Sleep Medicine Reviews 28 (2016) 18–31

Contents lists available at ScienceDirect

Sleep Medicine Reviews

journal homepage: www.elsevier.com/locate/smrv

CLINICAL REVIEW

Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): A systematic review

Inger L. Abma [a, *], Philip J. van der Wees [a], Vik Veer [b], Gert P. Westert [a], Maroeska Rovers [c, d]

[a] Radboud University Medical Center, Radboud Institute of Health Sciences, IQ Healthcare, Nijmegen, The Netherlands
[b] James Cook University Hospital, Middlesbrough, England, UK
[c] Radboud University Medical Center, Radboud Institute of Health Sciences, Department for Health Evidence, Nijmegen, The Netherlands
[d] Radboud University Medical Center, Radboud Institute of Health Sciences, Department for Operating Rooms, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 21 May 2015
Received in revised form
27 July 2015
Accepted 28 July 2015
Available online 7 August 2015

Keywords:
Patient-reported outcome measures
Obstructive sleep apnea
Measurement properties

SUMMARY

This systematic review summarizes the evidence regarding the quality of patient-reported outcome measures (PROMs) validated in patients with obstructive sleep apnea (OSA). We performed a systematic literature search of all PROMs validated in patients with OSA, and found 22 measures meeting our inclusion criteria. The quality of the studies was assessed using the consensus-based standards for the selection of health status measurement instruments (COSMIN) checklist. The results showed that most of the measurement properties of the PROMs were not, or not adequately, assessed. For many identified PROMs there was no involvement of patients with OSA during their development or before the PROM was tested in patients with OSA. Positive exceptions and the best current candidates for assessing health status in patients with OSA are the sleep apnea quality of life index (SAQLI), Maugeri obstructive sleep apnea syndrome (MOSAS) questionnaire, Quebec sleep questionnaire (QSQ) and the obstructive sleep apnea patient-oriented severity index (OSAPOSI). Even though there is not enough evidence to fully judge the quality of these PROMs as outcome measure, when interpreted with caution, they have the potential to add value to clinical research and clinical practice in evaluating aspects of health status that are important to patients.

Introduction

Obstructive sleep apnea (OSA) is characterized by repeated episodes of complete obstruction of the upper airway, resulting in oxygen desaturation and arousal from sleep. The prevalence of OSA is 2–5% in adult women and 3–7% in adult men [1]. The symptoms that these patients may experience are sleepiness, morning headaches, tiredness and fatigue, reduced vigilance and executive function, memory impairment, depression and impotence. Untreated OSA has been shown to be associated with cardiac pathologies (heart failure, arrhythmias, and ischemic heart disease) and stroke, as well as diabetes [1–3]. Specifically related to daytime sleepiness, the risk of road traffic accidents, near miss events and falling asleep at the wheel is significantly increased in severe OSA [4]. There is also evidence that untreated patients use more health services, take more medication, and are more often unemployed [4,5].

Successful treatment of OSA is often defined as demonstrating a reduction in the number of obstructive events occurring during each hour of sleep [6,7]. This is, however, weakly (or not at all) correlated with quality of life and daytime symptoms as experienced by patients with OSA [8–10]. To determine outcomes of treatment relevant to the experience of patients, patient-reported outcomes should be included for measuring the views of patients on their health and health-related quality of life [6,8,11].

* Corresponding author. Radboud University Medical Center, IQ Healthcare, PO box 9101, Huispost 114, 6500 HB, Nijmegen, The Netherlands. Tel.: +31 24 3616359.
E-mail address: Inger.abma@radboudumc.nl (I.L. Abma).

http://dx.doi.org/10.1016/j.smrv.2015.07.006

Abbreviations

BAI	Beck anxiety inventory
COSMIN	consensus-based standards for the selection of health status measurement instruments
CPAP	continuous positive airway pressure
EQ-5D	euroqol-5D
ESS	Epworth sleepiness scale
FLP	functional limitations profile
FOSQ	functional outcomes of sleep questionnaire
HADS	hospital anxiety and depression scale
ICC	intraclass correlation coefficient
MIC	minimal important change
MOSAS	Maugeri obstructive sleep apnea syndrome
NHP	Nottingham health profile
OSA	obstructive sleep apnea
OSAPOSI	obstructive sleep apnea patient-oriented severity index
PGI	patient-generated index
PROM	patient-reported outcome measure
QSQ	Quebec sleep questionnaire
SAQLI	sleep apnea quality of life index
SF-12	short-form 12
SF-36	short form 36
SNORE25	symptoms of nocturnal obstruction and related events-25
SOS	snore outcomes survey
SQS	sleep quality scale
SWIFT	sleepiness-wakefulness inability and fatigue test
ToDSS	time of day sleepiness scale
VAWS	visual analogical well-being scale

Glossary of terms

Patient-reported outcome measure	A questionnaire about health or functioning filled out by a patient
Construct	That which a questionnaire aims to measure (for example "sleepiness")
Reliability	The extent to which a measurement is free from measurement error. For questionnaires this is assessed with test–retest reliability: the questionnaire is filled out twice in a period of time when no true change has occurred in the construct.
Validity	The extent to which an instrument measures the construct that it aims to measure. Important aspects of validity are content validity (do the questions adequately cover the construct, as determined by the target population of the questionnaire) and construct validity (the degree to which scores are consistent with hypotheses about their correlation with scores of other instruments).
Responsiveness	The ability of an instrument to detect change over time in the construct to be measured.
Interpretability	The degree to which it is clear what the scores or change scores of an instrument mean.

These outcomes can be measured with patient-reported outcome measures (PROMs); questionnaires consisting of one or more multi-item scales, or single-item measures. These can be disease-specific, or generic. Disease-specific PROMs focus on the symptoms and/or impact on functioning related to a specific disease [12]. Generic PROMs aim to measure important general (aspects of) health-related quality of life or general functioning, such as mobility, or the degree to which the presence of health problems affects social functioning.

Initially, PROMs were developed for use in research, but in recent years their use has expanded to other areas, closer to clinical practice. That is, they can be used to assess the patient's health status prior to treatment and to support clinical decision-making. They may also be used after treatment to evaluate individual patient benefit by comparison with pre-treatment scores. When PROMs are operationalized as performance measures, they can be used to assess whether treatments by healthcare providers (and organizations) improve the health of patients [12,13].

For a valid and patient-centered evaluation of health status it is important that PROMs measure aspects of health status that are important to patients with OSA, and that their measurement characteristics are adequate for the specific patient population. Several literature reviews have assessed the measurement properties of different PROMs used in patients with OSA [14–18]. However, none of them provided an overview of the quality of all PROMs for outcome measurement in the specific target group of patients with OSA.

In this systematic review we therefore provide an overview of the quality of PROMs for health outcomes measurement which are validated in patients with OSA. This provides an evidence base for the choice of a PROM in clinical practice, for quality assessment, and in clinical research trials.

Methods

Identification of PROMs and validation studies

Literature search

A systematic search of the electronic databases MEDLINE, EMBASE and CINAHL from inception up to November 4th 2014 was conducted to identify all validation studies of PROMs assessed in patients with (suspected) OSA. Search terms used were "obstructive sleep apnea", "patient-reported outcome measure" and commonly used synonyms, acronyms, and related terms (Table S1). Additionally, we used the search filter for studies describing measurement properties developed by Terwee et al. [19] for our PubMed search, which has a sensitivity of 97.4%. For the other databases we developed a comparable filter with a similar approach to the PubMed version.

For each PROM identified in these studies we conducted an additional search to identify validation studies that our original search may have missed. We also performed a reference and related article search. Duplicate articles were manually filtered using the bibliographic EndNote database, version X5 (Thomas Reuters, New York City, NY, USA).

Selection of studies

Inclusion criteria for PROMs and validation studies

We included PROMs that have one or more eligible validation studies in adult patients with OSA and have outcome measurement as (one of) their aims. This means they are potentially suitable for use in evaluative situations. Furthermore, the PROMs needed to have been named, allowing identification. The aim of the PROM should be to capture general aspects of health status (such as

functional status, general health-related quality of life), OSA-related quality of life, or symptoms associated specifically with OSA, including sleepiness and fatigue, snoring and restless sleep, and anxiety and depression [20].

Validation studies were included if they studied the PROM in its original language of development, and if they were published as original and full text studies in English or Dutch. Furthermore, the findings needed to be presented for patients with OSA separately from any other study population, such as patients with other disorders causing sleepiness.

Two reviewers (IA and VV) independently assessed the eligibility of the identified PROMs and papers. Any disagreements were resolved by discussion with a third reviewer (PvdW). Where necessary we contacted study authors for clarification and additional information to inform study selection.

Measurement properties

We used the taxonomy of measurement properties as constructed by the COSMIN panel [21]. There are three domains of measurement properties: reliability, validity and responsiveness. We assessed all aspects of these domains, except cross-cultural validity, as we did not include translated PROMs. Additionally, we assessed interpretability, which is not a measurement property in itself but is an important characteristic of a measurement instrument.

Reliability

The reliability of a measurement instrument expresses to which extent scores are free from measurement error. It consists of three measurement properties:

- *Internal consistency:* measures to what extent items in a one-dimensional (sub)scale are related. It is commonly reported with the parameter Cronbach's α, which expresses the correlation between the items in the (sub)scale. A separate factor analysis (see construct validity) is needed to assess the dimensionality of a scale before Cronbach's α can be interpreted [22].
- *Reliability:* expresses the variance in the measurements which is due to true differences among patients, i.e., the score without measurement error. For PROMs this is usually assessed by test–retest reliability: the extent to which patients who have had no change in the construct have the same score at repeated measurements. This can be reported with the intraclass correlation coefficient (ICC) or weighted Kappa.
- *Measurement error:* All error (systematic and random) in a measurement that is not due to true differences in the construct that is measured. Whether the measurement error is acceptable is determined by comparing the minimally important change with the smallest detectable change or the limits of agreement.

Validity

Validity is the extent to which a measurement instrument measures what it purports to measure. In this domain three measurement properties can be distinguished:

- *Content validity*: the extent to which the content of the instrument adequately reflects the construct to be measured in a certain population. This involves a judgment by the target population itself on the relevance and comprehensiveness of the items of a PROM.
- *Construct validity:* the extent to which an instrument validly measures the construct it purports to measure. This includes:
  - *Structural validity*, which is the extent to which instrument scores are an adequate reflection of the dimensionality of the construct, as assessed by factor analysis.
  - *Hypothesis testing*: the degree to which a measurement instrument produces outcomes consistent with hypotheses. These hypotheses state expected outcomes when assuming that the instrument validly measures its construct. Hypothesis testing can be used to assess convergent validity (the degree to which scores on instruments with related constructs correlate), known-groups validity (the ability of an instrument to distinguish between groups that are expected to differ with respect to the construct to be measured) and discriminant validity (assessing whether instruments with unrelated constructs have low correlations).
  - *Criterion validity*: the extent to which a measurement instrument is an adequate reflection of a gold standard. For PROMs, a gold standard only exists when a shorter version of a PROM is created from a longer version, in which case the gold standard is the longer version of the PROM [23].

Responsiveness

Responsiveness is the ability of an instrument to detect change over time in the construct to be measured. To assess responsiveness, hypotheses should be constructed about the change scores of the instrument under study in correlation to the change scores of other instruments, as in hypothesis testing for construct validity [23].

Interpretability

Interpretability assesses to what extent qualitative meaning can be given to a score or change score of an instrument. Issues that can be considered in the context of interpretability are floor and ceiling effects (<15% of the respondents achieved the highest or lowest possible scores), scores and change scores in different (sub)groups, and the minimal important change (MIC) which expresses when a change score is clinically relevant.

Data extraction

We reviewed the included studies in duplicate (IA and PvdW) and extracted all reported aspects of reliability, validity and responsiveness, as well as interpretability of the PROMs.

Assessing the quality of the studies

We used the consensus-based standards for the selection of health status measurement instruments (COSMIN) checklist [24] to assess the methodological quality of the included studies. This checklist contains multiple questions to critically appraise the methods for each reported measurement property, and uses a four-point scale [16] ("poor", "fair", "good" and "excellent"). The lowest score counts as the overall score for that property. The quality assessment was performed by two independent reviewers (IA and PvdW). Any disagreements were resolved by discussion with a third reviewer (MR).

Assessing the quality of the PROMs

The reported results of the measurement properties of the PROMs were judged by criteria based on Terwee et al., 2007 [25] (Table 1).

For construct validity as well as responsiveness, the quality criteria call for a comparison of the findings with hypotheses constructed by the authors of the papers assessing these

Table 1
Quality criteria for measurement properties [25].

Property	Rating	Quality criteria
Reliability		
Internal consistency	+	(Sub)scale unidimensional AND Cronbach's α(s) ≥ 0.70
	?	Dimensionality not known OR Cronbach's α not determined
	−	(Sub)scale not unidimensional or Cronbach's α(s) < 0.70
Measurement error	+	MIC > SDC OR MIC outside the LOA
	?	MIC not defined
	−	MIC ≤ SDC OR MIC equals or inside LOA
Reliability	+	ICC/weighted Kappa ≥0.70
	?	ICC/weighted Kappa not determined
	−	ICC/weighted Kappa <0.70
Validity		
Content validity	+	The target population considers all items in the questionnaire to be relevant
	?	No target population involvement
	−	The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
Construct validity		
Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	−	Factors explain <50% of the variance
Hypothesis testing	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with hypotheses) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	−	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Criterion validity	+	Convincing arguments that gold standard is "gold" AND correlation with gold standard > 0.70
	?	No convincing arguments that gold standard is "gold"
	−	Correlation with gold standard < 0.70, despite adequate design and method
Responsiveness		
Responsiveness	+	(Correlation with an instrument measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with hypotheses OR AUC ≥ 0.70) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	−	Correlation with an instrument measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlation with related constructs is lower than with unrelated constructs

AUC = area under the curve; ICC = intraclass correlation coefficient; LOA = limits of agreement; MIC = minimal important change; SDC = smallest detectable change.
+: positive rating, ?: indeterminate rating, −: negative rating.

measurement qualities. However, such hypotheses appear to be scarce. We therefore decided to follow the strategy of a recent systematic review [16], in which the authors devised their own hypotheses where needed. We only devised hypotheses for the comparator instruments that we thought were suitable for adding valuable information to the evidence. We considered comparator instruments unsuitable if the expected relation with the construct of interest was unclear, or if the comparator instrument had a (very) different construct than the one under study. A detailed overview of the hypotheses can be found in Tables S2 and S3.

Data synthesis

The level of evidence, based on the number and the quality of the studies, as well as the consistency of the findings, was summarized for each measurement property based on the method of Schellingerhout et al. [26] (Table 2). The outcomes table provides positive, negative or indeterminate evidence scores based on the quality criteria for the measurement properties and the level of evidence. The COSMIN scores concerning the descriptions of (measurement properties of) comparator instruments, which are

Table 2
Levels of evidence for the overall quality of a measurement property [26].

Level	Rating	Criteria
Strong	+++ or −−−	Consistent findings in multiple studies of good methodological quality OR in one study of excellent quality
Moderate	++ or −−	Consistent findings in multiple studies of fair quality OR in one study of good methodological quality
Limited	+ or −	One study of fair methodological quality, or one or more studies with COSMIN score "poor" only due to poor quality of reporting[a]
Conflicting	±	Conflicting findings
Unknown	?	Only studies with a COSMIN score of "poor" due to doubtful design or method[a]

COSMIN = consensus-based standards for the selection of health measurement instruments.
+: positive result, −: negative result.
[a] The COSMIN scores of item 7 and 8 of hypothesis testing and of items 11 and 12 of responsiveness, concerning the descriptions of (measurement properties of) comparator instruments, assess in their "poor" scores only the quality of the reporting of background information. Therefore we approach "poor" scores on these items as "fair" scores for the purpose of determining the level of evidence.

addressed in “hypothesis testing” and “responsiveness”, assess the quality of the reporting of background information rather than the methodological quality of the study. When determining the level of evidence for these measurement properties, we therefore did not take “poor” scores for these descriptive items into account. Instead they were approached as “fair” scores.

There are no quality criteria for interpretability in the COSMIN checklist, which means the level of evidence cannot be determined with the method described above. The data on interpretability are presented in the text.

Statistical pooling was performed for all measurement properties which were assessed by more than one study with at least a COSMIN score of “fair”, or a score of “poor” due to a small sample size. For hypothesis testing and responsiveness the “poor” scores due to background information only were also included for pooling. Additionally, for hypothesis testing and responsiveness, we only pooled correlations between instruments measuring constructs that we considered suitable (Tables S2 and S3). In cases of high heterogeneity (>50%), we used a random effects model; for low heterogeneity (<50%) we used a fixed effects model [27]. A random effects model is not feasible if only two studies can be pooled. In cases of high heterogeneity and only two available studies, pooling was not performed.

Results

Selection of studies and PROMs

We identified 80 eligible validation studies in our primary search, which all assessed one or more measurement properties of a total of 39 PROMs (Fig. 1). Additional searches and the reference check resulted in six new validation studies.

After full-text screening of all the validation studies, 44 studies and 17 PROMs were excluded because they did not meet our inclusion criteria. This left a total of 42 included studies, assessing 22 PROMs (Table 3). PROMs were divided into three categories: OSA-related quality of life, single OSA-related symptoms, and generic health-related quality of life.

We identified eight OSA-related quality of life PROMs, which were assessed in 11 studies [28–37]. For all the PROMs in this category we identified and included the original development study, except for the symptoms of nocturnal obstruction and related events-25 (SNORE25).

We identified eight PROMs on single OSA-related symptoms which were (partly) validated for patients with OSA assessed in 27 studies [30,38–63] and six PROMs on generic health-related quality of life in nine studies [30,47,55,57,64–68]. The former group includes PROMs which aim to measure sleep propensity/fatigue, snoring, anxiety, and depression.

Quality of the included studies

The results of the quality assessment of the studies with the COSMIN checklist are presented in Table 4. The most common scores were “poor” and “fair”. For four of the measurement properties, most studies scored “poor”: internal consistency (10 out of 16 studies), content validity (7 out of 11 studies), criterion validity (3 out of 3 studies) and responsiveness (19 out of 26 studies). For structural validity, convergent validity, known-groups validity and discriminant validity, “fair” was the most common score. Only content validity and structural validity had one or more “excellent” scores.

The studies with poor methodological quality for internal consistency did not provide information on the factor structure of the PROM before calculating Cronbach's α, or calculated Cronbach's α for the whole PROM rather than separately for each subscale. For content validity, the “poor” scores were assigned because of a lack of patient involvement in the design of the PROM or a lacking

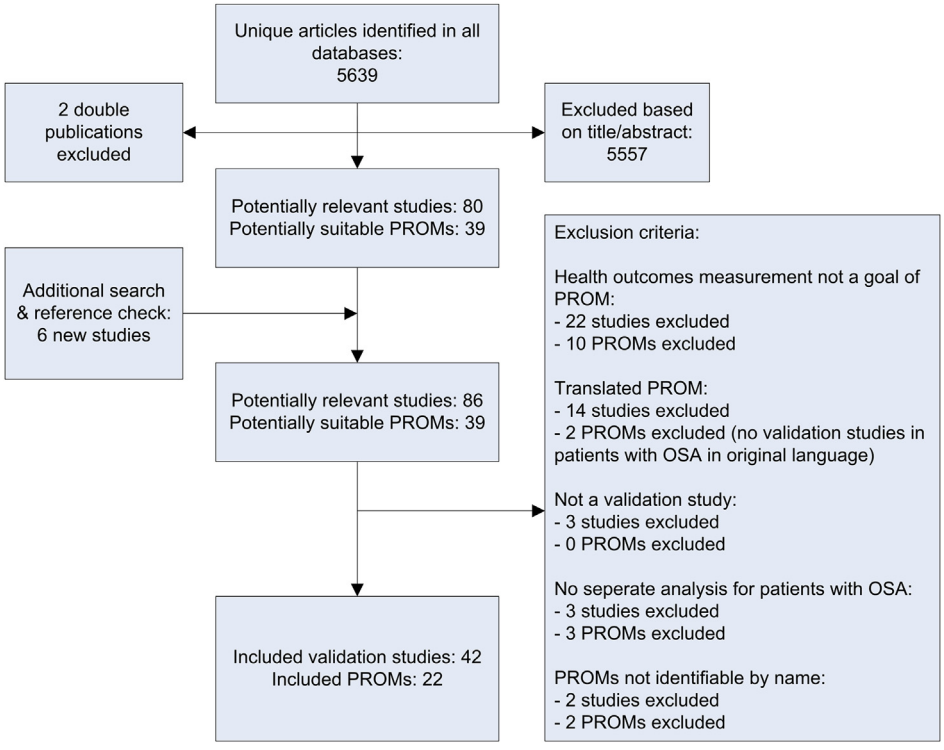

Fig. 1. Flow chart for identification of relevant PROMs and validation studies. OSA – obstructive sleep apnea; PROM – patient-reported outcome measure.

Table 3
Characteristics of the included PROMs.

Name of instrument	Year	Language	Domain(s)	Nr. of questions	Original target population	Target population involved in development?
OSA-related quality of life PROMs						
Functional outcomes of sleep questionnaire (FOSQ) [29]	1997	English	Activity level Vigilance Intimate and sexual relationships General productivity Social outcome	30	Patients with disorders of excessive sleepiness	no
Functional outcomes of sleep questionnaire-10 (FOSQ-10 – shorter version of the FOSQ) [36]	2009	English	Activity level Vigilance Intimate and sexual relationships General productivity Social outcome	10	Patients with disorders of excessive sleepiness	no
Maugeri obstructive sleep apnea syndrome (MOSAS) questionnaire [37]	2011	Italian	Sleep apnea psychological impact Sleep apnea physical Impact Discomfort and nuisance caused by CPAP	16 (OSA symptoms) +7 (CPAP discomfort)	Patients with OSA	yes
Obstructive sleep apnea patient-oriented severity index (OSAPOSI) [31]	1998	English	Sleep problems Awake problems Medical problems Emotional and personal problems Occupational impact	32	Patients with OSA	yes
Quebec sleep questionnaire (QSQ) [32]	2004	French	Sleepiness Diurnal symptoms Nocturnal symptoms Emotions Social interactions	32	Patients with OSA	yes
Sleep apnea quality of life index[a] (SAQLI) [34]	1998	English	Daily functioning Social interactions Emotional functioning Symptoms + Treatment related symptoms	56 + 28 treatment-related symptoms[b]	Patients with sleep-disordered breathing	yes
Symptoms of nocturnal obstruction and related events-25 (SNORE25 - shorter version of the OSAPOSI)	?[c]	English	Unclear	25	Patients with OSA	Yes (patients involved in development of OSAPOSI)
Visual analogical well-being scale (VAWS) [33]	2004	Spanish	Well-being status with regard to the symptoms which were the motive of the consultation	1	Patients with OSA	no
PROMs on single OSA-related symptoms						
Beck anxiety inventory (BAI) [71]	1988	English	Anxiety	21	Psychiatric outpatients	no
Epworth sleepiness scale (ESS) [43]	1991	English	Sleep propensity	8	Patients with sleep disorders	no
Hospital anxiety and depression scale (HADS) [72]	1983	English	Anxiety Depression	14	Non-psychiatric hospital patients	no
Rotterdam sleepiness scale [62]	1995	Dutch	Sleepiness	16	Patients with OSA	no
Sleepiness-wakefulness inability and fatigue test (SWIFT) [61]	2012	English	General wakefulness inability & fatigue (GWIF) Driving wakefulness inability & fatigue (DWIF)	12	Patients with OSA	no
Sleep quality scale (SQS) [63]	2006	Korean	Restoration after sleep Difficulty in falling asleep Difficulty in getting up Satisfaction with sleep Difficulty in maintaining sleep	28	General population	Yes (patients with various sleep disorders involved)
Snore outcomes survey (SOS) [59]	2002	English	(Problems related to) snoring	8	Patients with complaints of snoring and sleep-disordered breathing	no
Time of day sleepiness scale (ToDSS) [58]	2009	English	Sleep propensity at different times of the day (questions of ESS repeated for morning/afternoon/evening)	24	Patients with OSA, suspected OSA, or other primary sleep complaints	no

(*continued on next page*)

Table 3 (*continued*)

Name of instrument	Year	Language	Domain(s)	Nr. of questions	Original target population	Target population involved in development?
Generic quality of life PROMs						
Euroqol (EQ-5D) including Euroqol thermometer (EQ-T) [73]	1990	English	Mobility Self-care Usual activities Pain/discomfort Anxiety/depression + Global indication of health status	5 + 1	General population	no
Functional limitations profile (FLP) [74] (a British version of the (American) sickness impact profile (SIP) [75])	1976 (SIP) 1981 (FLP)	English	Ambulation, body care and movement Mobility Household management Recreation and pastimes Social interaction Emotional behavior Alertness behavior Sleep and rest Eating Communication Work	136	General population	yes (for SIP)
Nottingham health profile (NHP) part II [76]	1985	English	Paid employment Jobs around the house Social life Personal relationships Sex life Hobbies and interests Holidays	7	General population	yes
Patient-generated index (PGI) [77]	1994	English	Five most important areas/ activities in the patient's life affected by their condition (determined by the individual patient)	19	Patients with low back pain	yes
Short form 12 (SF-12) [78]	1996	English	Physical functioning Role limitations because of physical health problems Bodily pain Social functioning General mental health Role limitations because of emotional problems Vitality (energy/fatigue) General health perceptions	12	General population	no
Short form 36 (SF-36) [79]	1992	English	Physical functioning Role limitations because of physical health problems Bodily pain Social functioning General mental health Role limitations because of emotional problems Vitality (energy/fatigue) General health perceptions	36	General population	no

[a] The SAQLI is interviewer-administered.
[b] In the "symptoms" domain of the SAQLI, patients indicate for 21 symptoms whether they apply to them or not, with the option of adding symptoms which are not mentioned. Only the five most important symptoms are used for scoring. The same method is applied for the 28 treatment-related symptoms.
[c] No development article could be identified for the SNORE25.

description of the development of the PROM in its development article. For responsiveness, the most common methodological flaw was that none of the presented data was suitable for determining the validity of the change score, for example, when the results of comparator instruments were not presented in such a way that they could be related to the instrument under study. One of the studies on convergent validity scored "poor" solely because of a missing description of (measurement properties of) the comparator instrument. For criterion validity, all studies scored "poor" because they used the data of their criterion to calculate the scores of the short version of the PROM that was under study, rather than collecting the data for the latter separately. All other studies that scored "poor" for any of the measurement properties either had a study population of less than 30 patients, or suffered from a variety of other methodological flaws.

Measurement properties of the PROMs

The results for the measurement properties of the included PROMs in the light of their level of evidence can be found in Table 5. None of the studies in this review assessed measurement error, and therefore this property was removed from the results table. The results for all studied measurement properties with a score of "fair" or better are described below in more detail. The only data meeting our criteria for pooling were for convergent validity, the results of which are presented in Table S4.

Table 4
Methodological quality of each study per measurement property and questionnaire.[a]

Study	Internal consistency	Reliability (test–retest)	Content validity	Structural validity	Hypothesis testing (construct validity)			Criterion validity	Responsiveness
					Convergent validity	Known-groups validity	Discriminant validity		
OSA-related quality of life PROMs									
FOSQ									
Billings et al., 2014 [28]	Poor				Fair, poor[b,c]				Fair
Weaver et al., 1997 [29]	Fair[d]		Poor	Fair[d]	Fair, poor[e,f]	Poor			
Weaver et al., 2005 [30]					Fair				Poor
FOSQ-10									
Chasens et al., 2009 [36]	Poor		Poor			Fair		Poor	Poor
MOSAS questionnaire									
Moroni et al., 2011 [37]	Fair		Excellent	Fair	Fair	Poor			
OSAPOSI									
Piccirillo et al., 1998 [31]	Poor		Excellent		Poor				Poor
QSQ									
Lacasse et al., 2004 [32]	Poor	Poor	Excellent		Fair				Poor
SAQLI									
Billings et al., 2014 [28]	Poor				Fair, poor[b,c]				Fair
Flemons et al., 1998 [34]	Poor		Excellent		Poor				Poor
Flemons et al., 2002 [35]		Fair[d]			Fair, poor[c,g]				Fair, poor[c,g]
SNORE25									
Weaver et al., 2005 [30]					Fair				Poor
VAWS									
Masa et al., 2011 [33]		Good	Poor		Fair, poor[c]	Poor			Fair
PROMs on single OSA-related symptoms									
BAI									
Sanford et al., 2008 [52]	Fair[d]			Fair[d]	Poor[c]		Fair		
ESS									
Chervin et al., 1999 [38]					Fair, poor[c,h]	Poor			
Cowan et al., 2014 [39]						Fair			
Giudici et al., 2000 [40]					Fair				
Hardinge et al., 1995 [41]									Poor
Hesselbacher et al., 2012 [42]					Fair	Fair			
Johns, 1991 [43]			Poor		Fair	Fair, poor[i]			
Johns, 1992 [44]									Poor
Johns, 1993 [45]					Fair	Fair			
Johns, 1994 [46]	Fair[d]			Fair					
Kingshott et al., 1995[c] [48]					Fair				
Kingshott et al., 1998 [47]					Fair				
Olaite et al., 2013 [49]				Excellent					
Olson et al., 1998 [50]					Fair				
Osman et al., 1999 [51]					Fair				
Sangal et al., 1999 [53]					Fair				
Sil et al., 2012 [54]						Fair			
Smith et al., 2008 [69]	Fair[d]			Fair[d]					
Walter et al., 2002 [56]					Fair				
Weaver et al., 2004 [57]					Fair	Poor			
Weaver et al., 2005 [30]					Fair				Poor
HADS									
Law et al., 2014 [60]	Poor				Fair				
Kingshott et al., 1998 [47]					Fair				
Rotterdam sleepiness scale									
Van Knippenberg et al., 1995 [62]			Poor		Fair				
SQS									
Yi et al., 2009 [63]	Poor				Fair	Fair			
SWIFT									
Sangal, 2012 [61]					Fair	Fair			Poor
SOS									
Gliklich et al., 2002 [59]	Poor	Poor	Poor		Fair				Poor
ToDSS									
Dolan et al., 2009 [58]	Fair[d]		Poor	Fair[d]	Poor	Poor			Poor
Generic health-related quality of life PROMs									
EQ-5D									
Jenkinson et al., 1997 [67]									Poor
Jenkinson et al., 1998 [68]									Fair
FLP									
Jenkinson et al., 1997 [67]									Poor
NHP part II									
Kingshott et al., 1998 [47]					Fair				
PGI									
Jenkinson et al., 1998 [68]									Fair
SF-12									
Jenkinson et al., 1997 [66]								Poor	Poor
Jenkinson et al., 1997 [65]								Poor	

(*continued on next page*)

Table 4 (*continued*)

Study	Internal consistency	Reliability (test–retest)	Content validity	Structural validity	Hypothesis testing (construct validity)			Criterion validity	Responsiveness
					Convergent validity	Known-groups validity	Discriminant validity		
SF-36									
Bennett et al., 1999 [64]					Fair				Poor
Jenkinson et al., 1997 [67]									Poor
Jenkinson et al., 1998 [68]									Fair
Kingshott et al., 1998 [47]					Fair				
Smith et al., 1995 [55]	Poor					Fair,poor[j]			Poor
Weaver et al., 2004 [57]					Fair	Poor			
Weaver et al., 2005 [30]					Fair				Poor

BAI – Beck anxiety inventory; EQ-5D – euroqol-5D; ESS – Epworth sleepiness scale; FLP – functional limitations profile; FOSQ – functional outcomes of sleep questionnaire; HADS – hospital anxiety and depression scale; MOSAS – Maugeri obstructive sleep apnea syndrome; NHP – nottingham health profile; OSA – obstructive sleep apnea; OSAPOSI – obstructive sleep apnea patient-oriented severity index; PGI – patient-generated index; PROM – patient-reported outcome measure; QSQ – Quebec sleep questionnaire; SAQLI – sleep apnea quality of life index; SF-12 – short form 12; SF-36 – short form 36; SNORE25 – symptoms of nocturnal obstruction and related events-25; SOS – snore outcomes survey; SQS – sleep quality scale; SWIFT – sleepiness-wakefulness inability and fatigue test; ToDSS – time of day sleepiness scale; VAWS – visual analogical well-being scale.

[a] The measurement property "measurement error" was removed from this table because it was not assessed for any of the instruments.
[b] "Fair" for comparison with the ESS, "poor" for comparison with the SF-36.
[c] "Poor" score because of missing description of the questionnaire or its measurement properties.
[d] Rated "fair" because the percentage of missing items was not described – all other items were good or excellent.
[e] Rated "fair" due to missing items and/or description of measurement properties of comparator instrument – all other items were good or excellent.
[f] Hypothesis testing was performed in two groups of different sizes, one of which scored "poor".
[g] "Poor" for the comparison instrument "global quality of life rating", "fair" for the other comparison instruments.
[h] "Poor" for comparison with a question about problematic sleepiness, "fair" for comparison with the multiple sleep latency test.
[i] "Poor" for comparing snoring to the different severities of OSA, "fair" for known-groups validity comparing OSA patients and normal subjects.
[j] "Poor" for comparison of general population with patients with mild OSA, "fair" for comparison of general population with "OSA patients requiring treatment".

OSA-related quality of life PROMs

None of the OSA-related quality of life PROMs was fully validated. Content validity, convergent validity, internal consistency and responsiveness were assessed for most of these PROMs, whereas data on most other measurement properties is not available. The evidence that is available is often either indeterminate or of limited strength, due to low study quality. However, most of the PROMs in this category were developed specifically for OSA patients, and four out of eight PROMs have strong positive evidence for their content validity.

There is strong positive evidence of content validity for the Maugeri obstructive sleep apnea syndrome (MOSAS) questionnaire, the obstructive sleep apnea patient-oriented severity index (OSAPOSI), the Quebec sleep questionnaire (QSQ), and the sleep apnea quality of life index (SAQLI).

For the functional outcomes of sleep questionnaire (FOSQ) there is limited positive evidence of structural validity and internal consistency (Cronbach's $\alpha = 0.86–0.91$ for the five factors [29]). For the MOSAS questionnaire, there is limited negative evidence for these properties (factors explained 31% of the variance [37]).

Limited and moderate positive evidence of test–retest reliability is available for the SAQLI and the visual analogical well-being scale (VAWS), respectively; the ICC of the SAQLI being 0.92 [35] and that of the VAWS being 0.83 [33].

For the MOSAS questionnaire, SAQLI and VAWS there is limited positive evidence for convergent validity. For the FOSQ, evidence on convergent validity is conflicting. Weaver et al. [29] showed weaker than expected correlations with the short-form 36 (SF-36), while the correlations found in Billings et al. [28] matched our hypotheses. Due to high statistical heterogeneity, as well as the observation that all correlations were stronger in Billings et al. [28] than in Weaver et al. [29], we did not pool the correlations for five out of seven comparisons (Table S4). However, we could not identify a possible explanation for why there was a consistent discrepancy in these studies.

For the QSQ, less than 75% of the hypotheses for convergent validity were met. Many correlations did not meet the expectations stated in its validation article [32]. Therefore, there is limited negative evidence for convergent validity for this PROM.

There is limited positive evidence for known-groups validity of the FOSQ-10, as patients with OSA had a lower average score than normal subjects, as was expected.

For the FOSQ, SAQLI and VAWS there is limited positive evidence for responsiveness.

With regard to interpretability, for the SAQLI and VAWS no obvious floor or ceiling effects are reported [33–35]. However, for the QSQ, the distribution of scores indicates there might be floor and ceiling effects in several of its domains [32]. For the FOSQ, QSQ, and SAQLI, MICs are reported for the separate domains of the PROMs [28,32,35]. For the FOSQ and FOSQ-10, scores are presented for patients with OSA and normal subjects [29,36], for the VAWS of patients before and after treatment with continuous positive airway pressure (CPAP) [33], and for the MOSAS questionnaire for patients differing in CPAP adherence [37]. For the other PROMs, floor and ceiling effects, MIC, and subgroup scores were not reported.

PROMs on single OSA-related symptoms

None of the PROMs on single OSA-related symptoms was fully validated, but the Beck anxiety inventory (BAI) has the most evidence in its favor. Internal consistency and convergent validity were assessed for most of the PROMs in this category. However, for three PROMs convergent validity does not seem to be adequate, and for another three the evidence is indeterminate. Data on most other measurement properties is not available for these PROMs. The evidence that is available is often either indeterminate or of limited strength, due to low study quality.

For the Beck anxiety inventory (BAI) there is a limited positive level of evidence for structural validity and for internal consistency (one factor, Cronbach's $\alpha = 0.92$ [52]). For the ToDSS there is a limited negative level of evidence for structural validity and internal consistency, as the variance explained by the factors was below the required 50% for two of the three subscales [58]. There are conflicting findings about the factor structure of the Epworth sleepiness scale (ESS). Johns [46] found a one-factor structure, Smith et al. [69] reported that two items on low somnificity should be omitted for a sufficient one-factor fit, and Olaithe et al. [49] showed a sufficient one-factor fit as well as sufficient three-factor

Table 5
Quality of measurement properties per PROM.[a,b]

Instrument/patient group	Internal consistency	Reliability (test–retest)	Content validity	Structural validity	Hypothesis testing (construct validity)			Criterion validity	Responsiveness
					Convergent validity	Known-groups validity	Discriminant validity		
OSA-related quality of life PROMs									
FOSQ	+	na	?	+	±	?	na	[c]	+
FOSQ-10	?	na	na	na	na	+	na	?	?
MOSAS questionnaire	–	na	+++	–	+	?	na	[c]	na
OSAPOSI	?	na	+++	na	?	na	na	[c]	?
QSQ	?	?	+++	na	–	na	na	[c]	?
SAQLI	?	+	+++	na	+	na	na	[c]	+
SNORE25	na	na	na	na	?	na	na	[c]	?
VAWS	[d]	++	?	[d]	+	?	na	[c]	+
PROMs on single OSA-related symptoms									
BAI	+	na	na	+	+	na	+	[c]	na
ESS	±[e]	na	?	±	+	±	na	[c]	na
HADS	?	na	na	na	–	na	na	[c]	na
Rotterdam sleepiness scale	na	na	?	na	–	na	na	[c]	na
SQS	?	na	na	na	?	+	na	[c]	na
SWIFT	na	na	na	na	?	+	na	[c]	?
SOS	?	?	?	na	–	na	na	[c]	?
ToDSS	–	na	?	–	?	?	na	[c]	?
Generic health-related quality of life PROMs									
EQ-5D	na	na	na	na	na	na	na	[b]	–
FLP	na	na	na	na	na	na	na	[b]	?
NHP part II	na	na	na	na	?	na	na	[b]	na
PGI	na	na	na	na	na	na	na	[b]	–
SF-12	na	na	na	na	na	na	na	?	?
SF-36	?	na	na	na	+[f]	+	na	[b]	–

BAI – Beck anxiety inventory; EQ-5D – euroqol-5D; ESS – Epworth sleepiness scale; FLP – functional limitations profile; FOSQ – functional outcomes of sleep questionnaire; HADS – hospital anxiety and depression scale; MOSAS – Maugeri obstructive sleep apnea syndrome; NHP – Nottingham health profile; OSA – obstructive sleep apnea; OSAPOSI – obstructive sleep apnea patient-oriented severity index; PGI – patient-generated index; PROM – patient-reported outcome measure; QSQ – Quebec sleep questionnaire; SAQLI – sleep apnea quality of life index; SF-12 – short form 12; SF-36 – short form 36; SNORE25 – symptoms of nocturnal obstruction and related events-25; SOS – snore outcomes survey; SQS – sleep quality scale; SWIFT – sleepiness-wakefulness inability and fatigue test; ToDSS – time of day sleepiness scale; VAWS – visual analogical well-being scale.

[a] The scores in this table were constructed as described in Table 2. "na" – not available; no studies were performed on this measurement property for this PROM.

[b] The measurement property "measurement error" was removed from this table because it was not assessed for any of the instruments.

[c] Criterion validity is not relevant for this questionnaire.

[d] The VAWS is a one-item PROM, meaning that internal consistency and structural validity are not relevant for this PROM.

[e] Due to the conflicting results of the factor structure of the ESS in (suspected) OSA patients, evidence on internal consistency results cannot be clearly interpreted.

[f] The positive score is for the mental health component of the SF-36. The physical component was only compared with unsuitable comparator instruments so its validity in patients with OSA could not be determined.

fit. Due to the conflicting evidence on the factor structure, the evidence for the internal consistency of the ESS is also conflicting.

Moderate positive evidence for convergent validity is reported for the BAI and the ESS (see Table S4 for pooled correlations of the ESS). There is limited negative evidence for this property for the hospital anxiety and depression scale (HADS), the snore outcomes survey (SOS), and the Rotterdam sleepiness scale. The correlation of the overall HADS with an instrument that measures depression was stronger than the correlation with the HADS depression subscale only [60], which was not as expected. There is negative evidence for convergent validity for the Rotterdam sleepiness scale and the SOS because less than 75% of hypotheses were met.

There is a limited positive evidence base for known-groups validity of the sleepiness-wakefulness inability and fatigue test (SWIFT) and sleep quality scale (SQS), as patients with OSA had a higher average score on these PROMs than normal subjects, as expected. Known-groups validity for the ESS showed conflicting evidence. Of the six studies that compared ESS scores of different groups, four studies found expected differences [42,43,45,54], and two studies did not [39,57].

For the BAI, discriminant validity was assessed by determining whether the BAI could be distinguished from the depression score of the Beck depression inventory (BDI) by performing a factor analysis on all items of both questionnaires simultaneously. The items of the BAI and BDI were shown to load on different factors [52], providing limited positive evidence that they measure different constructs.

With regard to interpretability, no MIC for patients with OSA is reported for any of the PROMs in this category. The ESS does not show floor or ceiling effects, as can be concluded from the ranges of scores and their graphical presentation in many of the included studies [38,41,45,46,48,50,51,53,57]. For no other instruments there is information on floor or ceiling effects for patients with OSA. Scores of subgroups were presented for the BAI (male and female patients with OSA) [52], the time of day sleepiness scale (ToDSS) (patients with OSA before and after treatment with CPAP) [58], and the SQS and the SWIFT (normal subjects and OSA patients) [61,63]. Scores of subgroups are also available for the ESS (normal subjects and/or patients with different OSA severity [39,42–46,49,50,54,57], patients with OSA before and after treatment with CPAP [44], and for ethnicities and different genders [42]).

Generic health-related quality of life PROMs

Most measurement properties were not assessed in patients with OSA for general health-related quality of life PROMs, which means there is very little information available on their quality in this patient group. Only responsiveness was assessed for five out of six PROMs, but the evidence was either indeterminate or negative.

There is limited positive evidence for convergent validity and known-groups validity for the mental health component of the SF-36 (see Table S4 for pooled correlations of the SF-36). For the SF-36 and the patient-generated index (PGI) there is limited negative evidence for responsiveness because correlations with unrelated constructs were stronger than with related constructs. For the

euroqol-5D (EQ-5D) there is limited negative evidence for responsiveness because less than 75% of hypotheses were met.

With regard to interpretability, no information on the MIC or floor and ceiling effects is available for the PROMs in this category. Subgroups of patients with OSA (before and after treatment with CPAP) were presented for the EQ-5D [67,68], the functional limitations profile (FLP) [67], PGI [68], and the SF-36 [67,68]. For the short-form 12 (SF-12), scores were presented of the general population and OSA patients [65].

Discussion

In this review we determined the evidence base for PROMs for health outcomes measurement in patients with OSA. We identified 22 PROMs validated in patients with OSA, categorized into three domains: OSA-related quality of life, single OSA-related symptoms, and generic health-related quality of life. None of the identified PROMs has been fully validated, and many validation studies were of insufficient quality. Especially the lack of established content validity for most of the PROMs is problematic for a patient-centered approach to measuring health status, because the items of these PROMs might not address the issues that patients with OSA consider relevant or most important. Furthermore, it is important to note that measurement error, which is particularly relevant for the use of PROMs in clinical practice, i.e., for individual patients, was not assessed for any of the questionnaires. Therefore the results of all PROMs should be used with caution when interpreting scores for individual patients. Rather than relying on composite scores of the domains, the individual questions of the PROMs might be more suitable for alerting a healthcare professional to the most important problems of these patients.

The only PROMs with good content validity are four OSA-related quality of life PROMs: the OSAPOSI, MOSAS questionnaire, QSQ and SAQLI. Therefore, we consider these PROMs the most suitable for a patient-centered approach of health status and we consider all four potentially suitable for outcome measurement. Currently, the SAQLI has the most evidence for good quality, but its downside is that it contains many questions (n = 56, plus 28 treatment-related symptoms) and it is interview-administered, which makes it a less feasible option for use in clinical practice. The QSQ (n = 32) or MOSAS questionnaire (n = 16, plus 7 CPAP-related questions) might be more suitable for this purpose, as they can be filled out by the patient and are shorter. It should be noted that the MOSAS questionnaire does not contain any questions on nocturnal symptoms, a topic which is covered by the other three PROMs. Its CPAP-related questions may be relevant on an individual patient level, for those patients who get this treatment. The development article of the OSAPOSI (n = 32) reveals that this PROM contains some topics that were not covered in other PROMs (such as occupational impact, e.g., job loss), but the OSAPOSI is not publicly available or retrievable via the developer. Therefore our recommendation is to use the SAQLI for research purposes, when feasible, and either the QSQ or MOSAS questionnaire for use in clinical practice.

The PROMs on single OSA-related symptoms all focus on symptoms which are also addressed in the OSA-related quality of life PROMs. None of the PROMs on OSA-related symptoms has been well-validated or assessed for content validity. For the ESS this oversight is specifically surprising as it had the greatest number of validation articles devoted to it in OSA patients (n = 20 studies), and is frequently used in both research and practice to measure sleep propensity. Similar to a recent systematic review on the ESS [16], we conclude that the evidence regarding the quality of this PROM is modest at best. The other PROMs in this category measuring sleep propensity/sleepiness do not have more evidence for their quality, but one could consider using the ToDSS or the SWIFT. The ToDSS contains the same questions as the ESS but for three different times of day. This may be beneficial for clinical practice to identify the time of day that a patient feels most sleepy, though in terms of outcome measurement there does not seem to be a clear benefit compared to the ESS. The SWIFT measures sleepiness in combination with fatigue and is a possible alternative to the ESS for measuring the main complaints related to OSA. We would not recommend the Rotterdam sleepiness scale: it is similar to the ESS but contains mostly yes/no questions and therefore its scores are likely to be less sensitive. The main benefit of the ESS compared to the other sleepiness PROMs is that it is used all around the world in both clinical practice and research, and will be familiar to those involved with OSA.

The SQS (on subjective sleep quality) and SOS (on experienced problems due to snoring) measure complaints that can be relevant to OSA, but are not likely to be the main complaints. Since there is no evidence that they are of better quality than other PROMs, we would not recommend them for patients with OSA.

The BAI (measuring anxiety) has limited positive evidence for several measurement properties, and based on current evidence we would recommend it over the use of the HADS. The HADS was the only PROM in this review that measures depression. Since evidence for this PROM in OSA patients is either not available or negative, a possibility is to look outside the scope of this review for other PROMs measuring depression.

It should also be noted that if the use of a complete disease-specific quality of life PROM is not preferred (for example because of a preference for a short PROM, or because only a specific symptom needs to be measured), another option is to use one or more domains of such a PROM, for example the “daytime sleepiness” domain of the QSQ. The benefit is that content validity is good for this PROM and that some of the other measurement properties were assessed separately for each domain, even though we did not report our results at domain level in this review.

The main reason to use a generic health-related QoL PROM is to be able to compare PROM scores across diseases. These PROMs will by definition contain questions less relevant for the specific disease studied. Therefore we would not recommend the use of generic health-related quality of life PROMs for use in clinical practice, especially not when acceptable disease-specific PROMs are available, as these will provide more relevant information for the disease. Of the PROMs in this review the only exception is the PGI, which asks patients to write down and score the areas of their life most affected by the disease, allowing for a more disease-specific approach.

Very little evidence was found regarding the quality of generic health-related QoL PROMs for patients with OSA. The mental health component of the SF-36 is the only PROM with a positive score for any of the measurement properties, and as such could be considered the best option. However, whether a generic PROM is suitable for outcome measurement for any specific disease greatly depends on content validity – which in this case could be described as the degree to which the questions are relevant for this disease. The negative evidence that we found for responsiveness for the SF-36, EQ-5D and PGI is likely related to a lack of content validity of these PROMs for OSA, although this has not been assessed in the included studies. We did identify potential issues related to a lack of content validity when devising our hypotheses, for example for the SF-36. In this PROM, the questions about daily activities and social functioning are assessed by asking about limitations due to “physical health” or “emotional problems”. In our view, neither of these categories clearly covers the main reasons for reduced functioning that patients with OSA experience (i.e., sleepiness and fatigue). It needs to be investigated whether problems with daily activities or social functioning will be detected with this PROM in

patients with OSA. The only SF-36 domain that does address fatigue is the "vitality" domain, which is therefore most likely to be useful in measuring outcomes for patients with OSA.

We have also noticed problems with content validity for the SF-12 and the EQ-5D. The FLP contains items relevant for OSA patients, but it is very long (n = 136) and also contains a great many items which are irrelevant. The Nottingham health profile (NHP) part II allows only yes/no answers to questions about how health affects daily functioning, which is not likely to provide sensitive scores. Furthermore, the FLP and NHP are not used often and would be of limited use when the aim is to compare scores across diseases. Finally, it may be hard to make a meaningful comparison of PGI scores across diseases due to the wide range of items that can be created by the patient.

Summarizing, the mental health component of the SF-36, and in particular the "vitality" domain, is probably the best generic health-related QoL PROM for OSA patients, though we remain doubtful about its content validity and recommend the use of a disease-specific PROM alongside it.

We did not find many PROMs of which measurement properties could be statistically pooled. Studies on the same PROMs and properties were either of poor quality, or a given measurement property was only assessed in a single study. For the measurement properties of which we theoretically could pool data, heterogeneity appeared too high in about half of them to allow pooling. We did not find a plausible explanation for this high heterogeneity.

The main strength of this review is that we used the COSMIN checklist for a thorough evaluation of the quality of the included studies, and complemented this with our own critical assessment of which items on the COSMIN checklist assessed methodological quality, and which assessed quality of reporting. This allowed us to discriminate between studies of sufficient and insufficient methodological quality, when deciding which studies should contribute to the evidence base of the PROMs. Furthermore, we devised hypotheses for convergent validity and responsiveness where the authors of validation articles did not, which created the opportunity to use the available data to assess these measurement properties.

Limitations

Our study has a few limitations. First, we deviated slightly from our original protocol [70] in which we described two complementary search strategies, while we report only one. By broadening our original inclusion criteria for PROMs, no new PROMs were found with the second search strategy. We believe that this solution provides an article that is more easily readable, while being equally inclusive with regard to the PROMs suitable for outcome measurement in patients with OSA.

Second, the COSMIN checklist had very high demands in assessing the validation articles, resulting in low scores for many measurement properties. For example, the items about percentage and handling of "missing items" of the PROMs do not seem to follow current or historical standard practice. However, because the scores on these items had no impact on the evidence base, we did not change the way we handled these scores.

Third, the more subjective items on the COSMIN checklist may cause discrepancies between reviews. A recent systematic review [16] that assessed the measurement properties of the ESS in all populations, assigned higher COSMIN scores than we did to more than half of the measurement properties in the studies overlapping with our review. However, the differences on these items did only on a few occasions cause a different approach with regard to contribution to the evidence base for the ESS.

Fourth, we chose not to create hypotheses when the comparator instruments (or their domains) had a construct that was too different from the construct under study. Since some studies reported over 30 correlations between unrelated constructs, this would have resulted in many hypotheses predicting weak correlations. We consider hypotheses for related constructs more valuable than hypotheses for unrelated constructs, and decided to base our scores on only the former.

Finally, when discrepancies are found between hypothesized correlations and identified correlations for convergent validity and responsiveness, there is a possibility that the fault is not in the validity of the PROM, but in flawed hypotheses. This cannot be avoided, but by providing all of the hypotheses that we used to judge these measurement properties in the appendices, we do provide transparency into our results.

Conclusions

Our review found a lack of evidence for the quality of most measurement properties of the 22 included PROMs validated in patients with OSA. We identified four OSA-related quality of life PROMs with thorough patient involvement in their development: the OSAPOSI, MOSAS questionnaire, QSQ, and SAQLI. These are the current best candidates for assessing health status in patients with OSA. Our recommendation is to use the SAQLI for research purposes and either the QSQ or MOSAS questionnaire for use in clinical practice. Even though there is not enough evidence to fully judge the quality of these PROMs, they can potentially add value to outcome measurement or clinical practice, when they are interpreted with caution. Future research should focus on the further validation of these PROMs, to estimate their suitability as outcome measure. Of the PROMs measuring only sleepiness and fatigue, the ESS is the most widely used PROM. However, the quality of this PROM is moderate at best. The SWIFT could potentially serve as an alternative or addition, if future research shows that this PROM is of higher quality.

Practice points

This systematic review on the measurement properties of patient-reported outcome measures (PROMs) validated in patients with obstructive sleep apnea (OSA) shows that:

1) None of the PROMs are fully validated for patients with OSA, and there are few high-quality validation studies.
2) For many identified PROMs there was no involvement of patients with OSA during their development or before the PROM was tested in patients with OSA.
3) The PROMs which did have thorough patient involvement in their development are the obstructive sleep apnea patient-oriented severity index (OSAPOSI), Maugeri obstructive sleep apnea syndrome (MOSAS) questionnaire, Quebec sleep questionnaire (QSQ) and sleep apnea quality of life index (SAQLI), and these are the ones that we would recommend to use for patients with OSA.
4) The Epworth sleepiness scale (ESS) is of moderate quality at best; a possible alternative or addition might be the sleepiness-wakefulness inability and fatigue test (SWIFT) even though further validation studies are needed to confirm this.

Research agenda

The four PROMs with thorough patient involvement in their development (obstructive sleep apnea patient-oriented severity index (OSAPOSI), Maugeri obstructive sleep apnea syndrome (MOSAS) questionnaire, Quebec sleep questionnaire (QSQ) and sleep apnea quality of life index (SAQLI)) should be the focus of future high-quality validation studies. Additionally, the Epworth sleepiness scale (ESS) and the sleepiness-wakefullness inability and fatigue test (SWIFT), which can provide insight into the most common complaints of OSA patients, should be validated further.

- Structural validity and internal consistency need to be assessed for the OSAPOSI, QSQ and SAQLI; for the MOSAS questionnaire, previous results for these measurement properties should be replicated to see if adjustment of the proposed factor structure is necessary.
- For all four PROMs, we also recommend that their test–retest reliability, measurement error, construct validity (hypothesis testing) and responsiveness are (further) assessed.
- Additional high quality validation studies are needed to test the ESS in OSA patients, to clarify the conflicting evidence for this widely used PROM and assess its content validity.
- The potential of the SWIFT to serve as a good PROM to measure sleepiness and fatigue should be further explored by assessing its content validity and other measurement properties.

Conflict of interest

The authors do not have any conflicts of interest to disclose.

Acknowledgments

The first author is funded by a Radboudumc grant (project nr. R0002257).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.smrv.2015.07.006.

References

[1] Punjabi NM. The epidemiology of adult obstructive sleep apnea. Proc Am Thorac Soc 2008;5:136–43.
[2] Vanderveken OM, Boudewyns A, Ni Q, Kashyap B, Verbraecken J, De Backer W, et al. Cardiovascular implications in the treatment of obstructive sleep apnea. J Cardiovasc Transl Res 2011;4:53–60.
[3] Bradley TD, Floras JS. Obstructive sleep apnoea and its cardiovascular consequences. Lancet 2009;373:82–93.
[4] Leger D, Bayon V, Laaban JP, Philip P. Impact of sleep apnea on economics. Sleep Med Rev 2012;16:455–62.
[5] Jennum P, Kjellberg J. Health, social and economical consequences of sleep-disordered breathing: a controlled national study. Thorax 2011;66:560–6.
[6] Pang KP, Rotenberg BW. Redefining successful therapy in obstructive sleep apnea: a call to arms. Laryngoscope 2014;124:1051–2.
[7] Ravesloot MJ, de Vries N. Reliable calculation of the efficacy of non-surgical and surgical treatment of obstructive sleep apnea revisited. Sleep 2011;34:105–10.
*[8] Tam S, Woodson BT, Rotenberg B. Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index. Laryngoscope 2014;124:337–43.
[9] Macey PM, Woo MA, Kumar R, Cross RL, Harper RM. Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients. PloS One 2010;5:e10211.
[10] Dutt N, Janmeja AK, Mohapatra PR, Singh AK. Quality of life impairment in patients of obstructive sleep apnea and its relation with the severity of disease. Lung India Off Organ Indian Chest Soc 2013;30:289–94.
[11] Kezirian EJ, Weaver EM, Criswell MA, de Vries N, Woodson BT, Piccirillo JF. Reporting results of obstructive sleep apnea syndrome surgery trials. Otolaryngol Head Neck Surg Off J Am Acad Otolaryngol Head Neck Surg 2011;144:496–9.
*[12] Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.
*[13] Van der Wees P, Nijhuis-van der Sanden M, Ayanian Y, Black N, Westert G, Schneider E. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. Milbank Q 2014;92:754–75.
[14] Abrishami A, Khajehdehi A, Chung F. A systematic review of screening questionnaires for obstructive sleep apnea. Can J Anaesth 2010;57:423–38.
[15] Fedson AC, Pack AI, Gislason T. Frequently used sleep questionnaires in epidemiological and genetic research for obstructive sleep apnea: a review. Sleep Med Rev 2012;16:529–37.
*[16] Kendzerska TB, Smith PM, Brignardello-Petersen R, Leung RS, Tomlinson GA. Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review. Sleep Med Rev 2014;18:321–31.
[17] Ramachandran SK, Josephs LA. A meta-analysis of clinical screening tests for obstructive sleep apnea. Anesthesiology 2009;110:928–39.
[18] Stucki A, Cieza A, Schuurmans MM, Ustun B, Stucki G, Gradinger F, et al. Content comparison of health-related quality of life instruments for obstructive sleep apnea. Sleep Med 2008;9:199–206.
[19] Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual life Res Int J Qual life Aspects Treat Care Rehabil 2009;18:1115–23.
[20] Lacasse Y, Godbout C, Series F. Health-related quality of life in obstructive sleep apnoea. Eur Respir J 2002;19:499–503.
[21] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737–45.
[22] Cortina JM. What is coefficient alpha – an examination of theory and applications. J Appl Psychol 1993;78:98–104.
*[23] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
[24] Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res Int J Qual Life Aspects Treat Care Rehabil 2012;21:651–7.
*[25] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
[26] Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res Int J Qual Life Aspects Treat Care Rehabil 2012;21:659–70.
[27] Ryan R. Heterogeneity and subgroup analyses in cochrane consumers and communication review group reviews: planning the analysis at protocol stage. Cochrane Consumers and Communication Review Group; 2013.
[28] Billings ME, Rosen CL, Auckley D, Benca R, Foldvary-Schaefer N, Iber C, et al. Psychometric performance and responsiveness of the functional outcomes of sleep questionnaire and sleep apnea quality of life instrument in a randomized trial: the HomePAP study. Sleep 2014;37(12):2017–24.
[29] Weaver TE, Laizner AM, Evans LK, Maislin G, Chugh DK, Lyon K, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. Sleep 1997;20:835–43.
[30] Weaver EM, Woodson BT, Steward DL. Polysomnography indexes are discordant with quality of life, symptoms, and reaction times in sleep apnea patients. Otolaryngol Head Neck Surg Off J Am Acad Otolaryngol Head Neck Surg 2005;132:255–62.
*[31] Piccirillo JF, Gates GA, White DL, Schectman KB. Obstructive sleep apnea treatment outcomes pilot study. Otolaryngol Head Neck Surg Off J Am Acad Otolaryngol Head Neck Surg 1998;118:833–44.
*[32] Lacasse Y, Bureau MP, Series F. A new standardised and self-administered quality of life questionnaire specific to obstructive sleep apnoea. Thorax 2004;59:494–9.
[33] Masa JF, Jimenez A, Duran J, Carmona C, Monasterio C, Mayos M, et al. Visual analogical well-being scale for sleep apnea patients: validity and responsiveness. Sleep Breath 2011;15:549–59.

* The most important references are denoted by an asterisk.

*[34] Flemons WW, Reimer MA. Development of a disease-specific health-related quality of life questionnaire for sleep apnea. Am J Respir Crit care Med 1998;158:494–503.
[35] Flemons WW, Reimer MA. Measurement properties of the calgary sleep apnea quality of life index. Am J Respir Crit Care Med 2002;165:159–64.
[36] Chasens ER, Ratcliffe SJ, Weaver TE. Development of the FOSQ-10: a short version of the functional outcomes of sleep questionnaire. Sleep 2009;32:915–9.
*[37] Moroni L, Neri M, Lucioni AM, Filipponi L, Bertolotti G. A new means of assessing the quality of life of patients with obstructive sleep apnea: the MOSAS questionnaire. Sleep Med 2011;12:959–65.
[38] Chervin RD, Aldrich MS. The Epworth sleepiness scale may not reflect objective measures of sleepiness or sleep apnea. Neurology 1999;52:125–31.
[39] Cowan DC, Allardice G, Macfarlane D, Ramsay D, Ambler H, Banham S, et al. Predicting sleep disordered breathing in outpatients with suspected OSA. BMJ open 2014;4:e004519.
[40] Giudici S, Andrada T, Farmer W, Torrington K, Dollinger A, Rajagopal K. Lack of predictive value of the Epworth sleepiness scale in patients after uvulopalatopharyngopolasty. Ann Otol Rhinol Laryngol 2000;109:646–9.
[41] Hardinge FM, Pitson DJ, Stradling JR. Use of the Epworth sleepiness scale to demonstrate response to treatment with nasal continuous positive airways pressure in patients with obstructive sleep apnoea. Respir Med 1995;89:617–20.
[42] Hesselbacher S, Subramanian S, Allen J, Surani S, Surani S. Body mass index, gender, and ethnic variations alter the clinical implications of the Epworth sleepiness scale in patients with suspected obstructive sleep apnea. Open Respir Med J 2012;6:20–7.
[43] Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 1991;14:540–5.
[44] Johns MW. Reliability and factor analysis of the Epworth sleepiness scale. Sleep 1992;15:376–81.
[45] Johns MW. Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth sleepiness scale. Chest 1993;103:30–6.
[46] Johns MW. Sleepiness in different situations measured by the Epworth sleepiness scale. Sleep 1994;17:703–10.
[47] Kingshott RN, Engleman HM, Deary IJ, Douglas NJ. Does arousal frequency predict daytime function? Eur Respir J 1998;12:1264–70.
[48] Kingshott RN, Sime PJ, Engleman HM, Douglas NJ. Self assessment of daytime sleepiness: patient versus partner. Thorax 1995;50:994–5.
[49] Olaithe M, Skinner TC, Clarke J, Eastwood P, Bucks RS. Can we get more from the Epworth sleepiness scale (ESS) than just a single score? A confirmatory factor analysis of the ESS. Sleep Breath Schlaf Atmung 2013;17:763–9.
[50] Olson LG, Cole MF, Ambrogetti A. Correlations among Epworth sleepiness scale scores, multiple sleep latency tests and psychological symptoms. J Sleep Res 1998;7:248–53.
[51] Osman EZ, Osborne J, Hill PD, Lee BW. The Epworth sleepiness scale: can it be used for sleep apnoea screening among snorers? Clin Otolaryngol Allied Sci 1999;24:239–41.
[52] Sanford SD, Bush AJ, Stone KC, Lichstein KL, Aguillard N. Psychometric evaluation of the Beck anxiety inventory: a sample with sleep-disordered breathing. Behav Sleep Med 2008;6:193–205.
[53] Sangal RB, Sangal JM, Belisle C. Subjective and objective indices of sleepiness (ESS and MWT) are not equally useful in patients with sleep apnea. Clin EEG Electroencephalogr 1999;30:73–5.
[54] Sil A, Barr G. Assessment of predictive ability of Epworth scoring in screening of patients with sleep apnoea. J Laryngol Otol 2012;126:372–9.
[55] Smith IE, Shneerson JM. Is the SF 36 sensitive to sleep disruption? A study in subjects with sleep apnoea. J Sleep Res 1995;4:183–8.
[56] Walter TJ, Foldvary N, Mascha E, Dinner D, Golish J. Comparison of Epworth sleepiness scale scores by patients with obstructive sleep apnea and their bed partners. Sleep Med 2002;3:29–32.
[57] Weaver EM, Kapur V, Yueh B. Polysomnography vs self-reported measures in patients with sleep apnea. Arch Otolaryngol Head Neck Surg 2004;130:453–8.
[58] Dolan DC, Taylor DJ, Okonkwo R, Becker PM, Jamieson AO, Schmidt-Nowara W, et al. The time of day sleepiness scale to assess differential levels of sleepiness across the day. J Psychosomatic Res 2009;67:127–33.
[59] Gliklich RE, Wang PC. Validation of the snore outcomes survey for patients with sleep-disordered breathing. Arch Otolaryngol Head Neck Surg 2002;128:819–24.
[60] Law M, Naughton MT, Dhar A, Barton D, Dabscheck E. Validation of two depression screening instruments in a sleep disorders clinic. J Clin Sleep Med 2014;10:683–8.
[61] Sangal RB. Evaluating sleepiness-related daytime function by querying wakefulness inability and fatigue: sleepiness-wakefulness inability and fatigue test (SWIFT). J Clin Sleep Med JCSM Off Publ Am Acad Sleep Med 2012;8:701–11.
[62] van Knippenberg FC, Passchier J, Heysteck D, Shackleton D, Schmitz P, Poublon RM, et al. The Rotterdam daytime sleepiness scale: a new daytime sleepiness scale. Psychol Rep 1995;76:83–7.
[63] Yi H, Shin K, Kim J, Kim J, Lee J, Shin C. Validity and reliability of sleep quality scale in subjects with obstructive sleep apnea syndrome. J Psychosomatic Res 2009;66:85–8.
[64] Bennett LS, Barbour C, Langford B, Stradling JR, Davies RJ. Health status in obstructive sleep apnea: relationship with sleep fragmentation and daytime sleepiness, and effects of continuous positive airway pressure treatment. Am J Respir Crit Care Med 1999;159:1884–90.
[65] Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health survey). J Health Serv Res Policy 1997;2:14–8.
[66] Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med 1997;19:179–86.
[67] Jenkinson C, Stradling J, Petersen S. Comparison of three measures of quality of life outcome in the evaluation of continuous positive airways pressure therapy for sleep apnoea. J Sleep Res 1997;6:199–204.
[68] Jenkinson C, Stradling J, Petersen S. How should we evaluate health status? A comparison of three methods in patients presenting with obstructive sleep apnoea. Qual Life Res Int J Qual Life Aspects Treat Care Rehabil 1998;7:95–100.
[69] Smith SS, Oei TP, Douglas JA, Brown I, Jorgensen G, Andrews J. Confirmatory factor analysis of the Epworth sleepiness scale (ESS) in patients with obstructive sleep apnoea. Sleep Med 2008;9:739–44.
[70] Abma IL, Van der Wees PJ, Veer V, Westert GP, Rovers M. Measurement properties of patient-reported outcome measures in adults with obstructive sleep apnea: a systematic review. PROSPERO: International Prospective Register of Systematic Reviews; 2014. CRD42014014608.
[71] Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988;56:893–7.
[72] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67:361–70.
[73] EuroQol G. EuroQol – a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
[74] Patrick D. Standardisation of comparative health status measures: using scales developed in America in an English speaking country. In: Health survey research methods: third biennal conference. Hyattsville, MD: US Department of Health and Human Services; 1981.
[75] Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv Plan Adm Eval 1976;6:393–415.
[76] Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll General Pract 1985;35:185–8.
[77] Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life. The patient-generated index. Med Care 1994;32:1109–26.
[78] Ware Jr J, Kosinski M, Keller SD. A 12-Item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–33.
[79] Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.