global Initiative for Chronic Obstructive Lung Disease. (2020). GLOBAL STRATEGY FOR THE DIAGNOSIS, MANAGEMENT, AND PREVENTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE (2020 REPORT). *Global Initiative for Chronic Obstructive Lung Disease*. https://doi.org/10.1097/00008483-200207000-00004

Horsman, J., Furlong, W., Feeny, D., & Torrance, G. (2003). The Health Utilities Index (HUI): Concepts, measurement properties and applications. *Health and Quality of Life Outcomes*, *1*, 54. https://doi.org/10.1186/1477-7525-1-54

Karimi, M., & Brazier, J. (2016). Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? *PharmacoEconomics*, *34*(7), 645–649. https://doi.org/10.1007/s40273-016-0389-9

Kharroubi, S. A., Brazier, J. E., Roberts, J., & O’Hagan, A. (2007). Modelling SF-6D health state preference data using a nonparametric Bayesian method. *Journal of Health Economics*. https://doi.org/10.1016/j.jhealeco.2006.09.002

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuspinar, A., & Mayo, N. E. (2013). Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? *Health and Quality of Life Outcomes*. https://doi.org/10.1186/1477-7525-11-71

Laupacis, A., Feeny, D., Detsky, A. S., & Tugwell, P. X. (1992). How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, *146*(4), 473–481.

MacNee, W. (2006). Pathology, pathogenesis, and pathophysiology. *BMJ*. https://doi.org/10.1136/bmj.332.7551.1202

Mannino, D. M., & Buist, A. S. (2007). *Global burden of COPD: risk factors, prevalence, and future trends*.

Martin, F., Camfield, L., Rodham, K., Kliempt, P., & Ruta, D. (2007). Twelve years' experience with the Patient Generated Index (PGI) of quality of life: A graded structured review. *Quality of Life Research*, *16*(4), 705–715. https://doi.org/10.1007/s11136-006-9152-6

Mayo, N. E. (2015). *ISOQOL Dictionary of Quality of Life and Health Outcomes Measurement*. ISOQOL. https://books.google.ca/books?id=cKjksgEACAAJ

McCarthy, B., Casey, D., Devane, D., Murphy, K., Murphy, E., & Lacasse, Y. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews, 2*.

Miravitlles, M., & Ribera, A. (2017). Understanding the impact of symptoms on the burden of COPD. *Respiratory Research*, *18*(1), 67. https://doi.org/10.1186/s12931-017-0548-3

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2010.02.006

Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care*, S217–S232.

Ranu, H., Wilde, M., & Madden, B. (2011). Pulmonary function tests. *The Ulster Medical Journal*, *80*(2), 84–90.

Read, J. L., Quinn, R. J., & Hoefer, M. A. (1987). Measuring overall health: An evaluation of three important approaches. *Journal of Chronic Diseases*, *40*, 7S-21S.

Report from the Canadian Chronic Disease Surveillance System: Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada. (2018).

Richardson, J., Sinha, K., Iezzi, A., & Khan, M. A. (2014). Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. *Quality of Life Research*, *23*(8), 2395–2404.

Ruta, D. A., Garratt, A. M., Leng, M., Russell, I. T., & Macdonald, L. M. (1994). A new approach to the measurement of quality of life the patient-generated index. *Medical Care*, 1109–1126. https://doi.org/10.1097/00005650-199411000-00004

Seiber, W. J., Groessl, E. J., David, K. M., Ganiats, T. G., & Kaplan, R. M. (2008). Quality of well being self-administered (QWB-SA) scale. *San Diego: Health Services Research Center, University of California*.

Sintonen, H. (2001). The 15D instrument of health-related quality of life: Properties and applications. *Annals of Medicine*, *33*(5), 328–336.

Smetanin, P., Stiff, D., Briante, C., Ahmad, S., Wong, L., & Ler, A. (2011). Life and Economic Impact of Lung Disease in Ontario: 2011 to 2041. *RiskAnalytica, on Behalf of the Ontario Lung Association*.

Statistics Canada. (2019). *Table 13-10-0394-01 Leading causes of death, total population, by age group*. CANSIM.

Tang, J. A., Oh, T., Scheer, J. K., & Parsa, A. T. (2014). The current trend of administering a patient-generated index in the oncological setting: A systematic review. *Oncology Reviews*, *8*(1), 245. https://doi.org/10.4081/oncol.2014.245

Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). Multiattribute Utility Function for a Comprehensive Health Status Classification System Health Utilities Index Mark 2. *Medical Care*. https://doi.org/10.1097/00005650-199607000-00004

Welte, T., & Groneberg, D. A. (2006). Asthma and COPD. *Experimental and Toxicologic Pathology*, *57*, 35–40. https://doi.org/10.1016/j.etp.2006.02.004

World Health Organization. (2017). *Chronic obstructive pulmonary disease (COPD)*. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., Poissant, L., Johnson, J. A., & Group, C. E.-5D-5L V. S. (2016). A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. *Medical Care*, *54*(1), 98–105. https://doi.org/10.1097/MLR.0000000000000447

Zamzam, M. A., Azab, N. Y., El Wahsh, R. A., Ragab, A. Z., & Allam, E. M. (2012). Quality of life in COPD patients. *Egyptian Journal of Chest Diseases and Tuberculosis*, *61*(4), 281–289.

CHAPTER 2

Measurement Properties of Preference-Based Measures for Economic Evaluation in COPD: A Systematic Review

Measurement Properties of Preference-Based Measures for Economic Evaluation in COPD: A Systematic Review

Ava Mehdipour, BSc[1]; Marla K. Beauchamp, PT, PhD[1,2,3]; Joshua Wald, MD[3,4]; Nicole Peters, BSc[1]; Ayse Kuspinar, PT, PhD[1]

[1]School of Rehabilitation Science, McMaster University, 1400 Main St W, Hamilton, ON L8S 1C7, Canada
[2]Respiratory Research, West Park Healthcare Centre, Toronto, ON M6M 2J5, Canada
[3]Firestone Institute for Respiratory Health, 50 Charlton Ave E, Hamilton, ON L8N 4A6, Canada
[4]Department of Medicine, McMaster University, Hamilton, ON, Canada

Published in *Quality of Life Research.*

Reference:
Mehdipour, A., Beauchamp, M.K., Wald, J., Peters, N., Kuspinar, A. Measurement properties of preference-based measures for economic evaluation in COPD: a systematic review. Qual Life Res (2020). https://doi.org/10.1007/s11136-020-02569-4

Corresponding Author:
Ayse Kuspinar, PT, PhD
Assistant Professor
School of Rehabilitation Science
McMaster University
1400 Main St. W. Room 435, IAHS
L8S 1C7
Hamilton, Ontario
Canada
kuspinaa@mcmaster.ca

Abstract

Purpose: Preference-based measures can provide measurements of health-related quality of life and be utilized for cost-effectiveness analyses of interventions in individuals with chronic obstructive pulmonary disease (COPD). The purpose of this study is to evaluate whether generic preference-based measures are reliable, valid and responsive in COPD. *Methods:* A systematic review was performed using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) guidelines. Three databases were searched: MEDLINE, EMBASE and CINAHL. Studies were included if the sample represented individuals with COPD and the aim was to evaluate one or more psychometric properties or the interpretability of generic preference-based measures. *Results:* Six-hundred and sixty-seven abstracts were screened, 65 full-text articles were reviewed and 24 articles met the inclusion criteria. Measures which emerged from the search were: the EQ-5D, the SF-6D, the Quality of Well-being scale, the 15D and the Health Utilities Index 3. Evidence for the test-retest reliability of these measures was limited. Construct validity of the measures was well-supported with correlations with generic health profiles being 0.37-0.68, and correlations with COPD-specific health profiles being 0.53-0.75. Evidence for known-groups validity of these measures was poor and data on responsiveness was mixed. *Conclusion:* Generic preference-based measures' sensitivity to change and ability to discriminate between different disease severities in COPD was poorly supported. Future research may consider examining the development of COPD-specific preference-based measures that may allow for a more accurate detection of change and discrimination amongst disease severities to facilitate cost-effectiveness evaluations.

Keywords: 'Chronic Obstructive Pulmonary Disease', 'Health-Related Quality of Life', 'Psychometric Properties', 'Economic Evaluation'

Declarations:
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Conflicts of interest: There are no conflicts of interest to declare.
Availability of data and material: The data that supports the findings of this study are available in the supplementary material of this article.

Abbreviations list:

AUC, Area under the curve; COPD, Chronic obstructive pulmonary disease; COSMIN, Consensus-based standards for the selection of health measurement instruments; ES, Effect size; GPBM, Generic preference-based measure; GRADE, Grading of recommendations assessment, development, and evaluation; HUI2 & HUI3, Health utilities index mark 2 & 3; HRQoL, Health-related quality of life; MID, Minimal important difference; QWB, Quality of well-being; SRM, Standardized response mean

Introduction

Chronic obstructive pulmonary disease (COPD) is a highly prevalent and costly condition characterized by chronic airflow limitation due to a mixture of chronic bronchitis and emphysema, caused by exposure to noxious particles (e.g., cigarette smoke, air pollutants) [1]. Individuals with COPD experience symptoms such as dyspnea, cough, sputum production, wheezing, chest tightness, and fatigue [1]. These symptoms impact physical activity, mental health, and overall quality of life [2]. A variety of pharmacological and non-pharmacological interventions have been shown to reduce symptoms and increase quality of life [3, 4].

Generic preference-based measures (GPBMs) are health-related quality of life (HRQoL) measures developed using the general population's preferences for health states, with the intention of comparing quality of life across different interventions and different health conditions [5, 6]. GPBMs are anchored at 0.0 (death) and 1.0 (perfect-health) with some health state values being worse than death [6]. GPBM scores can be used to calculate quality-adjusted life years for an intervention by multiplying them by the number of years the intervention is predicted to extend life [5]. GPBMs can help identify interventions that are most cost-effective and have the highest impact on quality of life. They can be utilized by healthcare professionals and policymakers to make decisions about resource allocation and implementation of different treatment options [7]. GPBMs have also been used as quality indicators for hospitals and health care professionals, as well as measures of inequalities [8, 9].

Existing research evaluating health status for cost-effectiveness analyses in COPD have utilized GBPMs developed based on the general population. Since these measures are generic and were not developed specifically for individuals with COPD, it is important to assess their psychometric properties in this population [10, 11]. The aim of this systematic review is to examine the psychometric properties of GPBMs in people with COPD.

Methods

Search strategy

This review was performed following Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) guidelines. Three different electronic databases were searched: MEDLINE (1946 to July 8, 2019), EMBASE (1974 to July 8, 2019), and CINAHL (1981 to July 8, 2019). Search terms covered (1) the population (COPD), (2) recognized GPBMs: EQ-5D, SF-6D, Quality of Well-being (QWB) scale, the 15D, the Assessment of Quality of Life, the Health Utilities Index Mark 2 & 3 (HUI2, HUI3), and (3) measurement properties and characteristics (using the search filter developed by Terwee et al. [12]) (Online Resource 1, Table 1). Medical subject heading terms were first employed and if they were not available, keyword search terms were employed. Titles/abstracts and full-text were screened by two independent reviewers and reasons for exclusion were recorded, differences were discussed, and consensus was reached.

Studies were included if 1) the sample represented individuals with COPD (at least 80% had a clinical diagnosis of COPD); 2) they included one or more GPBMs, and 3) the aim was to evaluate one or more psychometric properties or the interpretability of GPBMs. Gray literature (e.g., meeting/conference proceedings/abstracts) and previous reviews were excluded, and only peer-reviewed articles in English were examined.

The review's protocol can be accessed on PROSPERO (registration number: CRD42019131061).

Data extraction and quality assessment

In addition to study characteristics (country, sample size, age, forced expiratory volume (FEV_1) % and utility value) and feasibility (% of completed data), the following measurement properties were extracted from the included studies:

- *Reliability*; *test-retest reliability*: the extent to which scores for stable individuals at different time points are the same [13].
- *Content validity*: the degree to which the content of an instrument reflects the intended construct [14].
- *Construct validity*
  - *Convergent validity*: the degree to which two instruments measuring a similar construct relate [15].
  - *Known-groups validity*: the degree to which an instrument can discriminate between two groups known to differ [16].
- *Predictive validity*: the ability of an instrument to measure an outcome in the future [10].
- *Responsiveness:* the ability of an instrument to detect change in a construct overtime [14].
- *Interpretability:* the qualitative meaning of scores on an instrument (i.e., distribution of scores (floor/ceiling effects) and minimal important difference (MID)) [14].

The methodological quality of each included study was assessed using the COSMIN risk of bias checklist [17]. The checklist consists of 10 boxes, one for each measurement property. The boxes examined for this review were Box 2. Content validity, Box 6. Reliability, Box 8. Criterion validity, Box 9. Hypothesis testing for construct validity, and Box 10. Responsiveness. Each box consisted of a few questions examining the methodological quality of the design and each aspect of the design was rated as very good, adequate, doubtful, or inadequate. The overall rating for each property was determined by taking the lowest rating out of all the items for the respective box. The methodological quality of each study was rated independently by two reviewers and any disagreements were addressed through discussion.

Subsequently, the result of each measurement property per study was rated against COSMIN's criteria for good measurement properties (Online Resource 1, Table 2) [18]. COSMIN's criteria rate results as sufficient, insufficient, or indeterminate based on whether they met previously defined hypotheses set by COSMIN or the research team. If a hypothesis was met, a sufficient rating was given, and if not, an insufficient rating was given. An indeterminate rating was given if hypotheses were not defined a priori or a psychometric value was not reported. For construct validity and responsiveness, the research team formed hypotheses about the results so that 1) all results were comparable to the same relevant hypotheses, and 2) studies that did not define hypotheses a priori did not receive an inadequate risk of bias rating (Online Resource 1, Tables 3-7) [18]. Reliability correlation coefficients were hypothesized to be greater or equal to 0.70 [18]. For predictive validity, areas under the curve (AUCs) were hypothesized to be greater or equal to 0.70 [18]. Hypotheses for correlations were that measures assessing similar constructs (e.g., HRQoL) should be ≥0.50, and measures assessing related but dissimilar constructs (e.g., performance/function/disease severity) should be 0.30-0.50 [18]. For known-groups validity, it was hypothesized for the AUC to be greater or equal to 0.70 or differences in means to be statistically significant (5% significance level) between groups of different pre-determined variables (e.g., GOLD stage severity). For responsiveness, a significant difference at 5% significance level was hypothesized between initial and follow-up means, over a period of expected change. Effect sizes (ESs) and standardized response means (SRMs) were interpreted using Cohen's *d* (0.2=small, 0.5=medium, 0.8=large) [19]. The rating for each result was also performed independently by the two reviewers.

Data synthesis

Results were either quantitatively pooled or qualitatively summarized (per measurement property per GPBM). If studies were homogenous in design, had at least adequate methodological quality, or did not have conflicting results, then they were quantitatively pooled [11, 18]. If these criteria were not met or studies could not be statistically pooled, then results were qualitatively summarized [18, 20], and either mean ranges, percentage of confirmed hypotheses, or both, were reported.

The pooled/summarized results for each measurement property per GPBM were rated against COSMIN's criteria for good measurement properties (Online Resource 1, Table 2) [18, 21]. The overall rating given was either sufficient, insufficient, or indeterminate. For construct validity and responsiveness, if ≥ 75% of hypotheses were consistent (sufficient or insufficient), then the overall rating was either sufficient or insufficient [18]. If results were inconsistent (e.g., both sufficient and insufficient), then the rating was based on the statistical cut-off (e.g., AUC) or the majority of the ratings (e.g., hypothesis testing). Moreover, each pooled/summarized result was graded using COSMIN's modified Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to evaluate the quality of evidence (Online Resource 1, Table 8) [18, 21]. The GRADE approach was based on four factors:

(1) Risk of bias (Online Resource 1, Table 9)

(2) Inconsistency (only for inconsistent ratings):

   a. Serious: if ≥ 50% of results were rated as sufficient according to COSMIN's criteria for good measurement properties [18].

   b. Very Serious: if < 50% of results were rated as sufficient according to COSMIN's criteria for good measurement properties [18].

(3) Imprecision

   a. Serious: if total sample size is between 50-100 [18].

   b. Very Serious: if total sample size is less than 50 [18].

(4) Indirectness

   a. Serious: if other populations were also examined or none of the comparison measures examined quality of life or HRQoL (for convergent validity and responsiveness) in the study.

   b. Very Serious: if other populations were also examined *and* none of the comparison measures examined quality of life or HRQoL (for convergent validity and responsiveness) in the study.

Results

Selection process

A total number of 908 articles were identified through the databases, and 231 articles were removed due to duplication. Six hundred and seventy-seven titles and abstracts were screened, and 612 were excluded because (1) the purpose of the study was not to evaluate psychometric properties or examine interpretability of the measures, (2) they were conference proceedings or abstracts, (3) they were not in English, (4) they were not examining GPBMs, or (5) the sample was not exclusive to people with COPD. From this, 65 articles were left for full-text screening, and out of these articles 41 were excluded because: (1) the purpose of the study was not to evaluate psychometric properties or examine interpretability of the measures, (2) study was not examining a GPBM, or (3) the sample was not exclusive to people with COPD. Figure 1 outlines the process from the initial records identified to the final number of full-text articles included in the review.

Out of the 24 included studies, 17 studies [22–38] examined the EQ-5D, 5 studies [22, 25, 26, 33, 39] examined the SF-6D, 3 studies [40–42] examined the QWB scale, 3 studies [37, 43, 44] examined the 15D, and 1 study [45] examined the HUI3. There were no studies that emerged examining the psychometric properties of the Assessment of Quality of Life or the HUI2. Online Resource 2 outlines the sample characteristics and measurement properties for each study.

Sample characteristics

A total of 9914 patients with COPD were included across the studies. There was wide variability in sample size with sample sizes as small as 18 to as large as 2291 in studies using the EQ-5D. Mean FEV1% in the individual studies ranged from a low of 32.7 % in studies using the EQ-5D and the SF-6D to a high of 58.6% in studies using the 15D. Sample characteristics and mean scores of individual studies can be found in Online Resource 2.

Results of measurement properties

Tables 1, 2 ,3, 4, and 5 provide a summary of the measurement properties reported and their overall rating for each measure. Online Resource 3 Tables 1-5 provide a detailed summary, with corresponding methodological quality and rating for each study. For each measure and property, studies varied in their methodological quality, ranging from inadequate to very good (Online Resource 3, Tables 1-5); therefore, the results were only qualitatively analyzed [18, 20].

The test-retest reliability interval varied between the studies, from one day to two years. Correlation coefficients for the QWB scale and the 15D met the acceptable cut-off of 0.70; however, correlation coefficients for the EQ-5D barely met the cut-off (0.67-0.73) [37, 38]. There were no studies that reported on the test-retest reliability of the SF-6D or the HUI3.

There were no studies evaluating the content validity of GPBMs.

For convergent validity, correlations between the EQ-5D and SF-6D and 15D were 0.40-0.75 [22, 25, 26, 33] and 0.65 [37] (respectively). Correlations between the EQ-5D and generic health profiles ranged from 0.37-0.68 [28, 32, 37] and correlations with disease-specific health profiles ranged from 0.53-0.70 [23–25, 27, 29, 32, 33]. Correlations between the SF-6D and disease-specific health profiles ranged from 0.57-0.75 [25, 33]. Correlation between the 15D and a generic health profile was approximately 0.60 [37], and with a disease-specific health profile was 0.71 [44]. Considering that GPBMs and health profiles vary in descriptive systems, it is important to consider these correlations with their respective methodological quality (Online Resource 3 Tables 1-5). Online Resource 4 Tables 1-3 outline the overlap of the descriptive systems between the GPBMs and health profiles. There were no studies that reported on the convergent validity of the HUI3.

For known-groups validity, 5 studies [22, 24, 27, 28, 36] found statistical differences in EQ-5D scores between GOLD stages and 6 studies [22–24, 32, 33, 36] reported no statistical differences in scores between GOLD stages. Among these 6 studies, 4 reported that the EQ-5D was not able to differentiate between GOLD stage 2 (moderate airflow obstruction) and 3 (severe airflow obstruction) [22–24, 36] and 2 studies [24, 33] reported the measure was not able to differentiate between GOLD stage 3 (severe airflow obstruction) and 4 (very severe airflow obstruction). For the SF-6D, evidence for differences in utility scores between GOLD stages was found in 2 studies; Thuppal et al. [22] found differences between very severe (GOLD stage 4) and other severities (GOLD stage 1-3) of airflow obstruction (p=0.0187), and Menn et al. [33] found differences between GOLD stages 3 and 4 (p=0.003). There were no studies that reported on the known-groups validity of the QWB, the 15D or the HUI3.

Predictive validity was only evaluated for 1 GPBM; the 15D. Koskela et al. [43] evaluated whether baseline 15D scores were able to predict future declines in HRQoL (over 5 years) by examining receiver operating characteristic curves. The AUC value was 0.83, above the acceptable cut-off of 0.70.

The responsiveness of the EQ-5D was evaluated in 8 out of 17 studies, in relation to events expected to improve individuals' health states and/or daily activities (e.g., pulmonary rehabilitation) or events expected to significantly reduce health states and activities (e.g., exacerbations). Thuppal et al. [22] evaluated EQ-5D scores in patients undergoing lung volume reduction surgery, an intervention proven to improve symptoms and exercise tolerance in selected patients, and reported a medium ES of 0.52. Nolan et al. [23] reported a SRM of 0.39 for EQ-5D after 8 weeks of pulmonary rehabilitation and correlations ranging from 0.14 to 0.40 with changes in COPD-specific health profiles. Ringbaek et al. [35] found a difference between utility scores after 7 weeks of pulmonary rehabilitation (p=0.034), but not after 3 months' post-rehabilitation (p=0.18). Two studies [31, 33] evaluated the responsiveness of the EQ-5D for exacerbation events and a medium ES of 0.69 and SRM of 0.65 were reported. Four studies [30, 31, 37, 38] used anchors of participant-perceived health change to assess responsiveness and there were no differences in the mean change in utility scores between the anchor-based categories (i.e., improving, staying the same, worsening). For the SF-6D, Thuppal et al. [22] reported a medium ES of 0.64 for patients undergoing a lung volume reduction surgery and Menn et al. [33] reported a small ES of 0.27 for an exacerbation event. For the QWB scale, Kaplan et al. [42] reported correlations ranging from 0.31 to 0.42 between change in QWB scores and exercise tolerance, self-efficacy, and walking compliance after 3 months. Stavem [37] evaluated the responsiveness of the 15D using a global rating of change scale and found differences in the mean change in utility scores between the three groups (better, unchanged, and worse) (p=0.004), and a large ES and responsiveness statistic for the 'better' group (ES=1.00, responsiveness statistic= 1.51). Puhan et al. [45] evaluated the responsiveness of the HUI3 in individuals receiving 12 weeks of respiratory rehabilitation and reported a SMR of 0.20.

Feasibility and interpretability

Feasibility of the EQ-5D (6 out of 17 studies) and the SF-6D (1 out 5 studies) was evaluated. Studies on the EQ-5D reported a completion rate of 92-100% [24, 28, 30, 33, 35, 38]. A study on the SF-6D reported a lower completion rate of 58-60% [33]. Ceiling effects were reported for the EQ-5D, by 5 out of 17 studies, with 17.9-43.1% reporting best health [24–26, 35, 36]. A MID of 0.051 and 0.010 was reported for the EQ-5D and the SF-6D, respectively, using anchor-based methods [23, 39], and a MID of 0.03 was reported for the QWB scale using statistical methods [40].

Quality assessment

Quality of evidence for each GPBM can be found in Tables 1, 2, 3, 4, and 5. Quality of evidence for test-retest reliability, known-groups validity, and responsiveness mainly ranged from very low to low, with the exception of moderate quality for the SF-6D's responsiveness (Table 1, 3, 5). Quality of evidence for convergent and predictive validity was generally moderate, with the exception of very low for the QWB scale's convergent validity (Table 2,4).

Discussion

The purpose of this review was to examine the measurement properties of GPBMs in people with COPD to evaluate whether these measures are appropriate for obtaining reliable and valid quality of life scores for economic decision-making. Overall, results from this review suggest limited and low-quality evidence supporting reliability and known-groups validity. Responsiveness, a property crucial for assessing the effects of interventions, which is the intended purpose of GPBMs, is poorly supported as current evidence is low quality and underlying methods of evaluating responsiveness are mainly incorrect. These findings highlight the need for rigorously designed studies evaluating the psychometric properties of GPBMs in COPD and/or the need to develop disease-specific preference-based measures that may be more sensitive in this population.

There was limited evidence for the test-retest reliability of these measures, with only three measures (EQ-5D, QWB, and 15D) examined, with low to very low quality. Values were around or above the expected cut-off; however, it is important to note that ICC/Spearman's correlation values for the EQ-5D were borderline. The reliability of GPBMs should be examined further with appropriate statistical tests (i.e., using ICC as opposed to Pearson's or Spearman's correlation coefficients) and more rigorous designs [11].

Even though convergent validity was sufficiently supported by GPBMs and moderate quality of evidence was reported, it is important to note that these values could have been affected by differences in descriptive systems. For example, when scores produced by a GPBM were compared against a health profile (such as the Chronic Respiratory Questionnaire), the former uses a preference-weighted scoring system, whereas the latter uses a summative scoring system (i.e., response levels are coded numerically and the sum is taken) [46, 47]. Furthermore, in terms of GPBMs, differences exist between the measures in terms of both descriptive systems (e.g., dimensions covered) and valuation methods (e.g., Time Trade-Off vs. Standard Gamble vs. Visual Analogue Scale), which in turn may affect comparability between them [6].

The ability of GPBMs to discriminate between different clinical states (e.g., disease severity) was not strongly supported by the literature. This property was only evaluated in two GPBMs: the EQ-5D and the SF-6D, and low quality of evidence was reported for both measures. This property was better supported in the SF-6D, with all studies providing evidence of the SF-6D's ability to discriminate, compared to the EQ-5D; however, it is important to note that the number of studies for the SF-6D was limited (4 vs. 12 for the EQ-5D). The EQ-5D had an adequate amount of studies (7/12) demonstrating that it was unable to differentiate between different disease severities. A limitation regarding studies reporting on known-groups validity was that the GOLD numerical staging was utilized to classify disease severity, as opposed to recent alphabetical staging which considers airflow obstruction, symptoms, and exacerbations, providing a more accurate classification of disease severity [1].

Evidence to support the responsiveness of GPBMs was weak with mainly low to very low quality. The EQ-5D was not responsive to rehabilitation or changes in health status over time, but was responsive to lung reduction surgery and exacerbation events. The responsiveness of the SF-6D was only assessed in response to post-lung volume reduction surgery and post-exacerbation, and was similarly found to be sensitive to change. Evidence to support the 15D's responsiveness was limited but findings showed that the 15D was able to capture improvements in health but not deteriorations in health [37]. The HUI3 lacked sensitivity to change as it was not able to fully capture improvements in health after respiratory rehabilitation in comparison to disease-specific measures of HRQoL [45].

The recommended guidelines for evaluating responsiveness of measures, according to COSMIN, are to examine correlations between changes in scores with a global rating scale or another measure known to be responsive in the same population [11]. Among the 13 studies that assessed responsiveness in our review, only 2 [23, 42] used this recommended approach and assessed correlations with other measures.

A systematic review was performed by Petrillo et al. [48] approximately 10 years ago that reported on the validity and responsiveness of condition-specific health profiles and multi-attribute preference-based measures in COPD. This review built on a review conducted in 2007 by Pickard et al. [49], examining the psychometric properties of the EQ-5D in asthma and COPD. While the review by Petrillo et al. [48] was an important contribution, it examined only 2 databases (PubMed and EMBASE) and did not follow COSMIN guidelines nor evaluated the quality of the studies. It solely evaluated the responsiveness and known-groups validity of these measures and highlighted studies concerned with exacerbations. Our review involved searching 3 databases using COSMIN's comprehensive search strategy for measurement studies, evaluated all types of measurement properties, and included new literature published since 2009. Similar to our review, both Pickard et al. [49] and Petrillo et al. [48] observed ceiling effects and limited known-groups validity for the EQ-5D. Reduced responsiveness when evaluating subtle but important changes in health was also observed by Petrillo et al. [48]. Moreover, broader reviews have been performed evaluating the psychometric properties of GPBMs and similar to our review, further rigorous testing has been recommended [50, 51]. However, the aims of these reviews were different; for example, Finch et al. [50] performed a review of reviews to evaluate the overall validity and responsiveness of 5 common

GPBMs and Qian et al. [51] sought to evaluate the construct validity, reliability, and responsiveness of GPBMs used in Asian countries. Our review is the first to provide a systematic and comprehensive evaluation of GPBMs' psychometric properties specifically in individuals with COPD.

Information that was missing in the literature on GPBMs was evaluation of content validity and predictive validity. Content validation, a fundamental component of validity [11], was not evaluated in any of the studies examining the psychometric properties of these measures. Future research should assess this property to examine whether items on GPBMs reflect areas of HRQoL affected in people with COPD. Only 1 study examined predictive validity and it reported very good predictive validity for the 15D; however, future research should examine this property in other GPBMs.

GPBMs possess many attributes that are useful for clinical and economic evaluations. Not only do they assess HRQoL, they provide a single index value which can be utilized by policymakers to make decisions regarding healthcare resources [7]. Selecting the intervention that increases good quality of life years while being cost-effective is beneficial for both patients and society. Our review revealed that the most widely used preference-based measure in COPD was the EQ-5D; however, this measure may not optimally reflect the particular disabilities and health concerns of people with COPD [52]. The EQ-5D demonstrated ceiling effects in COPD, which can leave less room for improvement when assessing response to treatment. There were also concerns about the EQ-5D's ability to discriminate between people with different levels of disease severity. In general, studies comparing the known-groups validity and responsiveness of GPBMs against disease-specific health profiles found COPD-specific health profiles to perform better than GPBMs [24, 30, 32, 33, 35, 38, 45]. These results suggest that a preference-based measure specific to people with COPD may be an area for further exploration in future studies [53].

Acknowledgements: All authors (AM, MKB, JW, NP, AK) contributed to the study's conception and/or design. Screening of articles and quality assessments were performed by AM and NP. The first draft of the manuscript was written by AM. Preliminary edits were made by AK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript, and AM is the guarantor of the paper.

References

1. *Global Strategy for Prevention, Diagnosis and Management of COPD*. (2019).
2. Miravitlles, M., & Ribera, A. (2017). Understanding the impact of symptoms on the burden of COPD. *Respiratory Research*, *18*(1), 67. https://doi.org/10.1186/s12931-017-0548-3
3. Perng, D. W., Tao, C. W., Su, K. C., Tsai, C. C., Liu, L. Y., & Lee, Y. C. (2009). Anti-inflammatory effects of salmeterol/fluticasone, tiotropium/fluticasone or tiotropium in COPD. *European Respiratory Journal*. https://doi.org/10.1183/09031936.00115308
4. Boueri, F. M. V., Bucher-Bartelson, B. L., Glenn, K. A., & Make, B. J. M. (2001). Quality of life measured with a generic instrument (Short Form-36) improves following pulmonary rehabilitation in patients with COPD. *Chest*. https://doi.org/10.1378/chest.119.1.77
5. Neumann, P. J., Goldie, S. J., & Weinstein, M. C. (2000). Preference-Based Measures in Economic Evaluation in Health Care. *Annual Review of Public Health*. https://doi.org/10.1146/annurev.publhealth.21.1.587
6. Brazier, J., Ara, R., Rowen, D., & Chevrou-Severac, H. (2017). A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. *PharmacoEconomics*. https://doi.org/10.1007/s40273-017-0545-x
7. Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin*. https://doi.org/10.1093/bmb/ldq033
8. Gutacker, N., Bojke, C., Daidone, S., Devlin, N., & Street, A. (2013). Hospital variation in patient-reported outcomes at the level of EQ-5D dimensions: Evidence from England. *Medical Decision Making*. https://doi.org/10.1177/0272989X13482523
9. Zhou, Z., Fang, Y., Zhou, Z., Li, D., Wang, D., Li, Y., … Chen, G. (2017). Assessing Income-Related Health Inequality and Horizontal Inequity in China. *Social Indicators Research*. https://doi.org/10.1007/s11205-015-1221-1
10. Bolarinwa, O. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*. https://doi.org/10.4103/1117-1936.173959
11. De Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. *Measurement in Medicine: A Practical Guide*. https://doi.org/10.1017/CBO9780511996214
12. Terwee, C. B., Jansma, E. P., Riphagen, I. I., & De Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*. https://doi.org/10.1007/s11136-009-9528-5
13. Vilagut, G. (2014). Test-Retest Reliability BT - Encyclopedia of Quality of Life and Well-Being Research. In A. C. Michalos (Ed.), (pp. 6622–6625). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_3001
14. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2010.02.006
15. Chin, C.-L., & Yao, G. (2014). Convergent Validity BT - Encyclopedia of Quality of Life and Well-Being Research. In A. C. Michalos (Ed.), (pp. 1275–1276). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_573
16. Davidson, M. (2014). Known-Groups Validity BT - Encyclopedia of Quality of Life and Well-Being Research. In A. C. Michalos (Ed.), (pp. 3481–3482). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_1581
17. Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1765-4
18. Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., De Vet, H. C. W., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of Patient - Reported Outcome Measures (PROMs). User Manual, (February), 1–78.
19. Cohen J. (1988). *Statistical power analysis for the behavioural science (2nd Edition).* Hillsdale, NJ: Lawrence Erlbaum Associates.

20. Deeks, J. J. (2011). *Chapter 9: Analysing data and undertaking meta-analyses. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)*.

21. Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. (2013). GRADE handbook for grading quality of evidence and strength of recommendations. *The GRADE Working Group*.

22. Thuppal, S., Markwell, S., Crabtree, T., & Hazelrigg, S. (2019). Comparison between the EQ-5D-3L and the SF-6D quality of life (QOL) questionnaires in patients with chronic obstructive pulmonary disease (COPD) undergoing lung volume reduction surgery (LVRS). *Quality of Life Research*, *28*(7), 1885–1892. https://doi.org/10.1007/s11136-019-02123-x

23. Nolan, C. M., Longworth, L., Lord, J., Canavan, J. L., Jones, S. E., Kon, S. S. C., & Man, W. D. C. (2016). The EQ-5D-5L health status questionnaire in COPD: Validity, responsiveness and minimum important difference. *Thorax*, *71*(6), 493–500. https://doi.org/10.1136/thoraxjnl-2015-207782

24. Wacker, M. E., Jörres, R. A., Karch, A., Wilke, S., Heinrich, J., Karrasch, S., … Holle, R. (2016). Assessing health-related quality of life in COPD: Comparing generic and disease-specific instruments with focus on comorbidities. *BMC Pulmonary Medicine*, *16*(1), 1–11. https://doi.org/10.1186/s12890-016-0238-9

25. Chen, J., Wong, C. K. H., M McGhee, S., Pang, P. K. P., & Yu, W. C. (2014). A comparison between the EQ-5D and the SF-6D in patients with chronic obstructive pulmonary disease (COPD). *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112389

26. Ferreira, L. N., Ferreira, P. L., & Pereira, L. N. (2014). Comparing the Performance of the SF-6D and the EQ-5D in Different Patient Groups. *Acta Médica Portuguesa*, *27*(2), 236. https://doi.org/10.20344/amp.4057

27. Kim, S. H., Oh, Y. M., & Jo, M. W. (2014). Health-related quality of life in chronic obstructive pulmonary disease patients in Korea. *Health and Quality of Life Outcomes*, *12*(1), 1–7. https://doi.org/10.1186/1477-7525-12-57

28. Lin, F. J., Pickard, A. S., Krishnan, J. A., M.J., J., D.H., A., S.S., C., … W.M., V. (2014). Measuring health-related quality of life in chronic obstructive pulmonary disease: properties of the EQ-5D-5L and PROMIS-43 short form. *BMC medical research methodology*, *14*, 78. http://dx.doi.org/10.1186/1471-2288-14-78

29. Manca, S., Rodriguez, E., Huerta, A., Torres, M., Lazaro, L., Curi, S., … Miravitlles, M. (2014). Usefulness of the CAT, LCOPD, EQ-5D and COPDSS scales in understanding the impact of lung disease in patients with Alpha-1 antitrypsin deficiency. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *11*(5), 480–488. https://doi.org/10.3109/15412555.2014.898030

30. Peters, M., Crocker, H., S., D., C., J., H., D., & R., F. (2014). Change in health status in long-term conditions over a one year period: A cohort survey using patient-reported outcome measures. *Health and Quality of Life Outcomes*, *12*(1), 1–10.

31. Goossens, L. M. A., Nivens, M. C., Sachs, P., Monz, B. U., & Rutten-Van Mölken, M. P. M. H. (2011). Is the EQ-5D responsive to recovery from a moderate COPD exacerbation? *Respiratory Medicine*, *105*(8), 1195–1202. https://doi.org/10.1016/j.rmed.2011.02.018

32. Pickard, A. S., Yang, Y., & Lee, T. A. (2011). Comparison of health-related quality of life measures in chronic obstructive pulmonary disease. *Health and Quality of Life Outcomes*. https://doi.org/10.1186/1477-7525-9-26

33. Menn, P., Weber, N., & Holle, R. (2010). Health-related quality of life in patients with severe COPD hospitalized for exacerbations - comparing EQ-5D, SF-12 and SGRQ. *Health and Quality of Life Outcomes*, *8*, 1–9. https://doi.org/10.1186/1477-7525-8-39

34. Polley, L., Yaman, N., Heaney, L., Cardwell, C., Murtagh, E., Ramsey, J., … McGarvey, L. (2008). Impact of cough across different chronic respiratory diseases: Comparison of two cough-specific health-related quality of life questionnaires. *Chest*, *134*(2), 295–302. https://doi.org/10.1378/chest.07-0141

35. Ringbaek, T., Brøndum, E., Martinez, G., & Lange, P. (2008). EuroQoL in assessment of the effect of pulmonary rehabilitation COPD patients. *Respiratory Medicine*, *102*(11), 1563–1567. https://doi.org/10.1016/j.rmed.2008.06.016

36. Rutten-Van Mölken, M. P. M. H., Oostenbrink, J. B., Tashkin, D. P., Burkhart, D., & Monz, B. U. (2006). Does quality of life of COPD patients as measured by the generic EuroQol five-dimension questionnaire differentiate between COPD severity stages? *Chest*, *130*(4), 1117–1128. https://doi.org/10.1378/chest.130.4.1117

37. Stavem, K. (1999). Reliability, validity and responsiveness of two multiattribute utility measures in patients with chronic obstructive pulmonary disease. *Quality of Life Research*, *8*(1–2), 45–54. https://doi.org/10.1023/A:1026475531996

38. Harper, R., Brazier, J. E., Waterhouse, J. C., Walters, S. J., N M B Jones, & P Howard. (1997). Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. *Thorax*, *52*(10), 879–887. https://doi.org/10.1136/thx.52.10.879

39. Walters, S. J., & Brazier, J. E. (2003). What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes*, *1*(4), 1–8.

40. Kaplan, R. M. (2005). The minimally clinically important difference in generic utility-based measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *2*(1), 91–97. https://doi.org/10.1081/COPD-200052090

41. Anderson, J. P., Kaplan, R. M., Berry, C. C., Bush, J. W., & Rumbaut, R. G. (1989). Interday Reliability of Function Assessment for a Health Status Measure : The Quality of Well-Being Scale, *27*(11), 1076–1084.

42. Kaplan, RM., Atkins, CJ., & R., T. (1984). Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *Journal of chronic diseases*, *37*(2 PG-85–95), 85–95.

43. Koskela, J., Kupiainen, H., Kilpeläinen, M., Lindqvist, A., Sintonen, H., Pitkäniemi, J., & Laitinen, T. (2014). Longitudinal HRQoL shows divergent trends and identifies constant decliners in asthma and COPD. *Respiratory Medicine*, *108*(3), 463–471. https://doi.org/10.1016/j.rmed.2013.12.001

44. Mazur, W., Kupiainen, H., Pitkaniemi, J., M., K., H., S., A., L., … Laitinen, T. (2011). Comparison between the disease-specific Airways Questionnaire 20 and the generic 15D instruments in COPD. *Health and Quality of Life Outcomes*, *9*, 4. http://dx.doi.org/10.1186/1477-7525-9-4

45. Puhan, M. A., Guyatt, G. H., Goldstein, R., Mador, J., McKim, D., Stahl, E., … Schünemann, H. J. (2007). Relative responsiveness of the Chronic Respiratory Questionnaire, St. Georges Respiratory Questionnaire and four other health-related quality of life instruments for patients with chronic lung disease. *Respiratory Medicine*, *101*(2), 308–316. https://doi.org/10.1016/j.rmed.2006.04.023

46. Brazier, J., Ratcliffe, J., Salomon, J. A., & Tsuchiya, A. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation*. OUP Oxford.

47. Chauvin, A., Rupley, L., Meyers, K., Johnson, K., & Eason, J. (2008). Outcomes in Cardiopulmonary Physical Therapy: Chronic Respiratory Disease Questionnaire (CRQ). *Cardiopulmonary physical therapy journal*, *19*(2), 61–67.

48. Petrillo, J., Van Nooten, F., Jones, P., & Rutten-Van Mölken, M. (2011). Utility estimation in chronic obstructive pulmonary disease: A preference for change? *PharmacoEconomics*. https://doi.org/10.2165/11589280-000000000-00000

49. Simon Pickard, A., Wilke, C., Jung, E., Patel, S., Stavem, K., & Lee, T. A. (2008). Use of a preference-based measure of health (EQ-5D) in COPD and asthma. *Respiratory Medicine*. https://doi.org/10.1016/j.rmed.2007.11.016

50. Finch, A. P., Brazier, J. E., & Mukuria, C. (2018). What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *European Journal of Health Economics*. https://doi.org/10.1007/s10198-017-0902-x

51. Qian, X., Tan, R. L. Y., Chuang, L. H., & Luo, N. (2020). Measurement Properties of Commonly Used Generic Preference-Based Measures in East and South-East Asia: A Systematic Review. *PharmacoEconomics*. https://doi.org/10.1007/s40273-019-00854-w

52. Fayers, P. M., Hays, R., & Hays, R. D. (2005). *Assessing Quality of Life in Clinical Trials: Methods and Practice*. Oxford University Press.

53. Brazier, J. E., Rowen, D., Mavranezouli, I., Tsuchiya, A., Young, T., Yang, Y., … Ibbotson, R. (2012). Developing and testing methods for deriving preferencebased measures of health from condition-specific measures (and other patient-based measures of outcome). *Health Technology Assessment*. https://doi.org/10.3310/hta16320

54. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000097

Table 1. Summary of test-retest reliability findings

Measure	Number of tests for test-retest	Summary or pooled result	Overall rating	Quality of evidence
EQ-5D	2	ICC & Spearman's correlation coefficient: 0.67-0.73 [37, 38]	Insufficient (inconsistent; based on statistical cut-off)	Low [serious risk of bias and serious inconsistency]
QWB	3	Pearson's correlation coefficient: 0.80-0.98 [41]	Sufficient	Very low [extremely serious risk of bias and serious indirectness]
15D	2	ICC & Spearman's correlation coefficient: 0.81-0.90 [37, 43]	Sufficient	Low [serious risk of bias and serious indirectness]

ICC=intra-class correlation coefficient

Table 2. Summary of convergent validity findings

Measure	Number of tests for convergent validity	Summary or pooled result (correlations)	Overall rating	Quality of evidence
EQ-5D	46	With generic preference-based measures; with SF-6D: 0.40-0.75[22, 25, 26, 33]; with 15D: 0.65 [37] With generic health profiles (SF-36 and PROMIS-43): 0.37-0.68 [28, 32, 37] With COPD-specific health profiles (SGRQ, CRQ, CAT, CCQ): 0.53-0.70 [23–25, 27, 29, 32, 33] With COPD-specific profile (LCOPD): 0.63-0.64 [29] With dyspnea measures (FACIT-Dyspnea, MRC, Borg scale): 0.28-0.58 [28, 32, 37] With COPD severity measures (BODE index and COPDSS): 0.33-0.71 [24, 29] With cough-specific health profiles (CQLQ, LQC): 0.30-0.60 [34] With performance measures (6MWT and Karnosfsky performance scale): 0.21-0.46 [28, 32, 37]	Sufficient	Moderate [serious indirectness]

Measure	Number of tests for convergent validity	Summary or pooled result (correlations)	Overall rating	Quality of evidence
		____________________ 83% of the correlations are in line with hypotheses.		
SF-6D	8	With generic preference-based measure (EQ-5D): 0.40-0.75 [22, 25, 26, 33] With COPD-specific health profile (SGRQ): 0.57-0.75 [25, 33] ____________________ 75% of the correlations are in line with hypotheses.	Sufficient	Moderate [serious indirectness]
QWB	4	With self-efficacy: 0.49 [42] With exercise tolerance: 0.41-0.54 [42] ____________________ 100% of the correlations are in line with hypotheses.	Sufficient	Very low [very serious risk of bias, serious imprecision, serious indirectness]
15D	8	With generic preference-based measure (EQ-5D): 0.65 [37] With generic health profile (SF-36): 0.60-0.61[37] With COPD-specific health profile (AQ20): 0.71 [44] With dyspnea measures (MRC, Borg scale): 0.59-0.60 [37] With performance measures (6MWT and Karnosfsky performance scale): 0.31-0.59 [37] ____________________ 100% of the correlations are in line with hypotheses.	Sufficient	Moderate [serious risk of bias]

6MWT= 6-minute walk test; Airway Questionnaire=AQ; BODE= BMI, Obstruction, Dyspnea, Exacerbation; CAT=COPD assessment test; CCQ= clinical COPD questionnaire; COPDSS=COPD severity score; CQLQ= cough quality of life questionnaire; CRQ= chronic respiratory questionnaire; FACIT= functional assessment of chronic illness therapy; LCOPD=living with COPD questionnaire; LCQ=Leicester cough questionnaire; MRC= the medical research council dyspnea scale; PROMIS=patient reported outcome measurement information system; SGRQ= St. George's respiratory questionnaire

Table 3. Summary of known-groups validity findings

Measure	**Number of tests for known-groups validity**	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
EQ-5D	41	Statistical differences between GOLD stages in 5 studies [22, 24, 27, 28, 36] No statistical differences between GOLD stages in 6 studies [22–24, 32, 33, 36]; Utility scores between GOLD stage 2 & 3 were not statistically different in 4 studies [22–24, 36]; Utility scores between GOLD stage 3 & 4 were not statistically different in 2 studies [24, 33] HRQOL decreased as breathlessness increased [23, 27, 38] Able to differentiate between ADO index scores, some EQ-VAS scores, SGRQ scores, and 6MWT scores [23, 25, 38] Mean differences between different medical conditions (p<0.001) [26] No mean differences within medical condition (p=0.72) [29] ____________________ 56% of hypotheses were able to confirm known-groups validity of the EQ-5D.	Sufficient (inconsistent; based on majority)	Low [serious inconsistency and serious indirectness]
SF-6D	14	Statistical differences between GOLD stages was found for 2 studies; between not very severe and very severe (p=0.0187) and between stage 3 and 4 (p=0.003) [22, 33] Able to differentiate most EQ-VAS cut-off scores and SGRQ cut-off scores (AUC>=0.70) [25] Mean differences between different medical conditions (p<0.001) [26] ____________________ 64% of hypotheses were able to confirm known-groups validity of the SF-6D.	Sufficient (inconsistent; based on majority)	Low [serious inconsistency and serious indirectness]

6MWT=6-minute walk test; ADO= Age, Dyspnea, Obstruction; AUC=area under the curve; EQ-VAS=EQ-visual analogue scale; HRQOL=health-related quality of life; SGRQ=St. George's respiratory questionnaire

Table 4. Summary of predictive validity findings

Measure	Number of tests for predictive validity	Summary or pooled result	Overall rating	Quality of evidence
15D	1	AUC=0.83 [43]	Sufficient	Moderate [serious imprecision]

AUC=area under the curve

Table 5. Summary of responsiveness findings

Measure	Number of tests for responsiveness	Summary or pooled result	Overall rating	Quality of evidence
EQ-5D	27	ES/SRM after treatment: 0.39-0.52, p=0.034-0.18, with change in COPD-specific health profiles (SGRQ, CRQ, CAT) r=0.14-0.40 [22, 23, 35] ES/SRM after exacerbation event: 0.65-0.69 [31, 33] Using perceived health change anchors: no statistical significant difference between groups (p=0.09-0.28) [30, 31, 37, 38] _____________________ 33 % of hypotheses were able to confirm responsiveness of the EQ-5D.	Insufficient (inconsistent; based on majority)	Very low [serious risk of bias, very serious inconsistency, serious indirectness]
SF-6D	3	ES after treatment: 0.64 [22] ES after exacerbation event: 0.27. [33] ____________________ 100% of hypotheses were able to confirm responsiveness of the SF-6D.	Sufficient	Moderate [serious risk of bias]
QWB	6	With change in performance capabilities: r=0.31-0.42 [42] With change in physiological/pulmonary capabilities: r=0.03-0.28 [42]	Sufficient (inconsistent; based on majority)	Very low [very serious risk of bias, serious inconsistency, serious imprecision, serious indirectness]

Measure	**Number of tests for responsiveness**	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
		____________________ 50% of hypotheses were able to confirm responsiveness of the QWB.		
15D	5	Using perceived health change anchors: statistical significant difference between groups (p=0.004) [37] ____________________ 80% of hypotheses were able to confirm responsiveness of the 15D.	Sufficient	Very low [extremely serious risk of bias and very serious imprecision]
HUI3	1	SRM after treatment: 0.20 [45] ____________________ 100% of hypotheses were able to confirm responsiveness of the HUI3.	Sufficient	Low [very serious risk of bias]

CAT=COPD assessment test; CRQ= chronic respiratory questionnaire; ES=effect size; r=Pearson’s correlation coefficient; SGRQ= St. George’s respiratory questionnaire; SRM=standardized response mean

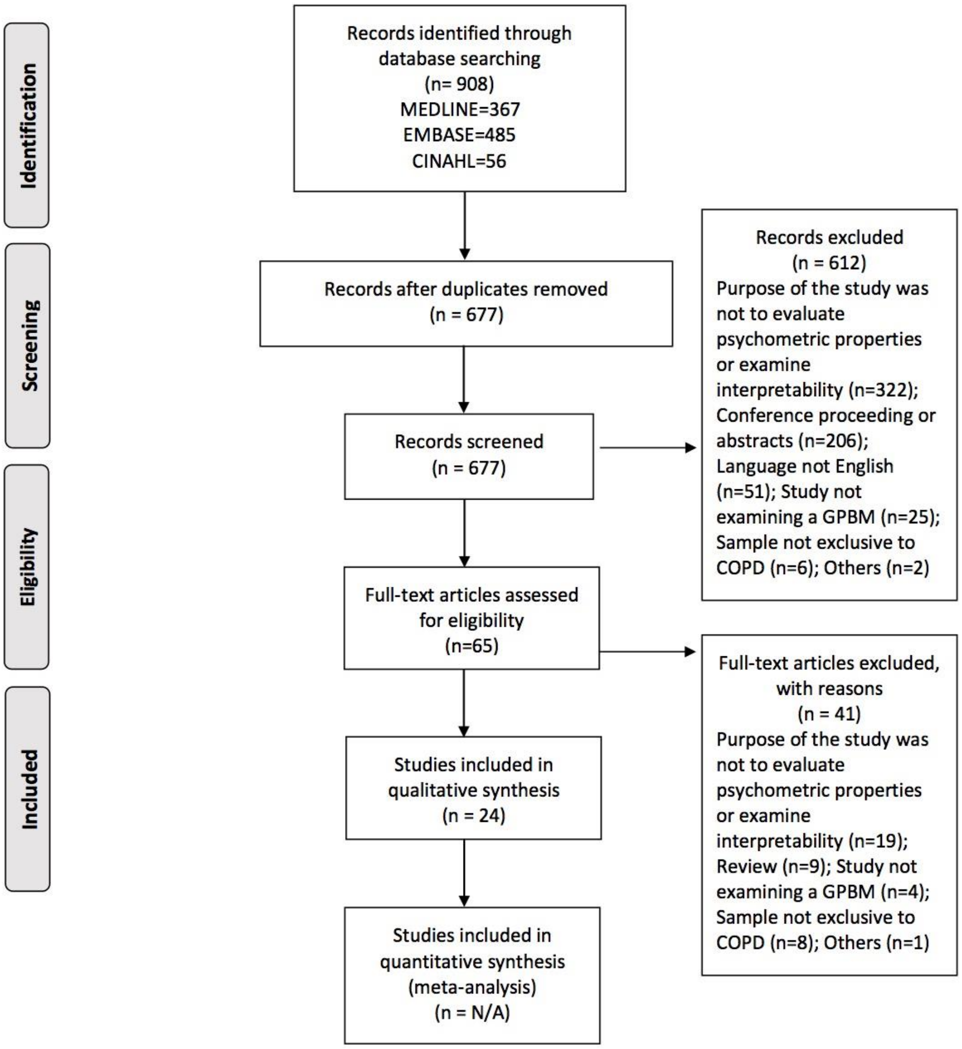
Identification
Screening
Eligibility
Included
Records identified through database searching (n= 908) MEDLINE=367 EMBASE=485 CINAHL=56
Records after duplicates removed (n = 677)
Records screened (n = 677)
Records excluded (n = 612) Purpose of the study was not to evaluate psychometric properties or examine interpretability (n=322); Conference proceeding or abstracts (n=206); Language not English (n=51); Study not examining a GPBM (n=25); Sample not exclusive to COPD (n=6); Others (n=2)
Full-text articles assessed for eligibility (n=65)
Full-text articles excluded, with reasons (n = 41) Purpose of the study was not to evaluate psychometric properties or examine interpretability (n=19); Review (n=9); Study not examining a GPBM (n=4); Sample not exclusive to COPD (n=8); Others (n=1)
Studies included in qualitative synthesis (n = 24)
Studies included in quantitative synthesis (meta-analysis) (n = N/A)

Fig. 1 Flow diagram of the article selection process. Adapted from the PRISMA statement [54]. COPD=Chronic Obstructive Pulmonary Disease; GPBM=Generic preference-based measure; N/A=not applicable

Supplementary Material:
Online Resource 1

Table 1. Search strategy for MEDLINE

	Terms
Population	1. exp Pulmonary Disease, Chronic Obstructive/ 2. copd.mp. 3. chronic obstructive pulmonary disease*.mp.
Types of Instruments	4. utility measure*.mp. 5. generic utilit*.mp. 6. EQ-5D*.mp. 7. EQ5D*.mp. 8. euroqol*.mp. 9. health utilit*.mp. 10. health utilit* index*.mp. 11. HUI*.mp. 12. short form 6 dimension*.mp. 13. SF-6D*.mp. 14. SF6D*.mp. 15. assessment of quality of life*.mp. 16. AQOL*.mp. 17. quality of well-being*.mp. 18. QWB*.mp. 19. 15-D*.mp. 20. 15D*.mp. 21. multi-attribute utilit*.mp.
Measurement Properties	22. (instrumentation or methods).fs. 23. (Validation Studies or Comparative Study).pt. 24. exp Psychometrics/ 25. psychometr*.ti,ab. 26. (clinimetr* or clinometr*).tw. 27. exp "Outcome Assessment (Health Care)"/ 28. outcome assessment.ti,ab. 29. outcome measure*.tw. 30. exp Observer Variation/ 31. observer variation.ti,ab. 32. exp Health Status Indicators/ 33. exp "Reproducibility of Results"/ 34. reproducib*.ti,ab. 35. exp Discriminant Analysis/ 36. (reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or internal consistency).ti,ab. 37. (cronbach* and (alpha or alphas)).ti,ab. 38. (item and (correlation* or selection* or reduction*)).ti,ab. 39. (agreement or precision or imprecision or precise values or test-retest).ti,ab. 40. (test and retest).ti,ab. 41. (reliab* and (test or retest)).ti,ab. 42. (stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intratester or intra-tester or interobserver or inter-observer or intraobserver or intraobserver or intertechnician or inter-technician or intratechnician or intra-technician or interexaminer or inter-examiner or intraexaminer or intra-examiner or interassay or interassay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-

	Terms
Measurement Properties	individual or interparticipant or inter-participant or intraparticipant or intra-participant or kappa* or repeatab*).ti,ab. 43. ((replicab* or repeated) and (measure or measures or findings or result or results or test or tests)).ti,ab. 44. (generaliza* or generalisa* or concordance).ti,ab. 45. (intraclass and correlation*).ti,ab. 46. (discriminative or known group or factor analysis or factor analyses or dimension* or subscale*).ti,ab. 47. (multitrait and scaling and (analysis or analyses)).ti,ab. 48. (item discriminant or interscale correlation* or error or errors or individual variability).ti,ab. 49. (variability and (analysis or values)).ti,ab. 50. (uncertainty and (measurement or measuring)).ti,ab. 51. (standard error of measurement or sensitiv* or responsive*).ti,ab. 52. ((minimal or minimally or clinical or clinically) and (important or significant or detectable) and (change or difference)).ti,ab. 53. (small* and (real or detectable) and (change or difference)).ti,ab. 54. (meaningful change or ceiling effect or floor effect or cross-cultural equivalence).ti,ab.

(These terms were modified for EMBASE and CINAHL, using their respective medical subject heading terms and search variables)

Table 2. Criteria for good measurement properties

Measurement Property	Rating [a]	Criteria
Reliability	+	ICC, weighted Kappa or correlations ≥ 0.70
	?	ICC, weighted Kappa, or correlations not reported
	-	ICC, weighted Kappa, or correlation <0.70
Construct Validity	+	Results in accordance with the hypothesis [b]
	?	Hypothesis not defined
	-	Result not in accordance with the hypothesis [b]
Predictive Validity	+	AUC ≥ 0.70
	?	No statistics reported
	-	AUC <0.70
Responsiveness	+	Results in accordance with hypothesis [b]
	?	Hypothesis not defined
	-	Results not in accordance with the hypothesis [b]

AUC=area under the curve; ICC= intra-class correlation coefficient

[a] +=sufficient, ?=indeterminate, -=insufficient

[b] If 75% of results are consistent (sufficient or insufficient), then the pooled rating will be determined based on the 75%

Hypotheses for Construct Validity and Responsiveness (Table A.3-7):

Table 3. Hypotheses for measurement properties of the EQ-5D

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
Thuppal et al. (2019)	A >=0.5 positive correlation with the SF-6D is expected at baseline and at 1 year.	An AUC>= 0.7 is expected for ROC curves between not very severe vs. very severe (at both baseline and 1 year); between moderate vs. severe/very severe. Mean differences between not very severe and very severe are expected to be statistically significant.	A >= 0.2 effect size is expected after 1 year post-surgery.
Nolan et al. (2016)	A >=0.5 negative correlation is expected with the SGRQ; the CAT; the CCQ, and a >=0.5 positive correlation is expected with the CRQ.	EQ-5D scores will significantly decrease with increasing GOLD stage; increasing MRC scores; and increasing ADO scores.	A >=0.2 effect size is expected after 8 weeks of PR. A >=0.5 positive correlation between the mean change in CRQ and EQ-5D is expected. A >=0.5 negative correlation between the mean change in SGRQ and EQ-5D; and CAT and EQ-5D is expected.
Wacker et al. (2016)	A >=0.5 negative correlation is expected with the SGRQ and the CAT; and a >=0.3 negative correlation is expected with the BODE.	GOLD grade means are expected to be significantly different from each other. >=0.5 effect sizes between grades is also expected.	
Chen et al. (2014)	A >=0.5 negative correlation Is expected with the SGRQ and a >=0.5 positive correlation is expected with the SF-6D.	AUCs >=0.7 are expected with EQ-VAS and SGRQ cut-offs.	
Ferreira et al. (2014)	A >=0.5 positive correlation is expected with SF-6D.	It is expected that means across different medical conditions are statistically different.	
Kim et al. (2014)	A >=0.5 negative correlation is expected with CCQ.	Expected for mean differences between GOLD stages to be statistically significant Expected that HRQL significantly decreases with increases in breathlessness.	

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
		>=0.5 effect sizes are expected between stages.	
Lin et al.(2014)	A >=0.3 negative correlation is expected with FACIT-dyspnea, MRC dyspnea, and Borg dyspnea. A >=0.3 positive correlation is expected with 6MWT. >=0.5 correlations are expected with PROMIS-43.	It is expected that mean differences across the 4 GOLD stages are statistically significant.	
Manca et al. (2014)	For both AATD and non-AATD COPD a >=0.3 negative correlation with the COPDSS and >=0.5 negative correlations with the LCOPD and CAT are expected.	Expected that the mean difference between AATD and non-AATD groups is statistically significant.	
Peters et al. (2014)			Expected for mean change from baseline to 1 year follow-up to be statistically significant. Expected mean change between health statuses to be statistically significant.
Goossens et al. (2011)			A >=0.2 SRM is expected after 6-weeks post-exacerbation. A change in SRM is expected to be statistically significant using PGI-C, CGI-C, sputum, cough, shortness of breath, expiratory peak flow, rescue medication use.
Pickard et al. (2011)	A >=0.3 negative correlation is expected with the Borg dyspnea scale. A >=0.5 negative correlation. is expected with the SGRQ. A >=0.3 positive correlation is expected with the 6MWT. A >=0.5 positive	A statistically significant mean difference is expected between the 4 GOLD stages for both the U.K. and the U.S. preferences.	

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
	correlation is expected with the SF-36 scales. (for both preferences; U.K. and U.S.).		
Menn et al. (2010)	A >=0.5 positive correlation is expected with the SF-6D. A >=0.5 negative correlation is expected with the SGRQ.	It is expected that the mean difference between stage 3 and 4 is statistically significant.	Expected for mean change from admission to discharge be statistically significant. A >=0.2 effect size is expected.
Polley et al. (2008)	A >=0.3 negative correlation is expected with the CQLQ. A >=0.3 positive correlation is expected with the LCQ.		
Ringbaek et al. (2008)			EQ-5D scores after 7-weeks of rehabilitation are expected to significantly improve.
Rutten-van Molken et al. (2006)		It is expected that both U.S. and U.K. mean scores are significantly different between GOLD stages 2-4. >=0.5 effect sizes are expected between stages for both U.S. and U.K. scores.	
Stavem (1999)	A >=0.5 positive correlation is expected with the 15-D. >=0.5 positive correlations are expected with the SF-36 scales. A >=0.3 positive correlation is expected with the Karnosfsky performance status and the 6MWT. A >=0.3 negative correlation is expected with MRC and the Borg scale.		EQ-5D scores are expected to be significantly different between subgroups determined by the GRC (better, unchanged, worse). >=0.2 absolute effect sizes and responsiveness statistics are expected for better and worse subgroups.
Harper et al. (1997)		Moderate to large effects sizes are expected for subgroups for breathlessness, 6MWT, VAS, and FEV1 %.	A significant difference between subgroups of perceived health change is expected. A >=0.2 SRM is expected between initial and 6

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
			months' assessment and 6 months' and 12 months' assessment.

Table 4. Hypotheses for measurement properties of the SF-6D

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
Thuppal et al. (2019)	A >=0.5 positive correlation with the EQ-5D is expected at baseline and at 1 year.	An AUC>= 0.7 is expected for ROC curves between not very severe vs. very severe (at both baseline and 1 year); between moderate vs. severe/very severe. Mean differences between not very severe and very severe are expected to be statistically significant.	A >=0.2 effect size is expected after 1 year post-surgery.
Chen et al. (2014)	A >=0.5 negative correlation is expected with the SGRQ using both Hong Kong and UK preferences and a >=0.5 positive correlation is expected with the EQ-5D.	AUCs >= 0.7 are expected with EQ-VAS and SGRQ cut-offs.	
Ferreira et al. (2014)	A >=0.5 positive correlation is expected with the EQ-5D.	It is expected that means across different medical conditions are statistically different.	
Menn et al. (2010)	A >=0.5 positive correlation is expected with the EQ-5D. A >=0.5 negative correlation is expected with the SGRQ.	It is expected that the mean difference between GOLD stage 3 and 4 is statistically significant.	Expected for mean change from admission to discharge to be statistically significant. A >=0.2 effect size is expected.

Table 5. Hypotheses for measurement properties of the QWB

Author (year)	**Convergent Validity** [a]	**Known-Groups Validity** [b]	**Responsiveness** [a,b]
Kaplan et al. (1984)	>=0.3 positive correlations are expected with self-efficacy and exercise tolerance (both at initial and 3-months follow-up assessments).		Change in scores on the QWB are expected to have a >=0.3 correlation with change in scores for exercise tolerance, self-efficacy, walking compliance, FVC, FEV, O2 saturation.

Table 6. Hypotheses for measurement properties of the 15D

Author (year)	**Convergent Validity [a]**	**Known-Groups Validity [b]**	**Responsiveness [a,b]**
Mazur et al. (2011)	A >=0.5 negative correlation is expected with the AQ20.		
Stavem (1999)	A >=0.5 positive correlation is expected with the EQ-TTO. >=0.5 positive correlations are expected with the SF-36 scales. A >=0.3 positive correlation is expected with the Karnosfsky performance status and the 6MWT. A >=0.3 negative correlation is expected with MRC and Borg scale.		15D scores are expected to be significantly different between subgroups determined by the global rating of change (better, unchanged, worse). >=0.2 absolute effect sizes and responsiveness statistics are expected for better and worse subgroups.

Table 7. Hypotheses for measurement properties of the HUI3

Author (year)	**Convergent Validity [a]**	**Known-Groups Validity [b]**	**Responsiveness [a,b]**
Puhan et al. (2007)			A >=0.2 SRM is expected after 12 weeks of rehabilitation.

6MWT= 6-minute walk test; AATD=alpha-1 antitrypsin deficiency; ADO= Age, Dyspnea, Obstruction; AQ=airway questionnaire; AUC=area under the curve; BODE= BMI, Obstruction, Dyspnea, Exacerbation; CAT=COPD assessment test; CCQ= clinical COPD questionnaire; CGI-C=clinician's global impression of change; COPDSS=COPD severity score; CQLQ= cough quality of life questionnaire; CRQ= chronic respiratory questionnaire; FACIT= functional assessment of chronic illness therapy; HRQL= health-related quality of life; LCOPD=living with COPD questionnaire; LCQ=Leicester cough questionnaire; MRC= the medical research council dyspnea scale; PGI-C= patient's global impressions of change; PR=pulmonary rehabilitation; PROMIS=patient reported outcome measurement information system; ROC=receiver operating characteristic; SGRQ= St. George's respiratory questionnaire; SRM=standardized response mean

[a] Pearson's correlation coefficient (r) or Spearman's correlation coefficient (rho) were utilized to assess correlations
[b] Statistical significance = p-value <0.05

Table 8. Modified GRADE approach for grading the quality of evidence

Quality of evidence	Lower if
High: confident that the true measurement property is close to the estimate (pooled/summarized result)	Risk of bias -1 Serious -2 Very serious -3 Extremely serious Inconsistency -1 Serious -2 Very serious Imprecision -1 total n=50-100 -2 total n<50 Indirectness -1 Serious -2 Very Serious
Moderated: moderately confident in the estimate measurement property; possibility that it substantially differs from the true measurement property	
Low: limited confidence in the estimate measurement property; may substantially differ from true measurement property	
Very low: very little confidence in the estimate; likely to differ from true measurement property	

n=sample size

Table 9. Downgrading Risk of Bias

Risk of bias	Downgrading for Risk of Bias
No	Multiple studies of at least ‘adequate’ quality, or one study of ‘very good’ quality
Serious	Multiple studies of doubtful quality, only one study of ‘adequate’ quality
Very serious	Multiple studies of ‘inadequate’ quality, one study of ‘doubtful’ quality
Extremely serious	Only one study of ‘inadequate’ quality

Online Resource 2. Study Characteristics

Author (Year)	**Country**	**Sample Characteristics (mean (SD))**	**Mean (SD) for Preference-based Measure**	**Properties Assessed**
EQ-5D (n=17)				
Thuppal et al. (2019)	USA	N=94, age= 66 (7.8), FEV_1 % pred. baseline = 26.7 (8.3), FEV_1 % pred. at 1 year = 37.5 (14.7)	Baseline: 0.66 (0.2), At 1yr: 0.77 (0.19)	Convergent Validity, Known-Groups Validity, Responsiveness
Nolan et al. (2016)	UK	(1) N=616, age= 70.4 (9.3), FEV_1 % pred.= 46.1 (19.6). (2) N=324, age= 70.2(69.2, 71.2) *, FEV_1 % pred.= 49.8 (47.5, 52.0) *	(1) 0.681 (0.236) (2) baseline: 0.697 (0.673, 0.720) *	Convergent Validity, Known-Groups Validity, Responsiveness, Interpretability
Wacker et al. (2016)	Germany	N=2291, age= 65.1 (8.4), FEV_1 % pred.= 52.5 (18.6)	0.82 (0.20)	Convergent Validity, Known-Groups Validity, Interpretability
Chen et al. (2014)	China	N= 154, age= 72.96 (8.1), post FEV_1 %= 32.7 (9.2)	0.644 (0.306)	Convergent Validity, Known-Groups Validity, Interpretability
Ferreira et al. (2014)	Portugal	N=72, age= 68.6 (9.5), FEV_1 % not available	0.86 (0.17)	Convergent Validity, Known-Groups Validity, Interpretability
Kim et al. (2014)	Korea	N=200, age = 68.5 (9.1), FEV_1 % pred.= 56.3	0.84 (0.16)	Convergent Validity, Known-Groups Validity
Lin et al. (2014)	USA	N=670, age=68.5 (10.4), FEV_1 % not available	0.79 (0.15)	Convergent Validity, Known-Groups Validity
Manca et al. (2014)	Spain	AATD: N=35, age= 56.5 (10.6), post FEV_1 % pred.= 48.7 (17.9); Non-AATD: N=61, age=70.3 (9.2), FEV_1 % pred.= 48.8 (16.5)	AATD: 0.74(0.23); Non-AATD: 0.72(0.22)	Convergent Validity, Known-Groups Validity
Peters et al. (2014)	UK	N=187, age & FEV_1 % not available	Baseline: 0.67, 1 yr. follow-up: 0.67	Responsiveness
Goossens et al. (2011)	USA	N=59, age=61.1 (10.4), FEV_1 % not available	Visit 1: 0.683 (0.209), Visit 4: 0.760 (0.181)	Responsiveness
Pickard et al. (2011)	USA	N=120, age=71.2 (10.3), FEV_1 %= 58.4 (24.8)	US index: 0.73 (0.19). UK index: 0.63 (0.27)	Convergent Validity, Known-Groups Validity
Menn et al. (2010)	Germany	N=117, age= 67 (8), FEV_1 % not available	At admission; stage 3: 0.62 (0.26), stage 4: 0.60 (0.26). At discharge; stage 3: 0.84 (0.20), stage 4: 0.75 (0.22)	Convergent Validity, Known-Groups Validity, Responsiveness
Polley et al. (2008)	Ireland	N=18, age= 64.4 (9.7), FEV_1 % pred.= 42.3 (16.9)	0.45 (0.31)	Convergent Validity
Ringbaek et al. (2008)	Denmark, UK	N=229, age= 69.1 (8.1), FEV_1 % pred.= 34.1 (12.2)	Pre-rehab: 0.759 (0.174), post-rehab: 0.778 (0.180), 3 mos. follow-up: 0.771 (0.192)	Responsiveness, Interpretability
Rutten-van Molken et al. (2006)	Multi-national	N=1235, age= 64.5 (8.4), post FEV_1 % pred.= 48.77 (12.19)	0.76 (0.21)	Known-groups Validity, Interpretability

Author (Year)	Country	Sample Characteristics (mean (SD))	Mean (SD) for Preference-based Measure	Properties Assessed
Stavem (1999)	Norway	N=59, age= 57.0(9.1), FEV$_1$ % pred.= 47.1 (15.3)	0.73 (0.62-0.81) **	Reliability; test-retest, convergent Validity, Responsiveness
Harper et al. (1997)	UK	N=156, age=67 (10.4), FEV$_1$ % pred.= 47	At initial assessment: 0.524 (0.157)	Reliability; test-retest, Known-Groups Validity, Responsiveness
SF-6D (n=5)				
Thuppal et al. (2019)	USA	N=94, age= 66 (7.8), FEV$_1$ % pred. baseline = 26.7 (8.3), FEV$_1$ % pred. at 1 yr. = 37.5 (14.7)	Baseline: 0.66 (0.11), At 1year: 0.74 (0.14)	Convergent Validity, Known-Groups Validity, Responsiveness
Chen et al. (2014)	China	N= 154, age= 72.96 (8.1), post FEV$_1$ %= 32.7 (9.2)	HK index: 0.591(0.147) UK index= 0.629 (0.133)	Convergent Validity, Known-Groups Validity, Interpretability
Ferreira et al. (2014)	Portugal	N=72, age= 68.6 (9.5), FEV$_1$ % not available	0.81 (0.12)	Convergent Validity, Known-Groups Validity, Interpretability
Menn et al. (2010)	Germany	N=117, age= 67 (8), FEV$_1$ % not available	At admission; stage 3: 0.61 (0.13) stage 4: 0.54 (0.08), At discharge; stage 3: 0.65 (0.12) stage 4: 0.58 (0.08)	Convergent Validity, Known-Groups Validity, Responsiveness
Walters & Brazier (2003)	UK	N=60, age & FEV$_1$ % not available	Mean (SD) not available	Interpretability
QWB (n=3)				
Kaplan (2005)	USA	Trial 1 N=119, Trial 2 N=164, Trial 3 N= 1215, age & FEV$_1$ % not available	Pre-rehab ranged from 0.537 to 0.666. Post-rehab ranged from 0.571 to 0.698.	Interpretability
Anderson et al. (1989)	USA	N=120, (1st interview N=84, 2nd N=63, 3rd N=45), age & FEV$_1$ % not available	Mean (SD) not available	Reliability; test-retest
Kaplan et al. (1984)	USA	N=75, age=64.79 (7.86), FEV$_1$ % pred. initial= 36.22 (23.84), FEV$_1$ % pred. follow-up=37.23 (24.36)	Initial (n=66) = 0.608 (0.08) Follow-up (n=67) = 0.603 (0.09)	Convergent Validity, Responsiveness
15D (n=3)				
Koskela et al. (2014)	Finland	N=548, age=68.1 (55.0, 81.1) *, FEV$_1$ % pred.= 58.6, (56.9, 60.2) *	Baseline: 0.799 (0.793, 0.811) *, 1 yr. follow-up: 0.792 (0.785, 0.804) *, 2 yr. follow-up: 0.788 (0.782, 0.801) *, 4 yr. follow-up: 0.783 (0.773, 0.794) *	Reliability; test-retest, Predictive Validity
Mazur et al. (2011)	Finland	N=739, age= 64 (6.8), FEV$_1$ % not available	0.79 (0.11)	Convergent Validity
Stavem (1999)	Norway	N=59, age= 57.0 (9.1), FEV$_1$ % pred.= 47.1 (15.3)	0.80 (0.73-0.88)**	Reliability; test-retest, convergent Validity, Responsiveness
HUI3 (n=1)				
Puhan et al. (2007)	Canada, USA	N=177, age=69 (8.7), FEV$_1$ % pred.=42.8 (19.2)	Mean (SD) not available	Responsiveness

* 95% Confidence Interval

** median (interquartile range)

FEV= forced expiratory volume; n=number of studies; N=sample size; SD= standard deviation

Online Resource 3

Table 1. Results of measurement properties for EQ-5D studies

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
EQ-5D (Thuppal et al., 2019)	USA	N/A	N/A	N/A	94	Inadequate	With SF-6D at baseline (rho=0.64), at 1 year follow-up (rho=0.75) (2+)	94	Doubtful	ROC curves for very severe vs. not very severe COPD (defined by GOLD) at baseline: AUC= 0.605, at 1 year: AUC = 0.645; for moderate vs. severe/very severe COPD: AUC=0.644. Statistically significant mean (SD) [at the end of 1year] difference between not very severe (0.80(0.2)) & very severe COPD (0.75 (0.16)) (p = 0.0137) (1+, 3-)	94	Doubtful	Baseline to 1 year post-LVRS: Effect size=0.52 (1+)
EQ-5D-5L (Nolan et al., 2016)	UK	N/A	N/A	N/A	616	Inadequate	With SGRQ (r= -0.623) With CRQ (r=0.704) With CAT (r=-0.528) With CCQ (r=-0.626)	616	Doubtful	EQ-5D significantly decreased with increasing GOLD stage (p=0.004), but was not able to differentiate between GOLD stages 1/2 (grouped together) and 3; EQ-5D significantly decreased with	324	Construct: inadequate. Intervention: doubtful	8 weeks of PR: SRM=0.39 Correlations with SGRQ (r=-0.14); With CRQ (r=0.40);

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
							(4+)			increasing MRC dyspnea score (p<0.001); EQ-5D significantly decreased with increasing ADO index (p<0.001) (2+, 1-)			With CAT (r=-0.14) (1+,3-)
EQ-5D (Wacker et al., 2016)	Germany	N/A	N/A	N/A	2291	Doubtful	With CAT (rho=-0.56) With SGRQ (rho=-0.56) With BODE (rho=-0.33) (3+)	2291	Very good	After adjusting & using regression (grade 1 as reference): grade 3 &4 means were significantly different than grade 1 [grade 2 (p=0.69), grade 3 (p=0.005), grade 4 (p<0.00001)] Effect size between grade 1 &2=0.03, between 2&3=0.17, between 3&4=0.41 (2+, 4-)	N/A	N/A	N/A
EQ-5D (Chen et al., 2014)	China	N/A	N/A	N/A	154	Very good	With SGRQ (r=-0.583) With SF-6D (r=0.677) (2+)	154	Doubtful	With EQ-VAS scores as cut-offs: >=50 vs. <50: AUC=0.724, >=60 vs. <60: AUC=0.649, >=70 vs. <70: AUC=0.652, >=80 vs. <80: AUC=0.687, >=90 vs. <90: AUC=0.755	N/A	N/A	N/A

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
										With SGRQ scores as cut-offs: >49 vs. <=49: AUC =0.826, >64 vs. <=64: AUC=0.850, >77 vs. <=77: AUC=0.846 (5+, 3-)			
EQ-5D (Ferreira et al., 2014)	Portugal	N/A	N/A	N/A	72	Doubtful	With SF-6D (r=0.40) (1-)	72	Inadequate	Statistically significant mean differences between different medical conditions (i.e. Asthma, COPD, Cataracts, Rheumatoid Arthritis) (p<0.001) (1+)	N/A	N/A	N/A
EQ-5D (Kim et al., 2014)	Korea	N/A	N/A	N/A	200	Very good	With CCQ (r= -0.69) (1+)	200	Very good	Statistically significant mean differences between the 4 GOLD stages (p<0.001) HRQL significantly worsened as severity of breathlessness increased (p<0.0001) Effect size between stages 2 & 3 = 0.47; between stages 3&4=1.18	N/A	N/A	N/A

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
										(4+)			
EQ-5D-5L (Lin et al., 2014)	USA	N/A	N/A	N/A	670	Inadequate	With FACIT-Dyspnea (rho=-0.58) With modified MRC dyspnea (rho=-0.48) With Borg dyspnea (at rest) (rho=-0.38) With Borg Dyspnoea (during 6MWT) (rho=-0.37) With 6MWT (rho=0.46) With PROMIS-43 domains (rho=0.37-	670	Very good	Mean differences between 4 GOLD stages were statistically significant using ANOVA (p =0.0004), and the Kruskal-Wallis test (p=0.002) (1+)	N/A	N/A	N/A

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
							0.68 (absolute values)) (5+, 1-)						
EQ-5D (Manca et al., 2014)	Spain	N/A	N/A	N/A	96	Doubtful	For AATD COPD: With COPDSS (r= -0.706); With LCOPD (r= -0.641); With CAT (r= -0.703) Non-AATD COPD: With COPDSS (r= -0.397); With LCOPD (r= -0.629); With CAT (r= -0.546) (6+)	96	Very good	Mean difference between AATD & non-AATD COPD scores was not statistically significant (p=0.72) (1-)	N/A	N/A	N/A
EQ-5D (Peters et al., 2014)	UK	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	177	Inadequate	Mean change from baseline & 1 year

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
													follow-up was not statistically significant (0.00, p=0.77) Mean change across change in health (improved, stable, deteriorated) was not statistically significant (p=0.23) (2-)
EQ-5D (Goossens et al., 2011)	USA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	59	Doubtful	6 weeks' post-exacerbation: SRM=0.653 Change in SRM between greater and less improvements after 6 weeks was not statistically significant using PGI-C (-0.413, p=0.128), CGI-C (-0.170, p=0.657), sputum (0.140, p=0.594),

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
													cough (-0.395, p=0.144), shortness of breath (0.518, p=0.051), expiratory peak flow (0.505, p=0.058); was statistically significant using rescue medication use (0.645, p=0.018) (2+, 6-)
EQ-5D (Pickard et al., 2011)	USA	N/A	N/A	N/A	120	Inadequate	Preference weights from the U.K.: With 6MWT (r=0.21); With Borg Dyspnea (r=-0.48); With SGRQ (r=-0.55); With SF-36 PCS (r=0.51); With SF-36 MCS (r=0.54)	120	Very good	Mean differences (for both U.K. and U.S. preferences) between the 4 GOLD stages was not statistically significant using ANOVA (p =0.26 & 0.25, respectively), and the Kruskal-Wallis test (p=0.079 & 0.069, respectively) (2-)	N/A	N/A	N/A

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
							Preference weights from the U.S.: With 6MWT (r=0.21); With Borg Dyspnoea (r=-0.48); With SGRQ (r=-0.57); With SF-36 PCS (r=0.51); With SF-36 MCS (r=0.56) (8+, 2-)						
EQ-5D (Menn et al., 2010)	Germany	N/A	N/A	N/A	117	Inadequate	With SF-6D (r=0.43), With SGRQ (r=-0.59) (1+,1-)	117	Very good	Mean difference between GOLD stage 3 (0.73) and 4 (0.68) was not statistically significant (p=0.180) (1-)	106	Doubtful	Mean change from exacerbation admission (mean (SD)=0.60(0.26)) to discharge (0.79(0.21)) was statistically significant (p<0.001) Admission to discharge:

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
													Standardized differences/effect size=0.69 (2+)
EQ-5D (Polley et al., 2008)	Ireland	N/A	N/A	N/A	18	Doubtful	With CQLQ (r=-0.30) With LCQ (r=0.60) (2+)	N/A	N/A	N/A	N/A	N/A	N/A
EQ-5D (Ringbaek et al., 2008)	Denmark & UK	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	229	Inadequate	After 7-weeks of PR statistically significant improvements were seen for EQ-5D scores (p=0.034) pre- vs. post-rehab, but not pre- vs. 3 months' follow-up (p=0.18) (1+, 1-)
EQ-5D (Rutten-van Molken et al., 2006)	Multi-national	N/A	N/A	N/A	N/A	N/A	N/A	1235	Very good	Statistically significant mean differences between GOLD stages 2-4 for U.K. & U.S. preference weights (p<0.001); all pairwise comparisons	N/A	N/A	N/A

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
										were statistically significant (p≤0.001) Effect size between stages 2 & 3: for U.K. preference=<0.2, for U.S. preference =0.2-0.3; between stages 3 &4: for U.K. preference=0.4-0.5, for U.S. preference =0.4-0.5 (4+, 2-)			
EQ-5D (Stavem, 1999)	Norway	49	Adequate	Spearman's rho=0.73 (1+)	59	Doubtful	With 15D (rho=0.65) With SF-36 PCS (rho=0.51) With SF-36 MCS (rho=0.45) With Karnosfsky performance status (rho=0.32) With MRC (rho=-0.28)	N/A	N/A	N/A	51	Inadequate	Using the global rating of change (obtained from the SF-36 question #2) after 1 year, the EQ-5D was only able to discriminate between better and unchanged and was not statistically significant between better, unchanged, and worse (p=0.09) Group 'better': Effect size= – 0.55

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
							With 6MWT (rho=0.21) With Borg scale (rho= -0.43) (4+, 3-)						Responsiveness statistic= –1.18, Group 'worse': Effect size= –0.07 Responsiveness statistic =–0.13 (2+,3-)
EQ-5D (Harper et al., 1997)	UK	156	Inadequate	ICC=0.67 (1-)	N/A	N/A	N/A	156	Adequate	Effect size for breathlessness groups: large (>/=0.8) Effect size for 6MWT & Visual Analogue Scale for breathlessness: moderate (>/=0.5-<0.8) Effect size for FEV_1% predicted: small (around 0.2) (3+,1-)	156	Inadequate	After 6 months: No significant difference between subgroups of perceived health change; worse vs. same vs. better (p=0.28) SRM between initial & 6 months' follow-up <0.2 SRM between and 6 & 12 months' follow-up <0.2 (3-)
Pooled or summary result (overall rating)		**205**		**0.67-0.73 (1+, 1-)**	**4507**		with generic preference-based	**5821**		**56% of hypotheses were able to confirm known-groups validity of the EQ-5D.**	**1196**		**33 % of hypotheses were able to confirm responsiveness**

Measure (author, year)	**Country**	**Reliability**			**Convergent Validity** [a]			**Known-Groups Validity** [b]			**Responsiveness** [a,b]		
		n	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Results (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]
							measures: 0.40-0.75 with generic health profiles: 0.37-0.68 with COPD-specific health profiles: 0.528-0.704 with dyspnea measures: 0.28-0.58 with COPD severity measures: 0.33-0.706 with cough-specific health profiles: 0.30-0.60			**(23+, 18-)**			**of the EQ-5D. (9+, 18-)**

Measure (author, year)	**Country**	**Reliability**			**Convergent Validity** [a]			**Known-Groups Validity** [b]			**Responsiveness** [a,b]		
		n	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Results (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]
							with performance measures: 0.21-0.46 **83% of the correlations are in line with the hypotheses. (38+, 8-)**						

Table 2. Results of measurement properties for SF-6D studies

Measure (author, year)	**Country**	**Convergent Validity** [a]			**Known-Groups Validity** [b]			**Responsiveness** [a,b]		
		n	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Results (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]
SF-6D (Thuppal et al., 2019)	USA	94	Inadequate	With EQ-5D at baseline (rho=0.64), at 1 year follow-up (rho=0.75) (2+)	94	Doubtful	ROC curves for very severe vs. not very severe COPD (defined by GOLD) at baseline: AUC = 0.625, at 1 year.: AUC = 0.661; for moderate vs. severe/very severe: AUC=0.696. Statistically significant mean (SD) [at the end of 1yr.] difference between not very severe (0.77(0.14)) & very severe COPD (0.70 (0.13)) (p = 0.0187)	94	Doubtful	Baseline to 1 year post-LVRS: Effect size=0.64 (1+)

Measure (author, year)	Country	Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
							(1+,3-)			
SF-6D (Chen et al., 2014)	China	154	Very good	With SGRQ (r=-0.745 using Hong Kong preference weights, r=-0.728 using U.K. preference weights) With EQ-5D (r=0.677) (3+)	154	Doubtful	With EQ-VAS scores as cut-offs: >=50 vs. <50: AUC=0.718, >=60 vs. <60: AUC=0.672, >=70 vs. <70: AUC=0.695, >=80 vs. <80: AUC=0.733, >=90 vs. <90: AUC=0.763 With SGRQ scores as cut-offs: >49 vs. <=49: AUC =0.864, >64 vs. <=64: AUC=0.835, >77 vs. <=77: AUC=0.867 (6+, 2-)	N/A	N/A	N/A
SF-6D (Ferreira et al., 2014)	Portugal	72	Doubtful	With EQ-5D (r=0.40) (1-)	72	Inadequate	Statistically significant mean differences between different medical conditions (i.e. Asthma, COPD, Cataracts, Rheumatoid Arthritis) (p<0.001) (1+)	N/A	N/A	N/A
SF-6D (Menn et al., 2010)	Germany	117	Inadequate	With EQ-5D (r=0.43) With SGRQ (r=-0.57) (1+, 1-)	117	Very good	Mean difference between GOLD stage 3 (0.62) and 4 (0.56) was statistically significant (p=0.003) (1+)	68	Doubtful	Mean change from exacerbation admission (mean (SD)=0.56(0.11)) to discharge (0.59(0.09)) was statistically significant (p=0.008) Admission to discharge: Standardized differences/Effect size=0.27

Measure (author, year)	Country	Convergent Validity [a]			Known-Groups Validity [b]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Results (rating) [c]	n	Methodological quality	Result (rating) [c]
										(2+)
Pooled or summary result (overall rating)		**437**		with generic preference-based measure: 0.40-0.75 with COPD-specific health profile: 0.57-0.745 **75% of the correlations are in line with the hypotheses. (6+, 2-)**	**437**		**64% of hypotheses were able to confirm known-groups validity of the SF-6D. (9+, 5-)**	**162**		**100% of hypotheses were able to confirm responsiveness of the SF-6D. (3+)**

Table 3. Results of measurement properties for QWB studies

Measure (author, year)	Country	Reliability			Convergent Validity [a]			Responsiveness [a,b]		
		n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]	n	Methodological quality	Result (rating) [c]
QWB (Anderson et al., 1989)	USA	196	Inadequate	Pearson's r: 1st interview range=0.84-0.98; 2nd interview = 0.81-0.95; 3rd interview = 0.80-0.98 (3+)	N/A	N/A	N/A	N/A	N/A	N/A
QWB (Kaplan et al., 1984)	USA	N/A	N/A	N/A	60	Doubtful	Initial assessment: With self-	75	Doubtful	After 3 months: With exercise

							efficacy (r=0.49) With exercise tolerance (r=0.41) After 3 months: With self-efficacy (r=0.49) With exercise tolerance (r=0.54) (4+)			tolerance (r=0.40) With self-efficacy (r=0.31) With walking compliance (r=0.42) With FVC (r=0.03) With FEV (r=0.11) With O2 saturation (r=0.28) (3+, 3-)
Pooled or summary result (overall rating)		**196**		**0.81-0.98 (3+)**	**60**		with self-efficacy: 0.49 with exercise tolerance: 0.41-0.54 **100% of the correlations are in line with the hypotheses. (4+)**	**75**		**50% of hypotheses were able to confirm responsiveness of the QWB. (3+, 3-)**

Table 4. Results of measurement properties for 15D studies

Measure (author, year)	**Country**	**Reliability**			**Convergent Validity** [a]			**Predictive Validity**			**Responsiveness** [a,b]		
		n	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]	**n**	**Methodological quality**	**Result (rating)** [c]
15D (Koskela et al., 2014)	Finland	548	Inadequate	Agreement between 0,1,2,4 years: ICC=0.81 (1+)	N/A	N/A	N/A	548	Very good	Predicting constant decliners within the next 5 years ROC curve: AUC=0.83 (1+)	N/A	N/A	N/A
15D (Mazur et al., 2011)	Finland	N/A	N/A	N/A	739	Adequate	With AQ20 (rho=-0.71) (1+)	N/A	N/A	N/A	N/A	N/A	N/A
15D (Stavem, 1999)	Norway	44	Adequate	Spearman's rho=0.90 (1+)	53	Doubtful	With EQ-5D (rho =0.65) With SF-36 PCS (rho=0.60) With SF-36 MCS (rho=0.61) With Karnosfsky	N/A	N/A	N/A	45	Inadequate	Using the global rating of change (obtained from the SF-36 question #2) after 1 year, the 15D was statistically significant between better, unchanged, and worse (p=0.004)

							performance status (rho=0.59) With MRC (rho=-0.59) With 6MWT (rho=0.31) With Borg scale (rho= -0.60) (7+)						Group ‘better’: Effect size= –1.00 Responsiveness statistic= –1.51, Group ‘worse’: Effect size= 0.15 Responsiveness statistic =0.38 (4+,1-)

Pooled or summary result (overall rating)	**592**		**0.81-0.90 (2+)**	**792**		with generic preference-based measure: 0.65 with generic health profile: 0.60-0.61 with COPD-specific health profile: 0.71 with dyspnea measures: 0.59-0.60 with performance measures: 0.31-0.59 **100% of the correlations are in line with the hypotheses. (8+)**	**548**		**AUC=0.83 (1+)**	**45**		**80% of hypotheses were able to confirm responsiveness of the 15D. (4+, 1-)**

Table 5. Results of measurement properties for HUI3 studies

Measure (author, year)	**Country**	**Responsiveness** [a,b]		
		n	**Methodological quality**	**Result (rating)**
HUI3 (Puhan et al., 2007)	Canada & USA	177	Doubtful	After 12 weeks of respiratory rehabilitation: SRM= 0.20 (1+)
Pooled or summary result (overall rating)		**177**		**100% of hypotheses were able to confirm responsiveness of the HUI3.** **(1+)**

6MWT= 6-minute walk test; AATD=alpha-1 antitrypsin deficiency; ADO= Age, Dyspnea, Obstruction; AUC=area under the curve; AQ=airway questionnaire; BODE= BMI, Obstruction, Dyspnea, Exacerbation; CAT=COPD assessment test; CCQ= clinical COPD questionnaire; CGI-C=clinician's global impression of change; COPDSS=COPD severity score; CQLQ= cough quality of life questionnaire; CRQ= chronic respiratory questionnaire; FACIT= functional assessment of chronic illness therapy; HRQL= health-related quality of life; LCOPD=living with COPD questionnaire; LCQ=Leicester cough questionnaire; LVRS=lung volume reduction surgery; MCS=mental component scale; MRC= the medical research council dyspnea scale; n=sample size; PCS= physical component scale; PGI-C= patient's global impressions of change; PR=pulmonary rehabilitation; PROMIS=patient reported outcome measurement information system; ROC=receiver operating characteristic; SD=standard deviation; SGRQ= St. George's respiratory questionnaire; SRM=standardized response mean

[a] Pearson's correlation coefficient (r) or Spearman's correlation coefficient (rho) were utilized to assess correlations
[b] Statistical significance = p-value <0.05
[c] +=sufficient, ?=indeterminate, -=insufficient

Online Resource 4

Table 1. EQ-5D descriptive system with measures of health

EQ-5D Dimensions	Generic Health Profiles		COPD-Specific Health Profiles			
	Corresponding SF-36 Domains	**Corresponding PROMIS-43 Domains**	**Corresponding SGRQ Domains**	**Corresponding CRQ Domains**	**Corresponding CAT items/components**	**Corresponding CCQ Domains**
Mobility	Physical functioning	Physical function	Activity limitation		Activities	Functional state (activities limitations)
Self-care						
Pain/discomfort	Bodily Pain	Pain				
Usual activities	-Social functioning - Role limitations due to emotional problems - Role limitations due to physical problems	Mental health (negative affect) -Satisfaction with participation in social roles and activities	- Social and emotional impact - Activity limitation			Functional state (activities limitations)
Anxiety/depression	Mental health		-Social and emotional impact	Emotion	Activities	Mental health
Components not covered by the EQ-5D	-Vitality -General Health	Fatigue	Symptoms	-Fatigue -Dyspnea -Mastery	-Dyspnea -Cough -Chest-tightness -Phlegm -Energy -Sleep -Confidence	Symptoms

*Domains/components may appear more than once if applicable to more than one EQ-5D dimension
CAT=COPD assessment test; CCQ= clinical COPD questionnaire; CRQ= chronic respiratory questionnaire; PROMIS=patient reported outcome measurement information system; SGRQ= St. George’s respiratory questionnaire

Table 2. SF-6D descriptive system with measures of health

SF-6D Dimensions	COPD-Specific Health Profiles
	Corresponding SGRQ Domains [4]
Physical functioning	Activity limitation
Role limitation	Activity limitation
Pain	

Social functioning	Social and emotional impact
Mental health	Social and emotional impact
Vitality	
Components not covered by the SF-6D	Symptoms

*Domains/components may appear more than once if applicable to more than one SF-6D dimension
SGRQ= St. George's respiratory questionnaire

Table 3. 15D descriptive system with measures of health

15D Dimensions	**COPD-Specific Health Profiles**
	Corresponding AQ20 Domains
Mobility	Activities
Discomfort/symptoms	Symptoms
Usual activities	Activities
Mental function	Emotional functioning
Vitality	
Speech	
Vision	
Elimination	
Breathing	Symptoms
Sleeping	
Hearing	
Depression	Emotional Functioning
Distress	Emotional Functioning
Eating	
Sexual activity	
Components not covered by the 15D	Environmental stimuli

*Domains/components may appear more than once if applicable to more than one 15D dimension
AQ=Airway Questionnaire

CHAPTER 3

Content Validity of Preference-Based Measures for Economic Evaluation in Chronic Obstructive Pulmonary Disease

Title: Content validity of preference-based measures for economic evaluation in chronic obstructive pulmonary disease

Short Title: Content validity of preference-based measures in COPD

Ava Mehdipour, BSc[1]; Sachi O'Hoski, PT[1,2]; Marla K. Beauchamp, PT, PhD[1,,2,3]; Joshua Wald, MD[3,4]; Ayse Kuspinar, PT, PhD[1]

[1]School of Rehabilitation Science, McMaster University, 1400 Main St W, Hamilton, ON L8S 1C7, Canada
[2]Respiratory Research, West Park Healthcare Centre, Toronto, ON M6M 2J5, Canada
[3]Firestone Institute for Respiratory Health, 50 Charlton Ave E, Hamilton, ON L8N 4A6, Canada
[4]Department of Medicine, McMaster University, Hamilton, ON, Canada

Submitted for publishing to *COPD: Journal of Chronic Obstrucive Pulmonary Disease.*

Corresponding Author:
Ayse Kuspinar, PT, PhD
Assistant Professor
School of Rehabilitation Science
McMaster University
1400 Main St. W. Room 435, IAHS
L8S 1C7
Hamilton, Ontario
Canada
kuspinaa@mcmaster.ca

Abbreviation List:

15D: 15-Dimensional
AQoL-8D: Assessment of Quality of Life 8-Dimensions
COPD: Chronic Obstructive Pulmonary Disease
EQ-5D: EuroQol 5-Dimensions
FEV1: Forced Expiratory Volume in 1 second
FVC: Forced Vital Capacity
GPBM: Generic Preference-Based Measure
HRQoL: Health-Related Quality of Life
HUI 2: Health Utilities Index Mark 2
HUI 3: Health Utilities Index Mark 3
ICF: International Classification of Functioning, Disability and Health
PGI: Patient-Generated Index
QALYs: Quality-Adjusted Life Years
QWB-SA: Quality of Well-Being Self-Administered
SF-6D: Six-Dimensional Short Form Survey

Abstract

Generic preference-based measures (GPBMs) are health-related quality of life (HRQoL) measures commonly used to evaluate the cost-utility of interventions in healthcare. However, the degree to which the content of GPBMs reflect the HRQoL of individuals with chronic obstructive pulmonary disease (COPD) has not yet been assessed. The purpose of this study was to examine the content and convergent validity of GPBMs in people with COPD. COPD patients were recruited from healthcare centers in Ontario, Canada. The Patient-Generated Index (PGI) (an individualized HRQoL measure) and the RAND-36 (to obtain SF-6D scores; a GPBM) were administered. Life areas nominated with the PGI were coded using the International Classification of Functioning Disability and Health and mapped onto GPBMs. We included 60 participants with a mean age of 70 and FEV1 % predicted of 43. The mean PGI score was 34.55/100 and the top three overarching areas that emerged were: ‘mobility’ (25.93%), ‘recreation and leisure’ (25.19%) and ‘domestic life’ (19.26%). Mapping of the nominated areas revealed that the Quality of Well-Being scale covered the highest number of areas (84.62%), Health Utilities Indices covered the least (15.38% and 30.77%) and other GPBMs covered between 46-62%. A correlation of 0.32 was calculated between the SF-6D and the PGI. The majority of GPBMs covered approximately half of the areas reported as being important to individuals with COPD. When areas relevant to COPD are not captured, HRQoL scores generated by these measures may inaccurately reflect patients’ values and affect cost-effectiveness decisions.

Keywords: COPD; HRQoL; Preference-based measures; Content validity

Introduction

Health-related quality of life (HRQoL) is “an individual’s perception of how an illness and its treatment affect the physical, mental and social aspects of his or her life” [1]. Different methods of measuring HRQoL have been developed and can be used in research to assign a value to one’s overall HRQoL. Among these methods are generic preference-based measures (GPBMs), which are patient-reported outcome measures of HRQoL that can be used for cost-utility analyses of different interventions [2]. Some well-known GPBMs are the EuroQol 5-Dimensions (EQ-5D), the Six-Dimensional Short Form Survey (SF-6D) and the Health Utilities Index Mark 3 (HUI3) [3]. They are typically scored from 0.0 (death) to 1.0 (perfect-health), and this value of HRQoL can be used to calculate quality-adjusted life years (QALYs) for an intervention by multiplying it by the number of years the intervention is predicted to extend life. QALYs can be used by healthcare professionals and policymakers to make decisions about resource allocation and implementation of interventions.

Individuals with chronic obstructive pulmonary disease (COPD) experience respiratory symptoms, such as cough, difficulty breathing and fatigue, which have been found to affect HRQoL [4,5]. Luckily, many treatments have shown to increase health status in people with COPD [6]. The use of GPBMs in COPD can help determine which treatments are more effective in terms of both quality and quantity of life. However, before a measure is used to make cost-effectiveness decisions for a specific population, its psychometric properties should be tested to ensure its reliability and validity [7]. Content validity of GPBMs in people with COPD has not yet been evaluated and is a fundamental step in establishing a measure’s validity as it assesses whether the measure reflects the construct under study [8]. Therefore, the primary objective of this study is to assess the content validity of GPBMs by estimating the extent to which GPBMs capture domains of quality of life that are important to individuals with COPD, as measured by the Patient-Generated Index (PGI). The secondary objective of this study is to examine the convergent validity of a well-known GPBM; the SF-6D [3], against the PGI.

Methods

Participants

Participants were recruited from outpatient clinics and pulmonary rehabilitation programs at two academic centers in Ontario. Ethics approval was obtained from both sites, from respective research ethics boards (Joint West Park Healthcare Centre-The Salvation Army Toronto Grace Health Centre Research Ethics Board #17-013WP; Hamilton Integrated Research Ethics Board #7661). Eligibility criteria for the study included: (1) over the age of 18, (2) a clinical physician-diagnosis of COPD, and (3) smoking history of at least 10 pack-years. Individuals who were not able to speak/understand English and those with a severe disability (caused by a musculoskeletal or neurological condition unrelated to their COPD) were excluded.

Outcome measures

Sociodemographic and clinical characteristics

Sociodemographic information, such as sex, age, number of pack years, oxygen use and mobility aid use, and clinical information, such as comorbidities and spirometry results (i.e., forced expiratory volume in one second (FEV1), forced vital capacity (FVC)), were obtained.

The Patient-Generated Index (PGI)

The PGI has been utilized in previous content validity studies to identify areas of quality of life important to individuals with chronic conditions [9–11]. This individualized measure of HRQoL was administered in three stages. First, participants were asked to list up to five most important areas of their life affected by their COPD, with the last/sixth item being: ‘all other areas of life that are not mentioned above’. Second, participants were asked to rate each area on a scale from 0 (the worst you could imagine) to 10 (exactly as you would like it to be), relative to the past month. Third, participants were given 12 imaginary points and asked to distribute these points among the areas which they would like to have improved; more points being allocated to areas with more hopes of improvement. The rating of each area and the proportion of complementary points allocated were multiplied and summed to produce a total score of HRQoL on a scale from 0 to 10; with higher scores indicating better HRQoL [12]. This score is typically reported as a percentage [13].

The Six-Dimensional Short Form Survey (SF-6D)

The SF-6D is a commonly-used GPBM, developed by Brazier et al. [14,15], from the SF-36 (generic health profile). The SF-6D defines 18,000 health states and items cover 6 dimensions: physical functioning, role limitation, social functioning, pain, mental health and vitality [16,17]. The RAND-36, a distributable version of the SF-36, was used to obtain SF-6D scores as recommended by the developers [18]. The RAND-36 is a 36-item questionnaire that covers various domains of HRQoL, across 8 scales, varying from physical functioning to mental health and social functioning, summed into 2 subscales (Physical and Mental Health) [19]. Scores obtained from the RAND-36 were transformed to SF-6D scores using an algorithm developed by Kharroubi et al. [20], using non-parametric Bayesian preference weights. The SF-6D produces a HRQoL score from 0.2 (worst possible health state) to 1.0 (perfect health state) [20]. Permission to use the SF-6D algorithm was obtained from the developers.

Procedure

Eligible participants who provided informed consent completed the PGI and the RAND-36 in person or over the phone. The areas reported from the PGI were coded independently by two reviewers (AM and SO) using the World Health Organization's International Classification of Functioning, Disability and Health (ICF) [21]. A third reviewer (AK) was consulted if agreement between the reviewers was not reached. The most specific code was selected for each reported area, and if the reported area covered more than one code, then all codes were stated. Similar codes were then pooled together (e.g., 'recreation and leisure, unspecified' and 'recreation and leisure, other specified').

Overarching domains were identified from the codes and mapped onto GPBMs: the EQ-5D, the SF-6D, the Health Utilities Index Mark 2 (HUI2), the HUI3, the Assessment of Quality of Life 8-Dimensions (AQoL-8D), the 15-Dimensional (15D) and the Quality of Well-Being Self-Administered (QWB-SA) scale [3]. Mapping was also performed independently by two reviewers (AM and SO) with a third reviewer (AK) for consultation, if needed. This methodology followed previous studies examining content validity of GPBMs using the PGI [9,10]. A flow diagram of the study's procedure is outlined in Figure 1.

Statistical analysis

All statistical analyses were performed using Stata, version 15.1 (StataCorp, College Station, TX, USA). Descriptive statistics (mean and standard deviation, or frequency and percentage) were calculated to analyze participants' sociodemographic/clinical information, ICF codes/domains identified and domains covered by GPBMs. A Pearson's correlation coefficient was calculated to assess the correlation between the SF-6D and PGI scores. A positive correlation coefficient of at least 0.5 was hypothesized between the PGI and the SF-6D [22].

Sample size

There are no specific sample size estimates for content validation; therefore, our sample size was based on the number needed to achieve saturation. Common saturation guidelines agree that saturation for qualitative analysis is achieved at small sample sizes (e.g., around 20-30) and usually do not need to be greater than 60 [23].

Results

Sample characteristics

Table 1 outlines the clinical and sociodemographic characteristics for the study sample. For our 60 participants, the mean age of the sample was 70 years and approximately 57% were males. On average, participants had a smoking history of 44 pack-years; 45% used supplemental oxygen and 50% used a mobility aid (e.g., walker, cane, wheelchair). The mean FEV1 % predicted of the sample was approximately 43, with the majority having severe to very severe airflow obstruction (GOLD stage 3-4) [6]. The most common comorbidities were cardiac and/or respiratory (e.g., asthma). The mean PGI score was approximately 35 out of 100, with 100 being the highest self-reported HRQoL. The mean SF-6D score was 0.57 out of 1, with 1 representing best HRQoL.

Life areas important to COPD

Nineteen overarching domains were identified and thirteen appeared more than once. Table 2 presents the thirteen domains. The top three overarching domains were ‘mobility’ (25.93%), ‘recreation and leisure’ (25.19%) and ‘domestic life’ (19.26%). Specifically, ‘mobility’ included walking and using transportation, ‘recreation and leisure’ included socializing, hobbies and sports, and ‘domestic life’ included housework, preparing meals and shopping.

Figure 2 outlines the mean severity rating (from 0 to 10, where 0 is the worst and 10 is the best one could imagine that area to be) of each overarching domain. Although, ‘work and employment’ was reported only 8 times, it was found to be the area most severely impacted by COPD with a mean score close to 2 out of 10 (very poor). ‘Mobility’, ‘recreation and leisure’, ‘domestic life’ and ‘interpersonal relationships’ were also severely affected with mean scores ranging from 3 (poor) to 4 (between poor and fair).

Figure 3 outlines the mean number of points (out of 12) that participants allocated to the overarching domains, indicating their desire for improvement in that area. With a frequency of 3, ‘respiratory system functions’ (e.g., breathing) was the area most desired for improvement (mean 6 points; 50% of their points), followed by ‘environmental factors’ (e.g., weather conditions) (mean 4.4 points; 37% of their points) and ‘mobility’ (mean 4 points; 33% of their points). Participants’ spent on average 2.5 points (21% of their points) on ‘recreation and leisure’, ‘domestic life’, ‘interpersonal relationships’ and ‘mental functions’ each.

Content validity

Table 3 presents the mapping of the overarching domains against items on the GPBMs. The QWB-SA covered the highest number of domains important to individuals with COPD (84.62%) and the HUIs covered the least (15.38% and 30.77%). The rest of the GPBMs covered between 46-62%. ‘Mobility’ and ‘mental functions’ domains were covered by all the measures, and ‘environmental factors’ and ‘looking after one’s health’ were not covered by any of the measures. ‘Recreation and leisure’ and ‘domestic life’, areas commonly reported by participants, were covered by the EQ-5D, SF-6D, AQoL-8D, 15D and QWB-SA, but not by HUI2 and HUI3.

‘Interpersonal relationships’ was covered by the AQoL-8D, 15D and QWB-SA, but not by EQ-5D, SF-6D, HUI2 and HUI3.

Convergent validity

A Pearson’s correlation coefficient of 0.32 was calculated between the PGI and the SF-6D. Figure 4 presents a scatter plot of SF-6D scores against PGI scores. Correlation values between the two measures did not fall around the line of best fit and were scattered, but did follow an upward trend, indicating a weak positive correlation between the measures [22].

Discussion

To our knowledge, this was the first study to evaluate the content validity of GPBMs in individuals with COPD. Areas of life most affected by COPD were identified by people with COPD, coded using the ICF and mapped onto GPBMs. A major finding of this study was that the majority of GPBMs covered only half of the areas reported as being important to individuals with COPD. In particular, several domains, such as respiratory problems, interpersonal relationships and work and employment, were missing from one or more of the GPBMs. We also found the SF-6D, a well-known GPBM, to be weakly associated with the PGI, an individualized measure of HRQoL capturing issues COPD patients consider important. Taken together, these findings suggest that GPBMs may not necessarily be suitable for assessing the HRQoL of COPD patients for cost-effectiveness analyses.

Many of the domains reported by patients with COPD were both severely affected and had a large proportion of points allocated to them, indicating their importance to participants. Mobility, for example, was not only an area that was severely impacted, but also an area that participants desired to improve notably. Without mobility, other aspects of life may become impaired. Being able to leave one's house can help expand one's social circle and allow for engagement in meaningful activities [24]. Similarly, physical movement is needed to engage in sports or perform chores around the house. This was evident in our findings as individuals with COPD highly reported social and participation restrictions in addition to mobility. Respiratory function was the second most impacted area by COPD and was given the highest amount of points in terms of desire for improvement. Even though this area was not highly reported, this finding suggests that among those listing it as important, they found it to be severely impacted by COPD and valued it highly by allocating, on average, half of their points to this area.

One of the biggest advantages of GPBMs is that they can be used for economic evaluation purposes to determine the cost-utility of alternative treatments and programs. They allow the different dimensions of health to be combined into a single index with anchors from 0 (death) to 1 (perfect health). GPBMs attach explicit weights to the various dimensions of health, allowing trade-offs to be made between them [17]. However, in the context of COPD, the majority of GPBMs, including the most widely used GPBM for cost-effectiveness analysis; the EQ-5D [3], only covered approximately half of the areas reported as being important to patients. Interpersonal relationships, a frequently-reported affected area, along with carrying/lifting objects, changing/maintaining body positions and respiratory problems were not covered by the majority of these measures. If such aspects are not captured by preference-based measures, then the overall HRQoL score may be inaccurate in terms of its reflection of patients' values, and thus, the cost-effectiveness of healthcare interventions and decisions made based on these results may also be inaccurate.

The HUIs covered less than one third of the areas nominated by COPD patients. The HUI3 evolved from the HUI1 and HUI2 [25], which were originally developed for infants and children [26,27]. Although the HUI2 has been applied in older populations (i.e., Alzheimer's disease) [28], its validity was not tested and some domains, such as 'fertility', remain relevant to younger populations. HUI2 and HUI3 focus on sensory difficulties, which is not necessarily relevant to a respiratory disease population. The HUIs were developed using the "within the skin" definition of health status, which focuses on impairments and excludes social interactions [25,29]. Therefore, frequently reported areas, such as recreation and leisure, domestic life and interpersonal relationships, that encompass social aspects of HRQoL were not covered by these measures.

The QWB-SA is a comprehensive measure of HRQoL encompassing 58 symptoms (mental, acute physical and chronic) [30]. Even though the QWB-SA covered many of the life areas reported by participants, it is not as widely used as other preference-based measures like the EQ-5D [3]. This may be because it consists of 71 items and has a 14-minute completion time in older adults [31], compared to the EQ-5D which consists of 5 items and only takes a few minutes to complete [32]. Furthermore, the QWB-SA is heavily focused on symptoms, which can be burdensome for respondents if they do not possess the listed symptoms. In our study, when asked about the important areas of life affected by COPD, none of the chronic symptoms and only 2 of the acute symptoms on the QWB-SA were mentioned by participants (i.e., shortness of breath and difficulty walking/standing). Having HRQoL measures with short administration times that target important areas affected by COPD may be valuable, providing accurate and easy to implement tools for cost-effectiveness analyses in clinical trials focused on patients with COPD.

A limitation of this study is that the sample comprised a low percentage of individuals with mild airflow limitations (5.17%). A recent study using data from the Canadian Cohort

Obstructive Lung Disease (CanCOLD) study found two-thirds of the cohort to be undiagnosed for COPD [33]. These individuals were not given a clinical diagnosis but had airflow obstruction according to spirometry tests [33]. Even though individuals with mild airflow limitations present fewer symptoms [34], they compose a large portion of the population and their perspectives may have not been completely captured in our study. However, the disease severity of our sample was comparable to other COPD samples in the measurement literature [35–39]. A second limitation of this study is the comparability of findings to other healthcare settings. Since recruitment was performed at tertiary care settings, findings may not be transferable to other settings (e.g., primary care settings). Last, for the PGI, participants were asked to list the most important areas of their life affected by their COPD. The phrasing of this question elicits reference to life activities and may result in less identification of the symptoms relevant to the disease. For example, respiratory system functions such as difficulty breathing, well-known to impact the COPD population [6], were not highly endorsed by this sample.

Conclusions

GPBMs form the basis for cost-effectiveness analysis and resource allocation decisions within the healthcare system, however, our findings showed that not a single measure covered all life areas important to those living with COPD and that their association with an individualized measure of HRQoL is weak. The content of preference-based measures should be reflective of the population's health concerns for accurate economic evaluation of treatments [40]. When GPBMs are used to evaluate the cost-utility of interventions in COPD, they may not always be sensitive to the concerns and values of individuals with COPD, which may result in inaccurate recommendations. Findings from this study suggest that a COPD-specific preference-based measure could be developed in order to more accurately reflect the health concerns of individuals living with COPD. Until such a measure is developed, researchers and policymakers can use these findings to make informed decisions when selecting a GPBM for cost-effectiveness analyses of interventions in the COPD population.

Author contributions: All authors (AM, SO, MKB, JW, AK) contributed to the study's conception and/or design. Data collection and analysis was performed by AM and SO under the supervision of AK. AM, MKB, JW and AK contributed to the study's interpretations. First draft of the manuscript was written by AM. Preliminary edits were made by AK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript, and AK is the guarantor of the paper.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Mayo NE. ISOQOL Dictionary of Quality of Life and Health Outcomes Measurement. ISOQOL; 2015.

[2] Neumann PJ, Goldie SJ, Weinstein MC. Preference-Based Measures in Economic Evaluation in Health Care. Annu Rev Public Health. 2000;21:587-611.

[3] Brazier J, Ara R, Rowen D, et al. A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. PharmacoEconomics. 2017;35:21-31.

[4] Breslin E, Van Der Schans C, Breukink S, et al. Perception of fatigue and quality of life in patients with COPD. Chest. 1998;114(4):958–964.

[5] Miravitlles M, Ribera A. Understanding the impact of symptoms on the burden of COPD. Respir Res. 2017;18:67.

[6] Global Initiative for Chronic Obstructive Lung Disease. GLOBAL STRATEGY FOR THE DIAGNOSIS, MANAGEMENT, AND PREVENTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE (2020 REPORT). Glob Initiat Chronic Obstr Lung Dis. 2020.

[7] Cláudia de Souza A, Maria Costa Alexandre N, de Brito Guirardello E. Psychometric properties in instruments evaluation of reliability and validity. Appl Epidemiol Epidemiol Serv Saude Brasília. 2017;26(3):649–659.

[8] De Vet HCW, Terwee CB, Mokkink LB, et al. Measurement in medicine: A practical guide. Meas. Med. Pract. Guide. 2011.

[9] Kuspinar A, Mayo NE. Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? Health Qual Life Outcomes. 2013;11:71.

[10] Kuspinar A, Mate K, Lafontaine AL, et al. Evaluating the content validity of generic preference-based measures for use in Parkinson’s disease. Parkinsonism Relat Disord. 2019;62:112-6.

[11] Mayo NE, Aburub A, Brouillette MJ, et al. In support of an individualized approach to assessing quality of life: comparison between Patient Generated Index and standardized measures across four health conditions. Qual Life Res. 2017;26(3):601-609.

[12] Patel KK, Veenstra DL, Patrick DL. A review of selected patient-generated outcome measures and their application in clinical trials. Value Health. 2003;6(5):595-603.

[13] Martin F, Camfield L, Rodham K, et al. Twelve years’ experience with the Patient Generated Index (PGI) of quality of life: A graded structured review. Qual Life Res. 2007;16(4):705–715.

[14] Brazier J, Usherwood T, Harper R, et al. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol. 1998;51(11):1115-1128.

[15] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271–292.

[16] Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14(6):1523-1532.

[17] Brazier J, Ratcliffe J, Saloman J, et al. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press; 2017.

[18] The University of Sheffield [Internet]. Available from: https://www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d/faqs.

[19] Hays RD, Sherbourne CD, Mazel RM. The rand 36-item health survey 1.0. Health Econ. 1993;2(3):217-227.

[20] Kharroubi SA, Brazier JE, Roberts J, et al. Modelling SF-6D health state preference data using a nonparametric Bayesian method. J Health Econ. 2007;26(3):597-612.

[21] World Health Organization. World Health Organisation. (2001). International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organisation. Int Classif. 2001.

[22] Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.

[23] Mason M. Sample size and saturation in PhD studies using qualitative interviews. Forum Qual SozialforschungForum Qual Soc Res. 2010;11(3).

[24] Rosso AL, Taylor JA, Tabb LP, et al. Mobility, disability, and social engagement in older adults. J Aging Health. 2013;25(4):617-637.

[25] Horsman J, Furlong W, Feeny D, et al. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes. 2003;1:54.

[26] Torrance GW, Boyle MH, Horwood SP. Application of Multi-Attribute Theory to Measure Social Preferences for Health States. Oper Res. 1982;30(6):1043-69.

[27] Boyle MH, Torrance GW, Sinclair JC, et al. Economic Evaluation of Neonatal Intensive Care of Very-Low-Birth-Weight Infants. N Engl J Med. 1983;308(22):1330-7.

[28] Neumann PJ, Kuntz KM, Leon J, et al. Health utilities in Alzheimer's disease: a cross-sectional study of patients and caregivers. Med Care. 1999;37(1):27–32.

[29] Ware JE, Brook RH, Davies AR, et al. Choosing measures of health status for individuals in general populations. Am J Public Health. 1981;71(6):620-5.

[30] Seiber WJ, Groessl EJ, David KM, et al. Quality of well being self-administered (QWB-SA) scale. San Diego Health Serv Res Cent Univ Calif. 2008.

[31] Andresen EM, Rothenberg BM, Kaplan RM. Performance of a Self-Administered Mailed Version of the Quality of Well-Being (QWB-SA) Questionnaire among Older Adults. Med Care. 1998;36(9):1349–1360.

[32] EQ-5D-5L User Guide [Internet]. EuroQol Res. Found. 2015. Available from: https://euroqol.org/publications/user-guides.

[33] Labonté LE, Tan WC, Li PZ, et al. Undiagnosed Chronic Obstructive Pulmonary Disease Contributes to the Burden of Health Care Use. Data from the CanCOLD Study. Am J Respir Crit Care Med. 2016;194(3):285–298.

[34] Martinez CH, Mannino DM, Jaimes FA, et al. Undiagnosed obstructive lung disease in the United States. Associated factors and long-term mortality. Ann Am Thorac Soc. 2015;12(12):1788–1795.

[35] Polley L, Yaman N, Heaney L, et al. Impact of cough across different chronic respiratory diseases: Comparison of two cough-specific health-related quality of life questionnaires. Chest. 2008;134(2):295–302.

[36] Rutten-Van Mölken MPMH, Oostenbrink JB, Tashkin DP, et al. Does quality of life of COPD patients as measured by the generic EuroQol five-dimension questionnaire differentiate between COPD severity stages? Chest. 2006;130(4):1117–1128.

[37] Stavem K. Reliability, validity and responsiveness of two multiattribute utility measures in patients with chronic obstructive pulmonary disease. Qual Life Res. 1999;8(1-2):45–54.

[38] Harper R, Brazier JE, Waterhouse JC, et al. Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. Thorax. 1997;52(10):879–887.

[39] Puhan MA, Guyatt GH, Goldstein R, et al. Relative responsiveness of the Chronic Respiratory Questionnaire, St. Georges Respiratory Questionnaire and four other health-related quality of life instruments for patients with chronic lung disease. Respir Med. 2007;101(2):308–316.

[40] Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. J. Health Serv. Res. Policy. 1999;4(3):174-84.

Table 1. Clinical and sociodemographic characteristics of sample (N=60).

Characteristic	N (%) [unless specified otherwise]
Mean age (SD)	69.7 (7.99)
Males	34 (56.67)
Mean pack-years (SD)	43.71 (16.82)
Oxygen Use	27 (45.00)
Mobility Aid Use	30 (50.00)
Mean FEV1 % predicted (SD)	42.98 (21.66)[a]
Mean FEV1/FVC % (SD)	45.84 (15.65)[b]
GOLD 1	3 (5.17)[a]
GOLD 2	17 (29.31)[a]
GOLD 3	18 (31.03)[a]
GOLD 4	20 (34.48)[a]
Cardiac comorbidities	41 (68.33)
Respiratory comorbidities	33 (55.00)
Rheumatology comorbidities	16 (26.67)
Gastro-intestinal comorbidities	16 (26.67)
Cancer comorbidities	13 (21.67)
Vascular comorbidities	11 (18.33)
Other co-morbidities	49 (81.67)
Mean PGI score (SD) [0-100]	34.55 (20.19)
Mean SF-6D score (SD) [0-1]	0.57 (0.09)

FEV1=forced expiratory volume in one second, FVC=forced vital capacity, N=sample size, PGI=Patient-Generated Index, SD=standard deviation

[a] Missing data (N=58), [b] Missing data (N=54)

Table 2. Overarching domains identified more than once from the Patient-Generated Index (total n=270).

Frequency n (%)	Overarching Domain	ICF Component	ICF Codes	Code Frequency n (%)
70 (25.93)	Mobility	Activities and participation	Walking	17 (6.30)
			Mobility	11 (4.07)
			Using transportation	10 (3.7)
			Walking long distances	8 (2.96)
			Climbing	6 (2.22)
			Swimming	5 (1.85)
			Moving around outside the home and other buildings	5 (1.85)
			Walking on different surfaces	3 (1.11)
			Running	2 (0.74)
			Driving motorized vehicles	2 (0.74)
			Driving human-powered transportation	1 (0.37)
68 (25.19)	Recreation and leisure	Activities and participation	Socializing	22 (8.15)
			Hobbies	17 (6.30)
			Sports	12 (4.44)
			Play	8 (2.96)
			Recreation and leisure	5 (1.85)
			Community, social and civic life, other specified	3 (1.11)
			Arts and culture	1 (0.37)
52 (19.26)	Domestic life	Activities and participation	Housework	19 (7.04)
			Preparing meals	9 (3.33)
			Cleaning living area	7 (2.59)
			Shopping	5 (1.85)
			Taking care of plants, indoors and outdoors	3 (1.11)
			Maintaining dwelling and furnishings	2 (0.74)
			Washing and drying clothes and garments	2 (0.74)

			Domestic life	2 (0.74)
			Taking care of animals	1 (0.37)
			Caring for household objects	1 (0.37)
			Maintaining domestic appliances	1 (0.37)
28 (10.37)	Interpersonal relationships	Activities and participation	Family relationships	13 (4.81)
			Informal relationships with friends	5 (1.85)
			Sexual relationships	4 (1.48)
			Interpersonal interactions and relationships	3 (1.11)
			Informal social relationships	2 (0.74)
			Parent-child relationships	1 (0.37)
10 (3.7)	Mental functions	Activities and participation	Emotional functions	6 (2.22)
			Energy level	2 (0.74)
			Openness to experience	1 (0.37)
			Confidence	1 (0.37)
8 (2.96)	Work and employment	Activities and participation	Remunerative employment	7 (2.59)
			Non-remunerative employment	1 (0.37)
6 (2.22)	Carrying/lifting objects	Activities and participation	Lifting and carrying	3 (1.11)
			Lifting	2 (0.74)
			Carrying in the hands	1 (0.37)
5 (1.85)	Self-care	Activities and participation	Washing whole body	5 (1.85)
4 (1.48)	Changing/maintaining body position	Activities and participation	Maintaining a standing position	2 (0.74)
			Bending	1 (0.37)
			Standing	1 (0.37)
4 (1.48)	Environmental factors	Environmental factors	Climate	4 (1.48)
4 (1.48)	Carrying out daily routine	Activities and participation	Carrying out daily routine	3 (1.11)
			Managing one’s own activity level	1 (0.37)

3 (1.11)	Respiratory system functions	Body functions	Respiratory functions	3 (1.11)
2 (0.74)	Looking after one’s health	Activities and participation	Maintaining one’s health	2 (0.74)

ICF=World Health Organization’s International Classification of Functioning, Disability and Health, n=number of appearances

Table 3. Mapping of overarching domains, identified by COPD patients, onto GPBMs.

Overarching Domains	Generic Preference-Based Measure						
	EQ-5D	SF-6D	HUI2	HUI3	AQoL-8D	15D	QWB-SA
Mobility	Y	Y	Y	Y	Y	Y	Y
Recreation and leisure	Y	Y	N	N	Y	Y	Y
Domestic life	Y	Y	N	N	Y	Y	Y
Interpersonal relationships	N	N	N	N	Y	Y	Y
Mental functions	Y	Y	Y	Y	Y	Y	Y
Work and employment	Y	Y	N	N	N	Y	Y
Carrying/lifting objects	N	N	Y	N	N	N	Y
Self-care	Y	Y	Y	N	Y	N	Y
Changing/maintaining body position	N	N	N	N	N	N	Y
Environmental factors	N	N	N	N	N	N	N
Carrying out daily routine	Y	Y	N	N	N	Y	Y
Respiratory system functions	N	N	N	N	N	Y	Y
Looking after one’s health	N	N	N	N	N	N	N
% of Yes	53.85%	53.85%	30.77%	15.38%	46.15%	61.54%	84.62%

Y=yes, it is covered by the measure. N=no, it is not covered by the measure
EQ-5D=EuroQol 5-Dimensions, SF-6D=Six-Dimensional Short Form Survey, HUI 2=Health Utilities Index Mark 2, HUI 3= Health Utilities Index Mark 3, AQoL-8D=Assessment of Quality of Life 8-Dimensions, 15D=15-Dimensional, QWB-SA=Quality of Well-Being Self-Administered

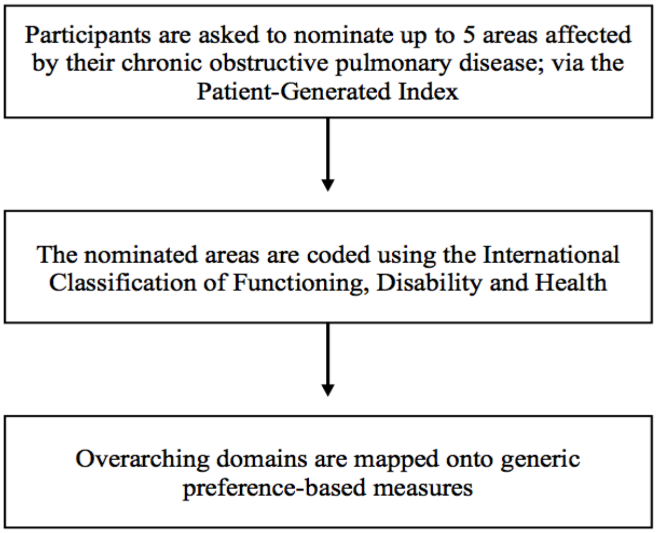
Participants are asked to nominate up to 5 areas affected by their chronic obstructive pulmonary disease; via the Patient-Generated Index
The nominated areas are coded using the International Classification of Functioning, Disability and Health
Overarching domains are mapped onto generic preference-based measures

Figure 1. Flow diagram outlining the study's procedure.

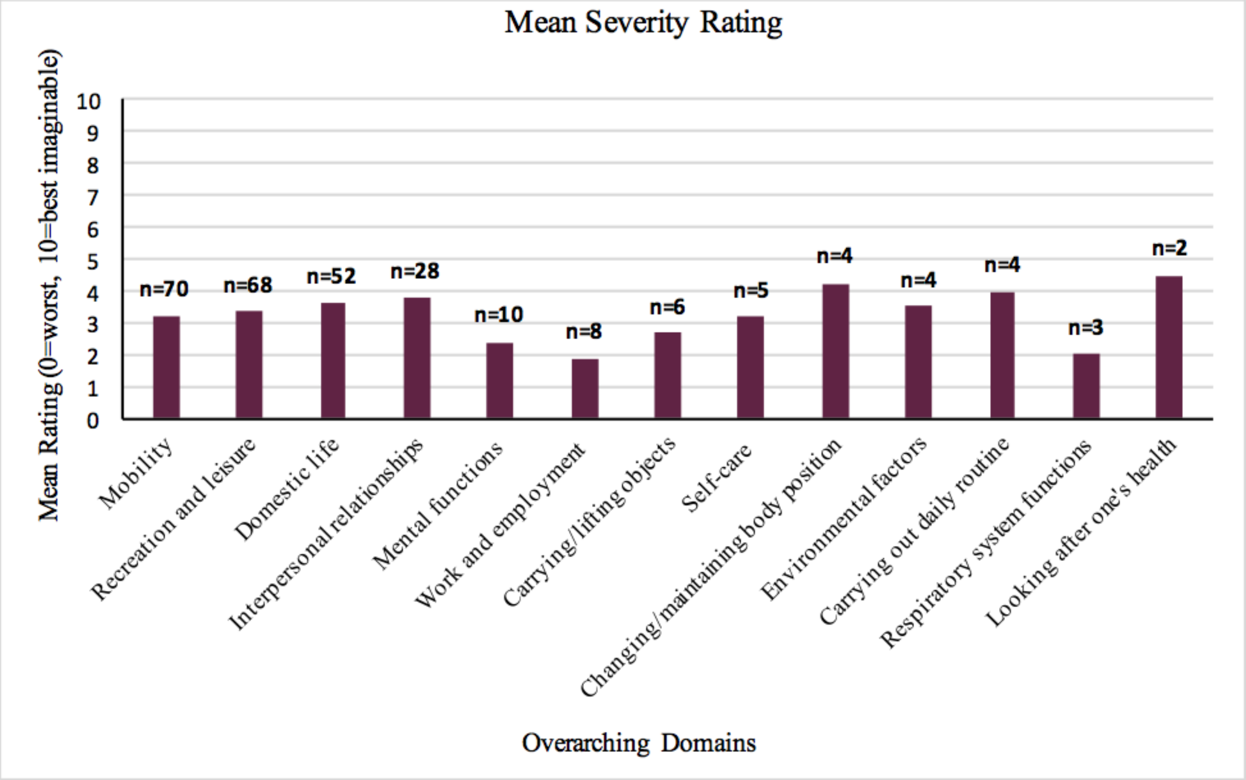

Figure 2. Mean severity rating given to each overarching domain appearing more than once, scaled from 0 (the worst one could imagine) to 10 (exactly as one would like it to be). n=number of appearances

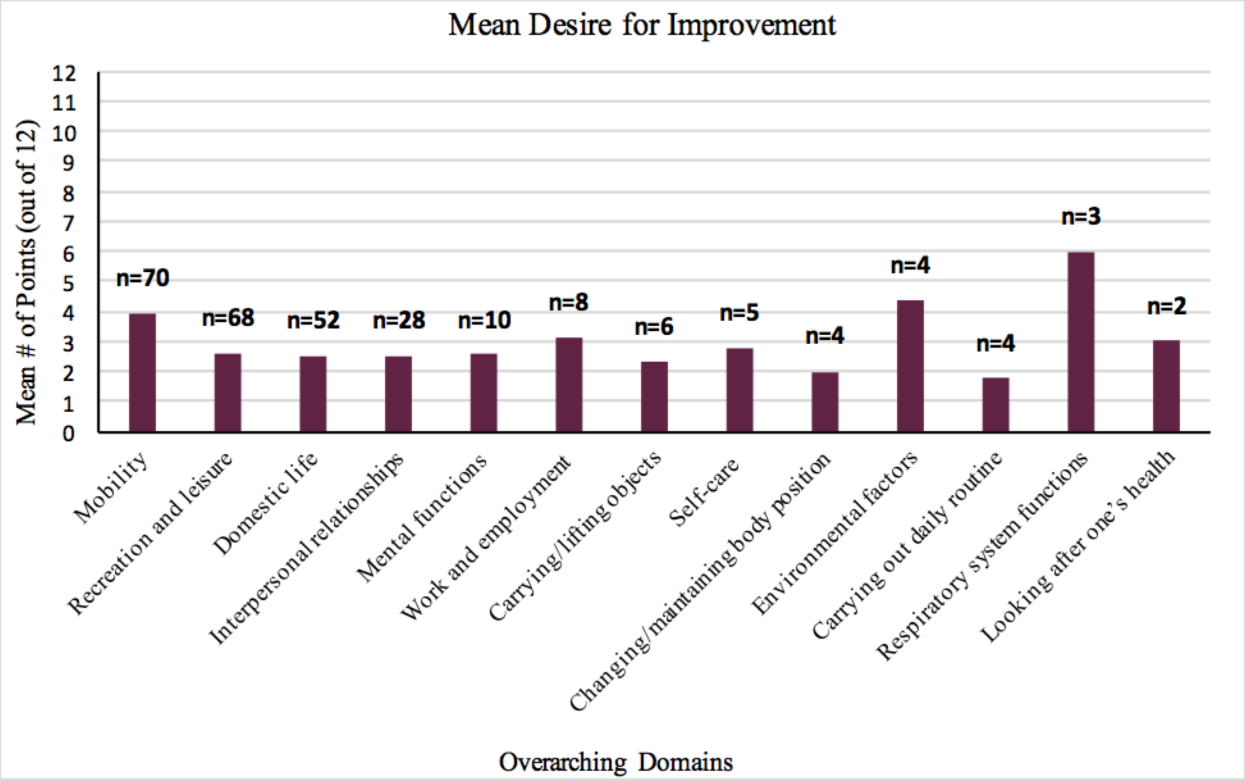

Figure 3. Mean number of points (out of 12) for improvement desires allocated to each overarching domain appearing more than once.
n=number of appearances

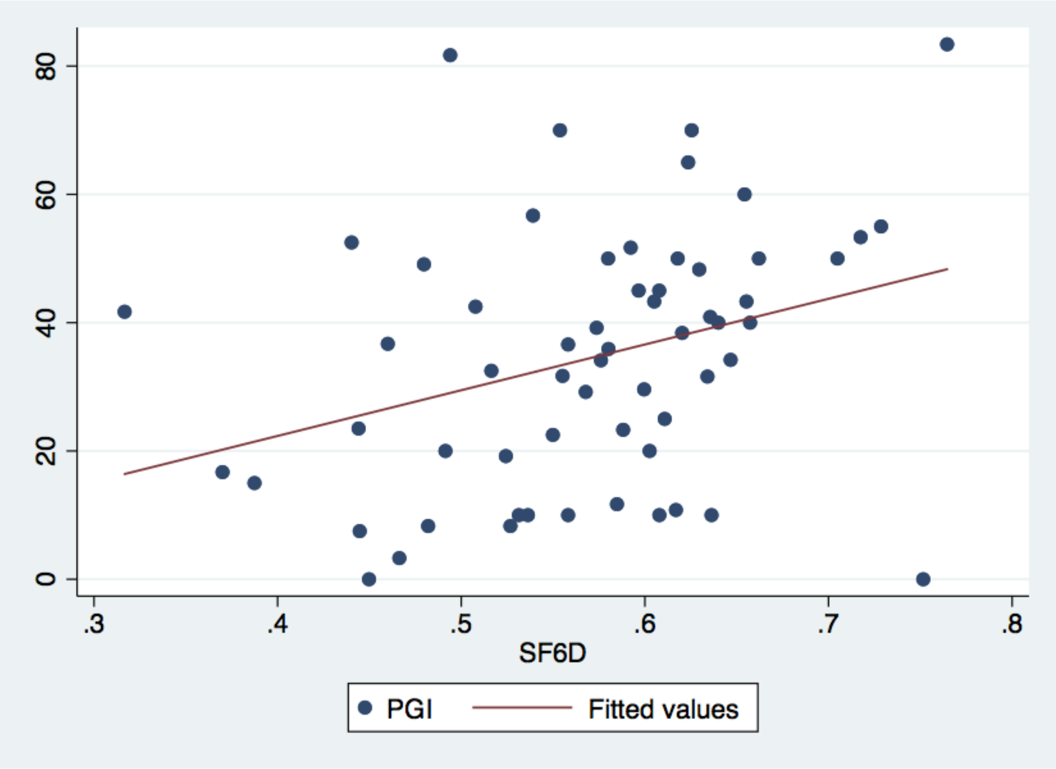

Figure 4. Scatter plot of SF-6D scores against Patient-Generated Index scores with a line of best fit.

CHAPTER 4

Discussion

1.0 Summary of Findings

GPBMs are HRQoL measures used to determine the cost-effectiveness of healthcare interventions (Brazier et al., 2017). The purpose of this thesis was to evaluate the psychometric properties of these measures in COPD. To do this, we conducted a systematic review of the available literature examining the psychometric properties of GPBMs in COPD (Chapter 2) and performed a cross-sectional study assessing the content validity of GPBMs in COPD (Chapter 3). Taken together, these two studies suggest that GPBMs may not necessarily be suitable for assessing the HRQoL of patients with COPD for cost-utility analyses.

Chapter 2 revealed that a large proportion of measurement studies involving GPBMs in COPD did not demonstrate responsiveness and were low in methodological quality. Our results showed that the effects of pulmonary rehabilitation were not always captured by GPBMs, which is concerning when it comes to cost-utility analyses as this intervention has been shown to provide many health benefits for patients with COPD (McCarthy et al., 2015). Our systematic review also revealed limitations in terms of the known-groups validity of GPBMs in COPD. A measure with adequate known-groups validity should be able to discriminate between different disease severities (e.g., moderate versus severe airflow obstructions). However, the majority of EQ-5D studies indicated that this GPBM lacked discriminatory ability; making cost-utility comparisons between different disease severities difficult.

Our systematic review also revealed an important gap in the literature; there were no studies that reported on the content validity of GPBMs in COPD. Thus, in Chapter 3, we conducted a cross-sectional study to examine the content validity of these measures in COPD. Content validity is a fundamental step in validity testing; if a measure's items do not reflect its construct then further evaluations (such as convergent validity) are unnecessary (De Vet et al., 2011). In order to reduce further errors and have an accurate representation of the construct under study, content validity needs to be established first (Haynes et al., 1995). Therefore, the need to evaluate the content validity of GPBMs in COPD was essential. Our results demonstrated that the content of GPBMs was not fully reflective of the areas of life important to people with COPD. Commonly used GPBMs, such as the EQ-5D (Brazier et al., 2017), covered approximately half of the areas important to individuals with COPD, suggesting that the content of GPBMs do not strongly support the construct of HRQoL in COPD.

1.1 Implications for Policymakers and Researchers

GPBMs have been endorsed by different national agencies around the world (e.g., Canadian Agency for Drugs and Technologies in Health and the United Kingdom's National Institute of Health and Care Excellence) for economic evaluation purposes (Canadian Agency for Drugs and Technologies in Health, 2017; National Institute for Health and Care Excellence, 2013; Rowen, Zouraq, et al., 2017). However, findings from this thesis suggest that researchers should be aware of potential limitations in using GPBMs in patients with COPD as we found GPBMs to inadequately detect changes in health status, discriminate between disease severities and capture areas of life important to patients. For example, the EQ-5D, a commonly recommended GPBM for cost-utility analysis (Rowen, Zouraq, et al., 2017), had weak known-groups validity, responsiveness and content validity in COPD. This suggests that policymakers and researchers should be cautious when making decisions as to which interventions and programs to implement in individuals with COPD based on the EQ-5D. Furthermore, conclusions drawn from studies that used the EQ-5D to assess quality of life in COPD or for cost-utility analyses should be considered carefully.

2.0 Future Research: Condition-Specific Preference-Based Measures (CSPBMs)

Findings from Chapters 2 and 3 highlighted a gap in the literature; the need for a HRQoL measure which can be used for cost-effectiveness analysis that is sensitive to change and that captures areas of life important to individuals with COPD. Brazier et al. (2012) performed psychometric analyses on nine existing data sets concerning condition-specific preference-based measures (CSPBMs) in different disease populations. They compared the performance of CSPBMs to GPBMs and found that CSPBMs had better known-groups validity and lower ceiling effects. They also found a respiratory disease CSPBM (i.e., a CSPBM for asthma) to have better responsiveness compared to a GPBM. A CSPBM specific to COPD may allow for more accurate cost-utility assessments by specifically targeting HRQoL domains specific to individuals living with COPD.

There are two methods for developing CSPBMs: (1) from an existing condition-specific measure (e.g., COPD-specific health profile) or (2) ‘de novo’; a new measure. The University of Sheffield developed a 6-stage process for developing a CSPBM from an existing condition-specific measure (Brazier et al., 2012). First, factor analysis is used to determine dimensionality; either to confirm existing dimensions, propose different dimensions or establish dimensions. Then, Rasch and classical psychometric analyses are employed to eliminate and select item(s) to reflect each dimension, and item level reductions are considered and explored. The classification system is then validated on another dataset. After validation, health states are valued by the general population or the condition group. The argument for using general population weights is that society is the payer of these interventions (i.e., tax-payers), therefore, society’s values of health states should matter (Stamuli, 2011). Whereas, the argument for patient weights is that they are the one's experiencing the health states and the condition's impact on health (Stamuli, 2011).

The advantage to using an existing measure is that utilities can be generated for existing data sets (Brazier et al., 2012). However, these measures may not capture the entirety of individuals' HRQoL as they may be disease and symptom focused (Brazier et al., 2012). Therefore, new CSPBMs can be developed to better capture the holistic nature of HRQoL. The US Food and Drug Administration (FDA) (2009) outlines guidelines for the development of new patient-reported outcomes. Methods for developing CSPBMs using the 'de novo' method can also be found in the literature (e.g., the development of a preference-based stroke index) (Poissant et al., 2003). This approach involves adequate participation from the target population in the item generation and development stages. Similarly, after validation of the classification system, items are valued by the general population or condition group.

3.0 Overall Conclusions

The goal of this thesis was to evaluate the measurement properties of GPBMs in patients with COPD, in order to understand the performance and suitability of these measures for cost-utility analyses. Our findings showed that GPBMs may not be sensitive to and/or fully reflective of COPD patients' health concerns; hence, weakening the accuracy of cost-utility analyses of healthcare interventions for this population. Moreover, these studies were able to identify gaps in the literature pertaining to preference-based measures that can be addressed in future measurement work. Conclusions drawn from the two manuscripts suggest a need for the development of CSPBMs for people with COPD. Future studies should focus on the development of a COPD-specific preference-based measure, as it may be more sensitive and relevant to patients with COPD compared to GPBMs (Rowen, Brazier, et al., 2017).

References

Brazier, J., Ara, R., Rowen, D., & Chevrou-Severac, H. (2017). *A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models*.

Brazier, J. E., Rowen, D., Mavranezouli, I., Tsuchiya, A., Young, T., Yang, Y., Barkham, M., & Ibbotson, R. (2012). Developing and testing methods for deriving preferencebased measures of health from condition-specific measures (and other patient-based measures of outcome). In *Health Technology Assessment*. https://doi.org/10.3310/hta16320

Canadian Agency for Drugs and Technologies in Health. (2017). Guidelines for the Economic Evaluation of Health Technologies: Canada 4th Edition. *CADTH Methods and Guidelines*.

De Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in medicine: A practical guide. In *Measurement in Medicine: A Practical Guide*. https://doi.org/10.1017/CBO9780511996214

FDA, & HHS. (2009). Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. In *Guidance for Industry*.

Global Initiative for Chronic Obstructive Lung Disease. (2020). GLOBAL STRATEGY FOR THE DIAGNOSIS, MANAGEMENT, AND PREVENTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE (2020 REPORT). *Global Initiative for Chronic Obstructive Lung Disease*. https://doi.org/10.1097/00008483-200207000-00004

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238.

McCarthy, B., Casey, D., Devane, D., Murphy, K., Murphy, E., & Lacasse, Y. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews, 2*.

National Institute for Health and Care Excellence. (2013). Guide to the methods of technology appraisal 2013. *National Institute for Health and Care Excellence*. https://doi.org/10.2165/00019053-200826090-00002

Poissant, L., Mayo, N. E., Wood-Dauphinee, S., & Clarke, A. E. (2003). The development and preliminary validation of a Preference-based Stroke Index (PBSI). *Health and Quality of Life Outcomes*. https://doi.org/10.1186/1477-7525-1-43

Rowen, D., Brazier, J., Ara, R., & Azzabi Zouraq, I. (2017). The Role of Condition-Specific Preference-Based Measures in Health Technology Assessment. *PharmacoEconomics*, *35*(1), 33–41. https://doi.org/10.1007/s40273-017-0546-9

Rowen, D., Zouraq, I. A., Chevrou-Severac, H., & van Hout, B. (2017). International regulations and recommendations for utility data for health technology assessment. *Pharmacoeconomics*, *35*(1), 11–19.

Stamuli, E. (2011). Health outcomes in economic evaluation: Who should value health? *British Medical Bulletin*, *97*(1), 197–210. https://doi.org/10.1093/bmb/ldr001