PATIENT-REPORTED OUTCOME MEASURES (PROMS) IN CLINICAL PRACTICE FOR PATIENTS WITH OBSTRUCTIVE SLEEP APNEA

Inger L. Abma

PATIENT-REPORTED OUTCOME MEASURES (PROMS) IN CLINICAL PRACTICE FOR PATIENTS WITH OBSTRUCTIVE SLEEP APNEA

Inger L. Abma

The work presented in this thesis was carried out within the Radboud Institute for Health Sciences

ISBN
978-94-028-1520-7

Design/lay-out
ProefschriftOntwerp.nl, Nijmegen

Print
Ipskamp printing

PATIENT-REPORTED OUTCOME MEASURES (PROMS) IN CLINICAL PRACTICE FOR PATIENTS WITH OBSTRUCTIVE SLEEP APNEA

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 25 juni 2019
om 14:30 uur precies

door
Inger Lucia Abma
geboren op 31 mei 1987
te Nijmegen

Promotoren:

Prof. dr. P.J. van der Wees

Prof. dr. M.M. Rovers

Prof. dr. G.P. Westert

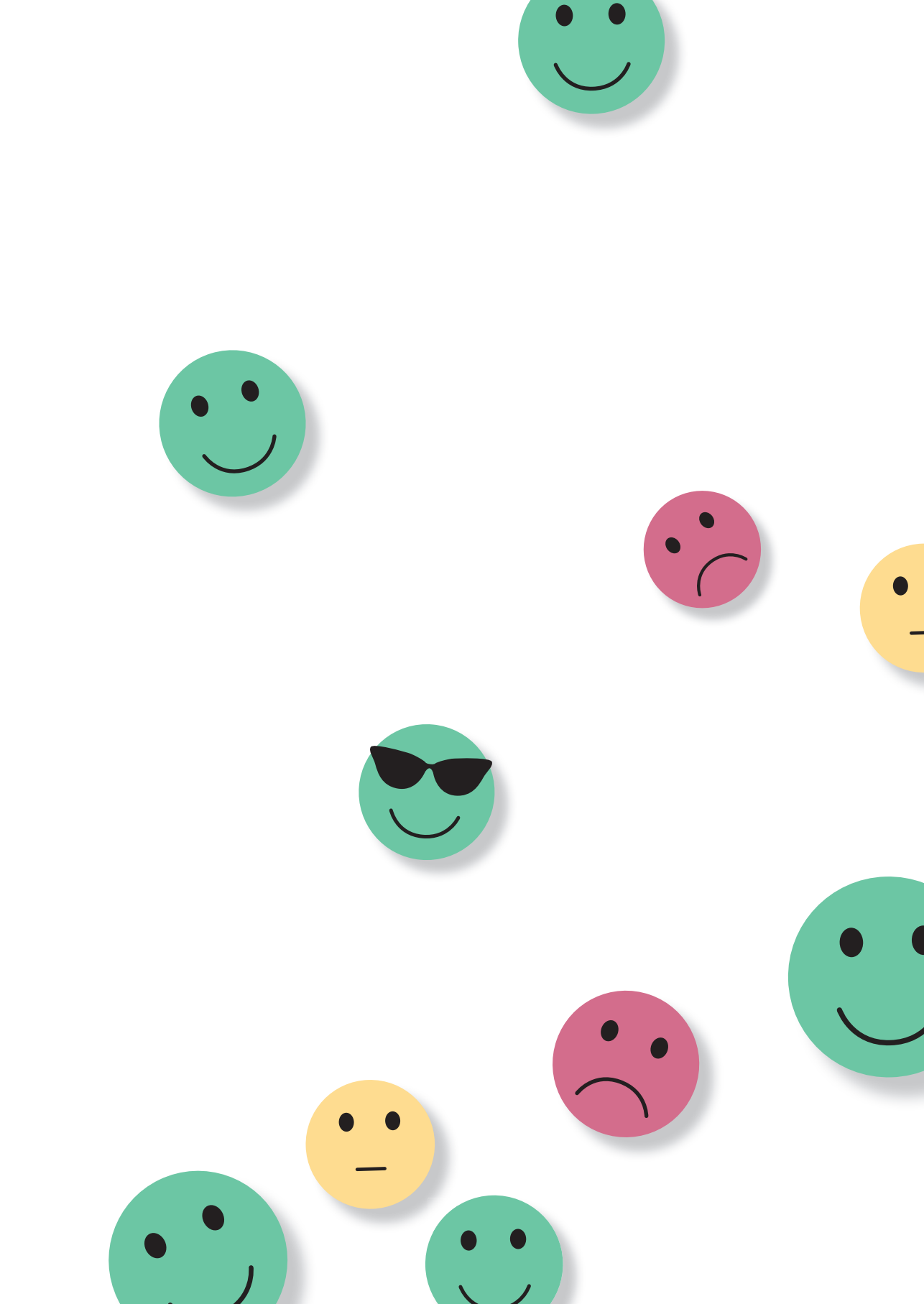
Manuscriptcommissie:

Prof. dr. J.B. Prins

Prof. dr. H.A.M. Marres

Dr. C.B. Terwee (Amsterdam UMC)

TABLE OF CONTENTS



CHAPTER 1

General introduction

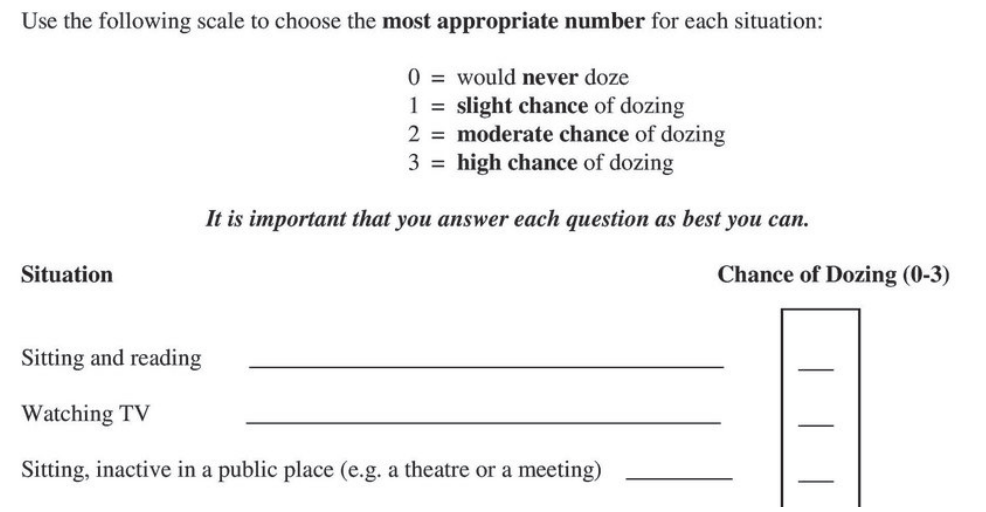
GENERAL INTRODUCTION

Healthcare generally aims to reduce patients' symptoms, minimize disability, and improve quality of life. These aspects of patients' health can be measured with patient-reported outcome measures (PROMs): questionnaires completed by a patient about any aspect of their health status [1]. It is increasingly being recognized that patient-reported outcomes are important outcomes to evaluate in addition to clinical outcomes, for example in clinical trials. In the past decade there has also been increasing interest in using PROMs in clinical practice [2]. There, individual PROM results can be used in the care for individual patients: for example, they can help bring forward the patients' perspective of their health during consultations, and can be used to evaluate individual treatment effects and adapt the treatment if necessary [3]. PROM data collected in clinical practice can also be used on an aggregated level to study quality of care, e.g. by comparing outcomes between healthcare providers [4]. This thesis focuses on the use of PROM data for individual patients with obstructive sleep apnea (OSA). Many studies have been conducted in which a PROM is implemented in the regular clinical practice of different healthcare professionals, but their results are inconsistent regarding the usefulness of PROMs [5-10]. There is especially much to learn still about how and why PROMs do or do not result in improvements in patient care. In this thesis, I describe the development of a PROM for patients with OSA, implement this PROM in clinical practice, and study whether this PROM is beneficial to the care of these patients – and particularly why or why not.

In this chapter I will first describe what different kinds of PROMs exist, how they can be used in regular clinical practice, and what is known about their effectiveness to improve care. Then, I will introduce obstructive sleep apnea. Lastly, I will describe the main objective and the outline of this thesis.

1 Patient-reported outcome measures

The increased interest in PROMs has led to development of many kinds of PROMs in the past decades, measuring different things in different patient groups. There are PROMs that aim to assess one symptom, for example sleepiness, which can be used in patients with different conditions for which this symptom is important. The Epworth Sleepiness Scale (ESS) [11] is an example of this type of PROM (Figure 1). There are also PROMs that aim to measure patients' broader well-being, for example "health-related quality of life" (HRQoL). This is defined as quality of life relative to one's health or disease status [12]. PROMs measuring HRQoL usually contain questions about several topics (called "domains") that assess physical, emotional and social aspects of a patient's functioning. They can either be tailored to specifically fit the symptoms and impact on function of a certain patient group ("disease-specific" PROMs) or aim to be relevant for all patients, irrespective of their disease ("generic" PROMs). Well-known examples of generic PROMs are the EQ-5D and the SF-36 [13, 14]. Generic PROMs are especially useful for comparing outcomes across different patient groups.



Use the following scale to choose the **most appropriate number** for each situation:

0 = would **never** doze
1 = **slight chance** of dozing
2 = **moderate chance** of dozing
3 = **high chance** of dozing

It is important that you answer each question as best you can.

Situation	**Chance of Dozing (0-3)**
Sitting and reading	___
Watching TV	___
Sitting, inactive in a public place (e.g. a theatre or a meeting)	___

Figure 1. Part of the Epworth Sleepiness Scale (ESS), measuring sleep propensity [11]

When selecting a PROM to measure an outcome in a certain situation, it is important that this PROM has been validated for the intended purpose and preferably also for the specific patient group in which the PROM will be used. Validation of a PROM means assessment of its validity, reliability and (for PROMs measuring change over time) responsiveness. Validity means that the PROM measures what it is supposed to measure. A reliable PROM is accurate in its measurements and has a low measurement error. Responsiveness means that the PROM is able to accurately show change over time. These kinds of characteristics of a PROM are called its measurement properties.

2 Individual and aggregate PROM data in clinical practice: possibilities and impact

Individual PROM data can be used in clinical practice to communicate information about a patient's symptoms or (health-related) quality of life to a healthcare professional. For this purpose, patients are usually asked to complete a digital questionnaire before their consultation, either at home or in the waiting room. In the literature, three main ways in which this individual PROM data can be employed have been described [15]: as a screening tool, e.g. to detect patients suffering from depression in primary care; as a monitoring tool, e.g. to assess whether the current treatment plan for a patient with a chronic condition is working; and/or as a tool to increase the patient-centeredness of care. When used to increase patient-centeredness of care, the goal of the PROM is to bring the patient's HRQoL to the forefront of the discussion during a consultation, and facilitate patient involvement in care planning and decision-making [16].

There are many studies in different patient groups that have looked at the potential impact of the use of individual PROM data in clinical care, which have been summarized in a number of systematic reviews [5-10]. Their general conclusions: there is evidence for improvements in communication between patient and healthcare professional, the detection of problems, as well as the patient-doctor relationship. However, the available evidence regarding changes to patient management and a positive impact on health outcomes is weak.

Aggregate PROM data collected in clinical practice can also be used to directly or indirectly improve patient care. For example, by benchmarking the data of individual professionals within one department, or benchmarking the data of hospitals, and feeding back this information to these parties. The idea behind this kind of feedback is that those with relatively low scores will want to work towards improving their care, for instance by learning from those with high scores. So far, no impact of this type of feedback of outcomes data in general (not necessarily PROM data) has been found on patient outcomes [17-21]. This may be partly due to a lack of timeliness and low interpretability of the data, and also because of a lack of clarity on how to go about improving care based on the outcomes data. Benchmarked PROM data could also be shared publically, in which case the healthcare providers with relatively bad outcomes will likely experience more pressure to work towards improving their care and their outcomes. A recent review on whether the public reporting of outcomes data (not necessarily PROM data) improves patient outcomes shows mixed results, with more positive results in more competitive contexts [16].

Another way in which aggregated PROM data collected in clinical practice can be used to improve care is by increasing scientific knowledge. For example by comparing the results of different treatments, to increase knowledge on the effectiveness of treatments in practice. Aggregated PROM data can also be used to help make better choices regarding whether patients should be selected for a certain procedure [4]. Studying baseline PROM results of a large group of patients could, for instance, provide information on whether a patient with hip osteoarthritis and a certain level of impaired function is likely to benefit from a hip replacement. In this latter case, both individual PROM data and aggregated PROM data is used to come to improvement of care.

This thesis is focused on the use of individual PROM data in clinical practice. In our view, implementing PROMs of which the data is meaningful in clinical practice, for individual patients, is the basis for a good PROM measurement system. If patients are able to see their own results and personally benefit from their effort, this is likely to increase their motivation to complete the PROM. This might lead to higher response rates than when patients are asked to complete a PROM that does not potentially benefit their own care. Furthermore, physicians will get more of a 'feel' for the PROM data. The PROM data that is collected for the care of individual patients can, as a second step, then also be interpreted on an aggregate level.

3 Individual PROM data in clinical practice – how do they work?

The reviews on the impact of individual PROM results on the care for patients, mentioned in section 2, have looked only at quantitative evidence. These are often health outcomes. However, rather than setting up studies to only assess *whether* PROMs in clinical practice work, it is also important to study *how* and *why* PROMs (do or do not) work. An important reason for this is that what is considered as one intervention in these reviews (PROM results fed back to a healthcare provider) is actually a diverse group of interventions: studies use different kinds of PROMs, methods of results feedback, settings and patient groups, and type of healthcare provider to which results are fed back. Finding out which aspects of the intervention and the setting in which it is implemented influence its usefulness is key in understanding how to go forward with implementing PROMs in clinical practice.

The ways in which individual PROM data can potentially benefit clinical practice are complex. To illustrate this complexity, Greenhalgh et al. [3] developed a model out of the (often implicit) assumptions that are made in empirical studies (Figure 2). It shows how it is assumed in studies that provision of PROM data to clinicians helps them to monitor treatment response and to detect unrecognized problems, and should lead to differences in doctor-patient communication. This will then potentially effect changes in the clinician's management of patients, as well as cause changes in a patient's own health behavior. This could result in improved patient satisfaction and health outcomes.

A more recent framework, focused on how PROMs can influence the care specifically of patients with chronic conditions, was created by Santana et al. [22] (Figure 3). A difference with the previous model is that it assumes that the PROM results are not only shared with and used by clinicians, but also patients and family members, who are a more explicit actor in the working mechanisms of PROMs. Furthermore, Santana's framework places a larger focus on the role of communication as an essential intermediate factor to bring about changes in patient management and health outcomes. The model shows how the completion of PROMs influences communication between patients, clinicians and family members, which, through increased patient engagement, could lead to changes in the decision-making process, patient management, and adherence to treatment. This has the potential to result in better patient outcomes.

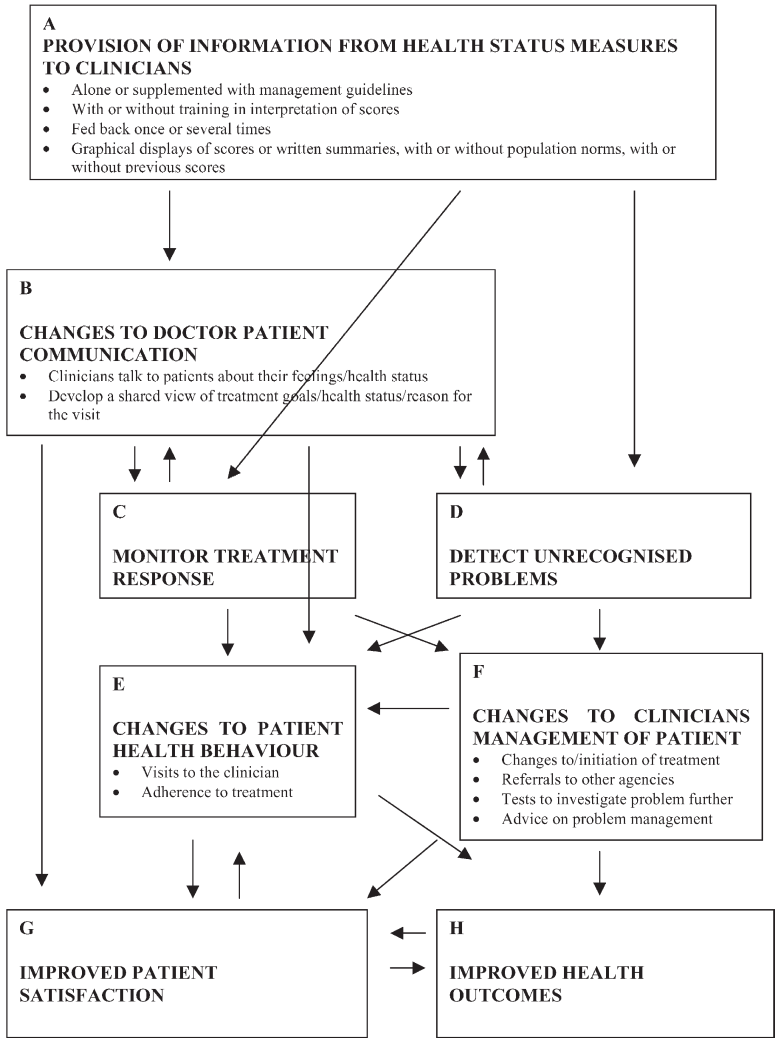


Figure 2 Model of hypotheses and outcomes in the trials evaluating the impact of health status measures on clinical decision making [3]

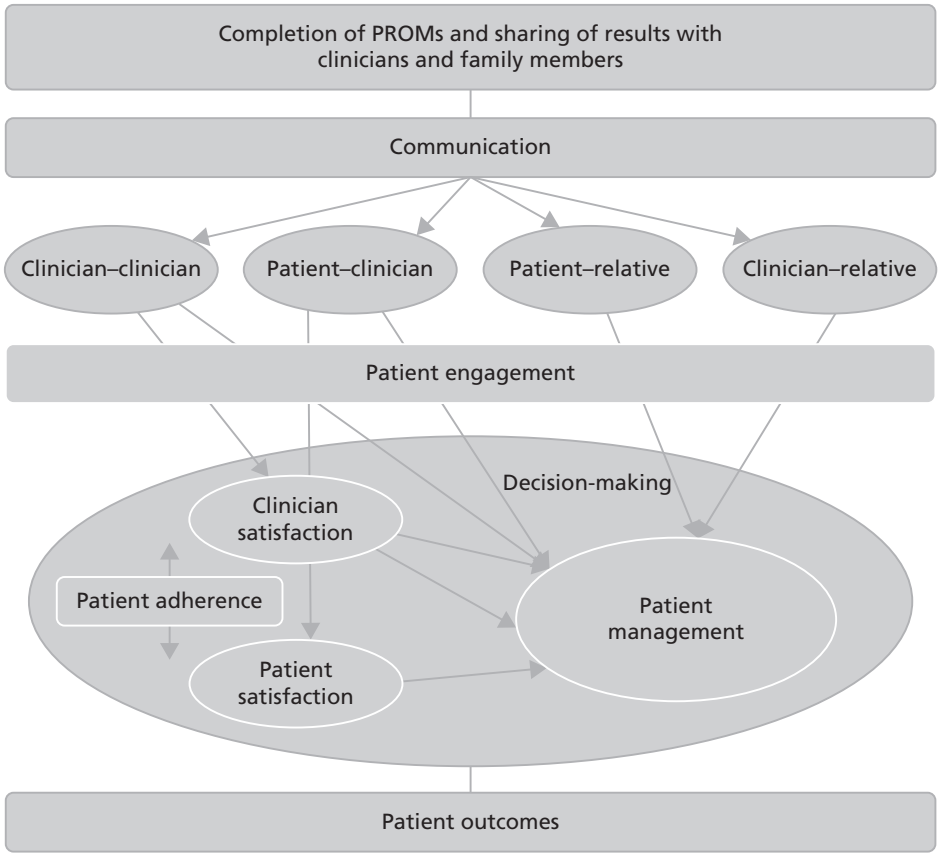


Figure 3 Framework to assess the effects of using patient-reported outcome measures in chronic care management [22]

4 Obstructive sleep apnea

One of the conditions for which a PROM can be potentially useful in clinical practice is obstructive sleep apnea (OSA). Patients with OSA experience collapses of the upper airway during their sleep, resulting in breathing stops that can last from ten seconds up to one minute. Breathing resumes once the brain triggers an arousal from deep sleep to light sleep. This disruption of deep sleep can happen up to hundreds of times per night, and can cause severe sleepiness and exhaustion during the day. Patients often also experience problems in their social lives, due to their loud snoring and their lack of energy, as well as at work, where they may not function optimally [23-25]. OSA has also been related to mental problems such as depression

and anxiety [26, 27]. Furthermore, OSA has been shown to be related to comorbidities such as high blood pressure, heart failure, diabetes, and stroke. Prevalence of OSA has been reported to be 6% to 38%, depending on the exact definition of OSA and the population studied [28].

Severity of OSA and necessity for treatment has historically been based on the number of (partial) breathing stops per hour: the apnea-hypopnea index (AHI) [29, 30]. However, there is no linear association between AHI and severity of symptoms or the presence of comorbidities [31-35]. There is also little evidence that treating patients with mild OSA (based on AHI) or patients with low sleepiness is useful in preventing cardiovascular disease or incidents [36-39]. In the past few years there has therefore been international discussion regarding new approaches to diagnose "clinically relevant" OSA [40, 41]. This discussion has also made its way into recent Dutch guidelines for OSA, [42] in which it is recommended that there should be a greater focus on the presence of potentially related comorbidities, as well as the experienced burden of disease for individual patients. The goal of treatment is the improvement of these aspects of OSA.

A PROM could help with shifting focus to a patient's experienced burden of disease, and - as presented in the previously mentioned models – could potentially lead to changes in communication and patient management, improve patient adherence to treatment, and improve patient satisfaction and health outcomes.

5 General and specific objectives

The general objective of this thesis is to provide more insight into how individual PROM results work when implemented in routine clinical practice. We studied this with OSA as an empirical example. In this context we had the following specific aims:

- To study whether a PROM of sufficient quality is available for patients with OSA, which measures OSA-related quality of life;
- If no existing PROM, measuring OSA-related quality of life, is available and of sufficient quality: to develop a new PROM for patients with OSA specifically for use in clinical practice, with the goal to be suitable for use on both an individual patient and aggregate level;
- To develop a 'patient-friendly' way of presenting the results of the PROM, in order to make them easy to interpret;
- To assess the validity, reliability and responsiveness of the PROM in a Dutch setting;
- To study the impact of individual results of the PROM on the care of patients with OSA, and study *why* this impact is or is not found.

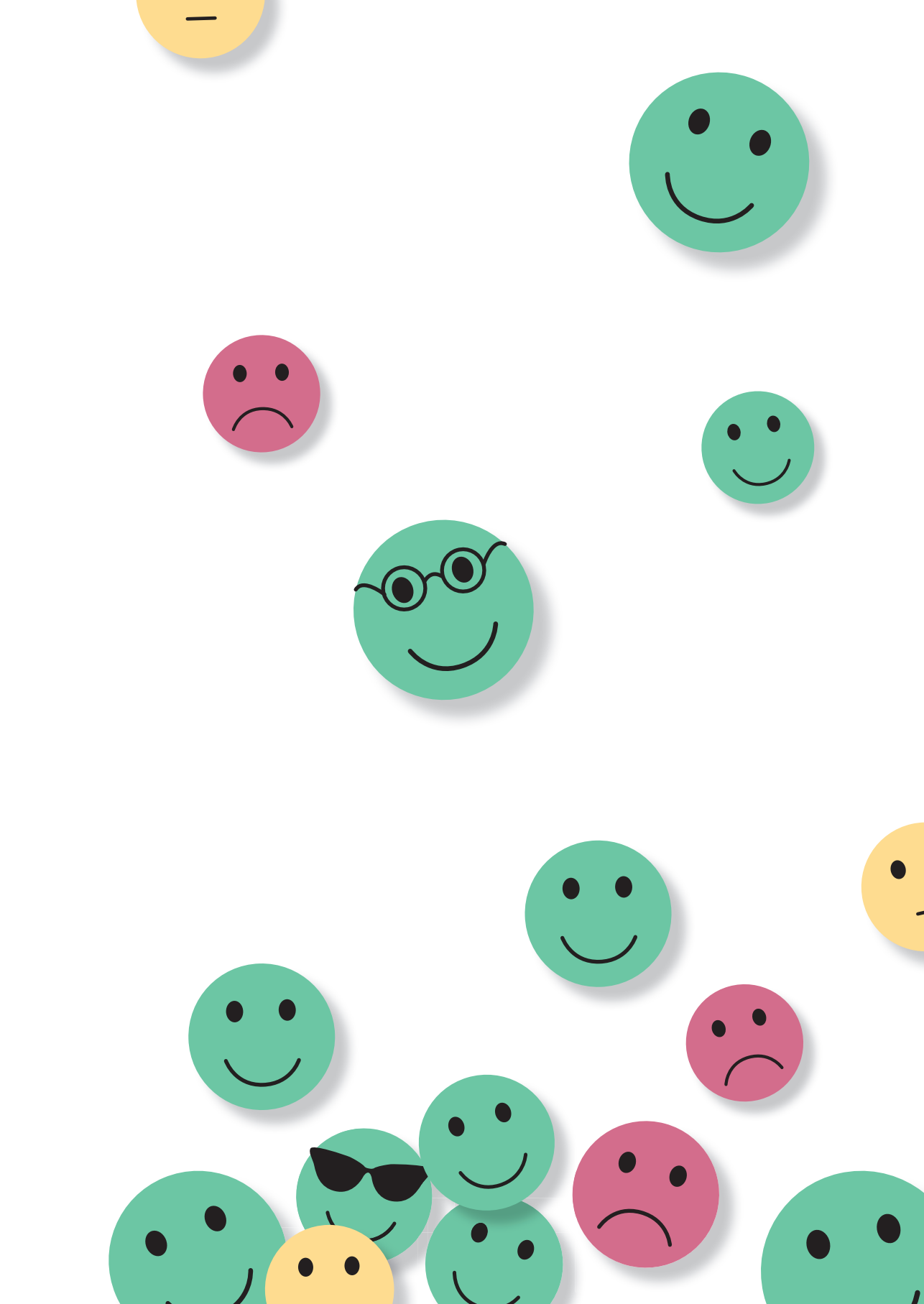
6 Thesis outline

Chapter two provides and overview of the different PROMs that have been validated in patients with OSA, and their measurement properties. **Chapter three** is a reflection on the assessment of validity of PROMs in systematic reviews, and offers suggestions for future reviewers of measurement properties. After concluding that none of the PROMs in our review was suitable in its current form for our purpose, we describe in **chapter four** how we developed a new PROM together with patients and healthcare professionals, specifically designed for use in clinical practice. In **chapter five**, a validation study is presented which shows the measurement properties of the newly developed PROM. Finally, in **chapter six**, the new PROM is implemented in clinical practice and its effects are studied by means of interviews, a patient survey, and a patient record study. In this way, we can study both the impact of individual PROM results on clinical practice, and why this impact is or is not found. In **chapter seven**, we describe and reflect on our main findings, and offer future perspectives for PROMs in the clinical practice of OSA and for evaluating the impact of PROMs in general.

REFERENCES

1. FDA. (2009). *Guidance for industry - Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Silver Spring, Maryland, USA.
2. Valderas, J.M., J. Alonso, and G.H. Guyatt, *Measuring patient-reported outcomes: moving from clinical trials into clinical practice*. Med J Aust, 2008. **189**(2): p. 93-4.
3. Greenhalgh, J., A.F. Long, and R. Flynn, *The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory?* Soc Sci Med, 2005. **60**(4): p. 833-43.
4. Greenhalgh, J., et al., *Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care*. 2017: Southampton (UK).
5. Marshall, S., K. Haywood, and R. Fitzpatrick, *Impact of patient-reported outcome measures on routine practice: a structured review*. J Eval Clin Pract, 2006. **12**(5): p. 559-68.
6. Valderas, J.M., et al., *The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature*. Qual Life Res, 2008. **17**(2): p. 179-93.
7. Boyce, M.B. and J.P. Browne, *Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? A systematic review*. Qual Life Res, 2013. **22**(9): p. 2265-78.
8. Greenhalgh, J. and K. Meadows, *The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review*. J Eval Clin Pract, 1999. **5**(4): p. 401-16.
9. Chen, J., L. Ou, and S.J. Hollis, *A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting*. BMC Health Serv Res, 2013. **13**: p. 211.
10. Knaup, C., et al., *Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis*. Br J Psychiatry, 2009. **195**(1): p. 15-22.
11. Johns, M.W., *A new method for measuring daytime sleepiness: the Epworth sleepiness scale*. Sleep, 1991. **14**(6): p. 540-5.
12. Bakas, T., et al., *Systematic review of health-related quality of life models*. Health Qual Life Outcomes, 2012. **10**: p. 134.
13. Devlin, N.J. and R. Brooks, *EQ-5D and the EuroQol Group: Past, Present and Future*. Appl Health Econ Health Policy, 2017. **15**(2): p. 127-137.
14. Ware, J.E., Jr. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection*. Med Care, 1992. **30**(6): p. 473-83.
15. Greenhalgh, J., *The applications of PROs in clinical practice: what are they, do they work, and why?* Qual Life Res, 2009. **18**(1): p. 115-23.
16. Greenhalgh, J., et al., *How do aggregated patient-reported outcome measures data stimulate health care improvement? A realist synthesis*. J Health Serv Res Policy, 2018. **23**(1): p. 57-65.
17. Taylor, A., et al., *How is feedback from national clinical audits used? Views from English National Health Service trust audit leads*. J Health Serv Res Policy, 2016. **21**(2): p. 91-100.
18. van der Veer, S.N., et al., *Improving quality of care. A systematic review on how medical registries provide information feedback to health care providers*. Int J Med Inform, 2010. **79**(5): p. 305-23.
19. Varagunam, M., et al., *Impact on hospital performance of introducing routine patient reported outcome measures in surgery*. J Health Serv Res Policy, 2014. **19**(2): p. 77-84.
20. Boyce, M.B., J.P. Browne, and J. Greenhalgh, *The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research*. BMJ Qual Saf, 2014. **23**(6): p. 508-18.

21. Boyce, M.B. and J.P. Browne, *The effectiveness of providing peer benchmarked feedback to hip replacement surgeons based on patient-reported outcome measures--results from the PROFILE (Patient-Reported Outcomes: Feedback Interpretation and Learning Experiment) trial: a cluster randomised controlled study.* BMJ Open, 2015. **5**(7): p. e008325.
22. Santana, M.J. and D. Feeny, *Framework to assess the effects of using patient-reported outcome measures in chronic care management.* Qual Life Res, 2014. **23**(5): p. 1505-13.
23. O'Donoghue, N. and E. McKay, *Exploring the impact of sleep apnoea on daily life and occupational engagement.* Br J Occup Ther, 2012. **75**(11): p. 609-516.
24. Reishtein, J.L., et al., *Sleepiness and relationships in obstructive sleep apnea.* Issues Ment Health Nurs, 2006. **27**(3): p. 319-30.
25. Rodgers, B., *Breaking through limbo: experiences of adults living with obstructive sleep apnea.* Behav Sleep Med, 2014. **12**(3): p. 183-97.
26. Bjornsdottir, E., et al., *The Prevalence of Depression among Untreated Obstructive Sleep Apnea Patients Using a Standardized Psychiatric Interview.* J Clin Sleep Med, 2016. **12**(1): p. 105-12.
27. Gupta, M.A., F.C. Simpson, and D.C. Lyons, *The effect of treating obstructive sleep apnea with positive airway pressure on depression and other subjective symptoms: A systematic review and meta-analysis.* Sleep Med Rev, 2016. **28**: p. 55-68.
28. Senaratna, C.V., et al., *Prevalence of obstructive sleep apnea in the general population: A systematic review.* Sleep Med Rev, 2017. **34**: p. 70-81.
29. Medicine, A.A.o.S. (2014). *International Classification of Sleep Disorders. Diagnostic and Coding Manual. .*
30. Committee, D.C.S. (1990). *International Classification of Sleep Disorders: Diagnostic and Coding Manual.* Rochester.
31. Macey, P.M., et al., *Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients.* PLoS One, 2010. **5**(4): p. e10211.
32. Tam, S., B.T. Woodson, and B. Rotenberg, *Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index.* Laryngoscope, 2014. **124**(1): p. 337-43.
33. Kingshott, R.N., et al., *Does arousal frequency predict daytime function?* Eur Respir J, 1998. **12**(6): p. 1264-70.
34. Turnbull, C.D. and J.R. Stradling, *To screen or not to screen for obstructive sleep apnea, that is the question.* Sleep Med Rev, 2017. **36**: p. 125-127.
35. Van Dongen, H.P., et al., *Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability.* Sleep, 2004. **27**(3): p. 423-33.
36. Abuzaid, A.S., et al., *Meta-Analysis of Cardiovascular Outcomes With Continuous Positive Airway Pressure Therapy in Patients With Obstructive Sleep Apnea.* Am J Cardiol, 2017. **120**(4): p. 693-699.
37. Yu, J., et al., *Association of Positive Airway Pressure With Cardiovascular Events and Death in Adults With Sleep Apnea: A Systematic Review and Meta-analysis.* JAMA, 2017. **318**(2): p. 156-166.
38. Marin, J.M., et al., *Association between treated and untreated obstructive sleep apnea and risk of hypertension.* JAMA, 2012. **307**(20): p. 2169-76.
39. Barbe, F., et al., *Effect of continuous positive airway pressure on the incidence of hypertension and cardiovascular events in nonsleepy patients with obstructive sleep apnea: a randomized controlled trial.* JAMA, 2012. **307**(20): p. 2161-8.
40. McNicholas, W.T., *Diagnostic criteria for obstructive sleep apnea: time for reappraisal.* J Thorac Dis, 2018. **10**(1): p. 531-533.
41. McNicholas, W.T., et al., *Challenges in obstructive sleep apnoea.* Lancet Respir Med, 2018. **6**(3): p. 170-172.
42. NVALT. (2017). *Richtlijn diagnostiek en behandeling van obstructief slaapapneu (OSA) bij volwassenen.* Richtlijnendatabase.nl: Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose.



CHAPTER 2

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

Inger L. Abma, Philip J. van der Wees, Vik Veer, Gert P. Westert, Maroeska Rovers

Published in *Sleep Medicine Reviews* 28:14–27 (2016)

ABSTRACT

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.

1 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].

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.

2 METHODS

2.1 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 (Appendix 1). 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).

2.2 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 (PW). Where necessary we contacted study authors for clarification and additional information to inform study selection.

2.3 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:* the extent to 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.

2.4 Data extraction

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

2.5 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 4-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).

2.6 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 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 Appendices 2 and 3.

2.7 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 used in 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, 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 is presented in the text.

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 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

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.

Statistical pooling was performed for all measurement properties which were assessed in more than one study with at least a COSMIN score of "fair", or a score of "poor" due to a small study population. 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 (Appendices 2 and 3). 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.

3 RESULTS

3.1 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 (Figure 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 assessed 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.

3.2 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 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.

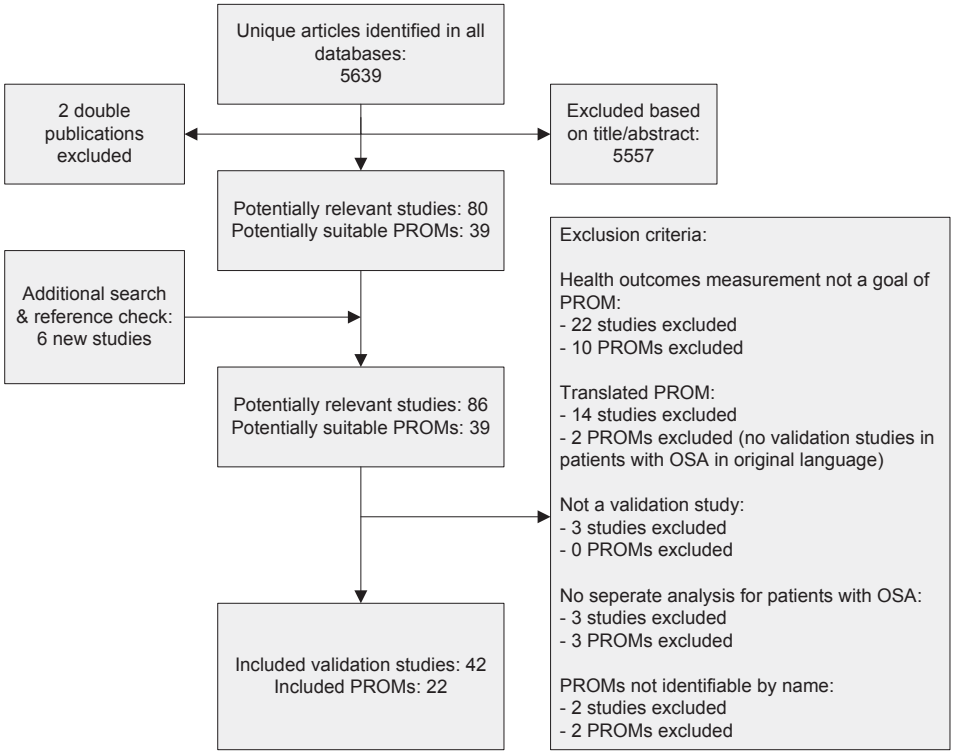


Figure 1 Flow chart for identification of relevant PROMs and validation studies
OSA=obstructive sleep apnea; PROM=patient-reported outcome measure

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)[69]	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) [70]	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

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)[71]	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)[72] (a British version of the (American) sickness impact profile (SIP)[73])	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[74]	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)[75]	1994	English	5 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)[76]	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)[77]	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

OSA=obstructive sleep apnea, PROM=patient-reported outcome measure, CPAP=continuous positive airway pressure

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 5 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.

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-gr. 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[3][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[78]	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		

Table 4 continued

Study	Internal consistency	Reliability (test-retest)	Content validity	Structural validity	Hypothesis testing (construct validity)			Criterion validity	Responsiveness
					Convergent validity	Known-gr. validity	Discriminant validity		
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	
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, SNORE=symptoms of nocturnal obstruction and related events, 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".

3.3 Measurement properties of the PROMs

The results for the measurement properties of the included PROMs considering 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 Appendix 4.

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 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 in favor 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 α=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]).

A 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 at 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 (Appendix 4). However, we could not identify a possible explanation for why there was a consistent discrepancy in these studies.

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-gr. 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	[2]	-
FLP	na	na	na	na	na	na	na	[c]	?
NHP part II	na	na	na	na	?	na	na	[c]	na
PGI	na	na	na	na	na	na	na	[c]	-
SF-12	na	na	na	na	na	na	na	?	?
SF-36	?	na	na	na	+[f]	+	na	[c]	-

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, SNORE=symptoms of nocturnal obstruction and related events, 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.

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 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 BAI there is limited positive level of evidence for structural validity and for internal consistency (one factor, Cronbach's α=0.92 [52]). For the Time of Day Sleepiness Scale (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. [78] 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 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 Appendix 4 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 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 Appendix 4 for pooled correlations of the SF-36). For the SF-36 and 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].

4 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. The Rotterdam Sleepiness Scale we would not recommend: 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 QoL 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 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 this latter type of PROM 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 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 added to this 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 [79] 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 standards regarding the assessment of 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.

5 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.

REFERENCES

1. Punjabi, N.M., *The epidemiology of adult obstructive sleep apnea.* Proc Am Thorac Soc, 2008. **5**(2): p. 136-43.
2. Vanderveken, O.M., et al., *Cardiovascular implications in the treatment of obstructive sleep apnea.* J Cardiovasc Transl Res, 2011. 4(1): p. 53-60.
3. Bradley, T.D. and J.S. Floras, *Obstructive sleep apnoea and its cardiovascular consequences.* Lancet, 2009. **373**(9657): p. 82-93.
4. Leger, D., et al., *Impact of sleep apnea on economics.* Sleep Med Rev, 2012. **16**(5): p. 455-62.
5. Jennum, P. and J. Kjellberg, *Health, social and economical consequences of sleep-disordered breathing: a controlled national study.* Thorax, 2011. **66**(7): p. 560-6.
6. Pang, K.P. and B.W. Rotenberg, *Redefining successful therapy in obstructive sleep apnea: a call to arms.* Laryngoscope, 2014. **124**(5): p. 1051-2.
7. Ravesloot, M.J. and N. de Vries, *Reliable calculation of the efficacy of non-surgical and surgical treatment of obstructive sleep apnea revisited.* Sleep, 2011. **34**(1): p. 105-10.
8. Tam, S., B.T. Woodson, and B. Rotenberg, *Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index.* Laryngoscope, 2014. **124**(1): p. 337-43.
9. Macey, P.M., et al., *Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients.* PLoS One, 2010. **5**(4): p. e10211.
10. Dutt, N., et al., *Quality of life impairment in patients of obstructive sleep apnea and its relation with the severity of disease.* Lung India, 2013. **30**(4): p. 289-94.
11. Kezirian, E.J., et al., *Reporting results of obstructive sleep apnea syndrome surgery trials.* Otolaryngol Head Neck Surg, 2011. **144**(4): p. 496-9.
12. Black, N., *Patient reported outcome measures could help transform healthcare.* BMJ, 2013. **346**: p. f167.
13. Van der Wees, P., et al., *Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries.* The Milbank Quarterly, 2014. **92**(4): p. 754-75.
14. Abrishami, A., A. Khajehdehi, and F. Chung, *A systematic review of screening questionnaires for obstructive sleep apnea.* Canadian Journal of Anaesthesia-Journal Canadien D Anesthesie, 2010. **57**(5): p. 423-438.
15. Fedson, A.C., A.I. Pack, and T. Gislason, *Frequently used sleep questionnaires in epidemiological and genetic research for obstructive sleep apnea: a review.* Sleep Med Rev, 2012. **16**(6): p. 529-37.
16. Kendzerska, T.B., et al., *Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review.* Sleep Med Rev, 2014. **18**(4): p. 321-31.
17. Ramachandran, S.K. and L.A. Josephs, *A meta-analysis of clinical screening tests for obstructive sleep apnea.* Anesthesiology, 2009. **110**(4): p. 928-39.
18. Stucki, A., et al., *Content comparison of health-related quality of life instruments for obstructive sleep apnea.* Sleep Med, 2008. **9**(2): p. 199-206.
19. Terwee, C.B., et al., *Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments.* Qual Life Res, 2009. **18**(8): p. 1115-23.
20. Lacasse, Y., C. Godbout, and F. Series, *Health-related quality of life in obstructive sleep apnoea.* Eur Respir J, 2002. **19**(3): p. 499-503.
21. Mokkink, L.B., 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**(7): p. 737-45.

22. Cortina, J.M., *What Is Coefficient Alpha - an Examination of Theory and Applications.* Journal of Applied Psychology, 1993. **78**(1): p. 98-104.
23. Mokkink, L.B., 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**: p. 22.
24. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Qual Life Res, 2012. **21**(4): p. 651-7.
25. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
26. Schellingerhout, J.M., et al., *Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review.* Qual Life Res, 2012. **21**(4): p. 659-70.
27. Ryan, R., *Heterogeneity and subgroup analyses in Cochrane Consumers and Communication Review Group reviews: planning the analysis at protocol stage.* 2013, Cochrane Consumers and Communication Review Group.
28. Billings, M.E., 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.
29. Weaver, T.E., et al., *An instrument to measure functional status outcomes for disorders of excessive sleepiness.* Sleep, 1997. **20**(10): p. 835-43.
30. Weaver, E.M., B.T. Woodson, and D.L. Steward, *Polysomnography indexes are discordant with quality of life, symptoms, and reaction times in sleep apnea patients.* Otolaryngol Head Neck Surg, 2005. **132**(2): p. 255-62.
31. Piccirillo, J.F., et al., *Obstructive sleep apnea treatment outcomes pilot study.* Otolaryngol Head Neck Surg, 1998. **118**(6): p. 833-44.
32. Lacasse, Y., M.P. Bureau, and F. Series, *A new standardised and self-administered quality of life questionnaire specific to obstructive sleep apnoea.* Thorax, 2004. **59**(6): p. 494-9.
33. Masa, J.F., et al., *Visual analogical well-being scale for sleep apnea patients: Validity and responsiveness.* Sleep and Breathing, 2011. **15**(3): p. 549-559.
34. Flemons, W.W. and M.A. Reimer, *Development of a disease-specific health-related quality of life questionnaire for sleep apnea.* Am J Respir Crit Care Med, 1998. **158**(2): p. 494-503.
35. Flemons, W.W. and M.A. Reimer, *Measurement properties of the calgary sleep apnea quality of life index.* Am J Respir Crit Care Med, 2002. **165**(2): p. 159-64.
36. Chasens, E.R., S.J. Ratcliffe, and T.E. Weaver, *Development of the FOSQ-10: a short version of the Functional Outcomes of Sleep Questionnaire.* Sleep, 2009. **32**(7): p. 915-9.
37. Moroni, L., et al., *A new means of assessing the quality of life of patients with obstructive sleep apnea: the MOSAS questionnaire.* Sleep Med, 2011. **12**(10): p. 959-65.
38. Chervin, R.D. and M.S. Aldrich, *The Epworth Sleepiness Scale may not reflect objective measures of sleepiness or sleep apnea.* Neurology, 1999. **52**(1): p. 125-31.
39. Cowan, D.C., et al., *Predicting sleep disordered breathing in outpatients with suspected OSA.* BMJ Open, 2014. **4**(4): p. e004519.
40. Giudici, S., et al., *Lack of predictive value of the Epworth Sleepiness Scale in patients after uvulopalatopharyngopolasty.* Annals of Otology, Rhinology & Laryngology, 2000. **109**(7): p. 646-649.
41. Hardinge, F.M., D.J. Pitson, and J.R. Stradling, *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**(9): p. 617-20.
42. Hesselbacher, S., et al., *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**: p. 20-7.
43. Johns, M.W., *A new method for measuring daytime sleepiness: the Epworth sleepiness scale.* Sleep, 1991. **14**(6): p. 540-5.

44. Johns, M.W., *Reliability and factor analysis of the Epworth Sleepiness Scale.* Sleep, 1992. **15**(4): p. 376-81.
45. Johns, M.W., *Daytime sleepiness, snoring, and obstructive sleep apnea. The Epworth Sleepiness Scale.* Chest, 1993. **103**(1): p. 30-6.
46. Johns, M.W., *Sleepiness in different situations measured by the Epworth Sleepiness Scale.* Sleep, 1994. **17**(8): p. 703-10.
47. Kingshott, R.N., et al., *Does arousal frequency predict daytime function?* European Respiratory Journal, 1998. **12**(6): p. 1264-1270.
48. Kingshott, R.N., et al., *Self assessment of daytime sleepiness: patient versus partner.* Thorax, 1995. **50**(9): p. 994-5.
49. Olaithe, M., et al., *Can we get more from the Epworth Sleepiness Scale (ESS) than just a single score? A confirmatory factor analysis of the ESS.* Sleep Breath, 2013. **17**(2): p. 763-9.
50. Olson, L.G., M.F. Cole, and A. Ambrogetti, *Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms.* J Sleep Res, 1998. **7**(4): p. 248-53.
51. Osman, E.Z., et al., *The Epworth Sleepiness Scale: can it be used for sleep apnoea screening among snorers?* Clin Otolaryngol Allied Sci, 1999. **24**(3): p. 239-41.
52. Sanford, S.D., et al., *Psychometric evaluation of the Beck anxiety inventory: a sample with sleep-disordered breathing.* Behav Sleep Med, 2008. **6**(3): p. 193-205.
53. Sangal, R.B., J.M. Sangal, and C. Belisle, *Subjective and objective indices of sleepiness (ESS and MWT) are not equally useful in patients with sleep apnea.* Clin Electroencephalogr, 1999. **30**(2): p. 73-5.
54. Sil, A. and G. Barr, *Assessment of predictive ability of Epworth scoring in screening of patients with sleep apnoea.* J Laryngol Otol, 2012. **126**(4): p. 372-9.
55. Smith, I.E. and J.M. Shneerson, *Is the SF 36 sensitive to sleep disruption? A study in subjects with sleep apnoea.* J Sleep Res, 1995. **4**(3): p. 183-188.
56. Walter, T.J., et al., *Comparison of Epworth Sleepiness Scale scores by patients with obstructive sleep apnea and their bed partners.* Sleep Med, 2002. **3**(1): p. 29-32.
57. Weaver, E.M., V. Kapur, and B. Yueh, *Polysomnography vs self-reported measures in patients with sleep apnea.* Arch Otolaryngol Head Neck Surg, 2004. **130**(4): p. 453-8.
58. Dolan, D.C., et al., *The Time of Day Sleepiness Scale to assess differential levels of sleepiness across the day.* J Psychosom Res, 2009. **67**(2): p. 127-33.
59. Gliklich, R.E. and P.C. Wang, *Validation of the snore outcomes survey for patients with sleep-disordered breathing.* Arch Otolaryngol Head Neck Surg, 2002. **128**(7): p. 819-24.
60. Law, M., et al., *Validation of two depression screening instruments in a sleep disorders clinic.* Journal of Clinical Sleep Medicine, 2014. **10**(6): p. 683-688.
61. Sangal, R.B., *Evaluating sleepiness-related daytime function by querying wakefulness inability and fatigue: Sleepiness-Wakefulness Inability and Fatigue Test (SWIFT).* J Clin Sleep Med, 2012. **8**(6): p. 701-11.
62. van Knippenberg, F.C., et al., *The Rotterdam Daytime Sleepiness Scale: a new daytime sleepiness scale.* Psychol Rep, 1995. **76**(1): p. 83-7.
63. Yi, H., et al., *Validity and reliability of Sleep Quality Scale in subjects with obstructive sleep apnea syndrome.* J Psychosom Res, 2009. **66**(1): p. 85-8.
64. Bennett, L.S., et al., *Health status in obstructive sleep apnea: relationship with sleep fragmentation and daytine sleepiness, and effects of continuous positive airway pressure treatment.* Am J Respir Crit Care Med, 1999. **159**(6): p. 1884-90.
65. Jenkinson, C. and R. Layte, *Development and testing of the UK SF-12 (short form health survey).* J Health Serv Res Policy, 1997. **2**(1): p. 14-8.

66. Jenkinson, 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**(2): p. 179-86.
67. Jenkinson, C., J. Stradling, and S. Petersen, *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**(3): p. 199-204.
68. Jenkinson, C., J. Stradling, and S. Petersen, *How should we evaluate health status? A comparison of three methods in patients presenting with obstructive sleep apnoea.* Qual Life Res, 1998. **7**(2): p. 95-100.
69. Beck, A.T., et al., *An inventory for measuring clinical anxiety: psychometric properties.* J Consult Clin Psychol, 1988. **56**(6): p. 893-7.
70. Zigmond, A.S. and R.P. Snaith, *The hospital anxiety and depression scale.* Acta Psychiatr Scand, 1983. **67**(6): p. 361-70.
71. EuroQol, G., *EuroQol--a new facility for the measurement of health-related quality of life.* Health Policy, 1990. **16**(3): p. 199-208.
72. 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*. 1981, US department of Health and Human Services: Hyattsville, MD.
73. Bergner, M., et al., *The sickness impact profile: conceptual formulation and methodology for the development of a health status measure.* Int J Health Serv, 1976. **6**(3): p. 393-415.
74. Hunt, S.M., J. McEwen, and S.P. McKenna, *Measuring health status: a new tool for clinicians and epidemiologists.* J R Coll Gen Pract, 1985. **35**(273): p. 185-8.
75. Ruta, D.A., et al., *A new approach to the measurement of quality of life. The Patient-Generated Index.* Med Care, 1994. **32**(11): p. 1109-26.
76. Ware, J., Jr., M. Kosinski, and S.D. Keller, *A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity.* Med Care, 1996. **34**(3): p. 220-33.
77. Ware, J.E., Jr. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.* Med Care, 1992. **30**(6): p. 473-83.
78. Smith, S.S., et al., *Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnoea.* Sleep Med, 2008. **9**(7): p. 739-44.
79. Abma, I.L., et al. *Measurement properties of patient-reported outcome measures in adults with obstructive sleep apnea: a systematic review.* 2014:CRD42014014608

Appendix 1: search strategy

Category	Search terms[a]
Patient-reported outcome measure	patient-reported outcome measure*[tiab] OR PROM[tiab] OR PROMs[tiab] OR quality of life[tiab] OR QoL[tiab] OR HRQoL[tiab] OR HRQL[tiab] OR questionnaire*[tiab] OR survey[tiab] OR surveys[tiab] OR instrument[tiab] OR instruments[tiab] OR scale[tiab] OR scales [tiab] OR checklist*[tiab] OR assessment*[tiab] OR computer adaptive test*[tiab] OR self-report*[tiab] OR diary[tiab] OR diaries[tiab] OR log[tiab] OR logs[tiab] OR interview*[tiab] OR questionnaires [Mesh] OR Quality of life[Mesh] OR self report [Mesh] OR interviews as topic [Mesh]
Obstructive sleep apnea	Sleep apnea[tiab] OR sleep apnoea[tiab] OR Obstructive sleep disorder*[tiab] OR OSA[tiab] OR OSAS[tiab] OR "Sleep Apnea Syndromes"[Mesh]
Filter validation studies[1]	instrumentation[sh] OR methods[sh] OR Validation Studies[pt] OR Comparative Study[pt] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR outcome assessment[tiab] OR outcome measure*[tw] OR "observer variation"[MeSH] OR observer variation[tiab] OR "Health Status Indicators"[Mesh] OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tiab] OR precision[tiab] OR imprecision[tiab] OR "precise values"[tiab] OR test–retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tiab] OR ((replicab*[tiab] OR repeated[tiab]) AND (measure[tiab] OR measures[tiab] OR findings[tiab] OR result[tiab] OR results[tiab] OR test[tiab] OR tests[tiab])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR factor analysis[tiab] OR factor analyses[tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR item discriminant[tiab] OR interscale correlation*[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR meaningful change[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab] NOT ("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms]

a. Only the PubMed search strategy is presented in the table. The different categories were combined with AND.

Appendix 2: hypotheses for convergent validity

PROM under study	Comparator instrument	Expected correlation strength (direction)	Details
BAI	Question about frequency of anxiety	Moderate (+)	The BAI asks the respondent how bothersome (mild/moderate/severe) symptoms and emotional aspects of anxiety have been in the past month. The constructs of the two questions are similar to the construct of the BAI, and we expect a moderate correlation with both questions.
	Question about severity of anxiety	Moderate (+)	
ESS	MWT	Moderate (-)	The ESS measures sleep propensity. The ESS, MWT and MSLT all measure sleep propensity, but the ESS asks after sleep propensity in a range of situations, whereas the two objective tests only measure sleep propensity in a sleep laboratory setting. Therefore we expected at maximum a moderate correlation. We expect a stronger correlation with MWT than with MSLT, because the MWT measures a patient's ability to stay awake, which is more closely related to the ESS than the MSLT's measurement of the ability to fall asleep. A moderate correlation between the ESS and PS is expected because of related constructs. The SF-36 mental health component is expected to have a weak to moderate correlation with the ESS because the constructs partially overlap. Self-rated health is a much broader construct than sleepiness, and as such we consider it too different to be a suitable comparator instrument. The objective severity of OSA, as measured by polysomnographic measures, can influence sleep propensity but it is unclear to what extent, so we consider it unsuitable.
	MSL	Moderate (-)	
	PS (subjective fre-quency of problematic sleepiness)	Moderate (+)	
	SF-36 mental health component	Weak to moderate (-)	
	Unsuitable comparator instruments		
	Self-rated health		
	Polysomnographic measures[a]		
FOSQ	ESS	Moderate (-)	The FOSQ measures the impact of sleepiness on functional status. We expect a moderate correlation between the FOSQ and ESS, as constructs of both questionnaires are related to sleepiness. We expect the strongest correlation with the "vigilance" domain of the FOSQ because the questions cover the ability to enjoy several somnific activities, which is mostly closely related to the ESS questions. Because there were a total of 48 correlations measured of the total score and domains of the FOSQ with the SF-36, we were somewhat stricter in determining which constructs were similar. We expect moderate correlations between the "vitality" domain of the SF-36 and all domains and the total score of the FOSQ, as the vitality domain asks about the energy level of the respondent. We also expect a moderate correlation between the SF-36's "social functioning" and the FOSQ domains "activity level" and "general productivity", as these domains assess the impact of mental health upon activities. The "physical functioning" domain of the SF-36 should be moderately correlated to the "activity level" of the FOSQ. The comparisons between domains of the FOSQ and the domains of the SF-36 that are not mentioned above, we do not consider to be suitable comparisons. We also do not consider the overall scores of the FOSQ and the SIP to be suitable comparator instruments, because only a limited number of items ask after similar problems.
	SF-36	Moderate (+)	
	Unsuitable comparator instruments		
	SIP		
	Some SF-36 domains (see description)		

HADS	BDI-FS	Moderate to Strong (+)	The HADS measures anxiety and depression. The HADS and BDI-FS have overlapping constructs and are expected to have a moderate to strong correlation. The HAD-D (measuring only depression) should have a stronger correlation with the BDI-FS than the overall HAD. Severity of OSA, as measured by the polysomnographic measures, and sleepiness, as measured by MSLT and MWT, we do not consider to be suitable comparator instruments for the HADS as their constructs are too different.
	Unsuitable comparator instruments		
	Polysomnographic measures[a]		
	MSLT		
	MWT		
MOSAS questionnaire (physical impact (phyImp) and psychological impact (psyImp) + CPAP discomfort)	ESS (phyImp only)	Moderate to strong (+)	The MOSAS questionnaire assesses quality of life for patients with OSA. It's psychological impact domain measures emotions and to what extent they influence the life of the patients. The physical impact domain measures sleepiness. We expect a moderate to strong correlation between the physiological impact domain of the MOSAS questionnaire and the ESS, as they both assess sleep propensity/sleepiness. We expect moderate correlations between the psychological impact domain of the MOSAS questionnaire and the STAI-X3 and QD-R, because all assess anxiety and depression. We expect a moderate correlation between the "CPAP nuisance/discomfort" domain of the MOSAS questionnaire and CPAP adherence after 6 months, as those who experience more discomfort are likely to use it less. Severity of OSA, as measured by polysomnographic measures, can influence experienced quality of life but it is unclear to what extent, so we do not consider it to be a suitable comparator instrument. The STAI-X3 and QD-R we do not consider to be suitable comparator instruments for the physical impact domain or the CPAP discomfort domain of the MOSAS questionnaire as their constructs are too different. We do also not consider the ESS to be suitable as comparator for the psychological impact domain of the MOSAS questionnaire as their constructs are too different.
	STAI-X3 (psyImp only)	Moderate (+)	
	QD-R (psyImp only)	Moderate (+)	
	CPAP adherence (CPAP discomfort only)	Moderate (+)	
	Unsuitable comparator instruments		
	Polysomnographic measures at baseline (with psyImp, phyImp and CPAP discomfort) STAI-X3 (with phyImp and CPAP discomfort) QD-R (with phyImp and CPAP discomfort) ESS (with psyImp and CPAP discomfort)		
NHP part II	*Unsuitable comparator instruments*		The NHP part II measures generic health-related quality of life. The objective severity of OSA, as measured by the polysomnographic measures, and sleepiness, as measured by MSLT and MWT, we do not consider to be suitable comparator instruments for the NHP part II as their constructs are too different.
	Polysomnographic measures[a]		
	MSLT		
	MWT		
OSAPOSI	Single rating of "bother/ disturbance" due to OSA	Moderate (?)	The OSAPOSI measures quality of life specifically for patients with OSA. The OSAPOSI severity score for each item is multiplied with the importance of the item to the patient, to reach a final score. We expect a moderate correlation with overall bother due to OSA as "bother" also implies a judgment of importance to the patient. It is not stated in the paper whether a higher or a lower score on this scale indicates more bother due to OSA, so we cannot predict the direction of the correlation.

Appendix 2 ***continued***			
QSQ	ESS	Moderate (-)	The QSQ measures quality of life specifically for patients with OSA. Domains of the QSQ are compared to (domains of) other instruments that the authors of the paper deem relevant. They hypothesize that all correlations will be between 0.4 and 0.7.[2] To add more detail to these comparisons, we added some additional hypotheses based on the similarities and differences of the compared domains. We expect that the QSQ domain "daytime sleepiness" will correlate strongest with the ESS, the "vitality" domain of the SF-36, and all domains of the FOSQ, as they have related constructs. A weaker but still moderate correlation is expected for the "role-physical" domain of the SF-36 and even weaker for the "physical functioning" domain of the SF-36. The QSQ domain "diurnal symptoms" is expected to correlate strongest with the "vitality" domain of the SF-36, the "general productivity" and "activity level" domains of the FOSQ. Again, a weaker correlation is expected with the "role-physical" domain of the SF-36, and the weakest correlation with the "physical functioning" domain of the SF-36. The correlation of the QSQ domain "social interactions" with all three comparison domains of the FOSQ, SF-36 and SCL-90 are expected to be on the weaker end of moderate as they mostly contain items with a different focus than those of the QSQ domain. We made no additional hypotheses for the "nocturnal symptoms" and "emotions" domains of the QSQ.
	FOSQ	Moderate (+)	
	SF-36	Moderate (+)	
	SCL-90	Moderate (-)	
	BDI	Moderate (-)	
Rotterdam Sleepiness Scale	POMS (fatigue and vigor domains)	Moderate	The Rotterdam Sleepiness Scale measures sleepiness, its behavioral impact, and the affected life domains due to sleepiness. We expect a moderate correlation of all Rotterdam Sleepiness Scale domains with the "fatigue" and "vigor" domains of the POMS. We do not consider the other domains of the POMS, as well as the somatization domain of the SCL-90, to be suitable comparator instruments, because their constructs are too different.
	Unsuitable comparator instruments		
	POMS (other domains) SCL-90 somatization domain		

SAQLI (overall score and 4 domains)	SF-36	Weak/Moderate (+)	The SAQLI measures quality of life specifically for patients with OSA. Flemons et al.[3] has constructed hypotheses of exact correlation strengths between the total SAQLI score and several instruments. Since it would be very unlikely if their exact predictions were true, we have translated these predictions into the categories that we have used for our own hypotheses. We excluded hypotheses for correlations with constructs that we consider non-similar. Moderate to strong correlations are then expected between the SAQLI total score and the "health and functioning" and "psychological and spiritual" domains of the F&P index, and moderate correlations between the SAQLI total score and the total score of the F&P index, the "general health", "vitality" and "social functioning" domains of the SF-36, and the global QoL rating. Weak to moderate correlations are expected between the total SAQLI score and the ESS and the "mental health" domain of the SF-36, and weak correlations are expected between the total SAQLI score and the "role emotional functioning" domain of the SF-36. Because there were a total of 32 correlations assessed between the domains of the SAQLI and the SF-36, we were stricter in determining which constructs were similar. We expect moderate correlations between the following domains: the correlation of the SAQLI's "symptoms" domain and the SF-36's "vitality" domain; the SAQLI's "social interactions" domain and the SF-36's "social functioning" domain; the SAQLIs "daily functioning" and the SF-36's "role physical functioning" domain; and the SAQLI's "emotional functioning" domain and the SF-36's "mental health" and "social functioning" domains. It is expected that these correlations are stronger than the other SF-36 domain correlations with the respective SAQLI domains. We expect a weak to moderate correlation of the SAQLI domains "symptoms" and "daily functioning" with the ESS. The ESS we consider to be unsuitable as a comparator instrument for all SAQLI subscales not mentioned above, because the constructs are too different. All comparisons with domains of the SF-36 and the F&P index that were not mentioned above, we do consider suitable comparator instruments/domains for the remaining domains of the SAQLI because the constructs are too different. The objective severity of OSA, as measured by polysomnographic measures, will likely influence quality of life, but it is not clear to what extent and we consider this an unsuitable comparator.
	Global QoL rating	Moderate (+)	
	F&P index	Moderate/ Moderate to strong (+)	
	ESS	Weak to Moderate (-)/ Moderate (-)	
	Unsuitable comparator instruments Some SF-36 domains (see description) Some Ferrans & Powers index domains (see description) Polysomnographic measuresa ESS (comparison with some SAQLI domains)		

Appendix 2 *continued*

SF-36 (mental health (MH) and physical health (PH) components)	ESS	Moderate (MH) (-)	The SF-36 measures the mental and physical aspects of quality of life, as well as health change compared to the previous year. The ESS is expected to correlate moderately with the mental health component of the SF-36. The correlation of the SF-36's "vitality" domain with the ESS is stronger than the one with the overall mental health component and the ESS. The MSLT and MWT we expect to be weakly to moderately correlated to (the domains of) the SF-36's mental health component and the "health change" domain. We expect a stronger correlation with MWT than with MSLT, because the MWT measures a patient's ability to stay awake, which is more closely related to sleepiness than the MSLT's measurement of the ability to fall asleep. Self-rated health is expected to have a weak to moderate correlation with the SF-36's mental health component, as it covers a broader construct than the SF-36, and has only one question (there was no comparison with the physical health component). The objective severity of OSA, as measured by polysomnographic measures, can influence experienced quality of life but we do consider it a suitable comparator for either the physical or mental health component of the SF-36 as the constructs are too different. The MSLT and MWT we do not consider to be suitable comparator instruments for the physical health component of the SF-36 and the domains within this component because the constructs are too different.
	MSLT	Weak to moderate (MH) (-)	
	MWT	Weak to moderate (MH) (-)	
	Self-rated health	Weak to moderate (MH) (-)	
	Unsuitable comparator instruments		
	Polysomnographic measuresa (MH and PH) MSLT (PH) MWT (PH)		
SNORE25	*Unsuitable comparator instruments*		The SNORE25 measures quality of life specifically for patients with OSA. The objective severity of OSA, as measured by polysomnographic measures, will likely influence quality of life, but the extent of this is not clear and we consider it an unsuitable comparator.
	Polysomnographic measuresa		
SOS	SBPS	Moderate (+)	The SOS subjectively measures frequency and severity of snoring as well as tiredness due to snoring. A moderate correlation is expected with the SPBS, as this questionnaire measures the patient's snoring as perceived by their partner. We expect a weak to moderate correlation of the SOS with the ESS and the "vitality" domain of the SF-36 because of partly overlapping constructs. A weak to moderate correlation is expected with the "sleep disturbances" domain of the PSQI, as it contains two questions about snoring and stopping breathing, and with the "overall sleep quality" of the PSQI. We expect a weaker correlation with the overall score of the PSQI. All other comparisons with domains of the SF-36 and the PSQI we do not consider suitable. It is unclear to what extent the severity of OSA is related to the severity of snoring and the subjective complaints due to snoring, therefore we consider this comparison unsuitable.
	SF-36	Weak to moderate (+)	
	PSQI	Moderate (-)	
	ESS	Weak to moderate (-)	
	Unsuitable comparator instruments		
	Polysomnographic measuresa Some SF-36 domains (see description) Some PSQI domains (see description)		

SQS	*Unsuitable comparator instruments* Polysomnographic measuresa		The SQS measures subjective sleep quality and also focuses for a great part on daytime complaints and reduced functioning due to poor sleep. The objective severity of OSA, as measured by polysomnographic measures, we do not consider a suitable comparator because its relation with subjective daytime complaints is unclear.
SWIFT	*Unsuitable comparator instruments* Polysomnographic measuresa		The SWIFT measures wakefulness inability and fatigue. The objective severity of OSA, as measured by polysomnographic measures, will likely influence wakefulness inability and fatigue, but we do not consider it a suitable comparator.
ToDSS	ESS	Strong (+)	The ToDSS measures sleepiness at different times of the day. We expect strong correlations for morning/afternoon/evening ToDSS scores compared to ESS scores, because the ToDSS questions are the same as the ESS questions apart from the "time of day" addition.
VAWS	FOSQ	Moderate (+)	The VAWS measures well-being status with regard to the symptoms which were the motive of the consultation. The strongest moderate correlations are expected with FOSQ and ESS (because the symptoms which were the motive for the consultation are likely related to sleepiness and fatigue) and EQ-T (because of a related construct and similar outcome scale). A somewhat weaker correlation with the SF-36 mental health component and the EQ-5D are expected due to only partly overlapping constructs.
	SF-36 mental health component	Weak to moderate (+)	
	EQ-5D (+EQ-T)	Weak to moderate (+)	
	ESS	Moderate (-)	
	ASDA sleepiness	?	
	Unsuitable comparator instruments Polysomnographic (PSG) measures SF-36 physical health		The objective severity of OSA, as measured by polysomnographic measures, will likely influence quality of life, but it is unclear to what extent and we consider it an unsuitable comparator. The content of the ASDA questionnaire could not be identified.

ASDA=American sleep disorders association; BAI=Beck anxiety inventory; BDI – Beck depression inventory; CPAP=continuous positive airway pressure; EQ-5D – euroqol-5D; EQ-T – euroqol thermometer; ESS=Epworth sleepiness scale; FLP=functional limitations profile; FOSQ – functional outcomes of sleep questionnaire; F&P index – Ferrans & Powers index; HADS=hospital anxiety and depression scale; MOSAS=Maugeri obstructive sleep apnea syndrome; MSLT – multiple sleep latency test; MWT – maintenance of wakefulness test; 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; PSQI – pittsburgh sleep quality index; QD-R=depression questionnaire-reduced form; QoL – quality of life; QSQ=Quebec sleep questionnaire; SAQLI=sleep apnea quality of life index; SBPS=spouse bed partner survey; SCL-90 – Hopkins symptom checklist 90; SF-12 – short-form 12; SF-36 – short form 36; SIP – sickness impact profile; SNORE25 – symptoms of nocturnal obstruction and related events-25; SOS=snore outcomes survey; SQS=sleep quality scale; STAI-X3=state anxiety inventory-reduced form; SWIFT=sleepiness-wakefulness inability and fatigue test; ToDSS=time of day sleepiness scale; VAWS=visual analogical well-being scale.

Correlations: weak: r<0.3; weak to moderate: 0.2<r<0.5; moderate: 0.3<r<0.7; moderate to strong: 0.5<r<0.8; strong: r>0.7

a. Polysomnographic measures are determined during a sleep study to measure the severity of OSA. The measures which are reported vary per study.

Appendix 3: hypotheses for responsiveness

PROM under study	Comparator instrument	Expected correlation (direction of correlation)	Details
ESS	None	-	No data was suitable for making hypotheses about the validity of the change score.
EQ-5D	SF-36 PCS	Moderate (+)	The EQ-5D measures generic quality of life. The correlation with the SF-36 physical component score (PCS) is expected to be strongest, because the domains of the EQ-5D are all (mostly) focused on physical health. The correlation with the SF-36 mental component score (MCS) is expected to be weaker. We consider the ESS and the PGI to be an unsuitable comparator instruments for the EQ-5D because the constructs are too different.
	SF-36 MCS	Moderate (+)	
	Unsuitable comparator instruments		
	PGI ESS		
FLP	None	N/A	No data was suitable for making hypotheses about the validity of the change score.
FOSQ	None	N/A	The FOSQ measures the impact of sleepiness on functional status. Change scores of patients with OSA who used CPAP for more than four hours per night, and patients with OSA who used CPAP less than 4 hours per night, are compared. We expect a larger increase in score for the total FOSQ score and all the FOSQ domains for those who used more than 4 hours of CPAP per night compared to those who used less than 4 hours of CPAP per night.
FOSQ-10	FOSQ	Strong (+)	The FOSQ-10 measures the impact of sleepiness on functional status. Since the FOSQ-10 is a shorter version of the FOSQ but aims to measure the same thing, we expect a strong correlation.
OSAPOSI	"overall assessment of response to treatment"	Strong (?)	The OSAPOSI measures quality of life specifically for patients with OSA. The correlation between the OSAPOSI's change scores and the "overall assessment of response to treatment" (much improved, somewhat improved, no change, somewhat worse, much worse) is expected to be strong as the OSAPOSI was developed as an outcome measure for a clinical trial and should be able to assess response to treatment. We cannot predict the direction of the correlation as no information is given on the scoring of the overall assessment of response to treatment.
PGI	ESS	Moderate (-)	The domains in the PGI are determined by the patients who fill out this PROM. The patient population of this study is selected for having OSA and its related subjective complaints (like sleepiness), so this is likely to be reflected in the PGI. Therefore a moderate correlation is expected with the ESS, and a weak to moderate correlation with the SF-36 mental score. The physical component score of the SF-36 and the EQ-5D (+EQ-T) we consider unsuitable comparator instruments because their constructs are too different from the PGI, as neither of them address OSA-specific complaints.
	SF-36 mental health component	Weak to moderate (+)	
	Unsuitable comparator instruments		
	SF-36 physical health component EQ-5D (+EQ-T)		

QSQ	ESS FOSQ SF-36 SCL-90 BDI	Weak to moderate (-) Weak to moderate (+) Weak to moderate (+) Weak to moderate (-) Weak to moderate (-)	The QSQ measures quality of life specifically for patients with OSA. Domains of the QSQ are compared to (domains of) other instruments that the authors of the paper deem relevant. They hypothesize that all correlations will be between 0.4 and 0.7.[2] To add some more detail to these comparisons, we added some additional hypotheses based on the similarities and differences of the compared domains. We expect that the QSQ domain "daytime sleepiness" will correlate strongest with the ESS, the "vitality" domain of the SF-36, and all domains of the FOSQ, as they have related constructs. A weaker but still moderate correlation is expected for the "role-physical" domain of the SF-36 and even weaker for the "physical functioning" domain of the SF-36. The QSQ domain "diurnal symptoms" is expected to correlate strongest with the "vitality" domain of the SF-36, the "general productivity" and "activity level" domains of the FOSQ. Again, a weaker correlation is expected with the "role-physical" domain of the SF-36, and the weakest correlation with the "physical functioning" domain of the SF-36. The correlation of the QSQ domain "social interactions" with all three comparison domains of the FOSQ, SF-36 and SCL-90 are expected to be on the weaker end of moderate as they mostly contain items with a different focus than those of the QSQ domain. We made no additional hypotheses for the "nocturnal symptoms" and "emotions" domains of the QSQ.
SF-36 (mental health (MH) and physical health (PH) components)	EQ-5D (+EQ-T) PGI ESS	Moderate (+) Weak to moderate (MH) (+) Weak to moderate (MH) (+)	The SF-36 measures the mental and physical aspects of quality of life. The physical and mental component of the SF-36 are expected to correlate moderately with the EQ-5D and the EQ-T. The mental component of the SF-36 is expected to have a weak to moderate correlation with the ESS and PGI because the constructs are less closely related. Both components are expected to correlate weakly to moderately with the PGI and ESS as they focus on more specific complaints.
	Unsuitable comparator instruments ESS (PH) PGI (PH)		Known-groups approach Additionally, change scores were compared of snorers, patients with mild OSA who received no treatment, and patients who were on CPAP at baseline and at follow-up, with patients with OSA severe enough to warrant treatment who received CPAP. We expect larger positive change scores in patients with severe OSA who got treatment than in the other groups. For the "vitality" domain this difference is largest. No effect is expected for the SF-36 domains "physical functioning", "role physical" or "bodily pain".
SF-12	None1	N/A	No data was suitable for making hypotheses about the validity of the change score.
SNORE25	Polysomnographic measuresa	Moderate (+)	The SNORE25 measures quality of life specifically for patients with OSA. The objective severity of OSA, as measured by polysomnographic measures, will likely influence quality of life, but it is unclear to what extent and we consider it an unsuitable comparator.

Appendix 3 *continued*			
SOS	Nonea	N/A	No data was suitable for making hypotheses about the validity of the change score.
SAQLI	SF-36 (overall score and domains)	Weak to Moderate/ Moderate (+)	The SAQLI measures quality of life specifically for patients with OSA. Flemons et al.[3] has constructed hypotheses of exact correlation strengths between the total SAQLI change score and the change score of several other instruments. Since it would be very unlikely if their exact predictions were true, we have translated these predictions into the categories that we have used for our own hypotheses. We excluded hypotheses for correlations with constructs that we consider non-similar. A strong correlation is expected between the SAQLI total score and the global QoL rating. A moderate correlation is expected between the SAQLI total scores and "vitality", "social functioning, "role emotional" and "mental health" domains of the SF-36, and the "health and functioning" domain of the F&P index. A weak to moderate correlation is expected with the "general health" domain of the SF-36 and the F&P index total score. A weak correlation is expected with the "psychological and spiritual" domain of the F&P index. The objective severity of OSA, as measured by polysomnographic measures, will likely influence quality of life, but it is unclear to what extent and we consider it an unsuitable comparator. We consider the domains of the SF-36 and the F&P index which are not mentioned above unsuitable comparator instruments because the constructs are too different. Known-groups approach Additionally, change scores of patients with OSA who used CPAP for more than four hours per night, and patients with OSA who used CPAP less than 4 hours per night, are compared. We expect a larger increase in score for the total SAQLI score and the SAQLI domains "daily functioning", "social functioning" and "symptoms" for those who used more than 4 hours of CPAP per night compared to those who used less than 4 hours of CPAP. For the domain "emotional functioning" we expect this effect to be less pronounced, as it is less directly influenced by OSA treatment.
	Global QoL rating	Strong (+)	
	F&P index	Moderate (+)	
	Unsuitable comparator instruments		
	Polysomnographic measures[a] Some SF-36 domains (see description) F&P total score and some domains (see description)		
SWIFT	None	N/A	No data was suitable for making hypotheses about the validity of the change score.
ToDSS	None	N/A	No data was suitable for making hypotheses about the validity of the change score.

VAWS	FOSQ	Moderate (+)	The VAWS measures well-being status with regard to the symptoms which were the motive of the consultation. The strongest moderate correlations are expected with FOSQ (because of the related construct) and EQ-T (because of the similar outcome scale). A weaker correlation with SF-36 "mental health" domain is expected due to only part overlap in the constructs, followed by the EQ-5D, which asks least specifically about complaints related to OSA. The SF-36 "physical health" domain we consider an unsuitable comparator instrument because its construct is too different.
	SF-36 (mental health component)	Moderate (+)	
	EQ-5D (+EQ-T)	Moderate/Weak to moderate (-)	
	Unsuitable comparator instruments		
	SF-36 (physical component)		

BDI – Beck depression inventory; CPAP=continuous positive airway pressure; EQ-5D – euroqol-5D; EQ-T – euroqol thermometer; ESS=Epworth sleepiness scale; FLP=functional limitations profile; FOSQ – functional outcomes of sleep questionnaire; F&P index – Ferrans & Powers index; OSA=obstructive sleep apnea; OSAPOSI=obstructive sleep apnea patient-oriented severity index; PGI=patient-generated index; PROM – patient-reported outcome measure; QoL – quality of life; QSQ=Quebec sleep questionnaire; SAQLI=sleep apnea quality of life index; SCL-90 – Hopkins symptom checklist; SF-12 – short form 12; SF-36 – short form 36; SNORE25 – symptoms of nocturnal obstruction and related events-25; SOS=snore outcomes survey; SWIFT=sleepiness-wakefulness inability and fatigue test; ToDSS=time of day sleepiness scale; VAWS=visual analogical well-being scale.

Correlations: weak: r<0.3; weak to moderate: 0.2<r<0.5; moderate: 0.3<r<0.7; moderate to strong: 0.5<r<0.8; strong: r>0.7

a. Polysomnographic measures are determined during a sleep study to measure the severity of OSA. The measures which are reported vary per study.

Appendix 4: pooling of convergent validity

PROM	Comparator instrument/ domain	Correlation 1	Correlation 2	Pooled score[a]	Hetero-geneity	Concurrent with hypothesis
ESS	MWT	-0.39[4]	-0.48[5]	-0.46	0%	Yes
SAQLI	ESS	-0.26[3]	-0.243[6]	-0.24	0%	No
SF-36 MHC	ESS	-0.33[7]	-0.51[8]	-0.42	30%	Yes
FOSQ general productivity	SF-36 vitality	0.479[6]	0.19[9]	na[b]	77%	Yes/no
FOSQ social outcomes	SF-36 vitality	0.321[6]	0.20[9]	0.31	0%	Yes
FOSQ activity level	SF-36 vitality	0.645[6]	0.32[9]	na[b]	87%	Yes/yes
FOSQ vigilance	SF-36 vitality	0.266[6]	0.14[9]	0.25	0%	No
FOSQ sexual intimacy	SF-36 vitality	0.362[6]	0.03[9]	na[b]	73%	Yes/no
FOSQ activity level	SF-36 social functioning	0.612[6]	0.27[9]	na[b]	87%	Yes/no
FOSQ general productivity	SF-36 social functioning	0.547[6]	0.20[9]	na[b]	85%	Yes/no

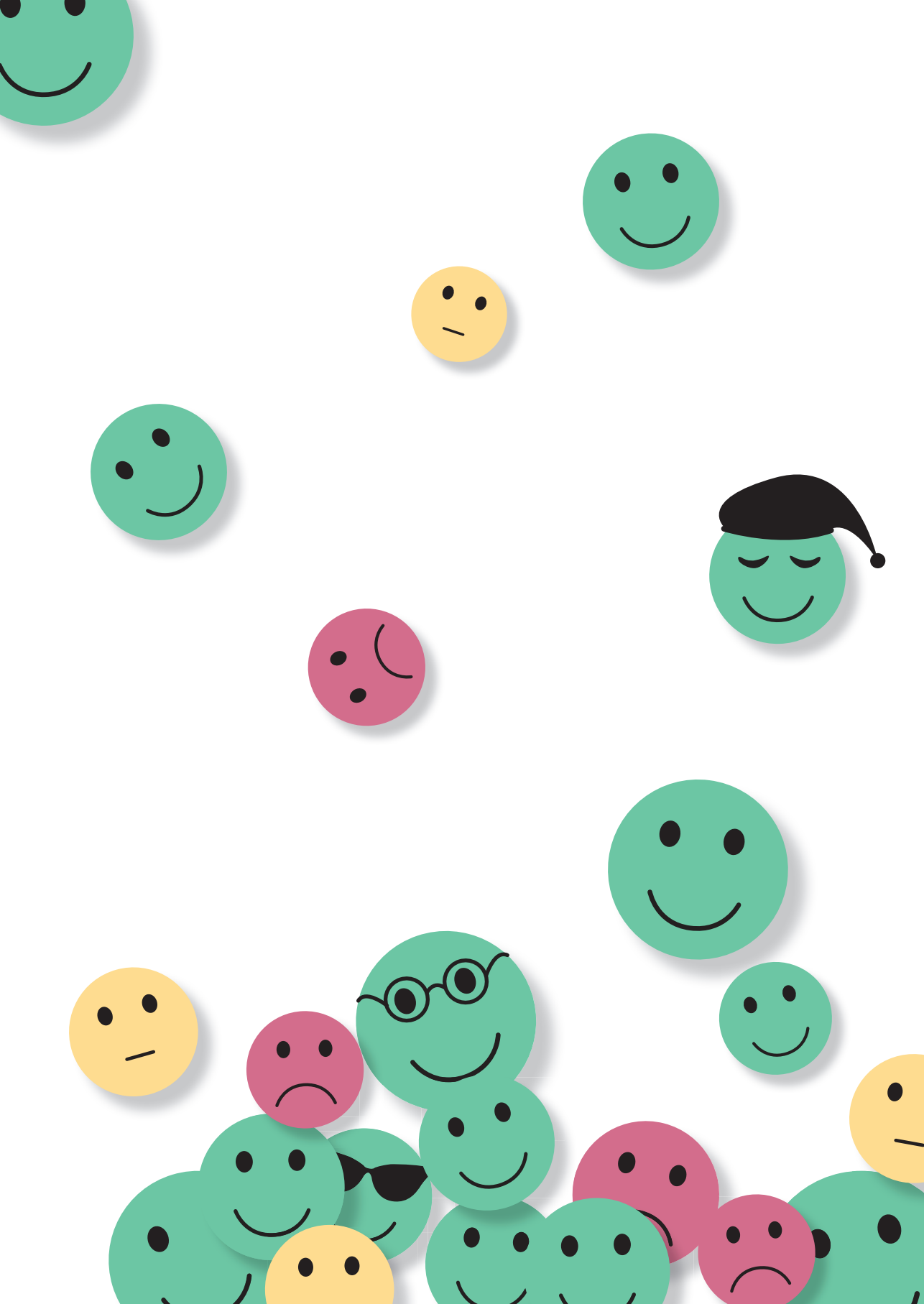
ESS=Epworth sleepiness scale; FOSQ=functional outcomes of sleep questionnaire; MWT=maintenance of wakefulness test; PROM=patient-reported outcome measure; SAQLI=sleep apnea quality of life index; SF-36 MHC=short form 36 mental health component

a. Fisher z-transformation and a fixed effects model was used for pooling of the correlations.

b. Correlations with a heterogeneity of >50% were not pooled. The original correlations are separately compared to the hypotheses in Appendix 2.

REFERENCES

1. Terwee, C.B., et al., *Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments.* Qual Life Res, 2009. **18**(8): p. 1115-23.
2. Lacasse, Y., M.P. Bureau, and F. Series, *A new standardised and self-administered quality of life questionnaire specific to obstructive sleep apnoea.* Thorax, 2004. **59**(6): p. 494-9.
3. Flemons, W.W. and M.A. Reimer, *Measurement properties of the calgary sleep apnea quality of life index.* Am J Respir Crit Care Med, 2002. **165**(2): p. 159-64.
4. Sangal, R.B., J.M. Sangal, and C. Belisle, *Subjective and objective indices of sleepiness (ESS and MWT) are not equally useful in patients with sleep apnea.* Clin Electroencephalogr, 1999. **30**(2): p. 73-5.
5. Kingshott, R.N., et al., *Does arousal frequency predict daytime function?* European Respiratory Journal, 1998. **12**(6): p. 1264-1270.
6. Billings, M.E., 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.
7. Weaver, E.M., V. Kapur, and B. Yueh, *Polysomnography vs self-reported measures in patients with sleep apnea.* Arch Otolaryngol Head Neck Surg, 2004. **130**(4): p. 453-8.
8. Bennett, L.S., et al., *Health status in obstructive sleep apnea: relationship with sleep fragmentation and daytine sleepiness, and effects of continuous positive airway pressure treatment.* Am J Respir Crit Care Med, 1999. **159**(6): p. 1884-90.
9. Weaver, T.E., et al., *An instrument to measure functional status outcomes for disorders of excessive sleepiness.* Sleep, 1997. **20**(10): p. 835-43.



CHAPTER 3

Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes

Inger L. Abma, Maroeska Rovers, Philip J. van der Wees

Published in *BMC Research Notes*, 9:226 (2016)

ABSTRACT

Purpose

Convergent validity is one type of validity that is commonly assessed for patient-reported outcome measures (PROMs). It is assessed by means of 'hypothesis testing': determining whether the scores of the instrument under study correlate with other instruments in the way that one would expect. Authors of systematic reviews on measurement properties for PROMs may encounter validation articles which do not state hypotheses by which convergent validity can be tested. The information in these articles can therefore not be readily used to determine the adequacy of convergent validity. We suggest that in these cases, reviewers construct their own hypotheses. However, constructing hypotheses and interpreting outcomes is not always straightforward, and we wish to aid reviewers based on our own recent experiences with a systematic review on measurement properties.

Recommendations

We have the following recommendations for authors of a systematic review on measurement properties who wish to construct hypotheses for convergent validity: take an active role in judging the suitability of the comparator instruments of validation articles; be transparent about which hypotheses were constructed, the underlying assumptions on which they are based, and whether they were constructed by the authors of the validation article or by the reviewer; discuss unmet hypotheses, especially if convergent validity is judged to be inadequate; and when synthesizing data, add up the results of all hypotheses for one instrument, rather than judging convergent validity per study.

1 INTRODUCTION

Questionnaires about patients' health and functioning filled out by the patient, also known as patient-reported outcome measures (PROMs), should be validated to ensure that they measure the topic ('construct') that they aim to measure (validity), and that they do this in a reliable way (reliability). There are several different aspects of validity and reliability that can be assessed to determine the quality of a PROM. The international Delphi panel of COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) reached consensus on a comprehensive terminology of these measurement properties, as well as on the content of the first user-friendly quality checklist for validation studies [1-3]. The COSMIN checklist and guidelines are frequently utilized: a search in PubMed for COSMIN shows 50 systematic reviews on measurement properties using the COSMIN checklist in 2015 alone.

One aspect of validity is construct validity, which is the degree to which the scores of a PROM are consistent with hypotheses, based on the assumption that the PROM validly measures the construct to be measured [4-7]. Convergent validity, a subtype of construct validity, verifies whether the scores of the instrument under study 'make sense' in relation to the scores of other, related instruments. Scores should correlate with scores of other instruments to the degree that one would expect. Assessing convergent validity is an iterative process: the more hypotheses are tested, the stronger the evidence towards the instrument being valid. Convergent validity is generally considered adequate if >75% of hypotheses are correct, or if a correlation with an instrument measuring the same construct is >0.50. The exact values of these cut-off points may be arbitrary, but they provide guidance when judging whether convergent validity is adequate. Furthermore, correlations with related constructs should be higher than with unrelated constructs [4, 8].

When performing a systematic review on measurement properties, assessing and summarizing the data for convergent validity is often less straightforward than for many other measurement properties. Authors of validation studies do not always construct hypotheses when studying convergent validity: many studies present only correlation sizes, without interpreting these or using them to test expectations. Based on the COSMIN guidelines, this data cannot be readily used in a systematic review. Therefore, the authors of a recent systematic review [9] decided to construct their own hypotheses for convergent validity. In our own recent systematic review on the measurement properties of PROMs for obstructive sleep apnea (OSA) [10], we followed their example.

We believe that constructing hypotheses for convergent validity should become more common in systematic reviews for measurement properties in which the included studies do not present their own hypotheses. However, there are certain issues that will arise when approaching hypothesis testing this way, which include: how to deal with unsuitable or low-quality comparator instruments; the different ways in which hypotheses can be constructed; interpreting the results and synthesizing the evidence, which are a general issues regardless of

the approach. These issues have not yet been discussed in the literature. The aim of this paper is to provide an overview of these issues regarding convergent validity, and to start a discussion on how they can best be handled in future systematic reviews. Additionally, we believe that the considerations of this paper will aid authors of future validation studies, who will be faced with many similar issues.

2 QUALITY OF COMPARATOR INSTRUMENTS

Ideally, it should be clear that comparator instruments validly and reliably measure what they should measure. In practice however, comparator instruments are often not extensively validated, or not validated in the target population. Furthermore, it is unclear when exactly a comparator instrument is 'valid' - there are no rules or suggestions about which measurement properties should be of sufficient quality for comparator instruments (and due to a sometimes limited availability of suitable comparator instruments, this may also not be desirable). The most practical approach for reviewers may be to exclude comparator instruments for insufficient quality only if there is no development or validation article available at all.

However, there is one situation in which the quality of an instrument or scale may clearly limit its value as comparator instrument for convergent validity: when the questions of a scale do not all tap into the same construct. Sometimes scales claim to measure a rather 'diffuse' topic, such as social functioning. In practice, the questions that comprise one 'social functioning' scale often differ greatly from the items of other similarly named scales, and one cannot necessarily assume their scores correlate to a great extent – which is problematic when trying to determine the validity of the instrument under study (for examples of this phenomenon, see Kemmler 1999 [11] or Lacasse 2004 [12]). It may be that 'social functioning' is simply not the right construct label for (one of) these scales, or not a precise enough description of the construct, or that the scales have different underlying theories about how to measure social functioning. Another possibility is that they are a collection of questions with different topics around the same general theme rather than one coherent construct. If factor analysis has been performed for the comparator scale, and/or if internal consistency of the scale has been determined, this can help identify scales for which this is the case. We would recommend to look at both the content of the scale and the available information on the measurement properties before deciding to disqualify a comparator scale or instrument due to problems with the coherency of the construct. In all cases, we would recommend to (briefly) discuss the quality of comparator instruments, as this may help put the results of convergent validity in perspective.

3 SUITABILITY OF COMPARATOR INSTRUMENTS

The construct of the comparator instruments is important for convergent validity: its construct should ideally have a clear relation with the construct under study. This clear relation is not always present for the comparator instruments used in validation studies. Correlation sizes may therefore be hard to predict. An example from our review is the relation between subjective sleepiness and the objective severity of sleep apnea, which is not straightforward [13-15]. If the results from a study disprove any constructed hypothesis, this would do more to illustrate the confusion around the relation between these two constructs, than to provide information about the validity of the instrument. We recommend excluding comparisons with these 'unsuitable' constructs from the evidence base.

Furthermore, sometimes comparator constructs are only vaguely related to the construct under study. An example from our review is the relation between sleepiness and quality of life. These constructs are likely somewhat related in patients that suffer from sleep apnea, a condition for which sleepiness is often the main complaint. However, hypotheses of low correlations for weakly related constructs are often correct, and reduce the impact of the hypotheses for more strongly related constructs – which is especially problematic in cases where the former outnumber the latter, and no clear rationale is provided for the choice of these weakly related comparator instruments or domains. We recommend using expected weak correlations only for the requirement that correlations with related constructs are higher than with unrelated constructs.

Sometimes two instruments are employed to validate each other. This is not ideal, as it is unclear which instrument is 'at fault' if a hypothesis is not met. However, since it can be quite hard to interpret results either way (see the section 'Interpreting outcomes'), reviewers may decide to include these studies and discuss unmet hypotheses in the context of the validation study in question.

4 CONSTRUCTING THE HYPOTHESES

COSMIN recommends constructing hypotheses for *relative correlation sizes* of the different comparator instruments. I.e. the correlation of the instrument of interest with instrument A is expected to be higher than its correlation with instrument B. However, the constructs of the comparator instruments may not always be suitable for making meaningful relative hypotheses. To be able to make hypotheses for each comparator instrument, it can be desirable to also construct hypotheses for the absolute magnitude of the correlations. In our review we put each comparator instrument in one of the following categories: either a weak (<0.3), weak to moderate (>0.2<0.4), moderate (>0.3<0.7), moderate to high (>0.6<0.8) or high correlation (>0.7). The overlap between these categories was on purpose, to allow more flexibility in hypotheses. For each correlation we also noted the expected direction of the correlation -

positive or negative. Note that we did not focus on the common requirement that convergent validity is adequate if the correlation with an instrument measuring the same construct is >0.50. We studied the instruments in detail, rather than relying only on the description of the comparator instruments, and when two instruments really measured the same construct we considered a more challenging hypothesis (correlation above >0.70) more adequate.

If an included validation study does use hypotheses to appraise convergent validity, these hypotheses can be integrated with those of reviewers. If the original hypotheses are stricter than as constructed by the reviewers, they can be adjusted. For example, in our review we adjusted a prediction of exactly 0.3 to fit within our 'weak to moderate' (>0.2<0.4) category.

5 INTERPRETING OUTCOMES

When a hypothesis is correct, this contributes to the evidence that the instrument under study measures what it is supposed to measure. However, when a hypothesis is wrong, this can have several causes: 1) the instrument does not measure what it is supposed to measure, 2) the comparator instrument does not measure what it is supposed to measure, or 3) the theory or the assumptions underlying the hypothesis are incorrect [5]. It is not always clear which of these possibilities is true in any given situation, though authors may have their own ideas about the most likely cause. Ideally, possible reasons why a hypothesis was not met are discussed by authors.

Hypotheses about correlations, especially when they measure different but related constructs, are to some extent a best educated guess. A different team of authors or reviewers will likely construct (slightly) different hypotheses, possibly leading to different conclusions. Therefore, we suggest a thorough reporting of the hypotheses that are tested.

6 EVIDENCE SYNTHESIS

Many systematic reviews about measurement properties report results by synthesis of the evidence. To determine the strength of the evidence for each measurement property, often the number of validation studies studying that measurement property is taken into account, as well as the quality of the studies [8, 16, 17]. The quality of the measurement properties themselves can have evidence that is positive, negative, indeterminate (when only studies of poor quality are available), or conflicting (when results of validation studies are mixed). While this approach makes sense for some measurement properties, for hypothesis testing, the number of studies may be less relevant than the number of hypotheses tested. For example: if there are two studies measuring convergent validity, and one study has only one hypothesis which was found to be inaccurate (negative evidence), and the other has three different hypotheses which are accurate (positive evidence), the scoring method would lead to a 'conflicting' overall score. However, 75% of hypotheses overall are accurate. As such, adding up the hypotheses of the

different studies would lead to a more sensible estimation of the convergent validity of an instrument. To incorporate the methodological quality of the different studies in the score one could assign more weight to the hypotheses of better studies, or one could simply decide that the studies all need to be of at least acceptable quality.

7 DISCUSSION

Constructing hypotheses for convergent validity in a systematic review requires effort, but is the only way to assess this measurement property if no hypotheses were previously constructed. This article has provided an overview of the issues that can arise in systematic reviews assessing measurement properties of PROMs. Our recommendations are summarized in Box 1. These may be useful for future reviewers and for authors of validation articles with regard to convergent validity as well as other measurement properties that are determined by means of hypothesis testing, such as known-groups validity and discriminant validity (both also subtypes of construct validity) and responsiveness.

The importance of hypothesis testing lies in its ability to help understand the construct the PROM measures. A PROM labeled with an inaccurate construct is a problem which may otherwise remain unrecognized as it does not necessarily affect other measurement properties. Inadequate construct validity leads to the question which construct the PROM *does* measure, and one will have to look again at the content of the questionnaire, and put this in the context of its comparator instruments. Depending on the situation, either the items of the PROM can be adapted, or it can be decided to re-label the construct the PROM aims to measure.

Establishing convergent validity is prone to several problems: its results depend to an important extent on the choice of comparator instruments and which, and how many, hypotheses are constructed. However, because 75% of hypotheses need to be accurate rather than all of them, a single inadequate comparator instrument or hypothesis will not immediately prohibit a positive judgment of convergent validity. Furthermore, if results are interpreted critically, we are convinced that an accurate judgment of convergent validity is possible.

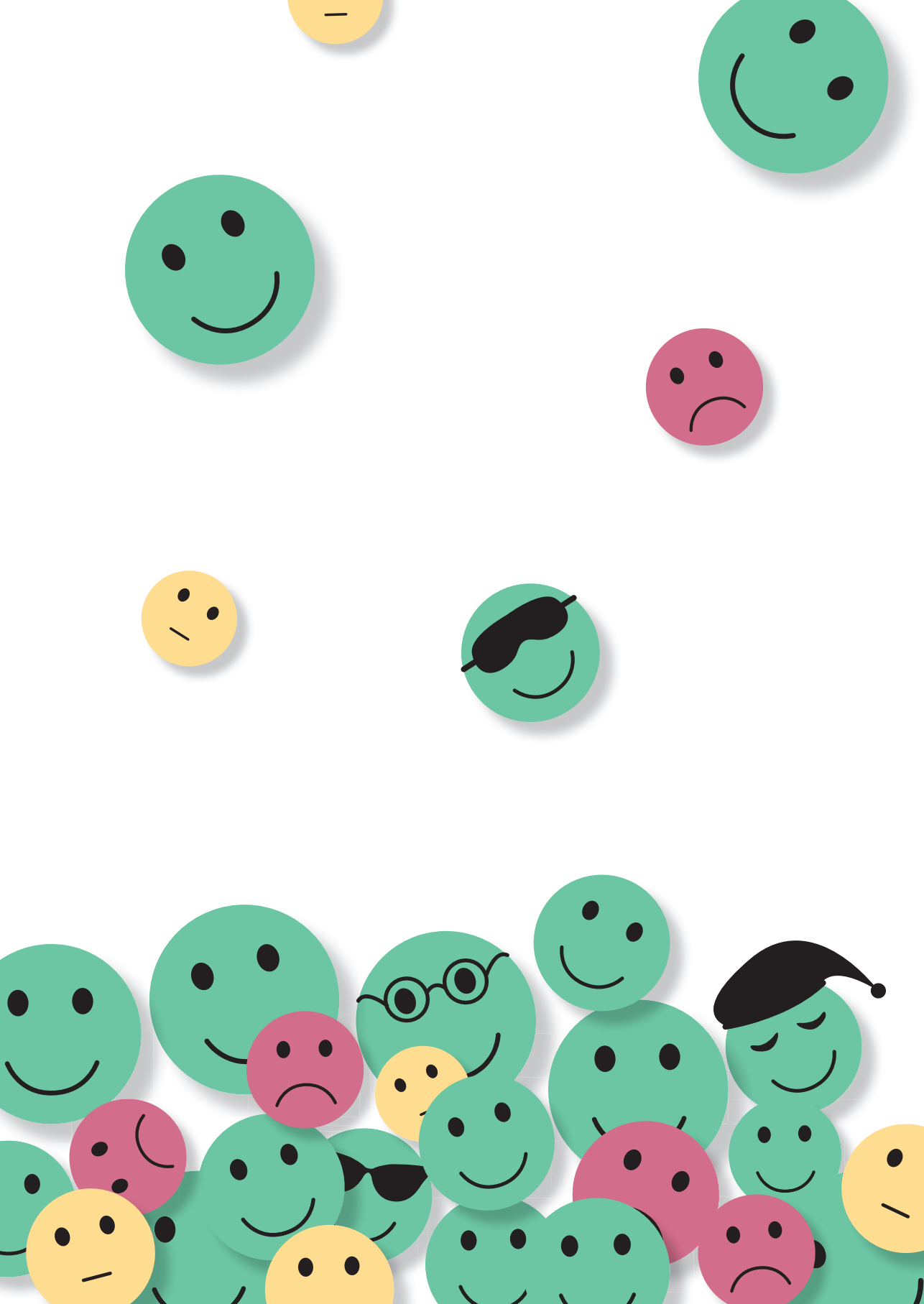
Box 1

Recommendations for reviewers

- Take an active role in judging the suitability of the comparator instruments:
  - Exclude comparator instruments which have an unclear relation with the construct under study, or which do not have a validation article.
  - Do not construct hypotheses for comparator instruments with expected weak correlations with the instrument under study, but use them as 'unrelated constructs' for the requirement that correlations with related constructs are higher than those with unrelated constructs.
- Be transparent about the constructed hypotheses and their underlying assumptions, and about whether hypotheses were constructed by the reviewers or the authors of the validation study.
- Discuss unmet hypotheses in the light of the comparator instruments and their quality, especially if convergent validity is judged to be inadequate.
- For data synthesis: add up the results of all hypotheses for one instrument, rather than judging convergent validity per study.

REFERENCES

1. Mokkink, L.B., 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**(7): p. 737-45.
2. Mokkink, L.B., et al., *The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study.* Qual Life Res, 2010. **19**(4): p. 539-49.
3. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Qual Life Res, 2012. **21**(4): p. 651-7.
4. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
5. Streiner, D. and G. Norman, *Health measurement scales: a practical guide to their development and use.* 4th ed. 1995, Oxford, UK: Oxford University Press.
6. De Vet, H.C., et al., *Measurement in Medicine.* 2011, Cambridge, UK: Cambridge University Press.
7. Fayers, P.M. and D. Machin, *Quality of Life: Assessment, Analysis and Interpretation.* 2nd ed. 2000, Chichester, England: Wiley.
8. Schellingerhout, J.M., et al., *Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review.* Qual Life Res, 2012. **21**(4): p. 659-70.
9. Kendzerska, T.B., et al., *Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review.* Sleep Med Rev, 2014. **18**(4): p. 321-31.
10. Abma, I.L., et al., *Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): A systematic review.* Sleep Med Rev, 2015. **28**: p. 14-27.
11. Kemmler, G., et al., *Comparison of two quality-of-life instruments for cancer patients: the functional assessment of cancer therapy-general and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-C30.* J Clin Oncol, 1999. **17**(9): p. 2932-40.
12. Lacasse, Y., M.P. Bureau, and F. Series, *A new standardised and self-administered quality of life questionnaire specific to obstructive sleep apnoea.* Thorax, 2004. **59**(6): p. 494-9.
13. Macey, P.M., et al., *Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients.* PLoS One, 2010. **5**(4): p. e10211.
14. Tam, S., B.T. Woodson, and B. Rotenberg, *Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index.* Laryngoscope, 2014. **124**(1): p. 337-43.
15. Weaver, E.M., V. Kapur, and B. Yueh, *Polysomnography vs self-reported measures in patients with sleep apnea.* Arch Otolaryngol Head Neck Surg, 2004. **130**(4): p. 453-8.
16. Noben, C.Y., et al., *Quality appraisal of generic self-reported instruments measuring health-related productivity changes: a systematic review.* BMC Public Health, 2014. **14**: p. 115.
17. Wigham, S. and H. McConachie, *Systematic review of the properties of tools used to measure outcomes in anxiety intervention studies for children with autism spectrum disorders.* PLoS One, 2014. **9**(1): p. e85268.



CHAPTER 4

The development of a patient-reported outcome measure for patients with obstructive sleep apnea: the Patient-Reported Apnea Questionnaire (PRAQ)

Inger L. Abma, Maroeska Rovers, Marijke IJff, Bernard Hol,
Gert P. Westert, Philip J. van der Wees

Published in *Journal of Patient Reported Outcomes* 1:14 (2017)

ABSTRACT

Background

Obstructive sleep apnea (OSA) is a chronic condition that can have a wide range of consequences for a patient's health-related quality of life. Monitoring aspects of quality of life in clinical practice has the potential to improve the patient-centeredness of care for patients with OSA. The aim of this article is to describe the development of the Patient-Reported Apnea Questionnaire (PRAQ), a patient-reported outcome measure (PROM) that is designed for use in clinical practice on an individual patient level, as well as subsequent outcome measurement on an aggregate level.

Methods

We used the items of available PROMs for OSA to create a new PROM with focus on its applicability in clinical practice. We used a tailored development process to come to a selection of domains and items. Patients and healthcare professionals were intensively involved in the development of the PRAQ via membership of the development team, online surveys and focus groups, as well as two rounds of cognitive validation.

Results

This first version of the PRAQ consists of 43 items and 10 preliminary domains, and covers the aspects of quality of life that healthcare professionals and patients wish to discuss in clinical practice. Patients indicate that PRAQ is comprehensive and that its length is acceptable. Comprehensive patient involvement has ensured good content validity for the PRAQ.

Conclusions

This article shows how a PROM can be developed with a method tailored towards the PROM's applicability in clinical practice.

1 INTRODUCTION

OSA is a highly prevalent, chronic condition in which temporary obstructions of the upper airway cause breathing stops while asleep [1]. Arousal of the brain in patients with OSA results in continuation of breathing, which often goes unnoticed by the patient but can happen up to hundreds of times per night. This causes fragmented sleep and can result in severe sleepiness, fatigue and impaired mood during the day, which in turn can affect a patient's relationships, psychological well-being, cognitive functioning, and participation in work and other activities [2-7]. Furthermore, OSA has been recognized as an independent risk factor for hypertension, heart failure, and diabetes [8-10]. The general population prevalence of OSA has been reported to be 13% to 33% in men and 6% to 19% in women [11], but in practice OSA goes undiagnosed in many patients [3, 12, 13].

The wide range of consequences and the chronic nature of OSA make focus on health-related quality of life (HRQoL) during the care process highly relevant. HRQoL is quality of life relative to one's disease status [14] and has been captured in several models [14-16]. Patient-reported outcome measures (PROMs) are questionnaires which are filled out by patients with the aim of measuring symptoms, daily functioning or HRQol. Most of the currently existing PROMs were developed for research purposes, to measure the impact of interventions on perceived health in clinical trials [17]. In recent years, the use of these existing PROMs has also expanded to areas closer to daily clinical practice [18-20]. There, individual PROM scores are used for the detection of problems with HRQoL, monitoring a patient's response to treatment, and to improve patient-centeredness of care by directing more attention to a patient's quality of life during consultations with healthcare professionals [21]. Furthermore, PROMs can be used as outcome measures to assess the quality of treatments or providers [22]. The integrated use of PROMs for these different purposes, which includes PROM measurements at both intake and during follow-up, is expected to stimulate meaningful use in clinical practice and quality improvement [23].

A recently published systematic review [24] identified three available PROMs that were developed specifically for and with patients with OSA, and which aim to measure quality of life. However, the focus during their development was only on outcome measurement. Furthermore, because of either practical reasons (the PROM has to be administered by an interviewer) or content reasons (omission of important aspects of quality of life, and unclear phrasing of some of the items) these PROMs did not seem suitable for use in clinical practice. Therefore, we decided to develop a new PROM that covers the topics that patients and clinicians find relevant to discuss with regard to apnea-related quality of life, and which is also suitable for outcome measurement.

The aim of this article is to describe the development of a new PROM, the Patient-Reported Apnea Questionnaire (PRAQ), that measures the different aspects of OSA-related quality of life. This PROM can help focus clinical practice on the HRQoL of an individual patient, and can subsequently be used as an outcome measure for quality assessment.

2 METHODS

In developing the new PROM we used a set of steps that would ensure thorough patient and clinician involvement. These steps follow the general PROM development process as described in the literature [25-27]: item generation based on patient interviews or focus groups, selecting the items, developing scales and scoring method, and pilot testing the items (cognitive validation). Our approach to item generation phase was different from that described in the literature, as we pooled the items of existing PROMs rather than generating items ourselves. However, the item generation of these PROMs was based on patient input [28-30]. Additionally, during the item selection process of the PRAQ, we also gathered information specifically on the suitability of the domains and items for a PROM which will be used in clinical practice. We undertook the following steps:

1) forming a working group with different stakeholders;
2) creating a preliminary pool of items from existing PROMs and sorting these items into preliminary domains based on the topics of the items;
3) using a patient survey and healthcare professional survey to gather input for domain and item selection;
4) selecting domains and items with the working group;
5) discuss and adapt this selection in patient focus groups;
6) performing two phases of cognitive validation.

Each of these steps is explained in further detail in the following paragraphs. The definite sorting of the PRAQ items into domains with the help of psychometric methods will be conducted after a follow-up study and is outside the scope of this article.

2.1 Forming a working group

A working group was formed consisting of two researchers (IA and PW), a board member (MI) of the patient organization for OSA in The Netherlands (ApneuVereniging), and a pulmonologist specialized in OSA (BH), based at the Albert Schweitzer Hospital, The Netherlands. The working group made the necessary decisions for the PROM development throughout the development process, based on the input from patients and healthcare professionals whenever possible.

2.2 Creating preliminary pool of items

Three available PROMs which were previously developed for patients with OSA used patient input for the creation and/or selection of items [24]: the Sleep Apnea Quality of Life Index (SAQLI) [28], the Quebec Sleep Questionnaire (QSQ) [29, 31], and the Maugeri Obstructive Sleep Apnea Syndrom (MOSAS) questionnaire [30]. In the opinion of the working group, the QSQ and the MOSAS questionnaire appear to miss some important topics, e.g. items about emotions or symptoms, respectively. Furthermore, the phrasing of some items was deemed suboptimal. The SAQLI appears unfeasible for use in clinical practice because it is interviewer-administered,

but it does cover a very broad range of topics identified by patients in its development phase. Therefore, the working group decided to create a pool of items consisting of the items of these three PROMs, and use these items to create a new PROM which covers all relevant issues and which also suits our different purposes. We decided to use a 7-point Likert scale similar to that used in the QSQ and SAQLI. 7 response options have shown to be more reliable than 5 response options, possibly because raters do not like to choose the two most extreme response options of a scale [27]. After discussion in the working group we also decided to keep the 4-week recall period, because patients with OSA generally struggle with symptoms over longer periods of time and our patient representative indicated that shortening this period may feel too restrictive for patients. 4 weeks is the maximum recall period that is recommended for this type of questionnaire [32], is suited to the effects of therapy on a chronic illness [33] and is also used in well-known PROMs such as the SF-36 [34].

The three PROMs were each translated into Dutch by two translators who are native Dutch speakers. The working group selected the translation considered optimal for each item. We did not perform a backwards translation because we did not aim to adhere to the exact phrasing of the items: we only wanted to keep the topics the same. The working group and particularly the patient representative paid specific attention to whether the translated items and topics made sense the context of measuring quality of life for patients with OSA, to ascertain that the translators had not made misinterpretations. Furthermore, the working group made sure that all items were suitable for patients that were suspected of having OSA, as well as patients already diagnosed with or treated for OSA, and that items were suitable to potentially measure change over time.

All items of these three PROMs together formed our pool of items. When items from different PROMs were highly similar in both phrasing and topic, only one of the items was kept in our item pool. The working group then grouped the items into preliminary domains according to their topic, keeping in mind the conceptual model of health-related quality of life developed by Wilson and Cleary [16], separating the items on symptoms from those on functional status.

2.3 Gathering information for item selection: patient and healthcare professional survey

An online patient survey was distributed to gain input for item and domain selection, covering how important the different items are for patients with OSA (on a scale of 1 to 9); whether any items or domains are missing; and which topics patients would like to discuss with an OSA physician or nurse. Patients were also asked to comment on the phrasing of the items and to indicate if they found any items hard to understand or confusing. The survey was sent out to patients with OSA and partners of patients with OSA who are volunteers of the Dutch patient organization for OSA. These volunteers have encountered many patients with OSA in their volunteer work, and were asked to base their importance ratings for the individual items on their expertise based on this broad experience.

An online survey for healthcare professionals was set up to gain the following information per domain: to what extent respondents would want to know if their patients had these kind of problems; to what extent they thought treatment for OSA would reduce these problems; and to what extent they considered themselves at least partially responsible for helping to solve these problems for their patients – which includes the option of referring patients to another healthcare professional, such as a psychologist. Furthermore, the respondents were asked if they thought any domains were missing.

2.4 Preliminary domain and item selection

The working group selected the domains and subsequently the items of the PRAQ based on the surveys. We considered a domain relevant if more than 50% of both patients and healthcare professionals answered positively on the questions regarding whether a patient would want to discuss this domain with their healthcare provider, or the other way around, i.e. whether a healthcare provider was interested to learn more on this domain topic form the patient (a score of 7, 8 or 9 was considered positive). If this criterion was not met, the domain was up for discussion in the working group and patient focus groups.

As a next step, the working group selected the items within the selected domains. We excluded all items considered important by less than 50% of patients, items considered important by 50 - 70% of patients were up for discussion in the working group. Additionally, items were adjusted and potentially included if there were specific comments explaining why the score of item was low, such as issues around comprehensibility of the item. Patient comments were also used to identify items which were considered highly similar, and we discarded those with the lowest importance score.

2.5 Discussing the preliminary item selection: patient focus groups

After the preliminary selection of items by the working group, two patient focus groups (n=9 and n=5 participants, with at least two women in each) were held to discuss the results and the choices of the working group, and to reaffirm the relevance of the items for patients with OSA. Participants were volunteers of the OSA patient organization who had completed the survey.

2.6 Cognitive validation

Two phases of cognitive validation [35, 36] were carried out, each involving six patients with OSA or suspected OSA attending a consultation at a sleep centre, and if present, their partners. Ages of the patients ranged from 42 to 74, and highest education level ranged from primary school to undergraduate college. Half of the included patients were women. For one of the patients, Turkish was their first language.

The aim of the cognitive validation was to check whether all selected items were understood by patients as intended, and whether the answering options were complete and made sense. All patients were asked to think aloud while completing the PROM, and were

asked additional questions (probing) about their interpretation of the items. Items that were unclear were either removed or adjusted. Subsequently, a second phase of cognitive validation was carried out with the adjusted PROM.

3 RESULTS

The development of the PRAQ is summarized in the flow chart in Figure 1.

3.1 Creating preliminary pool of items

Our preliminary pool of items consisted of 63 items, which the working group sorted into 10 preliminary domains: symptoms at night, sleepiness, tiredness, memory & concentration, unsafe situations, concerns about health, daily functioning, direct effect of apnea on others (e.g. bothering others due to snoring), social interactions (with sexuality as a subtopic), and emotions.

3.2 Patient and healthcare professional surveys

The patient survey was sent out to 85 volunteers of the Dutch OSA patient organization, of which 35 people completed the survey (41%). The characteristics of the respondents can be found in Table 1.

Most of the individual items were considered important (7 or higher on a scale of 1-9) by a majority of respondents (70-90%). The items in the 'social interactions' domain had generally lower scores than the items in other domains, with a range of 29-71% of patients regarding them as important. Within this domain, the question about sexuality was considered important by the most respondents.

There was a general desire to be able to discuss the ten domains in a consultation with an OSA healthcare professional. Looking at the percentage of patients that scored their desire to discuss a certain domain with at least 7, the highest scores were for the domains 'daily functioning' (88%) and 'symptoms at night' (87%). The lowest scores were for direct effect of apnea on others (65%), social interactions (60%) and sexuality (55%).

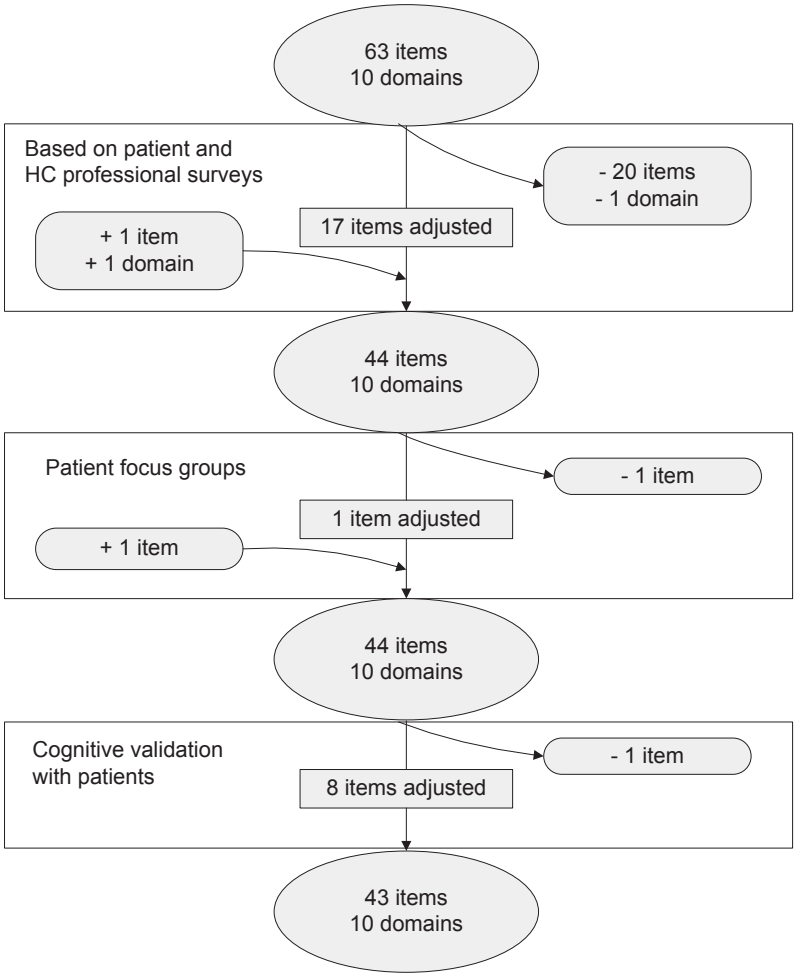


Figure 1 Flow chart of the PRAQ development process
HC=healthcare professional

Table 1 Patient survey respondent characteristics

Total nr of respondents	35
Current patients with OSA	30
Former patients with OSA	3
Partners	2
Gender	29% female
Median age category	60-69
Patients only (n=30)	
Treatment by CPAP/BPAP	100%
CPAP with additional MRA or operation	13%

BPAP=bilevel positive airway pressure, CPAP=continuous positive airway pressure, MRA=mandibular repositioning appliance, OSA=obstructive sleep apnea

The healthcare professional survey was sent out to 55 OSA professionals of whom 30 completed the survey (55%). The characteristics of the respondents can be found in Table 2.

For each of the domains, the majority of healthcare professionals indicated that they would want to know if their patients had problems with this topic (82%-100%). Most of them also felt at least partially responsible for helping to solve these problems, either by treating the patients themselves or referring the patient to another healthcare professional (72%-100%). Most healthcare professionals felt that treating their patients' OSA would improve complaints about sleepiness, symptoms at night, and the direct effect of apnea on others (89-96%). For the other domains, opinions were more diverse. Problems with sexuality were considered 'likely to improve' by the fewest survey respondents (46%).

Table 2 Healthcare professional survey respondent characteristics

Respondents (n)	30
Employed at n sleep centers	26
Median age category	50-59
Gender	53% female
Physician (n)	16
Pulmonologist (n)	10
Otolaryngologist (n)	2
Neurologist (n)	4
OSA nurse (n)	12
OSA nurse practitioner (n)	2

OSA=obstructive sleep apnea

3.3 Preliminary selection of topics and items

A majority of both healthcare professionals and patients wanted to be able to discuss each of the ten preliminary domains, so we considered all domains relevant for the PRAQ.

We decided to keep the item asking about anxiety in the domain 'emotions' despite its relatively low importance scores on the patient survey (60% of patients considered this an important item to add, versus over 70% for other items), because anxiety is more common in female patients with OSA [37, 38] and women were slightly underrepresented in the sample.

After item selection, only one item remained as part of the domain 'direct effect of OSA on others', so we decided to move this item to 'social interactions'.

Patients indicated that additional items about sleep problems should be part of the PRAQ, which was supported by the results of the healthcare professionals. Therefore, we added the domain 'quality of sleep', covering the suggested sleep problems. This resulted in a total of ten preliminary domains and 44 items.

3.4 Patient focus groups

During the patient focus groups, the preliminary selection of domains and items was discussed. One item was adapted, and one item was added to the domain 'emotions' about experiencing sudden, intense emotions. One other item for this domain ('how often did you feel you were unreasonable?') was removed after discussion in the group. The participants felt that a patient was unlikely to admit to being unreasonable, and that this type of emotion would be sufficiently covered by the items about feeling irritable and losing one's temper.

The number of answering categories was discussed, as several patients preferred to have ten answering categories rather than the proposed seven options, because this be similar to the scores of the Dutch version of a report card in school and thus would be more intuitively understandable. However, there was no consensus about this in the focus groups. We decided with the working group to maintain the 7-point Likert scale, for two reasons: scoring the PRAQ like a report card might give patients the idea that they are being judged on how well they are 'performing', which is not desirable; and as stated before, in the literature seven answering categories are often thought to be optimal [27].

Patients also commented on the recall period of the items: recalling symptoms of the past four weeks was generally seen as too short a time period. Newly diagnosed patients have often been experiencing symptoms for years, and choosing a long recall period, e.g. six months, would be more relevant for this particular group. However, follow-up appointments for CPAP users can be as early as four to six weeks after initiating treatment. Since we would like our PROM to be a useful addition to follow-up appointments as well as the intake appointment, using a recall period of more than four weeks is not desirable. To address the wishes of the patients, we therefore added an open text field at the end of each domain in which patients get the opportunity to describe past symptoms.

We also discussed the acceptable number of items for the new PROM. Patients of both focus groups felt that all remaining items were relevant and important, and that the length of the PRAQ was acceptable. The exception was the domain 'sleepiness' (containing eight items), which patients said could likely be further reduced without information loss. As there was no patient preference for which items should remain in the selection, we will perform the final selection of items for this domain with psychometric methods after a pilot study has taken place.

After the focus groups, there were 44 items left for the PRAQ.

3.5 Cognitive validation

Twenty-one patients were interviewed, aged between 42 and 74 years and with different education levels. There were several items in the PROM that were confusing to all or most of the interviewed patients, or that they understood in a way that was not intended, which were subsequently adjusted or removed. One example of a misunderstood item was 'Were you concerned about your safety or that of others in traffic or while operating machinery?'. Several patients indicated concern about their safety in traffic because they thought *other people* were often bad drivers. We adjusted this question to include 'due to your sleepiness', to shift the focus of this item to the patient's own potential problems due to OSA. During the second phase of the cognitive interviews, the meaning of the newly adjusted items as well as the other items in the current item selection was clear to the patients.

3.6 Final result

Based on the development process described in this paper, the current PRAQ comprises ten (preliminary) domains and 43 items, and takes approximately 15 minutes to fill out. The official English translation of this preliminary PRAQ can be found in Table 3. In a next stage of development, which involves a validation study assessing reliability, validity and responsiveness, final item selection and domain construction will take place.

Table 3 English translation of preliminary PRAQ[a]

Symptoms at night
During the past 4 weeks, did you have a problem with: 1. Snoring loudly? 2. Waking up frequently to urinate? 3. Waking up at night with the feeling that you are choking? 4. A feeling that you are sleeping restlessly? 5. Having a dry or painful mouth when you wake up? 6. Waking up in the morning with a headache?
Sleepiness
During the past 4 weeks, did you have a problem with: 7. Fighting to stay awake during the day? 8. Suddenly falling asleep? 9. Difficulty staying awake during a conversation? 10. Difficulty staying awake while watching something? (concert, movie, television) 11. Falling asleep at inappropriate times or places? 12. Difficulty staying awake while reading? 13. Fighting to stay awake when you are driving? 14. Did you feel like you needed to take a nap in the afternoon?
Tiredness
During the past 4 weeks, did you have a problem with: 15. Feeling very tired? 16. Lacking energy? 17. Still feeling tired when you wake up in the morning?
Daily activities
During the past 4 weeks: 18. How difficult was it for you to do your most important daily activity? (such as your job, studying, caring for the children, housework) 19. How often did you use all your energy to accomplish only your most important daily activity? (such as your job, studying, caring for the children, housework) 20. Did you feel you have a decreased performance with regard to your most important daily activity? (such as your job, studying, caring for the children, housework) 21. How much difficulty did you have finding energy for your hobbies? 22. How difficult was it for you to get your chores done?
Unsafe situations
During the past 4 weeks: 23. Did you have problems while driving a car due to sleepiness? 24. Were you concerned about your safety or that of others due to your sleepiness? (for example in traffic, or when operating machinery)
Memory and concentration
During the past 4 weeks: 25. Were you sometimes forgetful? 26. Did you sometimes have difficulty concentrating?
Quality of sleep
During the past 4 weeks, did you have a problem with: 27. Falling asleep when you go to bed at night? 28. Getting back to sleep after you woke up at night?

Emotions
During the past 4 weeks: 29. How often did you feel depressed or hopeless? 30. How often did you feel anxious? 31. How often did you lose your temper? 32. How often did you feel that you could not cope with everyday life? 33. How often did you feel irritated? 34. How often did you have a strong emotional reaction to everyday events?
Social interactions
During the past 4 weeks: 35. Did you sometimes feel upset because others were disturbed by your snoring? 36. Was it a problem for you that you sometimes had no energy or no desire to do things with your family or your friends? 37. Did you feel guilty towards your family or friends? 38. Did you feel upset because you argued frequently? 39. Did you sometimes experience problems in the relationship with your partner? 40. Did you feel upset because you could (maybe) not sleep in the same room as your partner? 41. Did you sometimes think up excuses because you were tired or sleepy? 42. Did you have a problem with unsatisfying and/or too little sexual activity? (by yourself or with another)
Health concerns
43. Were you concerned about other conditions that may be related to sleep apnea? (such as diabetes, high blood pressure, cardiovascular disease, being overweight)

a. The PRAQ was translated into English by an official translator who is a native English speaker, and by IA. The translator, IA and PW together reached consensus on the translation of each item. The English PRAQ was translated back into Dutch by another official native Dutch translator, and IA and PW used input from this translator to adapt the English version where needed.

4 DISCUSSION

In this article we describe the development of a quality of life PROM for patients with OSA, the Patient-Reported Apnea Questionnaire (PRAQ). The PRAQ was developed with the goal of serving as a useful addition to daily clinical practice at an individual patient level, to help focus more attention on quality of life, and subsequently as an outcome measure for quality assessment. We developed the PRAQ by using the pooled items of existing PROMs for patients with OSA and subsequently adapted and selected items with the input of physicians and patients. This resulted in a preliminary PRAQ with 10 domains and 43 items.

Item selection for the PRAQ is not yet entirely complete: within the domain of 'sleepiness', patients felt that the number of items could be reduced. However, they had no opinion on which of the eight items should be removed, because they were all relevant. Psychometric methods will be used to reduce the number of items in this domain using the data of a pilot study on this preliminary version of the PRAQ.

Next to being used on an individual patient level in clinical practice, our aim for the PRAQ was to be able to use its outcomes for quality assessment at an aggregate level. One important PROM measurement quality specifically for quality assessment is that a PROM must be *responsive*, i.e. that it is able to measure changes in a patient's condition over time. We took this into account in the development of the PRAQ by asking healthcare professionals whether they expected that different aspects of quality of life, as covered by the preliminary domains, would improve after treatment for OSA. The actual responsiveness of the PRAQ will be assessed in a validation study.

When developing a PROM that can be used both on an individual patient level in daily clinical practice and as an outcome measure for quality, it is important to find a balance between the wishes of patients and the requirements for creating a feasible outcome measure. For example, the patients wished to communicate symptoms from as far as six months ago, which is not a feasible recall period for a a quality of life PROM [27, 32]. We believe that the solution the working group devised together with the patients – offering patients an open text field for each domain, which can be used in clinical practice, if not for the scoring of the PRAQ – is a reasonable compromise in this case.

Patient input is very important during the PROM development process [39, 40]. For our patient input, we made use of the knowledge and experience of volunteers of the OSA patient organization in The Netherlands. Such volunteers are a relatively engaged population, and might therefore differ slightly from regular patients with OSA, but we do believe that they were able to give an accurate representation of what is important to this patient group. Furthermore, because we also used the input of healthcare professional and available literature, we do not believe that any important domains are missing.

To develop the PRAQ we used the items of the SAQLI, QSQ, and the MOSAS questionnaire [28, 30, 31]. Even though the PRAQ contains many items that are similar to the items of these PROMs, it also differs from them substantially. Compared to the SAQLI, the PRAQ is shorter and easier to understand, allowing patients to complete the PROM without an interviewer. The PRAQ is more elaborate than the QSQ in its emotions and social functioning domains. Furthermore, quite a few items of the QSQ and MOSAS questionnaire were seen as unclear or less relevant by the patients with OSA in this study and were therefore not added to the PRAQ. Furthermore, the preliminary PRAQ has its items grouped into more domains than the other PROMs, because we split up the symptoms that patients can experience during the day in more separate domains than the other PROMs. We chose this approach because we wanted to create domains of which the scores are more easily interpretable by healthcare professionals. A future factor analysis, as part of the validation of the PRAQ, will have to show whether the way items are currently grouped into domains makes sense psychometrically. Based on the validation study, the items of the PRAQ will then be sorted into their final domains.

The method used for developing this PROM, which includes forming an item pool out of the individual items of existing PROMs, can be employed by others as is, or can be adjusted

to fit the needs of a specific situation. The thoroughness of prior research and the number of PROMs and items available will have to be taken into account when deciding on the exact approach of the development process. If it is suspected that currently available PROMs do not cover all aspects which are important to the patient population, additional research may be warranted to expand the item pool, for example in the form of patient interviews.

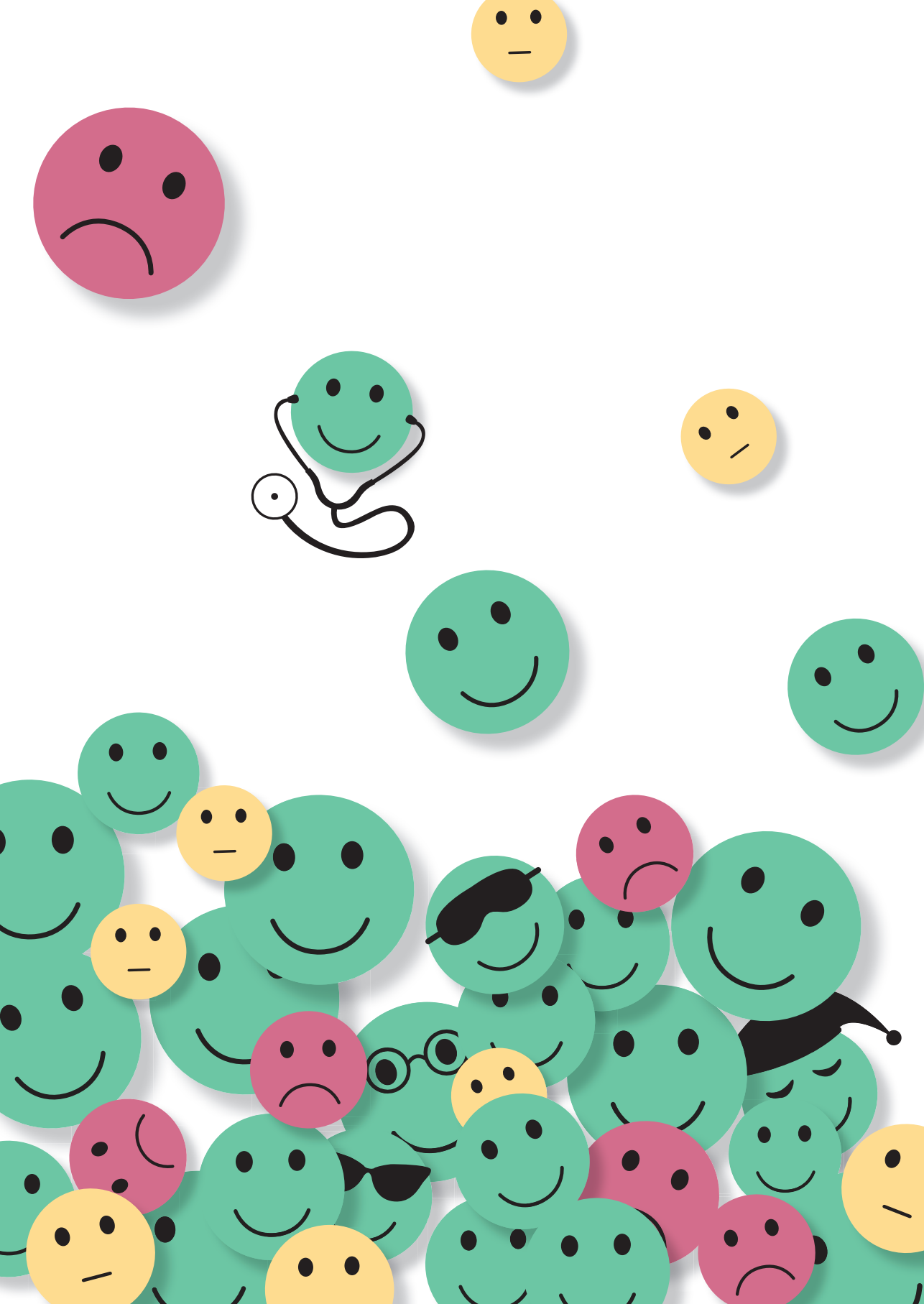
Future research

Our next step will be to perform a pilot study to finalize item selection and the formation of domains for the PRAQ. We subsequently study the measurement properties of the PRAQ, to estimate its suitability for measuring outcomes for individual patients and for quality assessment at an aggregate level. We will also develop a digital tool which summarizes the results of the PRAQ in a patient-friendly manner for use during consultations.

REFERENCES

1. Senaratna, C.V., et al., *Prevalence of obstructive sleep apnea in the general population: A systematic review.* Sleep Med Rev, 2016.
2. O'Donoghue, N. and E. McKay, *Exploring the impact of sleep apnoea on daily life and occupational engagement.* Br J Occup Ther, 2012. **75**(11): p. 609-516.
3. Rodgers, B., *Breaking through limbo: experiences of adults living with obstructive sleep apnea.* Behav Sleep Med, 2014. **12**(3): p. 183-97.
4. Reishtein, J.L., et al., *Sleepiness and relationships in obstructive sleep apnea.* Issues Ment Health Nurs, 2006. **27**(3): p. 319-30.
5. Bjornsdottir, E., et al., *The Prevalence of Depression among Untreated Obstructive Sleep Apnea Patients Using a Standardized Psychiatric Interview.* J Clin Sleep Med, 2016. **12**(1): p. 105-12.
6. Gupta, M.A., F.C. Simpson, and D.C. Lyons, *The effect of treating obstructive sleep apnea with positive airway pressure on depression and other subjective symptoms: A systematic review and meta-analysis.* Sleep Med Rev, 2016. **28**: p. 55-68.
7. Mokhlesi, B., S.A. Ham, and D. Gozal, *The effect of sex and age on the comorbidity burden of OSA: an observational analysis from a large nationwide US health claims database.* Eur Respir J, 2016. **47**(4): p. 1162-9.
8. Bradley, T.D. and J.S. Floras, *Obstructive sleep apnoea and its cardiovascular consequences.* Lancet, 2009. **373**(9657): p. 82-93.
9. Chan, A.S., C.L. Phillips, and P.A. Cistulli, *Obstructive sleep apnoea--an update.* Intern Med J, 2010. **40**(2): p. 102-6.
10. Young, T., P.E. Peppard, and D.J. Gottlieb, *Epidemiology of obstructive sleep apnea: a population health perspective.* Am J Respir Crit Care Med, 2002. **165**(9): p. 1217-39.
11. Senaratna, C.V., et al., *Prevalence of obstructive sleep apnea in the general population: A systematic review.* Sleep Med Rev, 2017. **34**: p. 70-81.
12. Finkel, K.J., et al., *Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center.* Sleep Med, 2009. **10**(7): p. 753-8.
13. Appleton, S.L., et al., *Undiagnosed obstructive sleep apnea is independently associated with reductions in quality of life in middle-aged, but not elderly men of a population cohort.* Sleep Breath, 2015. **19**(4): p. 1309-16.
14. Bakas, T., et al., *Systematic review of health-related quality of life models.* Health Qual Life Outcomes, 2012. **10**: p. 134.
15. Ferrans, C.E., et al., *Conceptual model of health-related quality of life.* J Nurs Scholarsh, 2005. **37**(4): p. 336-42.
16. Wilson, I.B. and P.D. Cleary, *Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes.* JAMA, 1995. **273**(1): p. 59-65.
17. Garratt, A., et al., *Quality of life measurement: bibliographic study of patient assessed health outcome measures.* BMJ, 2002. **324**(7351): p. 1417.
18. Guyatt, G.H., et al., *Exploration of the value of health-related quality-of-life information from clinical research and into clinical practice.* Mayo Clin Proc, 2007. **82**(10): p. 1229-39.
19. Valderas, J.M., J. Alonso, and G.H. Guyatt, *Measuring patient-reported outcomes: moving from clinical trials into clinical practice.* Med J Aust, 2008. **189**(2): p. 93-4.
20. Valderas, J.M., et al., *The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature.* Qual Life Res, 2008. **17**(2): p. 179-93.
21. Greenhalgh, J., A.F. Long, and R. Flynn, *The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory?* Soc Sci Med, 2005. **60**(4): p. 833-43.

22. Black, N., *Patient reported outcome measures could help transform healthcare.* BMJ, 2013. **346**: p. f167.
23. Van Der Wees, P.J., et al., *Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries.* Milbank Q, 2014. **92**(4): p. 754-75.
24. Abma, I.L., et al., *Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): A systematic review.* Sleep Med Rev, 2015. **28**: p. 14-27.
25. De Vet, H.C.W., et al., *Measurement in Medicine.* 2011, Cambridge, UK: Cambridge University Press.
26. Administration, F.a.D. (2009). *Guidance for industry - Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.* Silver Spring, Maryland, USA.
27. Streiner, D. and G. Norman, *Health measurement scales: a practical guide to their development and use.* 4th ed. 1995, Oxford, UK: Oxford University Press.
28. Flemons, W.W. and M.A. Reimer, *Development of a disease-specific health-related quality of life questionnaire for sleep apnea.* Am J Respir Crit Care Med, 1998. **158**(2): p. 494-503.
29. Lacasse, Y., C. Godbout, and F. Series, *Health-related quality of life in obstructive sleep apnoea.* Eur Respir J, 2002. **19**(3): p. 499-503.
30. Moroni, L., et al., *A new means of assessing the quality of life of patients with obstructive sleep apnea: the MOSAS questionnaire.* Sleep Med, 2011. **12**(10): p. 959-65.
31. Lacasse, Y., M.P. Bureau, and F. Series, *A new standardised and self-administered quality of life questionnaire specific to obstructive sleep apnoea.* Thorax, 2004. **59**(6): p. 494-9.
32. Stull, D.E., et al., *Optimal recall periods for patient-reported outcomes: challenges and potential solutions.* Curr Med Res Opin, 2009. **25**(4): p. 929-42.
33. Norquist, J.M., et al., *Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration.* Qual Life Res, 2012. **21**(6): p. 1013-20.
34. Ware, J.E., Jr. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.* Med Care, 1992. **30**(6): p. 473-83.
35. Farnik, M. and W.A. Pierzchala, *Instrument development and evaluation for patient-related outcomes assessments.* Patient Relat Outcome Meas, 2012. **3**: p. 1-7.
36. Willis, G., *Cognitive interviewing: a tool for improving questionnaire design.* 2005, Thousand Oaks, California: SAGE.
37. Kapsimalis, F. and M.H. Kryger, *Gender and obstructive sleep apnea syndrome, part 1: Clinical features.* Sleep, 2002. **25**(4): p. 412-9.
38. Sforza, E., et al., *Sex differences in obstructive sleep apnoea in an elderly French population.* Eur Respir J, 2011. **37**(5): p. 1137-43.
39. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
40. Paterson, C., *Seeking the patient's perspective: a qualitative assessment of EuroQol, COOP-WONCA charts and MYMOP.* Qual Life Res, 2004. **13**(5): p. 871-81.



CHAPTER 5

Instrument completion and validation of the Patient-Reported Apnea Questionnaire (PRAQ)

Inger L. Abma, Maroeska Rovers, Marijke IJff, Bernard Hol,
Gert P. Westert, Philip J. van der Wees

Published in *BMC Health and Quality of Life Outcomes* 16:158 (2018)

ABSTRACT

Background

We previously developed the preliminary version of the Patient-Reported Apnea Questionnaire (PRAQ), a questionnaire measuring health-related quality of life in patients with (suspected) obstructive sleep apnea (OSA). This questionnaire was developed for clinical practice, where it can potentially serve two goals: use on an individual patient level to improve patient care, and use on an aggregate level to measure outcomes for quality improvement at a sleep center. In this study we aim to finalize the PRAQ, make a subselection of items and domains specifically for outcome measurement, and assess the validity, reliability and responsiveness of the PRAQ.

Methods

Patients with suspected OSA were included and asked to complete the PRAQ and additional questionnaires one or more times. The collected data was used to perform the final item selection for clinical practice and for outcome measurement, create the domains for outcome measurement, and assess the measurement properties internal consistency, test-retest reliability, convergent validity and responsiveness.

Results

180 patients were included in the study. The final version of the PRAQ for use in clinical practice contains 40 items and 10 domains. A subselection of 33 items in 5 domains was selected for optimal outcome measurement with the PRAQ. The results for the outcome measurement domains were: Cronbach's α 0.88-0.92, ICC 0.81-0.88, and >75% of hypotheses correct for convergent validity and responsiveness.

Conclusions

The PRAQ shows good measurement properties in patients with (suspected) OSA.

1 INTRODUCTION

Patients with obstructive sleep apnea (OSA) experience breathing stops while asleep, causing symptoms during the day such as excessive sleepiness, tiredness, and irritability. This can have a large impact on daily functioning of patients, and often affects a patient's relationships and psychological wellbeing [1-3]. Furthermore, OSA is a known risk factor for comorbidities such as diabetes and heart failure [4-6], and is also associated with depression and anxiety [7-9]. Gaining an overview of the problems that OSA patients may experience, before, during, and in evaluating treatment, may be a challenge.

Patient-reported outcome measures (PROMs) are questionnaires for patients about symptoms or daily functioning. Most PROMs have been developed for use in clinical trials, but interest in their use in daily practice is growing [10, 11]. There, PROM scores can be used on an individual patient level to help bring patients' problems to the forefront during consultations and to monitor treatment response, or on an aggregate level across groups of patients for quality improvement purposes [12]. Use of a PROM on an individual patient level may be especially relevant when patient symptoms are multiple and complex. We therefore believe that it would be beneficial to employ a PROM for patients with OSA in daily clinical practice.

In a recently published article, we described the item generation and preliminary item selection of a PROM for patients with (suspected) OSA: the Patient-Reported Apnea Questionnaire (PRAQ) [13].We used the input of patients with OSA and healthcare professionals to select topics and items important for measuring quality of life for this patient group, which are also useful to discuss during an intake or follow-up consultation.

There are two ways in which the preliminary version of the PRAQ requires further development. First, the item reduction for the topic 'sleepiness' has not yet taken place. During the item selection process of the PRAQ, patients indicated that the number of items on the topic of sleepiness could be reduced. Since the patients had no preference for which items to exclude, we decided to perform the final item reduction after studying the psychometric properties of the items. Second, the factor structure of the PRAQ has not yet been studied, and we wanted to find the optimal way to group (a subset of) the items of the PRAQ into domains for the purpose of outcome measurement.

Our aim is for the PRAQ to be employed in the following way: patients complete all items of the PRAQ before their consultation, the results of which can be discussed with a healthcare professional; and the aggregate outcomes of groups of patients can then be studied by making use of a subset of the completed items. This is beneficial for patients, who get feedback from clinicians on their results; for physicians, who get a quick insight into their patients' main problems; and for sleep centers that wish to collect outcome data for quality improvement, because it ensures a steady stream of data due to integration in clinical practice.

In this article we describe the further development of the preliminary PRAQ. In addition, we aim to determine the reliability, validity and responsiveness of the PRAQ, with a focus on the domains that will be used for outcome measurement.

2 METHODS

2.1 Population & method of completion of the PROMs

Baseline measurement: Patients referred to the sleep center of the Albert Schweitzer Hospital in Dordrecht, The Netherlands for suspected OSA received an invitation by email to complete the PRAQ and additional PROMs, 2-3 weeks before their intake consultation. They were informed that the results of the PRAQ would be discussed during their intake consultation. A reminder was sent one week later if the PROMs were not yet completed at that time. Patients who had not completed the PRAQ at home were offered the option of completing the PRAQ at the sleep center before their consultation.

Retest measurement: In order to assess test-retest reliability, patients who had completed the baseline measurement at home were asked to complete it again immediately before their intake consultation, on a computer in a private area of the sleep center. Only patients who had completed the retest no less than 7, and no more than 21 days after the baseline measurement were included for assessment of test-retest reliability.

Follow-up measurement: A common measure to express the number of (partial) breathing stops experienced while asleep is the apnea-hypopnea index (AHI). We measured the responsiveness of the PRAQ in patients with an AHI≥15, which indicates moderate to severe sleep apnea [14], and who were prescribed continuous positive airway pressure (CPAP) after their intake consultation. CPAP is the preferred treatment for OSA [15]. If the patients were still using CPAP at the time of the first follow-up consultation (6-8 weeks after start of CPAP), they were included for responsiveness. They were asked to complete the PRAQ and the additional PROMs immediately before their follow-up consultation at the sleep center. Ideally, responsiveness should be determined in a patient group in which CPAP therapy is successful and therefore a substantial change is expected with regard to the patient's symptoms. CPAP therapy is generally considered successful when compliance is ≥4 hours nightly [16].

A secure website was used for the completion of the PROMs. For any of the measurements, patients who were unable or unwilling to use a computer were offered the option of completing a paper copy of the PROMs.

2.2 Final stage of PRAQ development

The development article of the preliminary PRAQ [13] shows how the initial 43 items were selected based on their relevance for clinical practice and were sorted into preliminary domains: symptoms at night (6 items), sleepiness (8 items), tiredness (3 items), daily activities (5 items), unsafe situations (2 items), memory and concentration (2 items), quality of sleep (2

items), emotions (6 items), social interactions (8 items), and health concerns (1 item) (Appendix 1). All items are scored on a 7-point Likert scale (higher scores indicate worse problems), and the average item scores in a domain form its domain score.

First, we performed item reduction on the sleepiness domain, as 8 items was deemed too much by patients. Then, we looked at how the PRAQ could be best used for outcome measurement. It is important that all items fit into a domain that is either 'coherent' in terms of clinical relevance, or (preferably) in terms of covariance matrix as determined by principal component analysis (PCA). Therefore, our aim was to identify which items of the PRAQ can be grouped into domains for outcome measurement after use of the results of the PRAQ for an individual patient. We describe below how we first reduced the number of items for the domain 'sleepiness', and then how from the remaining items a subset of items was selected for outcome measurement.

Item reduction of the sleepiness domain

During the development of the PRAQ, patients indicated that they felt that the number of items on the topic of sleepiness could be reduced. Because they had no preference for which items should be excluded, we took a statistical approach. We first looked for items with a high inter-item correlation (>0.9), indicating that one of these items can be removed without a substantial loss of information [17]. As a second step, we used exploratory factor analysis to identify potential items with lower factor loadings (<0.5), indicating that they do not cover the construct as well as the other items and are therefore more suitable for removal [17, 18].

Creating domains for the PRAQ-outcome

Two of our preliminary domains, 'symptoms at night' and 'social interactions', we considered formative rather than reflective domains: they do not aim to measure aspects of the same latent construct, but the items are grouped together based on clinical relevance. Grouping items in this way can be considered a 'clinimetric' approach, as opposed to a 'psychometric' approach which uses statistical methods to determine the dimensionality of a PROM [19]. We wanted to group these items together irrespective of their covariance matrix, because for content reasons we did not consider it desirable to combine these items with any of the other (potential) domains. Therefore, we excluded them from the PCA and kept these domains as they were.

We performed a PCA with oblique rotation (because correlations between the different patient complaints were expected) on the 26 items of the other preliminary domains. Items that did not load on any domain with a factor loading of at least 0.5 or that had a factor loading of >0.3 on more than one factor [17], were then one by one removed from the analysis, starting with those items that for content reasons did not seem to fit well with the items they were grouped with in the PCA. Additionally, since domains should ideally consist of at least three items, we used this as a requirement for the PRAQ-outcome domains [17]. The one-

dimensional domains that were identified by the analysis were added to the two clinimetric domains. Together, these domains form the subset of the PRAQ that can be used for outcome measurement.

2.3 Assessment of measurement properties

We studied the distribution of the individual items and the PRAQ domain scores at baseline to check for floor and ceiling effects (i.e. whether <15% of the respondents achieved the highest or lowest possible scores [20]). We assessed the reliability, validity and responsiveness of the PRAQ following the taxonomy of measurement properties as constructed by the COSMIN panel [21].

We calculated the internal consistency parameter Cronbach's α, which should have a value between 0.70 and 0.95 [20]. We assessed test-retest reliability by calculating the intraclass correlation coefficient ($ICC_{Cconsistency}$) for each PRAQ domain. ICC values of 0.7 are considered acceptable, but values of ≥0.8 are preferred [17]. Additionally, we calculated the standard error of measurement (SEM).

We used hypothesis testing to assess convergent validity, which involves studying the correlations of the scores of the PROM under study with the scores of other PROMs. We hypothesized on the size and direction of the (Spearman's) correlations of the PRAQ domains with the (subscales of) PROMs with similar constructs (Appendix 2). We also hypothesized which PROMs should have a lower correlation with the PRAQ domain. Good convergent validity means that 75% of hypotheses are correct [20]. We used the following (subscales of) PROMs for convergent validity in their official Dutch translations:

- The Epworth Sleepiness Scale (ESS) [22], measuring daytime sleep propensity. For eight situations, a patient indicates the likelihood that they would fall asleep while in that situation. The measurement properties of the ESS have been studied in a sleep apnea population [23].
- The 'vitality' domain of the RAND-36 [24]. The (freely available) RAND-36, which is the predecessor of the well-known SF-36, measures general quality of life in several domains. The vitality domain of the RAND-36 contains 4 items about a patient's perceived energy level. The items are identical to the items of the vitality domain of the SF-36, and the domain's measurement properties have been studied in a sleep apnea population in that context [23].
- The following short-forms of the Patient-Reported Outcomes Measurement Information System (PROMIS) databank [25-27]: sleep disturbance (5 items), sleep-related impairment (6 items), fatigue, satisfaction with participation in social roles, ability to participate in social roles, anger, anxiety and depression (the latter 6 all contained 4 items per short-form) [28-31]. For 'sleep disturbance' and 'anger' these were custom short-forms with fewer items than the standard short forms, in order to reduce the number of items that patients had to complete for this study.

To assess responsiveness, we constructed hypotheses about the change scores of PRAQ in correlation to the change scores of the same instruments that were employed for hypothesis testing in construct validity (Appendix 2).

3 RESULTS

3.4 Population characteristics

The baseline population consisted of 180 patients with suspected OSA who completed the baseline measurement. Of these patients, 105 completed the retest between 7 and 21 days (average 14 days), and 53 patients completed the follow-up measurement after 6-8 weeks of treatment with CPAP. Characteristics of these respective (sub)populations can be found in Table 1.

3.5 Missing data

Patients completing the online PRAQ were not allowed to leave any items open (no missings allowed). Eleven patients completed the PRAQ on paper one or more times, and in one of these completed PRAQs (for follow-up after CPAP), item 33 (Appendix 1) was missing from the domain 'social interactions'. We computed the domain score for this patient as the average of the remaining items.

Seven items allowed the response item 'not applicable' (see Appendix 1). Between 19% and 46% of respondents selected this response category for the respective items.

Table 1 Baseline characteristics of the study populations

	Baseline population (n=180)	Test-retest population (n=105)	Population with follow-up after CPAP (n=53)
Gender	31.7% female	38.1% female	25.0% female
Age (mean (SD))	50.1 (12.6)	50.4 (13.0)	55.8 (10.9)
Baseline AHI (mean (SD))	25 (23) (n=160a)	27 (25) (n=96[a])	41 (22)
BMI (mean (SD))	28.9 (4.7)	28.3 (4.6)	30.4 (4.2)
ESS score (mean (SD))	9.9 (4.7)	9.6 (4.4)	9.8 (4.7)
ESS score ≥ 11	43%	42%	40%
Sleep study (type)	43% PG /57% PSG (n=160*)	39% PG /61% PSG (n=96*)	43% PG/ 57% PSG
CPAP compliance (mean (SD))	N/A	N/A	6:46 hrs (1:40 hrs)
CPAP compliance ≥ 4hrs/night	N/A	N/A	96%
AHI with CPAP (mean (SD))	N/A	N/A	2.6 (3.4)

AHI=Apnea-hypopnea Index, BMI=Body Mass Index, CPAP=continuous positive airway pressure, ESS=Epworth Sleepiness Scale, PG=polygraphy, PSG=polysomnography, SD=standard deviation

a. 20 patients with suspected OSA of the total study population did for various reasons (choose to) not undergo a sleep study to determine their AHI.

3.6 Final stage of PRAQ development

Finishing the item selection of the sleepiness domain

None of the inter-item correlations in the preliminary 'sleepiness' domain was higher than 0.9. Principal component analysis showed that the lowest factor loading was 0.65, well above 0.5. Therefore, we took practical elimination decisions: the two items with a 'not applicable' option were removed (about sleepiness while reading, and while driving a car) as well as an item about napping in the afternoon that had a different answering scale than the other items. This improves the homogeneity of the domain for patients. The final version of the PRAQ consists of 10 domains and 40 items (Appendix 1).

Identification and grouping of items for outcome measurement

The results of the final PCA can be found in Table 2. The items of the PRAQ domains 'memory & concentration', 'sleep quality', and 'concerns about health' were removed because they did not have sufficient loading on any of the factors found in the PCA, or because the items loaded on more than one factor. The items of the PRAQ domains 'tiredness' and 'daily activities' loaded on a single factor rather than on two separate factors: these items were therefore combined in one domain called 'energy & daily activities' for the goal of outcome measurement. The items of the PRAQ domain 'unsafe situations' both loaded on one separate domain. However, since this domain contained only two items it was not added to the PRAQ-outcome.

The 19 remaining items in the PCA form three one-dimensional domains: sleepiness, energy & daily activities, and emotions, which together explain 73% of the variance. The PCA showed intercorrelations of these domains of 0.36-0.57. The domains are added to the two formative domains 'symptoms at night' and 'social interactions', resulting in subset of 33 items in five domains. Figure 1 illustrates how the items and domains of the PRAQ result in the subselection of PRAQ items for outcome measurement. The domains that are present in both the full 40-item PRAQ and in the 33-item outcome subset overlap to a great extent.

Table 2 Results of the principal component analysis[ab]

Items	Factor 1	Factor 2	Factor 3
During the past 4 weeks, did you have a problem with:			
Fighting to stay awake during the day?	.290	**.727**	
Suddenly falling asleep?		**.836**	
Difficulty staying awake during a conversation?		**.636**	.222
Difficulty staying awake while watching something? (concert, movie, television)		**.858**	
Falling asleep at inappropriate times or places?		**.802**	
Feeling very tired?	**.785**		
Lacking energy?	**.856**		
Still feeling tired when you wake up in the morning?	**.790**		
In the past 4 weeks:			
How difficult was it for you to do your most important daily activity? (such as your job, studying, caring for the children, housework)	**.841**		
How often did you use all your energy on only your most important, daily activity? (such as your job, studying, caring for the children, housework)	**.940**		
How often did you use all your energy to accomplish only your most important daily activity? (such as your job, studying, caring for the children, housework)	**.825**		
How much difficulty did you have finding energy for your hobbies?	**.770**		
How difficult was it for you to get your chores done?	**.849**		
How often did you feel depressed or hopeless?	.266		**.677**
How often did you feel anxious?			**.793**
How often did you lose your temper?			**.803**
How often did you feel that you could not cope with everyday life?			**.746**
How often did you feel irritated?			**.889**
How often did you have a strong emotional reaction to everyday events?			**.875**

a. The bold font numbers indicate the highest factor loading for that item.

b. Absolute factor loadings <0.2 are not shown in the table.

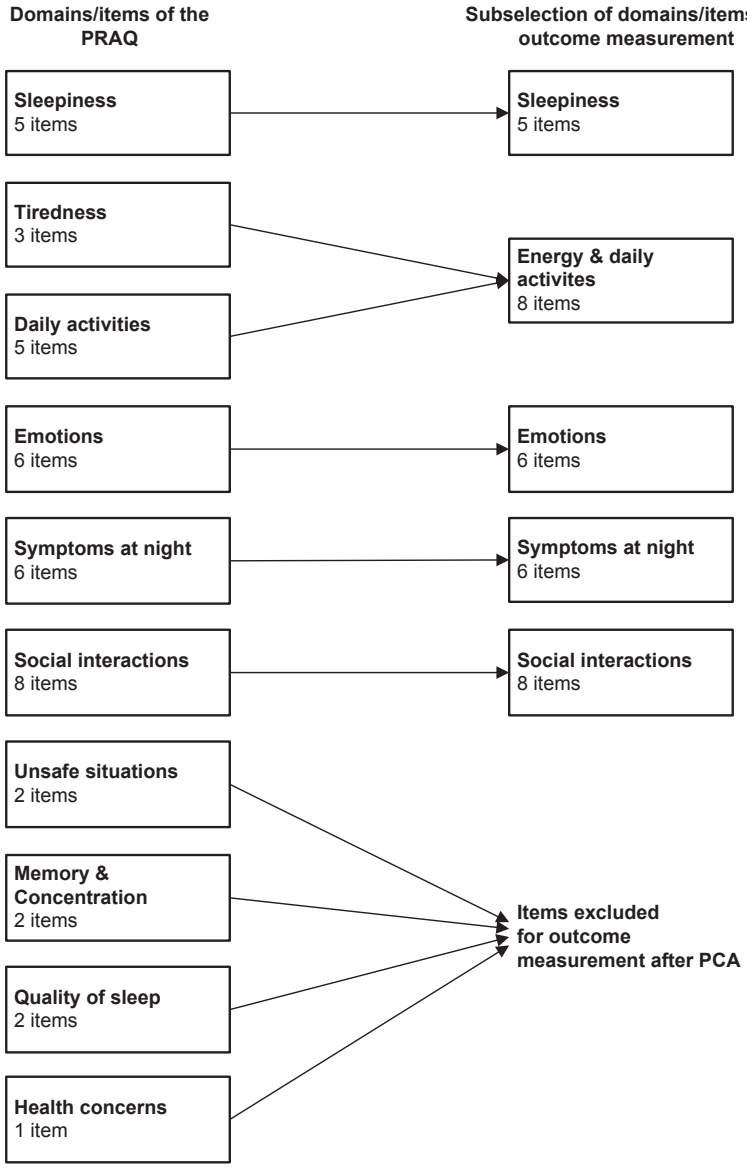


Figure 1 The PRAQ domains for clinical practice and outcome measurement

3.7 Measurement properties

In this section we describe the measurement properties of the domains that are used for outcome measurement; the results for the domains of the 40-item version can be found in Appendix 3. The average baseline scores, standard deviations, and percentages of lowest and highest scores of the five outcome domains can be found in Table 3. No floor- or ceiling effects were found, except for a floor effect in the 'sleepiness' domain (20% of subjects scored 1-1.5). The results of the different aspects of reliability (internal consistency with Cronbach's α, test-retest reliability with ICC, SEM) are also shown in Table 3. The values of Cronbach's α and the ICC values are all above 0.8, indicating that these measurement properties are of good quality.

The correlations of the outcome domains with comparator instruments, which were used to determine convergent validity, are presented in Table 4. The correlations with the (somewhat) similar constructs were all within the ranges that we hypothesized (n=14 hypotheses), and the correlations of selected PRAQ-domains with the dissimilar constructs were all lower than those with the similar constructs (n=3 hypotheses), as expected.

The absolute change scores of the PRAQ outcome domains after patients were treated with CPAP ranged from 0.76 (domain 'emotions') to 1.96 (domain 'energy & daily activities') (Appendix 4). The correlations of the change scores of the PRAQ and the change scores of the comparator instruments (Table 5) were generally in agreement with our hypotheses (n=17 hypotheses). The exception was the 'emotions' domain of the PRAQ, which did not correlate as strongly with the change scores of the PROMIS domains about emotions (anger, anxiety, depression; r=0.26 – 0.43) as we had expected. When a hypothesis is not met, it is important to identify why the results are different than expected [32]. To gain more insight into these unexpected scores, we therefore ran an additional analysis on the correlation of the PRAQ scores and the PROMIS scores at the follow-up measurement, showing results of r=0.62-0.71. This shows that the discrepancy lies with the change score itself and not the absolute score of the follow-up measurement.

Table 3 PRAQ outcome domains: scores and reliability parameters (n=180)

Domain name	Average (range 1-7)	Standard deviation	Lowest score (1-1.5)	Highest score (6.5-7)	Cronbach's α	ICC[a]	SEM[a]
Sleepiness	3.13	1.57	20%	2.2%	0.88	0.81	0.69
Energy&daily activities	4.52	1.59	4.4%	7.8%	0.95	0.86	0.60
Emotions	2.89	1.28	13.3%	0.0%	0.92	0.85	0.50
Symptoms at night	3.48	1.27	3.9%	1.1%	- [b]	0.88	0.44
Social Interactions	3.11	1.42	13.9%	0.6%	- [b]	0.86	0.53

a. n=105

b. These domains are formative, and Cronbach's α is only relevant when a domain is one-dimensional (36).

Table 4 Correlations PRAQ outcome domains and comparator instruments a (n=180)

	ESS	PROMIS Sleep-related impairment	RAND vitality	PROMIS fatigue	PROMIS Ability to participatie in Social Roles and Activities	PROMIS Satisfaction Social Roles	PROMIS anger/ anxiety/ depression	PROMIS sleep disturbance
Sleepiness	**0.67**	**0.60**	-.40	.52				
Energy & daily activities	0.45	**0.83**	**-.77**	**0.86**	**-0.78**	-0.60		
Emotions	0.28		-0.59	0.56	-0.60	-0.42	**0.69-0.76**	
Symptoms at night								**0.47**
Social interactions		**0.56**			**-0.52**	**-0.39**		

a. Correlations in bold are considered similar constructs, for which detailed hypotheses were created. The other correlations are of (somewhat) different constructs and are expected to be weaker than the bold font correlations for that PRAQ domain (for details, see Appendix 2). Correlations for which we had no specific expectations are not shown.

Table 5 Correlations between change scores PRAQ outcome domains and comparator instruments a (n=53)

	ESS	PROMIS Sleep-related impairment	RAND vitality	PROMIS fatigue	PROMIS Ability to participatie in Social Roles and Activities	PROMIS Satisfaction Social Roles	PROMIS anger/ anxiety/ depression	PROMIS sleep disturbance
Sleepiness	**.62**	**.55**	-.35	.35				
Energy&daily activities	.52	**.62**	**-.69**	**.70**	**-.74**	-.61		
Emotions	0.06	.23	-.32	.14	-.29		**.26-.43**	
Symptoms at night								**.58**
Social interactions		**.45**			**-.36**	**-.26**		

ESS=Epworth Sleepiness scale, PROMIS= Patient-Reported Outcomes Measurement Information System

a. Correlations in bold are considered similar constructs, for which detailed hypotheses were created. The other correlations are of (somewhat) different constructs and are expected to be weaker than the bold font correlations for that PRAQ-outcome domain (for details, see Appendix 2). Correlations for which we had no specific expectations are not shown.

4 DISCUSSION

In this article we present the finalized Patient-Reported Apnea Questionnaire (PRAQ). The PRAQ has a unique approach with regard to the integration of its use on an individual patient level and for aggregate outcome measurement: patients complete all items of the PRAQ before their consultation, the results of which can be discussed with a healthcare professional; and the aggregate outcomes of groups of patients can then be studied by making use of a subset of the completed items. The PRAQ contains all topics and items that patients and healthcare providers consider important to discuss in practice, and for this purpose includes 40 items in 10 domains. For outcome measurement, a subset of 33 items of the PRAQ were selected, divided into two formative domains (items grouped together based on what makes sense clinically) and three one-dimensional subscales. These five outcome domains generally have good measurement properties in terms of internal consistency, test-retest reliability, convergent validity and responsiveness.

PCA showed that items of the PRAQ domains 'tiredness' and 'daily activities' load on the same factor, which is why the items of these preliminary domains are combined into one domain for the purpose of outcome measurement. For use an individual patient level, however, we decided to keep the two domains separate. Even though we acknowledge that feeling tired (a symptom), and the extent to which daily activities can be performed normally (a consequence of that symptom), are closely related concepts, they may be relevant to discuss separately for an individual patient in clinical practice. We will test this assumption in future research, in which the PRAQ will be employed and studied empirically.

The domains that are used for outcome measurement show good responsiveness. The one exception is the domain 'emotions', the change score of which showed a much weaker correlation than expected with the change scores of PROMs with similar constructs. We hypothesize that the discrepancy between expectation and results caused by the low scores of this domain at baseline (average 2.89) and the subsequent relatively small improvement that is achieved after treatment with CPAP (average 0.76). We do not doubt the construct validity of the domain, because the comparator PROMs show the same pattern in terms of low scores and small change scores, and because the correlation of the absolute scores after treatment with CPAP shows good convergent validity. However, because the change scores are small, it is likely that measurement error plays a relatively large role in the change scores of both the PRAQ domain and the comparator instruments, reducing the accuracy of the change scores and therefore also diffusing the correlation size. This means that that in terms of outcome/quality measurement, emotional problems appear to be of less importance than the topics of the other domains and more difficult to accurately measure, because relatively few people with (suspected) OSA experience severe problems.

Surprisingly, 20% of the study population had low scores (1-1.5) on the domain 'sleepiness', while sleepiness is one of the main complaints of OSA. We think that this is

due to a relatively high difficulty of the sleepiness items of the PRAQ (such as falling asleep during a conversation) in combination with a generally low sleepiness in this population (average ESS<10). This reason for the low sleepiness in the population is probably twofold. First, the main complaint of some patients who were referred for suspected OSA in this study is probably (socially problematic) snoring rather than sleepiness or tiredness during the day. OSA treatment will reduce their snoring and is reimbursed by healthcare insurers, making it beneficial for these patients to visit the sleep center. Second, for logistical reasons some patients with suspected severe OSA were not included in the study. These patients followed a fast-track procedure to bypass the sleep center's the waiting list, which meant they were in practice not always asked to join the study. This is a limitation of the study. What we can derive from the current results is that the sleepiness domain of the PRAQ seems more useful to detect cases of severe sleepiness, which definitely requires treatment, than to distinguish mild and moderate sleepiness. However, future research should take place in a more representative patient group to study how the sleepiness domain performs in this population.

The PRAQ is designed for use in clinical practice, to help focus consultations on the problems that individual patients encounter. When using a PROM for this purpose, the ICC should preferably be very high (0.9-0.95 at individual level vs. 0.8 or higher at group level for aggregate outcome measurement [17]). The ICC values of the PRAQ are lower (0.81-0.88). However, the PRAQ is meant to open the conversation about a patient's symptoms and functioning, not to serve as a 'cut-off' score. Any elevated score could therefore result in conversation about this topic, and we believe that the PRAQ can serve its purpose despite the slightly lower ICCs.

Methodological considerations

The domains for outcome measurement were created with a combination of the 'clinimetric' approach, in which items are grouped together based on clinical relevance; and the 'psychometric' approach, which groups items together based on PCA [33-35]. The combination of these two approaches is uncommon. We believe that scores of psychometric domains, with a clear one-dimensional construct, are more meaningful than formative domains because they have a clear interpretation. However, this approach is not always feasible when items have been selected to be part of a quality-of-life or symptoms questionnaire based on their deemed importance by the target population [36]. Items which cover symptoms of the same disease or treatment will often share covariance and thus appear to be covering the same latent construct, even when looking at the content of the items this makes no apparent sense (e.g. lack of appetite and decreased sexual interest in patients undergoing cancer treatment [36]). Therefore, we considered the best approach grouping together the different symptoms patients experience at night, as well as the variety of different ways in which sleepiness, tiredness and emotions might influence a patients' social life, without subjecting them to PCA.

To aid the use of the PRAQ in clinical practice, we developed a patient-friendly digital report together with patients and healthcare professionals (Figure 2). When using the PRAQ in clinical practice, it can be useful to look at individual item scores as well as the domain scores, especially in the formative domains in which item scores will generally differ more from each other. Therefore, both domain and individual item scores are shown in the report.

5 CONCLUSIONS

In conclusion, we have shown that the PRAQ-practice and PRAQ-outcome generally have acceptable measurement properties and appear to be suitable PROMs for their respective purposes. However, further validation research is needed in patients who suffer from higher levels of sleepiness, to study the validity of the sleepiness domain. The applicability of a PROM for use in clinical practice and for measuring outcomes on aggregate level, may be of great importance for the further implementation of PROMs in healthcare.

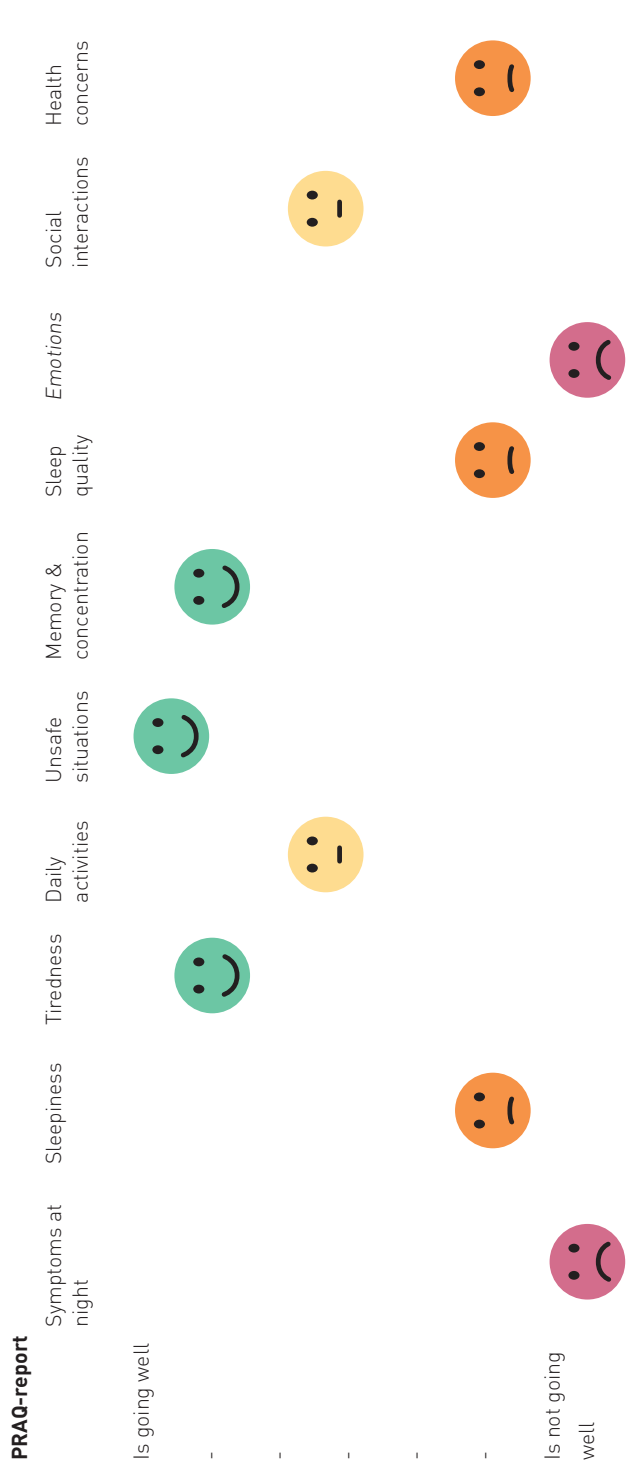


Figure 2 The PRAQ-report

REFERENCES

1. O'Donoghue, N. and E. McKay, *Exploring the impact of sleep apnoea on daily life and occupational engagement.* Br J Occup Ther, 2012. **75**(11): p. 609-516.
2. Reishtein, J.L., et al., *Sleepiness and relationships in obstructive sleep apnea.* Issues Ment Health Nurs, 2006. **27**(3): p. 319-30.
3. Rodgers, B., *Breaking through limbo: experiences of adults living with obstructive sleep apnea.* Behav Sleep Med, 2014. **12**(3): p. 183-97.
4. Bradley, T.D. and J.S. Floras, *Obstructive sleep apnoea and its cardiovascular consequences.* Lancet, 2009. **373**(9657): p. 82-93.
5. Chan, A.S., C.L. Phillips, and P.A. Cistulli, *Obstructive sleep apnoea--an update.* Intern Med J, 2010. **40**(2): p. 102-6.
6. Young, T., P.E. Peppard, and D.J. Gottlieb, *Epidemiology of obstructive sleep apnea: a population health perspective.* Am J Respir Crit Care Med, 2002. **165**(9): p. 1217-39.
7. Bjornsdottir, E., et al., *The Prevalence of Depression among Untreated Obstructive Sleep Apnea Patients Using a Standardized Psychiatric Interview.* J Clin Sleep Med, 2016. **12**(1): p. 105-12.
8. Gupta, M.A., F.C. Simpson, and D.C. Lyons, *The effect of treating obstructive sleep apnea with positive airway pressure on depression and other subjective symptoms: A systematic review and meta-analysis.* Sleep Med Rev, 2016. **28**: p. 55-68.
9. Mokhlesi, B., S.A. Ham, and D. Gozal, *The effect of sex and age on the comorbidity burden of OSA: an observational analysis from a large nationwide US health claims database.* Eur Respir J, 2016. **47**(4): p. 1162-9.
10. Black, N., *Patient reported outcome measures could help transform healthcare.* BMJ, 2013. **346**: p. f167.
11. Van Der Wees, P.J., et al., *Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries.* Milbank Q, 2014. **92**(4): p. 754-75.
12. Greenhalgh, J., et al., *Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care.* 2017: Southampton (UK).
13. I.L., A., et al., *The development of a patient-reported outcome measure for patients with obstructive sleep apnea: the Patient-Reported Apnea Questionnaire (PRAQ).* Journal of Patient-Reported Outcomes, 2017. **1**(14).
14. Mannarino, M.R., F. Di Filippo, and M. Pirro, *Obstructive sleep apnea syndrome.* Eur J Intern Med, 2012. **23**(7): p. 586-93.
15. Epstein, L.J., et al., *Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults.* J Clin Sleep Med, 2009. **5**(3): p. 263-76.
16. Grunstein, R.R., *Sleep-related breathing disorders. 5. Nasal continuous positive airway pressure treatment for obstructive sleep apnoea.* Thorax, 1995. **50**(10): p. 1106-13.
17. De Vet, H.C.W., et al., *Measurement in Medicine.* 2011, Cambridge, UK: Cambridge University Press.
18. Floyd, F.J. and K.F. Widaman, *Factor analysis in the development and refinement of clinical assessment instruments.* Psychological Assessment, 1995. **7**(3): p. 286-299.
19. De Vet, H.C.W., C.B. Terwee, and L.M. Bouter, *Clinimetrics and psychometrics: two sides of the same coin.* J Clin Epidemiol, 2003. **56**: p. 1146-1147.
20. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
21. Mokkink, L.B., 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**(7): p. 737-45.

22. Johns, M.W., *A new method for measuring daytime sleepiness: the Epworth sleepiness scale.* Sleep, 1991. **14**(6): p. 540-5.
23. Abma, I.L., et al., *Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): A systematic review.* Sleep Med Rev, 2016. **28**: p. 18-31.
24. Hays, R.D., C.D. Sherbourne, and R.M. Mazel, *The RAND 36-Item Health Survey 1.0.* Health Econ, 1993. **2**(3): p. 217-27.
25. DeWalt, D.A., et al., *Evaluation of item candidates: the PROMIS qualitative item review.* Med Care, 2007. **45**(5 Suppl 1): p. S12-21.
26. Riley, W.T., et al., *Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks.* Qual Life Res, 2010. **19**(9): p. 1311-21.
27. Terwee, C.B., et al., *Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS).* Qual Life Res, 2014. **23**(6): p. 1733-41.
28. Buysse, D.J., et al., *Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments.* Sleep, 2010. **33**(6): p. 781-92.
29. Yu, L., et al., *Development of short forms from the PROMIS sleep disturbance and Sleep-Related Impairment item banks.* Behav Sleep Med, 2011. **10**(1): p. 6-24.
30. Pilkonis, P.A., et al., *Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): depression, anxiety, and anger.* Assessment, 2011. **18**(3): p. 263-83.
31. Pilkonis, P.A., et al., *Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study.* J Psychiatr Res, 2014. **56**: p. 112-9.
32. Abma, I.L., M. Rovers, and P.J. van der Wees, *Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes.* BMC Res Notes, 2016. **9**: p. 226.
33. Marx, R.G., et al., *Clinimetric and psychometric strategies for development of a health measurement scale.* J Clin Epidemiol, 1999. **52**(2): p. 105-11.
34. Newcombe, P.A., et al., *Development of a parent-proxy quality-of-life chronic cough-specific questionnaire: clinical impact vs psychometric evaluations.* Chest, 2008. **133**(2): p. 386-95.
35. Ribera, A., et al., *Is psychometric scoring of the McNew Quality of Life after Myocardial Infarction questionnaire superior to the clinimetric scoring? A comparison of the two approaches.* Qual Life Res, 2006. **15**(3): p. 357-65.
36. Fayers, P.M. and D.J. Hand, *Factor analysis, causal indicators and quality of life.* Qual Life Res, 1997. **6**(2): p. 139-50.

Appendix 1: The Patient-Reported Apnea Questionnaire (PRAQ)

Symptoms at night
During the past 4 weeks, did you have a problem with: 1. Snoring loudly? 2. Waking up frequently to urinate? 3. Waking up at night with the feeling that you are choking? 4. A feeling that you are sleeping restlessly? 5. Having a dry or painful mouth when you wake up? 6. Waking up in the morning with a headache?
Sleepiness
During the past 4 weeks, did you have a problem with: 7. Fighting to stay awake during the day? 8. Suddenly falling asleep? 9. Difficulty staying awake during a conversation? 10. Difficulty staying awake while watching something? (concert, movie, television) 11. Falling asleep at inappropriate times or places?
Difficulty staying awake while reading?[a,b] Fighting to stay awake when you are driving?[a,b] Did you feel like you needed to take a nap in the afternoon?[a]
Tiredness
During the past 4 weeks, did you have a problem with: 12. Feeling very tired? 13. Lacking energy? 14. Still feeling tired when you wake up in the morning?
Daily activities
During the past 4 weeks: 15. How difficult was it for you to do your most important daily activity? (such as your job, studying, caring for the children, housework) 16. How often did you use all your energy to accomplish only your most important daily activity? (such as your job, studying, caring for the children, housework) 17. Did you feel you have a decreased performance with regard to your most important daily activity? (such as your job, studying, caring for the children, housework) 18. How much difficulty did you have finding energy for your hobbies? 19. How difficult was it for you to get your chores done?
Unsafe situations
During the past 4 weeks: 20. Did you have problems while driving a car due to sleepiness?[b] 21. Were you concerned about your safety or that of others due to your sleepiness? (for example in traffic, or when operating machinery)
Memory and concentration
During the past 4 weeks: 22. Were you sometimes forgetful? 23. Did you sometimes have difficulty concentrating?
Quality of sleep
During the past 4 weeks, did you have a problem with: 24. Falling asleep when you go to bed at night? 25. Getting back to sleep after you woke up at night?

Appendix 1 *continued*

Emotions

During the past 4 weeks:

26. How often did you feel depressed or hopeless?
27. How often did you feel anxious?
28. How often did you lose your temper?
29. How often did you feel that you could not cope with everyday life?
30. How often did you feel irritated?
31. How often did you have a strong emotional reaction to everyday events?

Social interactions

During the past 4 weeks:

32. Did you sometimes feel upset because others were disturbed by your snoring?
33. Was it a problem for you that you sometimes had no energy or no desire to do things with your family or your friends?
34. Did you feel guilty towards your family or friends?
35. Did you feel upset because you argued frequently?
36. Did you sometimes experience problems in the relationship with your partner?[b]
37. Did you feel upset because you could (maybe) not sleep in the same room as your partner?[b]
38. Did you sometimes think up excuses because you were tired or sleepy?
39. Did you have a problem with unsatisfying and/or too little sexual activity? (by yourself or with another)[b]

Health concerns

40. Were you concerned about other conditions that may be related to sleep apnea? (such as diabetes, high blood pressure, cardiovascular disease, being overweight)

a. The shaded items of the "sleepiness" domain were removed from this domain in the final version of the PRAQ.

b. These items had an additional response option "not applicable" or (for item 39) "no answer"

Appendix 2: Hypotheses for convergent validity and responsiveness

PRAQ domain	Comparator instrument	Expected correlation strength (direction)	Explanation (hypothesis nr between brackets)
Sleepiness	ESS	0.5-0.8 (+)	The ESS asks about current daytime sleep propensity, while the PRAQ domain asks to look back on the past month and indicate how much of a *problem* sleepiness or falling asleep was. The domains do not cover the exact same construct, but are relatively similar. We expect a moderately strong correlation (h1). "Sleep-related impairment" is a domain with questions covering both sleepiness and tiredness. We expect a moderate to strong correlation with the PRAQ domain "sleepiness" (h2) because of the overlapping items on sleepiness, but also because the concept of sleepiness itself correlates with the concept of tiredness and how tiredness affects daily activities in our study population (correlation strength 0.57 as found in our principal component analysis). *Dissimilar constructs* We expect the correlations with the PROMIS fatigue and RAND-36 vitality domains to be lower than the correlations with the similar constructs of the PROMs mentioned above (h3).
	PROMIS: Sleep-related impairment	0.5-0.8 (+)	
	Different constructs		
	PROMIS fatigue (+) RAND-36 vitality (-)		
Energy & daily activities	PROMIS: Sleep-related impairment	0.6-0.9 (+)	"Sleep-related impairment" is a domain with questions covering both sleepiness and tiredness, in the context of daily activities. Therefore we expect a strong correlation with the PRAQ domain (h1). The RAND-36 vitality domain and the PROMIS fatigue both contain questions about tiredness. We expect both to have a strong correlation with the PRAQ domain (h2+3). We expect a somewhat stronger (h4) correlation with the PROMIS fatigue domain because this domain is focused on to what extent one feels tired (similar to the PRAQ), while the RAND focuses on *how often* one feels tired. The PROMIS domain "ability to participate in social roles and activities" asks question about to what extent someone is able to do their usual daily or social activities. We expect a strong correlation with the PRAQ domain because the constructs are similar (h5). *Dissimilar constructs* We expect the correlations with the ESS and the "PROMIS satisfaction with social roles and activities" domain to be lower than the correlations with the similar constructs of the PROMs mentioned above (h6).
	RAND-36 vitality	0.6-0.9 (-)	
	PROMIS Fatigue	0.6-0.9 (+)	
	PROMIS ability to participate in social roles and activities	0.6-0.9 (-)	
	Different constructs		
	PROMIS satisfaction with social roles and activities (-) ESS (+)		

Emotions	PROMIS anger	0.6-0.9 (+)	The three PROMIS domains contain items asking about *how often* certain emotions are felt, which is the same approach as the PRAQ-domain. The PRAQ-domain also contains all three of these types of emotions. Therefore we expect a strong correlation with all three of these domains (h1-3). *Dissimilar constructs* We expect the correlations of the PRAQ "emotions" domain with the ESS, RAND-36, and the PROMIS domains ability to participate in social roles and activities" and "satisfaction with social roles and activities" to be lower than the correlations with the similar constructs of the PROMs mentioned above (h4).
	PROMIS anxiety	0.6-0.9 (+)	
	PROMIS depression	0.6-0.9 (+)	
	Different constructs		
	ESS (+)		
	RAND-36 vitality (-)		
	PROMIS ability to participate in social roles and activities (-)		
	PROMIS satisfaction with social roles and activities (-)		
Symptoms at night	PROMIS sleep disturbance	0.2-0.6 (+)	The most similar domain that we included for the PRAQ domain "symptoms at night" is the PROMIS "sleep disturbance" domain. This domain contains items about whether patients are sleeping well. Even though a majority of the items in the PRAQ domain "symptoms at night" will affect the quality of sleep, the content of the two domains is very different. We therefore will not make a very precise hypothesis, and expect at least a low to moderate correlation (h1). Dissimilar constructs are not included for this domain. Due to the domain's clinimetric nature it covers several topics that are related to severity of OSA and that might therefore have unpredictable correlations with other domains.
Social interactions	PROMIS: Sleep-related impairment	0.2-0.6 (+)	The "social interactions" domain of the PRAQ contains a collection of items about different social problems that apnea patients might experience due to their snoring, sleepiness, tiredness, or emotions. Because this domain does not clearly cover one single construct, we expect it to have no more than low to moderate correlations with any of the comparator PROMIS domains (h1-3). Dissimilar constructs are not included for this domain. Due to the domain's clinimetric nature it covers several topics that are related to severity of OSA and that might therefore have unpredictable correlations with other domains.
	PROMIS ability to participate in social roles and activities	0.2-0.6 (-)	
	PROMIS satisfaction with social roles and activities	0.2-0.6 (-)	

Appendix 3: Domain and reliability scores of the PRAQ clinical practice domains

Scores of the PRAQ clinical practice domains[a]

Domains	Average (range 1-7)	Standard deviation	Lowest score (1-1.5)	Highest score (6.5-7)
Sleepiness	3.13	1.57	20%	2.2%
Tiredness	4.85	1.71	4.4%	20.0%
Daily activities	4.33	1.64	5.0%	8.9%
Emotions	2.89	1.28	13.3%	0.0%
Symptoms at night	3.48	1.27	3.9%	1.1%
Social Interactions	3.11	1.42	13.9%	0.6%
Memory&concentration	3.85	1.34	4.4%	1.7%
Unsafe situations	2.13	1.56	57%	1.1%
Sleep quality	2.71	1.71	42.8%	6.7%
Health concerns	3.34	2.00	25.6%	8.9%

a. The clinical practice domains partially overlap with the outcome domains of the PRAQ, as shown in Figure 1.

Reliability scores of the PRAQ clinical practice domains[a]

Domains	Cronbach's α	ICC	SEM
Sleepiness	0.88	0.81	0.69
Tiredness	0.93	0.86	0.64
Daily activities	0.94	0.83	0.68
Emotions	0.92	0.85	0.50
Symptoms at night	-[b]	0.88	0.44
Social Interactions	-[b]	0.86	0.53
Memory&concentration	0.83	0.86	0.50
Unsafe situations	0.87	0.87	0.56
Sleep quality	0.72	0.83	0.71
Health concerns[c]	-	-	-

SEM = standard error of measurement

a. The clinical practice domains partially overlap with the outcome domains of the PRAQ, as shown in Figure 1.

b. These domains are formative, and Cronbach's α is only relevant when a domain is one-dimensional [1].

c. This domain contains only one item, meaning these measurement properties cannot be calculated.

Appendix 4: Change scores of the PRAQ outcome domains after treatment with CPAP (n=53)

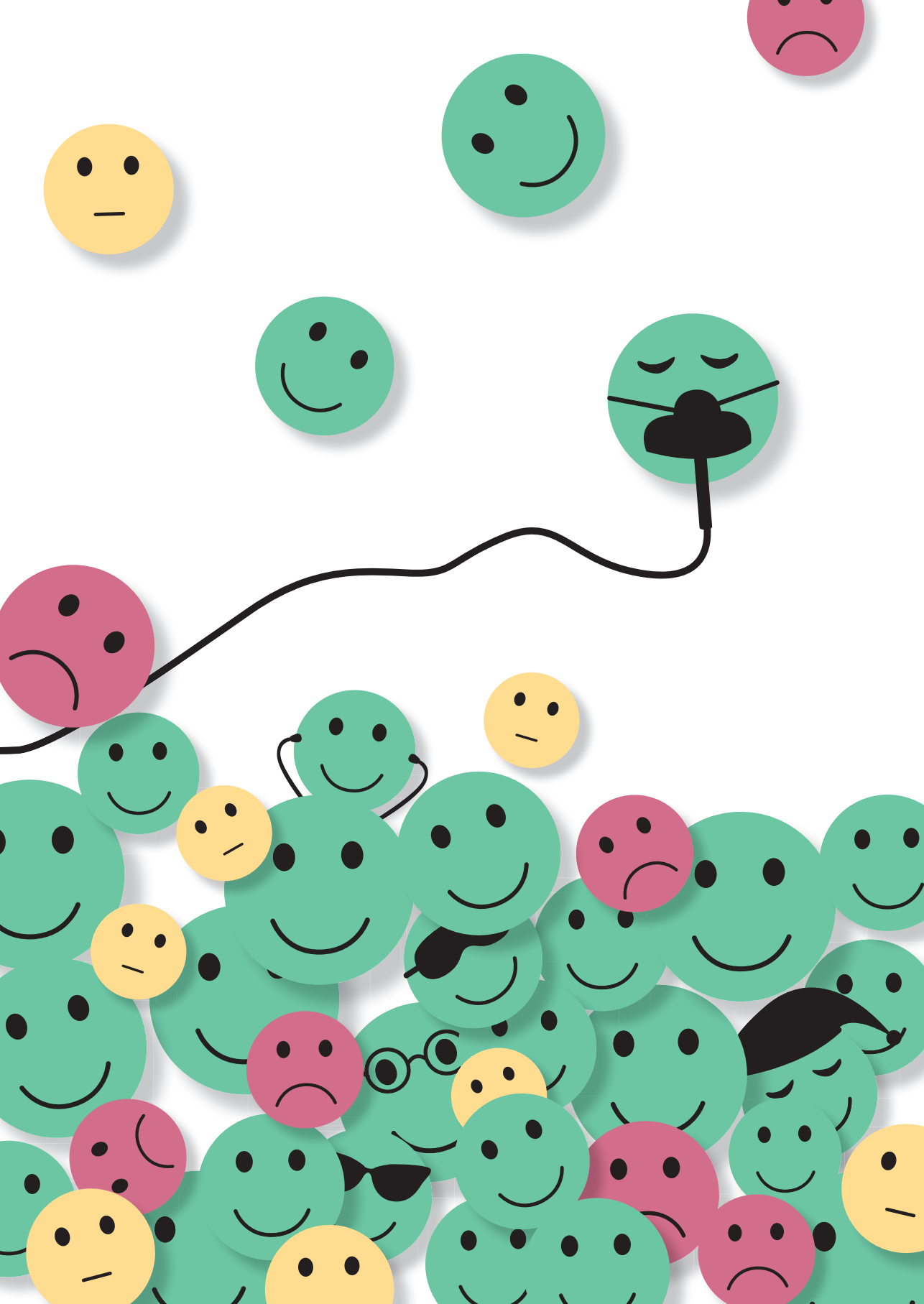
Domain name	Baseline score	Average score after treatment	Average change score [a,b]
Sleepiness	3.27	1.56	1.70
Energy & daily activities	4.66	2.70	1.96
Emotions	2.76	2.00	0.76
Symptoms at night	3.45	1.82	1.63
Social Interactions	2.94	1.79	1.15

a. A positive change score stands for a reduction in symptoms.

b. All change scores are significant, p-value <0.00

REFERENCE

1. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess. 2003; 80: 217-22



CHAPTER 6

Does the Patient-Reported Apnea Questionnaire (PRAQ) increase patient-centeredness in the daily practice of sleep centers? A mixed-methods study.

Inger L. Abma, Maroeska Rovers, Marijke IJff, Bernard Hol,
Masha E. Nägele, Gert P. Westert, Philip J. van der Wees
Submitted

ABSTRACT

Objectives

The objective of this exploratory study was to see how the Patient-Reported Apnea Questionnaire (PRAQ) may impact the daily clinical practice of sleep centers, and why it may or may not work as expected. The hypotheses were tested that this patient-reported outcome measure (PROM) makes patients more aware of which of their health complaints may be related to obstructive sleep apnea (OSA), and that it improves patient-centeredness of care by shifting the focus of care away from (only) medical problems towards the individual burden of disease and quality of life.

Design

Mixed methods. The quantitative study (surveys, patient records) was a before-and-after study.

Setting

Three sleep centres in The Netherlands (secondary care).

Participants

27 patients and 14 healthcare professionals were interviewed. 487 patients completed surveys pre-implementation, and 377 patients completed surveys post-implementation of the PRAQ. For the health records, 125 patients were included in the pre-implementation group, and 124 other patients in the post-implementation group.

Interventions

The PRAQ was used in clinical practice for six successive months.

Outcome measures

Scores on individual survey items, number of patients receiving non-medical treatment, adjustment of treatment at first follow-up, compliance with treatment.

Results

Patients were generally positive about the usefulness of the PRAQ before and during the consultation., as they felt more informed. Healthcare providers did not consider the PRAQ very useful, and they reported minor impact on their consultations. The surveys and health record study did not show an impact of the PRAQ on clinical practice.

Conclusions

Implementing the PRAQ may be beneficial to patients, but this study does not show much impact with regard to patient-centeredness of care. New Dutch guidelines for OSA care may lead to a greater emphasis on quality of life and value of care for patients, making its integration in clinical care potentially more useful.

1. INTRODUCTION

The integration of patient-reported outcome measures (PROMs) in clinical practice has been gaining popularity in the past decade [1-3]. PROM data collected in clinical practice can be aggregated and used for quality improvement purposes, or individual scores can be used in daily clinical practice to improve patient care. In this latter function PROMs can be used in different ways, e.g. as a screening tool, a monitoring or evaluation tool, a tool to inform and empower patients, and/or to increase the patient-centeredness of care by shifting the focus of care away from (only) medical problems towards the problems patients experience in their daily life [4]. When using PROMs in daily clinical practice, it may be sensible to combine the use of a PROM on an individual patient level with application on an aggregate level [5]. There have been a number of studies that aimed to evaluate the usefulness of PROMs in clinical practice in a variety of settings, of which the results are mixed [6-8]. Though qualitative research on this topic has been synthesized in a recent review [4, 9] including a list of hypotheses on how PROMs might work, there are still many questions regarding which PROMs can be potentially useful in which settings.

This study is focused on the application of individual PROM scores in sleep centers which diagnose and treat patients with obstructive sleep apnea (OSA), a condition for which a PROM could be a useful tool to improve patient-centeredness of care. OSA is a highly prevalent but often unrecognized condition in which frequent collapse of the upper airway causes breathing stops while asleep. The subsequent arousals can result in severe sleepiness and fatigue during the day, often affecting a patient's cognitive function, psychological well-being, relationships, and ability to work [10-12]. OSA has also been shown to be an independent risk factor for hypertension, heart failure and diabetes [13-15]. The prevalence of OSA has been reported to be 6% to 38%, depending on the exact definition of OSA and the population studied, and is higher in men [16].

Severity of OSA and necessity for treatment has historically been based on the number of (partial) breathing stops per hour: the apnea-hypopnea index (AHI) [17, 18]. However, there is no linear association between AHI and severity of symptoms or the presence of comorbidities [19-23]. There is also little evidence that treating patients with mild OSA (based on AHI) or patients with low sleepiness is useful in preventing cardiovascular disease or incidents [24-27]. In the past few years there has therefore been international discussion regarding new approaches to diagnose "clinically relevant" OSA [28, 29]. This discussion has also made its way into recent Dutch guidelines for OSA, in which it is recommended that there should be a greater focus on the presence of potentially related comorbidities, as well as the experienced burden of disease for individual patients. The goal of treatment is the improvement of these aspects of OSA [30].

We have developed and validated a PROM for use in clinical practice which may aid this new focus of care for patients with OSA: the Patient-Reported Apnea Questionnaire

(PRAQ) [31, 32], which measures OSA-related quality of life. The goal of this PROM is to improve patient-centeredness of care on an individual level by shifting the conversation away from the medical problems and towards and individual's burden of disease/quality of life, and also to measure quality of care on an aggregate level. To develop the PRAQ, the input from patients and healthcare professionals was used to select the topics that were considered most important to discuss in clinical practice [31]. The individual PRAQ scores of each patient with (suspected) OSA are captured in the 'PRAQ-report', which was designed together with patients and uses colored smileys to show the results for the 10 domains of the PRAQ. The advantage of the PRAQ compared to other commonly used PROMs in the care for patients with OSA (such as the Epworth Sleepiness Scale (ESS), Functional Outcomes of Sleep Questionnaire (FOSQ), etc) is that it provides a comprehensive overview of the possibly impacted aspects of quality of life that patients with OSA may experience. It is therefore potentially suitable for shifting the focus of care away from (only) medical problems towards the problems patients experience in their daily life.

This explorative study aims to study the impact of the PRAQ and PRAQ-report on the clinical practice of OSA, and explore *why* the PRAQ did or did not have an impact. A combination of both qualitative and quantitative methods is used that will add to the general knowledge on the circumstances under which PROMs do or do not work in clinical practice.

2. METHODS

This article describes an exploratory mixed methods study in which the PRAQ is implemented in the clinical practice of three sleep centers. Qualitative interviews and a patient survey were used to explore patients' and healthcare providers' experiences with the PRAQ, and to identify potential barriers and facilitators to its use. Additionally, data were collected from electronic health records to study whether the hypotheses about the potential impact of the PRAQ mentioned in the introduction are correct. For the patient survey and the patient record study we conducted a before-and-after study. The different methods are described in more detail in the next sections.

2.1 Hypotheses

We have several hypotheses regarding how the PRAQ may influence patients and healthcare professionals, and how this could impact clinical practice. First of all, completing the PRAQ could:

- Encourage patients to consider which problems they experience that might be related to OSA and that they might want to discuss
- Aid healthcare professionals in opening a conversation about an individual patient's burden of disease (apnea-related quality of life)

- Aid healthcare professionals to evaluate treatment and identify problems that are still present

We think that this may potentially lead to:

- Higher patient compliance with treatment
- More explicit choices regarding whether clinical treatment for OSA is (potentially) beneficial to the patient
- An increase in referrals to other healthcare providers, such as psychologists
- More 'holistic' care, in which there is increased attention for the well-being of patients, including the psychological and social effects of OSA and its comorbidities

2.2 The PRAQ and its implementation

The PRAQ and its complementary PRAQ-report were designed with the input of patients with OSA and healthcare professionals [31]. The questions of the PRAQ can be found in Appendix 1 of chapter 5. The PRAQ takes approximately 15 minutes to complete [31]. More information about the PRAQ-report and how the PRAQ was implemented in clinical practice can be found in Appendix 1.

2.3 Setting and subjects

Sleep centers of three Dutch hospitals took part in the study. The PRAQ was part of the clinical practice routine of these centers for six successive months. The PRAQ was distributed to patients attending an intake consultation for possible OSA (which takes place after a patient's diagnostic sleep study), and subsequently to the subselection of these intake patients diagnosed with OSA who returned for a follow-up consultation after starting treatment.

2.4 Interviews

In-depth, semi-structured interviews were conducted with patients and healthcare professionals in the last two months of the study. The interview guides contained broad, open questions as well as more specific questions informed by topics previously identified in the literature [4]. For patients the main goal was to assess whether completing the PRAQ was acceptable to them, and to find out the impact that the PRAQ and PRAQ report had for them on the (preparation for) the consultation. For healthcare providers, questions were mostly focused on how they used the PRAQ and why they used it this way, and the impact the use of the PRAQ has on their practice. This information can provide the basis for interpreting the results of the electronic health record study.

Patients were invited via email by the sleep center before their scheduled consultation, or by their healthcare professional directly after their consultation (for more information see Appendix 2). Only patients who had completed the PRAQ were invited. We interviewed 27

patients. Data saturation was reached. Characteristics of the interviewed patients and of the interviews can be found in Table 1.

All healthcare professionals of the three participating sleep centers that had had the option to work with the PRAQ were invited to participate. This resulted in interviews with 14 healthcare professionals: six pulmonologists, six physician assistants (PAs) and two nurses. Two pulmonologists refused an interview because they had not seen many patients for OSA, two others because they had not used the PRAQ at all, and one PA refused for personal reasons. At least four healthcare professionals were interviewed at each of the three sleep centers. More information on the (analysis of) the interviews can be found in Appendix 2.

Table 1 Characteristics of the interviewed patients and the interviews

Patient characteristics (n=27)	
Age (mean, range)	59 (31-82)
Gender (male)	18
Highest education level (range)	Primary school - PhD
Interview characteristics (n=27)	
Interview after intake consultation (n)	18
Interview after follow-up consultation (n)	9
Interview together with partner or other relative that attended the consultation	4
Patients who had not seen the PRAQ-report at the time of the interview[a]	5

a. Viewing the PRAQ-report before the consultation was optional, and not all healthcare providers showed the report to the patient during the consultation

2.5 Surveys

The patient survey was designed for this study to study potential differences in patient empowerment and patient-centeredness of care before and after the implementation of the PRAQ. The items of the survey covered how prepared patients felt for their consultation, whether there was discussion of the health problems that patients consider relevant during the consultation, and whether patients were motivated to start their treatment. Patients could indicate their agreement on several statements on these topics with the statement on a 7-point Likert scale. The survey was checked by the members of the research team, which included a patient, but was not pilot tested. A translated version of the survey can be found in Appendix 3.

Surveys were distributed by healthcare professionals to all of their patients attending either an intake or first follow-up consultation for (suspected) OSA. Distribution of the surveys took place in the two months before implementation of the intervention (control group), and in the last two months of the six months that the intervention was part of daily clinical practice (intervention group). For the intervention group, the survey also contained additional questions about the patient's opinion on the usefulness of the PRAQ. Participation was voluntary and anonymous.

2.6 Electronic health records

Electronic health records from one of the included sleep centers were studied to explore potential changes in treatment and compliance with treatment resulting from the use of the PRAQ. Data were collected from patients with an AHI≥5 attending an intake consultation during the final two months of the study period and during the same time period the previous year. Information was collected about treatment choice at intake, treatment adaptations and compliance with treatment at the first follow-up consultation, and patient characteristics. Compliance data is only available for patients who receive Continuous Positive Airway Pressure (CPAP), the most commonly prescribed treatment for patients with OSA. As part of standard care, hours of use are registered by the CPAP device and entered into the health record at follow-up consultations. CPAP compliance is expressed as average hrs CPAP use/night in the month before the follow-up consultation, with an average of 4 hrs/night generally being the minimum to be considered compliant [33].No identifying information was collected from the health records. The data collection procedure guaranteed that the records would at all times remain anonymous to the researchers.

2.7 Statistical analyses

Mann-Whitney U tests were conducted for each of the survey items that patients were asked to complete both pre- and post-implementation of the PRAQ. For the electronic health record study, treatment choice at intake was studied by aggregating the choice into two variables: medical treatment of OSA (e.g. CPAP, MRA, referral for surgery) and no or non-medical treatment (e.g. lifestyle advice), as these are the variables which we potentially expected the PRAQ to influence. A Chi-Square test was used to test for statistical significance. For the follow-up variables of the patient record study, Chi-Square tests (for dichotomous variables) and an independent samples T-test (for CPAP compliance in minutes) were conducted.

No correction for multiple testing was performed because this is an exploratory study. A p-value of <0.05 was therefore taken as a significant difference, which can be interpreted as an indication that this is a potentially interesting variable for a possible future study.

2.8 Patient and Public Involvement

A board member (author MI) of the Dutch patient organisation for OSA (Apneuvereniging) was involved with this study from its inception, including the research question and outcome measures and interpretation of the results. This author was also closely involved in the development of the intervention itself (the PRAQ and its complementary PRAQ-report), as were other members of the patient organization [31]. They also approved of the burden and time required for the intervention. Patients were not involved in the recruitment for the study.

3. RESULTS

3.1 Interviews

Patient perspective

Patients were generally willing to complete the PRAQ before their consultation, and patient response as reported by the healthcare professionals was high. About half of the interviewed patients indicated that completing the PRAQ helped them prepare for their intake consultation by giving them more insight into their complaints and functioning and how this might relate to OSA, and/or made them consider what they wanted to discuss with the healthcare professional. Many patients completed the PRAQ with a family member which instigated discussions patients often considered useful. A great majority of interviewed patients indicated that they did not mind taking the time to complete the PRAQ, and many also considered the smileys of the PRAQ-report a clear and easy way of communicating the results. Box 1 contains quotes illustrating the statements in this paragraph.

The interviews also revealed some unintended effects of the PRAQ. A majority of patients assumed that the main purpose of the PRAQ was to aid their healthcare professional in setting a diagnosis, by providing information about symptoms ahead of time. A few patients believed that discussion of patient complaints during the consultation was therefore no longer necessary after completing the PRAQ (Box 2), while healthcare professionals consider this discussion very important (see next section). What may have played a role here is that several interviewed patients seemed eager to hear their sleep study results, rather than (first) spend much time talking about their symptoms or problems.

Additionally, there were some issues around the interpretation of the smileys in the PRAQ-report. Several of the interviewed patients did not seem to view the PRAQ-report as merely a visualization of the answers they had given, but rather as a 'test result'. Some considered the number of 'unhappy' smileys as an indication of whether they were doing well or not, which made some patients reconsider the severity of their complaints (Box 2).

Box 1

"Look, it's just very insightful. You can see instantly where the problems are and on this other [page] you can see what the improvements are. Yes, it's kinda nice." (Centre 3, patient 10)

"Yes, you know I do find it useful, because you have so many... so many things that bother you, that you forget what it is that bothers you. Or because it has become part of you, so to say. So yeah in order [not] to forget things, a questionnaire like this comes in handy." (Centre 2, patient 1)

"But there were quite a lot of questions where I was like, oh, sometimes I'm like, how does that fit with [apnea]? But most did, but there were questions where I was like, is that related to sleep apnea? So. Yes. Apparently." (Centre 3, patient 7)

"Actually I liked [seeing it beforehand], because this way I can by myself... otherwise I would have gone into it timidly like, tell me, what did you see? And now I could ask specific questions."
(Centre 3, patient 2)

Box 2

"I think it's very good, because you can from the beginning very clearly indicate your problems. So it doesn't need to all be done during the short conversation you have with the specialist. [..] It's clear it doesn't need to be mentioned again, because it's clear to her as well what the problems are." (Centre 1, patient 4)

"Just that when you complete a questionnaire aimed at establishing something, then it's useful that you also get a sort of result. So a preliminary... not that you should instantly think like nothing is wrong, nothing needs to be done, let's get out of here. But, I did like it, yeah." (Centre 1, patient 3)

"Well, because there were only two orange [smileys], and the others were all green and then you think, well.... And then when you look at it again then I'm like, 'I can live with that'." (Centre 2, patient 7)

Healthcare professional perspective

Most healthcare professionals used the PRAQ during consultations (Table 2), but usually briefly. Several professionals mentioned that, especially during intake consultations, they used it for the sake of the study. Only a few tried to provide more holistic care with the PRAQ. Some professionals stated that their minimal use of the PRAQ was due to unwillingness to change their practice, while others mentioned a general aversion to questionnaires, and/or not being convinced that the PRAQ would offer new or useful information considering what was already discussed during a regular consultation. There were also practical issues that to some extent hindered the uptake of the PRAQ: most notably the (limited) time available for consultations, and the fact that the PRAQ was not embedded in the electronic health records which hindered the regular workflow. There were no notable differences in attitude towards the PRAQ between physicians, PAs and nurses.

Table 2 Use of PRAQ-report by interviewed healthcare professionals

Use of the PRAQ-report during intake consultations				
Discussed it with patients	Only looked it up	Did not look at it		N/A[a]
8	1	3		2
Use of the PRAQ-report during follow-up consultations				
Discussed it with patients	Only looked it up	Did not look at it	Want to use it[b]	N/A[a]
3	1	3	6	1

a. Not all healthcare professionals held both intake and follow-up consultations

b. Did not see (many) patients with follow-up PRAQ but are interested in using it in this setting

Most of the professionals that used the PRAQ did so at the end of their usual discussion of symptoms, to check whether all topics that were problematic had been discussed and potentially address more topics. As such they could still start the conversation in their usual way, allowing patients to explain their problems in their own words, and allowing the healthcare professionals to ask their standard diagnostic questions. Professionals indicated that most "symptoms" that are part of the PRAQ were already part of the standard diagnostic questions during an intake consultation (sleepiness, problems at night), and also overlapped with their usual (diagnostic) intake questionnaire. However, several professionals mentioned that the PRAQ-report increased discussion of the topic "health concerns", which was considered valuable. Furthermore, the few professionals that indicated that they valued offering more holistic care noticed that the PRAQ was useful in drawing the conversation away from medical facts and more towards the underlying emotions related to a patient's problems. However, many other professionals did not see much added value in actively bringing up topics like emotions and social interactions. They were potentially willing to discuss these issues but considered it up to the patient to raise them. If the PRAQ was used to identify problems, it was

more common for the professional to mention very briefly that these problems were likely to improve with treatment of OSA, without further discussing these problems. Professionals reported that they did not notice any increase in OSA-related knowledge in their patients, or a difference in whether or how patients raised health complaints or quality of life issues of their own accord. Box 3 contains quotes illustrating the statements in this paragraph.

With regard to treatment choice, the professionals mentioned that the severity of symptoms generally only plays a role in patients with an AHI<15, for which shared-decision making could potentially lead to a decision not to start clinical treatment for OSA. If the AHI is ≥15, professionals generally wish to treat a patient for health reasons irrespective of symptoms. Many patients also have a reason to opt for treatment: there is a motor vehicle driving ban for untreated patients with AHI≥15.

Use of the PRAQ during follow-up consultations could not be fully evaluated, because a limited number of patients had completed the PRAQ at follow-up at the time of the interviews. This was due to practical implementation issues in combination with the relatively short duration of the study. However, several healthcare professionals mentioned that they thought the PRAQ would be more useful during follow-up consultations than intake conversations, as it would be interesting to see which problems remained after starting treatment (Table 2). Those that had the opportunity to use the PRAQ in this setting mentioned that it was nice to show patients how their problems had improved, with the improvement sometimes being greater than the patients had realised. This could be used as encouragement to continue with treatment.

Box 3

"Well I myself don't ask 'are you worried about your [health]'? I won't ask that, but that is what it shows. So then... then it's like 'hey, I would otherwise not have discussed that'." (Centre 3, healthcare provider 4)

"Yes, but then in a solution-oriented way - then you will see someone with 30 apneas an hour and you see that and you say I hope that [your problem with emotions] will get a lot better with the therapy I will start for you." (Centre 1, healthcare provider 2)

"Especially I thought people were, uhm... that lack of initiative, not going out, right? So they don't do things because of their sleep problem, that was what [the PRAQ] often showed. And I didn't always get that from taking the patient history. So people maybe find that hard to tell me, or they have trouble indicating that it really does have an impact on them. And then they try to focus more on the fact than on the underlying emotion. And that would sometimes give added value." (Centre 1, healthcare provider 1)

3.2 Survey results

A total of 487 patients completed surveys pre-implementation, and 377 patients completed surveys post-implementation of the PRAQ. Characteristics of the survey populations pre-implementation and post-implementation can be found in Table 3.

Table 3 Characteristics of the survey population

Intake consultations		
	Pre-implementation (n=239)	Post-implementation (n=197)
Age (yrs)	53.9	55.4
Gender (% male)	68.4	69.5
Severity of symptoms[a]	6.50	6.44
Diagnosed with OSA (%)	82.8	83.2
CPAP[b] (%)	71.0	70.7
MRA[b,c] (%)	13.7	19.5
Other treatment[b] (%)	10.7	7.4
No treatment[b] (%)	1.0	2.4
Missing[b] (%)	3.6	0.0
Follow-up consultations		
	Pre-implementation (n=248)	Post-implementation (n=180)
Age	57.33	58.54
Gender (% male)	75.3	69.7
Severity of remaining symptoms or problems with treatment[a]	4.25	5.03
CPAP (%)	89.1	89.4
MRA[c] (%)	3.6	3.9
Other[d] or missing (%)	7	5.6

CPAP = continuous positive airway pressure, MRA = mandibular repositioning device

a. Scale 1-10, higher is more problems

b. Percentage of patients with this treatment of the total of patients diagnosed with OSA

c. Device worn over the teeth that pushes tongue and jaw forward to hold the airway open

d. Other possible treatments are surgery of the jaw or throat, and methods that will help a patient with positional OSA (who experiences breathing stops mainly when they lie on their backs) sleep on their side

Patients generally showed high agreement with the statements of the survey: 73.3% - 97.3% of patients indicated "agree" or "completely agree" per statement about the intake and follow-up consultations (Table 4). Follow-up patients post-implementation showed significantly less agreement with the statement "In my opinion, my treatment is worth it for me" (p=.005).

Table 4 Survey results[a]

Intake consultations										
	Pre-implementation PRAQ					**Post- implementation PRAQ**				
Scores	1-3 % (n)	4-5 % (n)	6-7 % (n)	N /A[b] n	Median	1-3 % (n)	4-5 % (n)	6-7 % (n)	N /A[b] (n)	Median
I knew which problems I wanted to discuss with the doctor	17 (7.2)	18 (7.7)	200 (85.1)	4	6	8 (4.2)	14 (7.3)	170 (88.5)	5	6
I discussed with the doctor the topics I wanted knew I wanted to discuss beforehand	22 (9.8)	28 (12.5)	174 (77.7)	15	6	16 (8.9)	22 (12.2)	142 (78.9)	17	6
Because of my conversation with the doctor, I understand better what causes my problems	14 (6.0)	14 (6.0)	206 (88.0)	5	6	6 (3.1)	10 (5.1)	178 (91.8)	3	6
The doctor and I chose the treatment together or together chose not to treat my apnea	14 (6.0)	12 (5.2)	206 (88.8)	7	6	7 (3.6)	14 (7.3)	172 (89.1)	4	6
Because of my conversation with the doctor, I understand how the treatment can benefit me	12 (5.6)	12 (5.6)	192 (88.9)	23	6	4 (2.2)	10 (5.4)	172 (92.5)	11	6
I think the treatment will be worth it for me	10 (4.7)	17 (8.0)	196 (87.3)	26	6	4 (2.2)	18 (9.8)	161 (88.0)	14	6
Follow-up consultations										
I knew which problems I wanted to discuss with the doctor	11 (5.1)	10 (4.7)	194 (90.2)	33	7	5 (3.4)	10 (6.9)	130 (89.7)	35	7
I discussed with the doctor the topics I wanted to discuss	10 (4.7)	14 (6.5)	190 (88.8)	34	6	8 (5.7)	9 (6.4)	123 (87.9)	40	6
There was enough attention for the complaints that I still have	7 (3.2)	6 (2.7)	206 (94.1)	29	7	2 (1.4)	2 (1.4)	144 (97.3)	32	7
My complaints have lessened since start of my treatment	21 (8.9)	36 (15.3)	179 (75.8)	12	7	19 (11.5)	25 (15.1)	121 (73.3)	15	6
In my opinion, my treatment is worth it for me*	7 (2.9)	10 (4.2)	222 (92.9)	9	7	6 (3.4)	12 (6.9)	157 (89.7)	5	7
Usefulness of the PRAQ										
The PRAQ-report was useful for preparing my consultation[c]	-	-	-	-	-	6 (6.5)	22 (23.9)	64 (69.6)	2	6
The PRAQ-report was useful during my consultation[d]	-	-	-	-	-	3 (2.3)	18 (13.6)	111 (84.1)	29	6

* Significant difference between pre- and post-implementation (p=.005, Mann-Whitney U test)
a. Scale 1-7 (1 = completely disagree, 2 = disagree, 3 = disagree a little, 4 = neither agree nor disagree, 5=agree a little, 6=agree, 7= completely agree)
b. "not applicable" (see supplementary file 4) or missing
c. Showing results for patients who indicated they had seen the PRAQ-report before their (intake or follow-up) consultation (n=94)
d. Showing results of patients who indicated the PRAQ-report was shown during their (intake or follow-up) consultation (n=161)

The main difference between pre- and post-implementation scores lies in distribution between scores 6 and 7 ('agree' and 'completely agree'), with 68.2% of pre-implementation patients giving a score of 7, and 54.3% of post-implementation patients giving a score of 7. The other statements showed no obvious or statistically significant differences in the level of agreement pre- and post-implementation.

Patients showed high agreement with the two statements about the usefulness of the PRAQ-report, particularly regarding its use during a consultation (Table 4). However, not all patients had completed the PRAQ and seen the PRAQ-report before or during their consultation. Patients who did not look up the PRAQ-report before their consultation may also have been the ones less interested in using the PRAQ-report, so the reported results may be somewhat biased towards are more positive evaluation (Table 5).

Table 5 Percentage of patients that completed and viewed the PRAQ, and patient opinion on usefulness PRAQ

	Intake (n=197)	**Follow-up (n=180)**
Completed PRAQ before consultation	77.7%	51.1%
Seen PRAQ-report before consultation[a]	40.0%	44.4%
Seen PRAQ-report during consultation[a]	74.1%	60.2%

a. This percentage is a sub-percentage of the patients who indicated they completed the PRAQ

3.3 Electronic health record results

125 patients were included in the pre-implementation group, and 124 other patients in the post-implementation group. Patient characteristics are described in Table 6. No differences were found with regard to how many patients with OSA received non-medical treatment (either no treatment at all or referral to a psychologist (Table 7)), or in the number of patients for whom treatment was adjusted at the first follow-up consultation after starting CPAP treatment (Table 8).

In both groups, 98 patients were prescribed CPAP. Patient characteristics did not differ between the two groups of patients with CPAP (data not shown). Compliance with CPAP treatment did not differ between the two groups (Table 8).

Table 6 Patient file study: patient characteristics

	Pre-implementation (n=125)	Post-implementation (n=124)
Age (SD)	55,4 (12,0)	56,6 (15,7)
Gender	68% male	67,7% male
BMI (SD)	31,4 (6,5)	30,8 (6,1)
AHI (SD)	23,1 (16,1)	25,0 (18,5)
AHI < 15	40,8%	33,9%
ESS (SD)	8,0 (4,8)	7,4 (5,0)
Start with CPAP at intake	78,4%	79,0%

Table 7 Treatment choice at intake[a]

	Pre-implementation (n=125)	Post-implementation (n=124)	Pre-implementation, AHI <15 (n=51)	Post-implementation, AHI <15 (n=42)
Medical treatment for OSA (incl CPAP) (n, %)	123 (98.4)	123 (99.2)	49 (96.1)	41 (97.6)
No medical treatment for OSA (n, %)	2 (1.6)	1 (0.8)	2 (3.9)	1 (2.4)
Referred to psychologist (n, %)	2 (1.6)	0 (0)	2 (3.9)	0 (0)
No treatment (n, %)	0 (0)	1 (0.8)	0 (0)	1 (2.4)

a. If nothing is indicated, no significant difference was found.

Table 8 Treatment adjustments and compliance in patients with CPAP at the first follow-up[a]

	Pre-implementation of PRAQ (n=98)	Post-implementation of PRAQ (n=98)	Missings
Adjustment of current treatment	45	36	N/A[b]
Switch to different treatment	5	9	N/A[b]
Referral to different specialization	6	2	N/A[b]
CPAP compliance[c] (SD)	5:47 hrs (2:11)	5:53 hrs (2:10)	Pre-impl.: 11 Post-impl: 7
CPAP compliance <4hrs	25.0%	27.5%	Pre-impl.: 11 Post-impl: 7
Stopped CPAP treatment	4.1%	5.1%	N/A[b]

a. If nothing is indicated, no significant difference was found.

b. If nothing was noted down in the patient health record, it was assumed this did not take place. Therefore missings are not applicable.

c. Hours of CPAP use by patients who had stopped treatment altogether (see "stopped CPAP treatment") are not included in this number.

4. DISCUSSION

This exploratory study showed limited success regarding the uptake of the PRAQ in the daily clinical practice of sleep centers, and the improvement of patient-centeredness of care. From the interviews it became clear that most patients were willing to complete the PRAQ and were generally positive about the usefulness of the PRAQ before the consultation (e.g. because of feeling more informed) and during the consultation (due to the clear visual representation of their problems). This may therefore have lead to some improvement of preparation for the consultation by patients, and better communication, though this is not reflected in the results of the patient survey. Amongst healthcare professionals the willingness to use the PRAQ-report in consultations differed, as the perceived need was minimal. Most of the professionals that used the PRAQ also reported that the impact on their consultations was minor. Therefore, it is not surprising that comparison of health records pre- and post-implementation of the PRAQ did not show any differences in treatment choice and CPAP compliance.

The interviews showed that the professionals mostly felt that they already sufficiently address the "symptom-like" topics of the PRAQ (sleepiness, problems at night) in their usual care, in the context of setting a diagnosis. The topics of the PRAQ that are not necessary for setting a diagnosis, but could potentially be used to motivate patients for their treatment, were not seen as essential to discuss by many professionals. The limited perceived benefit of the PRAQ is likely also mitigated by the fact that many steps of the care process have to be covered during the intake consultation, including discussing the sleep study results and choosing a treatment. This leaves little extra time to discuss a patient's quality of life and detailed treatment goals. Furthermore, burden of disease plays a limited role in setting a diagnosis when AHI≥15, due to views on strict medical necessity of treatment, but also due to the driving ban for untreated patients. Therefore, adding the PRAQ to the current practice for OSA does not appear to be a sufficient trigger to increase attention to quality of life issues.

Patients generally held a more positive view towards the usefulness of the PRAQ. From the interviews it became clear that completing the PRAQ has the potential to give patients more insight into their OSA-related health complaints and encourages communication between family members. Furthermore, the patient survey results indicated that patients thought the PRAQ-report was useful for their preparation for the consultation and (when it was used by the healthcare professional) during the consultation.

Agreement to the patient survey statement "I feel like my treatment is worth it for me" was significantly lower on the post-implementation survey than on the pre-implementation survey. The main difference was in the number of patients indicating "agree" versus "completely agree", meaning both pre-and post-implementation of the PRAQ patients were very positive about their treatment. This being an exploratory study, statistically significant results should be interpreted with caution, and we deem the relevance of this finding to be limited.

There appears to be room for improvement of communication around the PRAQ, as there was confusion for some patients around the necessity of still discussing symptoms

during the consultation. Whereas some patients seemed to be more interested in hearing their sleep study results than talk about their symptoms, for the healthcare professionals hearing about the patient's symptoms in their own words is an essential part of the diagnosis. It may be beneficial to communicate the purpose of the PRAQ more clearly in the invitation email, and/or to instruct professionals to, at the beginning of their consultation, mention the PRAQ to patients and how its results will be addressed. More in-depth discussion with the field about what is most suitable or desirable in this context is needed.

In the past few years, several similar initiatives involving PROMs have been introduced in The Netherlands, such as the Assessment of Burden of COPD (ABC) tool [34], the Nijmegen Clinical Screening Instrument for COPD [35], the QLIC-ON PROfile for children [36], and MyIBDcoach for patients with inflammatory bowel disease [37]. Studies into these applications show promising results regarding their benefits [37, 38] despite some resistance from professionals who do not believe in the added benefit or believe the tool would be more useful for different professionals within the care pathway [39, 40]. However, the healthcare professionals' skepticism about the potential benefits of the PRAQ seems to be more extensive. Potentially, professionals will see greater benefit of the PRAQ in the context of the recently released new guidelines for OSA [30] with their greater emphasis on (improving) burden of disease, which were not yet available at the time of this study. However, the question remains whether a more "holistic" approach to caring for OSA patients fits within the current setting of relatively short intake consultations which take place after the patients' diagnostic sleep study. It may be necessary to move towards a reorganization of care: for example to plan the intake consultations before the sleep study to allow for more focus on the individual patients' symptoms and problems, and to specifically evaluate the necessity of doing a diagnostic sleep study. Additionally, integrating the PRAQ in the electronic health record will help professionals fit the PRAQ-report better into their workflow.

Another option that can be explored is to adapt the PRAQ itself or the context in which it is used, in order to fit better to healthcare professionals' preferences. For example, an option would be to remove the domains of the PRAQ focused on symptoms that are (nearly) always discussed already, and instead put the focus on the additional domains. It is also possible to distribute the PRAQ to a more select group of patients, for example by moving the first measurement moment to the follow-up consultation, therefore targeting only patients with a diagnosis and treatment. It could then be used to identify those patients still experiencing problems. Downside to both of these adaptations is that they limit the option to monitor changes over time on all domains that are relevant for patients with OSA, while monitoring over time is what most interviewed healthcare professionals are interested in. Not having a baseline measurement would also limit the options to usefully study the PRAQ data on aggregate level. It may be most feasible to let sleep centers decide how they want to use the PRAQ in the context

of what is desirable to them, which may also evolve over time. It is hoped that they will also take into account the patient perspective when deciding how to use the PRAQ.

Strengths and limitations of the study

The major strength of this study is that we used mixed methods, which provides insight into the reasons why the PRAQ does not work as intended. Many other studies on PROMs study only *whether* a PROM works, rather than why or how.

There are also limitations to the study. First, the survey used for this study was not tested and maybe not discriminative enough to show differences between the groups pre- and post-implantation of the PRAQ. Potentially, patients who have not completed the PRAQ do not know that, for example, their preparation for the consultation could maybe have been better than it currently was. Second, electronic health records were only studied in one of the included sleep centers. However, considering the information we collected in the interviews, we do not expect that we would have found different results in either of the other two sleep centers. Third, though technically there was enough time in this study for professionals to also use the PRAQ during the first follow-up consultation, practical implementation issues as well as a lack of initiative from healthcare professionals to actively check whether a follow-up PRAQ was available meant that it was not used often at this time point. Therefore we did not gain much insight into the potential use of the PRAQ for follow-up consultations. Lastly, only patients who looked up the PRAQ-report could give an opinion on its usefulness for preparing the consultation in the survey. However, patients who did not look up the PRAQ-report may also be generally less interested in these kinds of tools and, if they had looked it up, may have experienced it as less useful. Additionally, patients who have a more positive opinion on the PRAQ may be more likely to complete the items on its usefulness.

Conclusions

Using the PRAQ in the daily clinical practice of OSA is viewed as useful by patients, but the enthusiasm of healthcare professionals differed per individual and was generally not very great. Implementation of the PRAQ does not seem a sufficient trigger to focus more attention to quality of life during consultations, and in current practice does not show impact on treatment choice or CPAP compliance. However, new Dutch guidelines for OSA care that have recently been published may lead to a greater emphasis on quality of life for patients, making the integration of the PRAQ in clinical care potentially more useful.

REFERENCES

1. Valderas, J.M., J. Alonso, and G.H. Guyatt, *Measuring patient-reported outcomes: moving from clinical trials into clinical practice.* Med J Aust, 2008. **189**(2): p. 93-4.
2. Fleischmann, M. and B. Vaughan, *The challenges and opportunities of using patient reported outcome measures (PROMs) in clinical practice.* Int J Osteopath Med, 2018: p. 1-6.
3. Howell, D., et al., *Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors.* Ann Oncol, 2015. **26**(9): p. 1846-58.
4. Greenhalgh, J., et al., *Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care.* 2017: Southampton (UK).
5. Van Der Wees, P.J., et al., *Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries.* Milbank Q, 2014. **92**(4): p. 754-75.
6. Valderas, J.M., et al., *The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature.* Qual Life Res, 2008. **17**(2): p. 179-93.
7. Boyce, M.B. and J.P. Browne, *Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? A systematic review.* Qual Life Res, 2013. **22**(9): p. 2265-78.
8. Holmes, M.M., et al., *The impact of patient-reported outcome measures in clinical practice for pain: a systematic review.* Qual Life Res, 2017. **26**(2): p. 245-257.
9. Greenhalgh, J., et al., *How do aggregated patient-reported outcome measures data stimulate health care improvement? A realist synthesis.* J Health Serv Res Policy, 2018. **23**(1): p. 57-65.
10. Bjornsdottir, E., et al., *The Prevalence of Depression among Untreated Obstructive Sleep Apnea Patients Using a Standardized Psychiatric Interview.* J Clin Sleep Med, 2016. **12**(1): p. 105-12.
11. O'Donoghue, N. and E. McKay, *Exploring the impact of sleep apnoea on daily life and occupational engagement.* Br J Occup Ther, 2012. **75**(11): p. 609-516.
12. Reishtein, J.L., et al., *Sleepiness and relationships in obstructive sleep apnea.* Issues Ment Health Nurs, 2006. **27**(3): p. 319-30.
13. Bradley, T.D. and J.S. Floras, *Obstructive sleep apnoea and its cardiovascular consequences.* Lancet, 2009. **373**(9657): p. 82-93.
14. Chan, A.S., C.L. Phillips, and P.A. Cistulli, *Obstructive sleep apnoea--an update.* Intern Med J, 2010. **40**(2): p. 102-6.
15. Young, T., P.E. Peppard, and D.J. Gottlieb, *Epidemiology of obstructive sleep apnea: a population health perspective.* Am J Respir Crit Care Med, 2002. **165**(9): p. 1217-39.
16. Senaratna, C.V., et al., *Prevalence of obstructive sleep apnea in the general population: A systematic review.* Sleep Med Rev, 2017. **34**: p. 70-81.
17. Medicine, A.A.o.S., *International Classification of Sleep Disorders. Diagnostic and Coding Manual.* , A.A.o.S. Medicine, Editor. 2014.
18. Committee, D.C.S., *International Classification of Sleep Disorders: Diagnostic and Coding Manual.*, A.S.D. Association, Editor. 1990: Rochester.
19. Macey, P.M., et al., *Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients.* PLoS One, 2010. **5**(4): p. e10211.
20. Tam, S., B.T. Woodson, and B. Rotenberg, *Outcome measurements in obstructive sleep apnea: beyond the apnea-hypopnea index.* Laryngoscope, 2014. **124**(1): p. 337-43.
21. Kingshott, R.N., et al., *Does arousal frequency predict daytime function?* Eur Respir J, 1998. **12**(6): p. 1264-70.

22. Turnbull, C.D. and J.R. Stradling, *To screen or not to screen for obstructive sleep apnea, that is the question.* Sleep Med Rev, 2017. **36**: p. 125-127.
23. Van Dongen, H.P., et al., *Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability.* Sleep, 2004. **27**(3): p. 423-33.
24. Abuzaid, A.S., et al., *Meta-Analysis of Cardiovascular Outcomes With Continuous Positive Airway Pressure Therapy in Patients With Obstructive Sleep Apnea.* Am J Cardiol, 2017. **120**(4): p. 693-699.
25. Yu, J., et al., *Association of Positive Airway Pressure With Cardiovascular Events and Death in Adults With Sleep Apnea: A Systematic Review and Meta-analysis.* JAMA, 2017. **318**(2): p. 156-166.
26. Marin, J.M., et al., *Association between treated and untreated obstructive sleep apnea and risk of hypertension.* JAMA, 2012. **307**(20): p. 2169-76.
27. Barbe, F., et al., *Effect of continuous positive airway pressure on the incidence of hypertension and cardiovascular events in nonsleepy patients with obstructive sleep apnea: a randomized controlled trial.* JAMA, 2012. **307**(20): p. 2161-8.
28. McNicholas, W.T., *Diagnostic criteria for obstructive sleep apnea: time for reappraisal.* J Thorac Dis, 2018. **10**(1): p. 531-533.
29. McNicholas, W.T., et al., *Challenges in obstructive sleep apnoea.* Lancet Respir Med, 2018. **6**(3): p. 170-172.
30. NVALT, *Richtlijn diagnostiek en behandeling van obstructief slaapapneu (OSA) bij volwassenen.* 2017, Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose: Richtlijnendatabase.nl.
31. Abma, I.L., et al., *The development of a patient-reported outcome measure for patients with obstructive sleep apnea: the Patient-Reported Apnea Questionnaire (PRAQ).* Journal of Patient-Reported Outcomes, 2017. **1**(14).
32. Abma, I.L., et al., *Instrument completion and validation of the patient-reported apnea questionnaire (PRAQ).* Health Qual Life Outcomes, 2018. **16**(1): p. 158.
33. Grunstein, R.R., *Sleep-related breathing disorders. 5. Nasal continuous positive airway pressure treatment for obstructive sleep apnoea.* Thorax, 1995. **50**(10): p. 1106-13.
34. Slok, A.H., et al., *Development of the Assessment of Burden of COPD tool: an integrated tool to measure the burden of COPD.* NPJ Prim Care Respir Med, 2014. **24**: p. 14021.
35. Peters, J.B., et al., *Development of a battery of instruments for detailed measurement of health status in patients with COPD in routine care: the Nijmegen Clinical Screening Instrument.* Qual Life Res, 2009. **18**(7): p. 901-12.
36. Engelen, V., et al., *Development and implementation of a patient reported outcome intervention (QLIC-ON PROfile) in clinical paediatric oncology practice.* Patient Educ Couns, 2010. **81**(2): p. 235-44.
37. de Jong, M.J., et al., *Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial.* Lancet, 2017. **390**(10098): p. 959-968.
38. Slok, A.H., et al., *Effectiveness of the Assessment of Burden of COPD (ABC) tool on health-related quality of life in patients with COPD: a cluster randomised controlled trial in primary and hospital care.* BMJ Open, 2016. **6**(7): p. e011519.
39. Schepers, S.A., et al., *Real-world implementation of electronic patient-reported outcomes in outpatient pediatric cancer care.* Psychooncology, 2017. **26**(7): p. 951-959.
40. Slok, A.H., et al., *'To use or not to use': a qualitative study to evaluate experiences of healthcare providers and patients with the assessment of burden of COPD (ABC) tool.* NPJ Prim Care Respir Med, 2016. **26**: p. 16074.

Appendix 1: The PRAQ-report and its implementation in clinical practice

In the PRAQ-report, the results of each of the ten PRAQ-domains are shown in the form of a colored smiley, ranging from green (patient indicated very few problems) to dark red (patient indicated a lot of problems). Domain scores over time and individual item scores are shown on subsequent pages of the PRAQ-report. The included domains were: symptoms at night, sleepiness, tiredness, daily activities, unsafe situations, memory and concentration, quality of sleep, emotions, social activities, and health concerns. The PRAQ also contains a set of "intake questions" that were designed together with the participating centers and aimed to replace the diagnostic intake questionnaires that the centers usually distribute to all their new patients. This involved more factual, broader questions to help professionals in setting a correct diagnosis.

The PRAQ was distributed via a secure online platform (VitalHealth QuestManager) which sent out email invitations to a patient to complete the PRAQ at ten and (if the PRAQ was not yet completed) three days before the patient's consultation. After completion of the PRAQ, patients and healthcare professionals both had the ability to access the PRAQ-report directly from the online platform.

Individual implementation plans for collecting email addresses of patients, creating patient accounts, and entering consultation dates were developed for each study center to optimally fit their usual work flow.

Healthcare professionals received information about the content of the PRAQ and PRAQ-report, and instructions and a short training in how to use QuestManager. They were then encouraged to integrate the PRAQ into their own workflow in whichever way each individual professional found most convenient. After approximately two months of using the PRAQ, the researchers organized a meeting in each sleep centre in which the healthcare professionals were invited to discuss how they were using the PRAQ-report in their practice, in order to exchange ideas and potentially adjust their way of using the PRAQ.

Appendix 2: information about the interviews and coding

Interviewers IA and MN held the patient interviews based on the interview guides, mostly together but MN did some patient interviews alone, and IA did some patient and some professional interviews alone. IA had some training in qualitative research/interviewing, and participated in qualitative study with interviews before. MN did not have official training but received some interview training from IA. IA was the developer of the PRAQ and PRAQ-report, and the healthcare professionals were aware of this, which may have lead to bias. However, this was specifically addressed before the start of the interviews, reminding the interviewees that this was scientific research and the researchers were looking for honest opinions in order to learn more about the application of PROMs in clinical practice, and negative opinions were also welcome. The patients were not told that IA was the developer of the PRAQ.

Patient recruitment took place in two different ways:

- Patients were approached via email by the sleep center before their scheduled consultation. The email was sent directly via the online platform as an added message to the invitation to complete the PRAQ, or by a team member of the sleep center.
- All patients scheduled on a certain specific day for a specific healthcare professional that had completed the PRAQ, were invited by their healthcare professional to participate directly after their consultation.

18 patients were interviewed face to face at the sleep center after their consultation in a private room; 9 patients were interviewed over the phone for convenience reasons. The patient interviews lasted on average 15 minutes. Healthcare provider interviews lasted on average 44 minutes and were all held at the sleep center. All interviews were audiotaped, transcribed verbatim and anonymised. All interviewees were provided with information about the study and signed an informed consent form or gave verbal informed consent on the audiotape. Transcripts of the interviews were not provided to the interviewees. Analysis of the interviews took place with a phenomenology approach via open coding, with different code books for patients and healthcare providers. IA and MN first coded five interviews independently for both patients and healthcare professional interviews. A researcher (IG) experienced in qualitative research and knowledgeable about PROMs, but not involved in the study, coded one of the healthcare professional interviews independently. IG, IA, MN and PW held a collaborative coding session in which the code books were constructed. MN analyzed all remaining interviews, which IA then read and checked to the code book. When there was a disagreement about the coding, IA and MN reached consensus.

Appendix 3: Patient survey. Version: intake consultation, post-implementation of the PRAQ.

When answering the questions, keep in mind the consultation that you just attended.

	completely disagree	disagree	disagree a litte	don't agree, don't disagree	agree a little	agree	completely agree	not applicable*
1. I knew which problems I wanted to discuss with the doctor	☐	☐	☐	☐	☐	☐	☐	☐
2. I discussed with the doctor the topics I knew I wanted to discuss beforehand	☐	☐	☐	☐	☐	☐	☐	☐
3. Because of my conversation with the doctor, I understand better what causes my health complaints or problems	☐	☐	☐	☐	☐	☐	☐	☐
4. The doctor and I chose the treatment together, or together chose not to treat my apnea	☐	☐	☐	☐	☐	☐	☐	☐
5. Because of my conversation with the doctor, I understand how the treatment can benefit me	☐	☐	☐	☐	☐	☐	☐	☐
6. I think the treatment will be worth it for me	☐	☐	☐	☐	☐	☐	☐	☐

Clarifications:
* I have no complaints or problems / I did not think of anything to discuss
** I don't have sleep apnea / no choice was made yet

Questions about the PRAQ-report

7. Your sleep center asked you to fill out a questionnaire before attending your consultation, about your health complaints and daily functioning (the PRAQ). Did you complete this questionnaire?

☐ Yes

☐ Partially (go to question 12 at the bottom of this page)

☐ No (go to question13, on the next page)

The results of the questionnaire are summarised in a report, with on the first page smileys for each topic. The questions below are about whether you looked at this report, and whether you thought this was useful. Answer each question in the way that fits you or your situation best. **Please answer the questions as well if you have not seen the report.**

8. I looked at the report before my consultations with the doctor.

☐ Yes, elaborately

☐ Yes, briefly

☐ No, not important or didn't get around to it

☐ No, I didn't know that there was a report or how to open it

9. The report was shown during the consultation with the doctor (for example, you looked at the smileys together and discussed your health complaints)

☐ Yes, elaborately

☐ Yes, briefly

☐ No, not at all

10. I thought the report was useful as preparation of my consultation with the doctor
(if you did not look at the report beforehand, you may answer "not applicable")

completely disagree	disagree	disagree a litte	don't agree, don't disagree	agree a little	agree	completely agree	not applicable
☐	☐	☐	☐	☐	☐	☐	☐

11. I thought the report was useful during my consultation with the doctor
(If the report was not shown during the consultation, you may answer "not applicable")

completely disagree	disagree	disagree a litte	don't agree, don't disagree	agree a little	agree	completely agree	not applicable
☐	☐	☐	☐	☐	☐	☐	☐

12. Is there anything else about the PRAQ-questionnaire or the report that you would like to share? We are happy to hear your opinion.

We would also like to know something about you.

13. What is your age? ☐☐☐ years old

14. Gender: ☐ man ☐ woman

15. What is the highest level of education you finished with a diploma?

☐ No education (did not finish primary school)

☐ Primary school

☐ Basic vocational education (LTS, LEAO)

☐ General secundary education (MAVO, VMBO)

☐ Intermediate vocational education (MTS, MEAO, MBO)

☐ Senior secundary general education or pre-university education (HAVO, VWO, grammar school)

☐ Higher professional education (HBO, HEAO, HTS)

☐ University

☐ Other, which is: _____

16. How bothered are you by the health complaints or problems for which you attended the sleep center today?

Not bothered at all ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 Very bothered

17. Did the doctor diagnose you with sleep apnea?

☐ Yes, I have sleep apnea ☐ No, I do not have sleep apnea ☐ Don't know (yet)

18. If you answered "yes" to the previous question, was a treatment chosen and if so, which one?

☐ No treatment

☐ Yes, CPAP (mask)

☐ Yes, MRA (device over teeth)

☐ Yes, lifestyle advice

☐ Other, which is: _____

CHAPTER 7

General Discussion

1 AIM AND OVERVIEW

The aim of this thesis was to provide more insight into how individual PROM results work when implemented in clinical practice. This was studied in the context of obstructive sleep apnea (OSA), a condition in which breathing stops during sleep cause a variety of symptoms and functional problems. The specific aims of this thesis were to study whether a PROM of sufficient quality is available for patients with OSA, which measures OSA-related quality of life; if no existing PROM measuring OSA-related quality of life is of sufficient quality: to develop a new PROM for patients with OSA specifically for use in clinical practice, with the goal to be suitable for use on an individual patient level and on aggregate level; to develop a 'patient-friendly' way of presenting the results of the PROM, in order to make them easy to interpret; to assess the validity, reliability and responsiveness of the Dutch PROM; and to study the impact of individual results of the PROM on OSA care as well as reasons for this (lack of) impact.

First, I will discuss the main findings of this thesis. Then I will provide insight into why some of our expected findings were not met, and the likely role that healthcare professionals, patients, the development of the PRAQ, the context, technology, and efforts of the research team to improve implementation may have played in this. Any methodological limitations will be discussed in the paragraphs about these topics. I then study whether our findings are in line with what is already known about PROMs in clinical practice. I finish with some recommendations and future directions regarding PROMs in clinical OSA practice, and the evaluation of the impact of PROMs in clinical practice.

2 MAIN FINDINGS

Chapters two to five of this thesis describe why and how we developed the PRAQ, a PROM measuring OSA-related quality of life. First, we studied the measurement properties of existing PROMs validated in patients with OSA, and concluded that none of these was suitable and of sufficient quality in its current form (chapter two). Therefore, we developed a new PROM, called the PRAQ, with the input of patients and healthcare professionals. During the development, we focused on the PROMs potential usefulness in clinical practice (chapter four). In chapter four we also describe the development of the PRAQ-report, an overview of an individual patient's PRAQ results in which colored smilies show which domains are more and less problematic for that patient. A final version of the PRAQ is presented in the validation article, where we also report the acceptable measurement properties of the PROM with regard to reliability, validity and responsiveness (chapter five).

In chapter 6 of this thesis we present the outcomes of an explorative pilot study for the impact of the PRAQ on clinical practice, for which the PRAQ was implemented in three Dutch sleep centers (chapter 6). We had several expectations with regard to how the PRAQ

could impact clinical practice. First, we expected that completing the PROM would educate and empower patients with suspected OSA by making them more aware of which of their symptoms and problems in daily life can be related to OSA. Interviews with patients showed that this expectation was met: completing the PRAQ has the potential to teach patients about the potential impact of OSA on their lives, and through discussing the questions of the PRAQ with their family it also provided more insight into their own situation. Second, we expected that the PRAQ would shift focus during a consultation from the medical facts towards the subjective experiences and problems of patients with OSA. The PRAQ-report could be used as the starting point of a conversation about aspects of health-related quality of life to achieve this goal. This could help determine the burden of disease for a patient, encourage healthcare providers to set personal treatment goals for patients and help motivate patients for their (medical or for example lifestyle-focused) treatment. Additionally, a broader view of a patient's health could lead to referrals to other healthcare professionals, such as psychologists. In that way, care could also get a more 'holistic' focus where solving health and related problems take precedence over (only) reducing experienced breathings stops. Third, we tested the hypotheses that through abovementioned mechanisms, patients feel more motivated for their treatment and therefore show a higher compliance with treatment.

The results of the pilot study show that these hypotheses about the impact of the PRAQ were not met for a majority of healthcare professionals. Interviews with healthcare professionals showed that they were not all willing to use the PRAQ, and those that did use the PRAQ, reported that they let the PRAQ play only a minor role during their consultations. We also performed a complimentary electronic health record study, before and after implementation of the PRAQ, where we studied: treatment choice; referrals to other healthcare professionals; and patients' compliance with treatment. We had expected that these aspects of care could change based on the changes in the consultations. However, we did not find any differences, which is in line with our qualitative findings that there was hardly any change in the consultations. The availability of individual OSA-related quality of life data did therefore not seem to impact care.

3 THE PRAQ IN CLINICAL PRACTICE: WHY DID IT NOT WORK AS INTENDED?

In chapter one, I mentioned two existing models that depict how PROMs feedback can impact patient care and patient health outcomes: a general model by Greenhalgh et al. [1], and a model specifically focused on the use of PROMs for patients with chronic conditions by Santana et al. [2]. In both models, changes in communication have an important (and in Santana's model even crucial) role in the chain of effects of PROMs feedback. Considering that in my pilot study the healthcare professionals mostly reported only minor changes in their communication with patients during consultations, it is clear that finding any impact of the PRAQ on treatment

choice, referrals to other healthcare professionals, or patients' compliance with treatment in the pilot study was unlikely. It remains uncertain whether the PRAQ could impact these process measures – and potentially also health outcomes - if healthcare professionals do change their communication.

The main question which remains, and which has also been deemed highly important in the literature to provide more insight into the working mechanisms of PROMs [1], is *why* the PRAQ did not change communication and patient care as expected. This may be related to the (development of) the innovation itself, and is likely also intertwined with the success of implementing it in clinical practice and getting the healthcare providers to use it. Therefore we have used the determinants of change that may influence implementation identified in Grol et al. [3] as framework for which factors are relevant to discuss: the OSA healthcare professionals and the context in which they work, the patients, the nature of OSA, technology, the PRAQ itself, and efforts of the research team to improve implementation. In the following paragraphs I will discuss for each of these topics whether or not they were a likely contributor to the lack of changes impact of the PRAQ and how or why this may be the case. The discussion of the topics is based on the data gathered in interviews and discussions with the sleep centers after the study.

3.1 Healthcare professionals

During the interviews with healthcare professionals it became clear that there are several interrelated explanations as to why the PRAQ had little impact on communication during consultations. The first potential explanation concerns the amount of available consultation time in relation to the actions that need to be taken during that consultation. The 20-30 minutes intake consultation (depending on the sleep center) for patients with OSA takes place after the diagnostic sleep study. This consultation is used for 1) taking a patient history focused on symptoms that together with the sleep study result leads to a diagnosis, 2) explaining the sleep study results to the patient, 3) explaining the diagnosis to the patient, and if OSA is diagnosed: 4) explaining the treatment options, and 5) choosing a treatment with the patient. In other words, allocating more time to discuss quality of life would mean that there is less time to explain important things to patients and for shared decision-making, while time is already tight.

Then there is a second reason: the PRAQ does not seem to fit with most healthcare professionals' ideas of the care they should be delivering. From the interviews it became clear that many healthcare professionals are not used to actively bringing up topics with patients (e.g. emotions) if they do not feel that they are useful for setting the diagnosis during an intake consultation, and they are also not necessarily convinced that this is useful. Focusing on the medical aspects (number of breathing stops per hour) and the most common symptoms is often considered sufficient: the main responsibility is to set the medical diagnosis and start a treatment. The impact of symptoms on a patient's life may surface during the consultation, but

is a secondary issue. The same seems to be true for the follow-up consultation after starting treatment: the treatment is considered successful if the number of breathing stops per hour is reduced, symptoms are generally improved and the treatment is tolerated well. The focus of the follow-up consultation is often on the tolerance and potential adjustments of treatment; discussion of remaining problems and whether they can be treated is – unless the patient brings them up actively – often minimal.

3.2 The context of OSA care

The broader context in which OSA care is provided is also relevant, for example in terms of reimbursement structures and regulations. Financially it is more beneficial for a sleep center if healthcare professionals reach a quick diagnosis and prescribe (medical) treatment, rather than spending time discussing quality of life. This reward for prescribing medical treatment also discourages the potential choice of opting for no treatment or non-medical treatment such as lifestyle coaching, as has been shown in previous research [4]. Furthermore, there is one current regulation that specifically impacts OSA care. It states that patients with ≥15 breathing stops per hour are not allowed to drive motorized vehicles until they have had medical treatment of their OSA for at least two months. For patients for whom driving is important, this in practice eliminates the option of choosing no treatment or non-medical treatment, even when they experience few symptoms. Therefore, even if the PRAQ would lead to changes in communication, this regulation limits the potential for changes in patient management.

3.3 The patients

The initiative to discuss the PRAQ results can come from the patient as well as the healthcare professional. During our patient interviews we did not specifically ask whether patients actively brought up certain topics during their consultation, triggered by the completion of the PRAQ or by looking at the PRAQ-report, so our information on this topic is not complete. Only one patient brought up, out of their own accord, that they felt the PRAQ supported them in asking more questions. However, what was striking about the interviews with other patients is that they did not seem overly eager to discuss their symptoms and problems in daily life with their healthcare professional. Some patients were under the impression that discussion of symptoms during the intake consultation was not necessary after completing the PRAQ: they seemed to think that the PRAQ results would provide the healthcare provider with enough information regarding their symptoms. Many healthcare professionals also reported this and mentioned resistance from patients to discuss the topics of the PRAQ-report again. Rather than leading to more discussion about quality of life, it may therefore be the case that the completion of the PRAQ had the opposite effect. Since the PRAQ is sent out during the diagnostic process, this may have led patients to believe that the PRAQ is diagnostic tool, though this was not what was communicated. Additionally, many intake patients seemed very focused on hearing about the results of their sleep study, and learning whether they suffered from OSA or not, whereas

they were less focused on discussing their personal problems. In other words, it seems that the way healthcare professionals view the intake consultation – namely, mostly focused on the diagnosis – may actually match the way patients view these consultations. Once the results of the sleep study are available, the 'medical' view of OSA seems to gain the upper hand for both parties.

This focus on the diagnosis may also have influenced the patient survey results, on which patients mostly indicated that they discussed all the topics they wanted to discuss with their healthcare professional. If patients expect their intake consultation to focus on discussing the diagnosis, then it seems likely that their expectations concerning discussion of their personal situation are low. The question is whether this means that more attention to quality of life is not useful from the patient perspective. It is understandable that patients want to know the results of a performed test, but in another context patients may consider it more desirable to discuss aspects of their quality of life. Separating the discussion of quality of life from the discussion about the sleep study results may change the patient perspective.

3.4 The development of the PRAQ

During the development of the PRAQ, we involved both patients and healthcare professionals. We included in our development team a board member of the Dutch patient organization for sleep apnea (ApneuVereniging) and a pulmonologist, and we also consulted larger groups of patients and healthcare professionals to determine which topics were relevant to discuss during a consultation. Our specific development goal was to optimize the PRAQ and PRAQ-report for its proposed use in clinical practice. This is an important step in making sure that the tool that is being developed for clinical practice will serve its needs. To our knowledge, no PROM has been developed with this particular purpose in mind, and in terms of serving the needs of patients this approach turned out successful.

However, even though we involved healthcare professionals during the development process, our main focus was on the needs of the patients and their views on what they would potentially wish to discuss with a healthcare professional. We involved healthcare professionals by asking them to complete a survey on which they indicated their agreement to several statements per potential domain of the PRAQ. These statements included that the professional would want to be aware if patients were experiencing problems regarding a certain topic (for example emotions), and that they felt at least partially responsible for helping to solve these problems. 72%-100% of healthcare professionals participating in the survey agreed to these statements for the ten domains of the PRAQ. However, considering the limited perceived usefulness of the PRAQ by professionals in our last study, it appears that expressing the wish to be aware of and help tackle certain problems may not be the same as wishing to actively bring up these topics, nor does it necessarily mean that it is feasible in practice to do so. Additionally, some of the responding healthcare professionals may have felt that they already address these problems sufficiently and do not need a tool to aid them, which was a view that some of our

interviewees expressed after implementation of the PRAQ. Here, we missed a chance to gain insight beforehand into how different healthcare professionals viewed our idea of using a tool like the PRAQ-report. Setting up a shared vision with healthcare professionals in an early development phase about how a PROM can and should aid clinical practice and what this could look like during a consultation, is a step is likely to benefit a successful implementation.

3.5 The nature of OSA as a 'different' chronic condition

Tools for clinical practice which include the results of PROMs have been developed for several conditions and target groups in the past decade, and they seem to have had a more positive reception amongst healthcare professionals than the PRAQ-report [5-10]. Conditions for which tools have been developed are for example for chronic obstructive pulmonary disease (COPD), irritable bowel syndrome (IBD), and for children with a wide range of chronic or longer-lasting conditions. What these conditions have in common is that they are chronic and involve problems over longer periods of time. PROMs can help detect and monitor problems that are relevant to these patients.

Differences in the characteristics and treatment of these diseases compared to OSA can help explain why the PRAQ was less successful. COPD and IBD differ from OSA in the sense that treatment is for a great part focused on relieving symptoms and improving quality of life, while the underlying condition (COPD or IBD) remains present. However, for OSA treatment is regularly successful in strongly reducing the number of breathing stops, in which case the main (physical) problem for which they attended the sleep centre can often be considered solved. A patient that is successfully treated for OSA will not necessarily continue to experience chronic symptoms or problems, even though OSA is considered a chronic disease. Therefore, usually OSA is a not a closely monitored disease once successful treatment has been established for an individual. The percentage of patients that continue to experience problems even when treatment appears successful is, however, not known; and the (low) estimation of this percentage by healthcare professionals may be an underestimation because they tend to focus on whether a patient is doing better on the whole without focusing on remaining problems.

From the healthcare professional perspective, this focus on only the general well-being of a patient could be viewed as sensible. If the number of breathing stops are sufficiently reduced with treatment and the patient is doing better, then it is possible that remaining problems either need more time to improve (the brain can potentially take up to two years to recover from the impact of OSA [11]) or that they are not directly caused by OSA. In neither scenario does this necessarily encourage healthcare professionals to act upon the remaining problems. However, there are options to give additional care to these patients: for example by exploring other sleep-related problems, considering whether there may be potential (undiscovered) comorbidities which may warrant referral back to the GP, or by referring the patient to a psychologist to learn how to cope better with their problems. Still, the (perceived) lack of tools for an OSA professional to act upon remaining problems within their own expertise

likely contributes to the lower experienced usefulness of this PROM for OSA compared to PROMs for other chronic diseases.

3.6 Technology

In existing literature about PROMs in clinical practice it is often mentioned that using the PROM results must fit well into the workflow of the healthcare professionals [12]. In the case of the PRAQ, this fit with the workflow was not optimal: the PRAQ-report was only available on a separate website which required the professionals to log in and bring up the patient's record in that system. The professionals did not consider this process complicated, but it did require some extra time and effort in each consultation. For some of the interviewed healthcare professionals, this was one of the factors that contributed to not (always) using the PRAQ. Being able to access the PRAQ-report directly from the electronic health record would potentially have encouraged these professionals to use the PRAQ more. However, considering how little impact the PRAQ had on consultations even when it was used, it seems unlikely that this would have had a big impact on the results of this study. Still, for potential future use of the PRAQ, and for PROMs in general, availability of the results in the patient record is an important aspect of fitting it into the workflow of the professionals.

We also made the PRAQ-report available to patients, which allowed them to see and reflect on their results (in comparison to their previous results, if available) before the consultation. Unfortunately, the way in which the website made its results available to patients was not very intuitive, which could not be solved in time for this study. This means that fewer patients viewed the PRAQ-report before their consultation than we would have hoped. Our survey in chapter six showed that approximately 40% of patients that had completed the PRAQ had seen the report before their consultation, of which approximately 70% indicated that they thought PRAQ-report was useful for preparing their consultation. Potentially, these results are biased because patients who do not find PROM information interesting or useful may not have looked at the PRAQ-report, and could therefore also not express their opinion on the report itself. Still, considering the positive opinion of those who did look at the report, the usefulness of sharing the PRAQ-report with patients before their consultation is worth exploring further.

3.7 Efforts of the research team to promote implementation

We provided a training and instructions to the healthcare professionals on how to use the website on which the PRAQ-report could be viewed, and how to interpret the PRAQ results based on the smileys. We did not aim to prescribe exactly how the healthcare professionals had to use the PRAQ-report, but we presented information and several suggestions on how to use the PRAQ to the professionals. We then asked the professionals at the sleep centers to set up their own, collective plan on how to start using the PRAQ in their consultations. Motivation of the professionals to work out this plan turned out to be low. They felt that they had to first see and try the PRAQ in their clinical practice before being able to estimate how

to best use it. Therefore, the research team set up a meeting 6-8 weeks after implementing the PRAQ with all involved healthcare professionals and several members of the research team, in the three participating sleep centers. In this meeting the professionals were invited to discuss their current use of the PRAQ, in order for them to learn from and inspire each other. However, attendance and success of the sessions varied, and no clear plans were formulated. The sessions did encourage some professionals who were not using the PRAQ to start using it.

Looking back on this process, we were too optimistic about the enthusiasm of the healthcare professionals to consider possibilities of using the PRAQ and help us in setting up a clear plan of use for the PRAQ. This is likely related to their ideas about the limited benefit of the PRAQ to their clinical practice. We had also expected professionals to be more willing to influence and learn from each other. In reality, our impression was that professionals had their own way of holding their consultations and while some were willing to try out changing their approach, others were not. The presence of a (more influential) clinical champion within the participating centers may have been beneficial in motivating the professionals to work with the PRAQ [3].

4 COMPARISON OF OUR FINDINGS TO EXISTING LITERATURE

Reviews that have tried to identify the factors that may hinder or benefit the effect of a PROM as intervention in clinical practice have mentioned many of the things that have been discussed in this chapter: the attitude of the healthcare professionals and the way they are instructed to use the PROM results [13, 14], whether the PROM provides new information, the opportunity costs for what professionals perceive to be more important aspects of care, and the technology and integration of PROM data in the electronic health record [12]. The content and type of the PROM has also been found to be a potentially hindering factor, for example when generic PROMs are employed (measuring general quality of life) that bear little connection to what patients and healthcare professionals consider important in their specific context [15]. In our study the PROM was closely aligned with the main problems of patients with OSA, but this was in itself not enough to make the intervention effective.

Apart from ideas about why PROM data may or may not work, there has also been an effort by Greenhalgh et al. [14] to summarize the currently available evidence on how PROMs may impact care. They have focused on the potential for PROMs to increase the discussion of patient concerns, either via the route of healthcare professionals bringing up these concerns more often, or because of patients bring up these concerns more often. The authors then came to a set of theories about how this increased discussion of concerns may be triggered, and 'probe' these theories by studying the results of empirical studies to test and refine them.

Our study has provided some insights into several of the theories from Greenhalgh's synthesis. One theory from the synthesis is that PROMs act as a tool to enable patients to raise

or share issues with clinicians. Three mechanisms could contribute to this according to the synthesis: patients engage in a process of self-reflection to identify what is important to them and what they want to talk about; the process of PROMs completion reminds patients to share or raise issues with clinicians; and the PROM also makes them feel that they have permission to raise these issues because the healthcare professional is interested. The synthesis found some (positive or negative) evidence for all three of these mechanisms. In our study, patients did reflect more on their situation, often together with their partner, but rather than planning to raise these issues, patients seemed to feel that sharing these issues via the PROM was sufficient. They expected that professionals were interested in the information they provided via the PROM, but often appeared to think this was for diagnostic purposes only (see also section 3.3).

Then, the synthesis presents a set of theories and a counter-theory regarding how healthcare professionals use PROMs. The theories state that PROMs feedback will lead to increased discussion of issues of health-related quality of life, which will then directly (because of changes to patient management) or indirectly (because of better communication) lead to improvement in patient well-being. Some studies were found to support improvement of patient well-being via the direct route, but not the indirect route where better communication itself will lead to better patient management. The counter-theory poses that healthcare professionals do not change their communication practices during the consultation. Studies are then presented in Greenhalgh's synthesis that supported this theory, to which we can add our own study. Our study supports the findings of the synthesis that healthcare professionals are more likely to discuss more medical issues, such as symptoms, than issues that reflect broader health-related quality of life, such as social functioning. Providing healthcare professionals with the information that there are issues with broader health-related quality of life for a patient, is not necessarily sufficient to trigger discussion of these issues.

5 THE FUTURE OF USING PROMS IN CLINICAL PRACTICE OF OSA

The main reasons of why the PRAQ did not have the expected impact seems related to perceptions of both professionals and patients about the necessity of discussing aspects of quality of life, in the context of short consultations and the nature of OSA as a chronic disease.

A separate intake consultation which takes place before the diagnostic sleep study would allow for more focus on determining which problems a patient experiences, without the instant focus on sleep study results. Interestingly however, these 'early' intake consultations were abolished relatively recently in some sleep centers (personal communication). There are benefits to merging the intake consultation with the consultation in which the diagnosis is set: it saves patients the time and effort of an extra visit to the hospital while they likely need to undergo a sleep study either way, and it makes the intake process more streamlined, which

for example allows healthcare professionals to see more new patients. This may save time and money for the sleep center. These benefits seem to have outweighed a potential loss of time to listen to a patient's perspective.

It will be interesting to see how this will develop in the future. In the past few years there has been an international discussion regarding new approaches to diagnose 'clinically relevant' OSA [16, 17], and how to achieve more personalized care for patients with OSA [18]. This discussion has also made its way into recent Dutch guidelines for OSA, in which it is recommended that there should be a greater focus on the presence of potentially related comorbidities, as well as the experienced burden of disease for individual patients. The goal of treatment is the improvement of these OSA aspects [19]. It is increasingly being recognized that having an in-depth conversation with the patient is crucial in this context [20]. The PRAQ could potentially fulfill a useful role in this new approach to OSA care.

In my view, the next step lies within healthcare practice rather than science. Since the effectiveness of the PRAQ in clinical practice has not been shown in this study, it will be up to sleep centers to decide whether they *consider* it to be potentially useful enough to warrant the effort and money of implementation and continuous measurement; and also whether they are willing and able to provide the context in which it can be employed in a useful way on an individual patient level. For example, by planning more time for consultations, planning a separate intake consultation before the sleep study, but also by deciding as a department to increase focus on the patient-centeredness of care. The sleep center will then have to evaluate whether it is worth continuing the collection and use of PROM data, by taking into account the healthcare provider perspective as well as the patient perspective. In addition use for individual patients, sleep centers may also find benefit of the PRAQ by using its results on an aggregated level to study outcomes of healthcare providers or different treatments.

Currently, one Dutch sleep center is using individual PRAQ data in its clinical practice, and has shown interest in also using this data on an aggregated level.

6 PROMS IN CLINICAL PRACTICE – CHALLENGES

Gathering information on the perspective of patients on their own health to monitor and evaluate provided care is gaining increasing interest. In the future, the use of PROMs in clinical practice will therefore likely grow. As discussed before, PROMs in clinical practice can be used on different levels. On the level of an individual patient, individual PROM results can be used to inform treatment choice and monitor outcomes of that patient. Additionally, also for individual patients, it is possible to use aggregated PROM data of similar patients to inform treatment choice (e.g. patients like you improve most when they undergo a certain treatment). On an aggregated level, PROM data can be used to improve quality of care, or for external transparency. Collecting PROM data for use only on an individual or only on an aggregated level

is possible, but collecting data for both purposes may have benefits [21]: it likely provides more motivation for patients to complete the PROM and it gives healthcare professionals a better 'feel' for the collected data. It also allows healthcare providers to get the maximum benefit and information out of the collected data.

In practice, this combination of purposes is not yet easy to achieve. Making use of a separate website and database to collect the PROM data, as in this thesis, offers more elaborate possibilities to create a patient-friendly way to present the results. However, the collected data cannot be readily analysed on an aggregate level because the PROM database and the electronic health record are not linked. This means that for example analysis of PROM results per treatment option or correction for patient characteristics is not possible. It may be possible to create this link in the future, but since there are several different patient record systems in use this may prove laborious and expensive. Another option that may be more feasible is to distribute PROMs directly via a patient record system. Ideally, this data is then stored in a database to allow for analyses on aggregate level, and also becomes available to the physicians and to patients so that the results can be used in clinical practice. There will likely be fewer options regarding the patient-friendly presentation of the results, but the benefits of this approach may still outweigh its limitations.

An important consideration regarding the increased future use of PROMs is also how to deal with increased patient burden. Patients with multiple chronic diseases may be asked to complete separate questionnaires before each of their consultations with different specialties, while the questions asked will most likely overlap. Frustrated patients and a decreased willingness to complete the PROMs could be the result. This could potentially be tackled by selecting generic measures for the domains that are relevant to measure, for example by making use of the Patient Reported Outcomes Measurement Information System (PROMIS). PROMIS contains item banks for many of the most frequently assessed patient-reported outcomes (domains), which can be used irrespective of patient's disease. A patient's social functioning or problems due to bad sleep could then be measured with one generic set of items, which can be used by different specialties in one hospital, or by interdisciplinary teams within a hospital or between organisations. It may still be needed to add some disease-specific domains, but this approach should still limit the patient burden.

7 FUTURE OF EVALUATING THE USE OF PROM DATA IN PRACTICE

The results of this thesis also raise the question whether approaching a PROM on its own as an intervention to bring about patient-centered care is the optimal approach. PROM information may be more useful when it is part of a larger intervention to increase patient-centeredness of care. This can involve a restructuring of the care process and/or an intervention that focuses on more than only PROM data. An example of this is the (effective) intervention

MyIBDcoach, a telemonitoring system including PROMs, promoting self-management for patients with irritable bowel disease [8]. Alignment of the goals of the intervention and those of the healthcare professionals and patients that will use the intervention is important for a successful implementation.

In future evaluations of the impact of PROM data on clinical practice, academia may be involved through participatory research [22] in which the health care organization and the involved professionals share ownership of the research. Participatory research involves a qualitative approach in which understanding of problems and (the embedding of) the intervention can change during the study in order to evaluate what works best [23]. It allows the users of the PROM data to be in the lead, which has been indicated as a success factor before [21]. For complex interventions like PROMs in clinical practice, participatory research is likely to benefit the understanding on whether and how it works.

REFERENCES

1. Greenhalgh, J., A.F. Long, and R. Flynn, *The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory?* Soc Sci Med, 2005. **60**(4): p. 833-43.
2. Santana, M.J. and D. Feeny, *Framework to assess the effects of using patient-reported outcome measures in chronic care management.* Qual Life Res, 2014. **23**(5): p. 1505-13.
3. R, G. and W. M, *Implementatie - effectieve verbetering van de patiëntenzorg.* 4th edition ed. 2011, Amsterdam: Reed Business.
4. Hensher, M., J. Tisdell, and C. Zimitat, *"Too much medicine": Insights and explanations from economic theory and research.* Soc Sci Med, 2017. **176**: p. 77-84.
5. Peters, J.B., et al., *Development of a battery of instruments for detailed measurement of health status in patients with COPD in routine care: the Nijmegen Clinical Screening Instrument.* Qual Life Res, 2009. **18**(7): p. 901-12.
6. Slok, A.H., et al., *Development of the Assessment of Burden of COPD tool: an integrated tool to measure the burden of COPD.* NPJ Prim Care Respir Med, 2014. **24**: p. 14021.
7. Slok, A.H., et al., *Effectiveness of the Assessment of Burden of COPD (ABC) tool on health-related quality of life in patients with COPD: a cluster randomised controlled trial in primary and hospital care.* BMJ Open, 2016. **6**(7): p. e011519.
8. de Jong, M.J., et al., *Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial.* Lancet, 2017. **390**(10098): p. 959-968.
9. Haverman, L., et al., *Monitoring health-related quality of life in paediatric practice: development of an innovative web-based application.* BMC Pediatr, 2011. **11**: p. 3.
10. Haverman, L., et al., *Effectiveness of a web-based application to monitor health-related quality of life.* Pediatrics, 2013. **131**(2): p. e533-43.
11. Rosenzweig, I., et al., *Sleep apnoea and the brain: a complex relationship.* Lancet Respir Med, 2015. **3**(5): p. 404-14.
12. Boyce, M.B., J.P. Browne, and J. Greenhalgh, *The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research.* BMJ Qual Saf, 2014. **23**(6): p. 508-18.
13. Greenhalgh, J. and K. Meadows, *The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review.* J Eval Clin Pract, 1999. **5**(4): p. 401-16.
14. Greenhalgh, J., et al., *Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care.* 2017: Southampton (UK).
15. Marshall, S., K. Haywood, and R. Fitzpatrick, *Impact of patient-reported outcome measures on routine practice: a structured review.* J Eval Clin Pract, 2006. **12**(5): p. 559-68.
16. McNicholas, W.T., *Diagnostic criteria for obstructive sleep apnea: time for reappraisal.* J Thorac Dis, 2018. **10**(1): p. 531-533.
17. McNicholas, W.T., et al., *Challenges in obstructive sleep apnoea.* Lancet Respir Med, 2018. **6**(3): p. 170-172.
18. Bonsignore, M.R., et al., *Personalised medicine in sleep respiratory disorders: focus on obstructive sleep apnoea diagnosis and treatment.* Eur Respir Rev, 2017. **26**(146).
19. NVALT. (2017). *Richtlijn diagnostiek en behandeling van obstructief slaapapneu (OSA) bij volwassenen.* Richtlijnendatabase.nl: Nederlandse Vereniging van Artsen voor Longziekten en Tuberculose.
20. Van Mechelen, P.H., *De kern van de diagnose: een diepgaand gesprek*, in *Apneu Magazine*. December 2018. p. 28-29.

21. Van Der Wees, P.J., et al., *Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries.* Milbank Q, 2014. **92**(4): p. 754-75.
22. Wright, M.T., et al. (2013). *Position Paper 1: What is Participatory Health Research?* Berlin.
23. Baum, F., C. MacDougall, and D. Smith, *Participatory action research.* J Epidemiol Community Health, 2006. **60**(10): p. 854-7.

Summary

The aim of this thesis was to provide more insight into how individual PROM results work when implemented in routine clinical practice. This was studied in the context of care for patients with obstructive sleep apnea (OSA), a condition in which patients experience breathing stops while asleep. This causes a variety of symptoms and functional problems. We had the following specific aims:

- To study whether a PROM of sufficient quality is available for patients with OSA, which measures OSA-related quality of life;
- If no existing PROM, measuring OSA-related quality of life, is available and of sufficient quality: to develop a new PROM for patients with OSA specifically for use in clinical practice, with the goal to be suitable for use on both an individual patient and aggregate level;
- To develop a 'patient-friendly' way of presenting the results of the PROM, in order to make them easy to interpret;
- To assess the validity, reliability and responsiveness of the PROM in a Dutch setting;
- To study the impact of individual results of the PROM on the care of patients with OSA, and study *why* this impact is or is not found.

Below, we summarize the chapters of this thesis, in which these aims are addressed.

Chapter two presents a systematic review which summarizes the evidence regarding the quality of PROMs validated in patients with OSA. We identified 22 PROMs in the literature: eight measuring OSA-related quality of life, eight measuring OSA-related symptoms (e.g. sleepiness), and six generic measures of quality of life. 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 any check of content validity before the PROM was tested in patients with OSA. Positive exceptions are four PROMs measuring OSA-related quality of life that were developed with input from patients: 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). For none of these PROMs there was sufficient evidence to properly judge their quality, and all of them need to be validated further.

We considered which of these PROMs we would recommend for research and clinical practice. The SAQLI was the PROM with the most positive evidence for its measurement properties. Because this PROM is interview-administered, we did not recommend it for use in clinical practice, but it can potentially be used in research. The QSQ, MOSAS questionnaire and OSAPOSI may potentially be feasible for use in research and clinical practice, if interpreted with caution. However, these PROMs were not developed with clinical practice in mind and their face validity was not studied in this review.

In **chapter three**, we make recommendations for problems regarding the interpretation of convergent validity that we encountered while conducting the systematic review of chapter two. Convergent validity is a type of validity that is assessed by means of "hypothesis testing": determining whether the scores of the instrument under study correlate with other instruments that measure related constructs to the extent that one would expect. Authors of systematic reviews on measurement properties for PROMs may encounter validation articles in which correlations are presented without hypotheses to which they should be tested. We suggest that in these cases, reviewers construct their own hypotheses to determine the adequacy of convergent validity for the PROM under study. However, constructing hypotheses and interpreting outcomes is not always straightforward. We made the following recommendations for authors of reviews: take an active role in judging the suitability of the comparator instruments; be transparent about which hypotheses were constructed, the underlying assumptions on which they are based, and whether they were constructed by authors of the validation article or the reviewer; discuss unmet hypotheses, especially if convergent validity is judged to be inadequate based on these hypotheses; and when synthesizing data, add up the results of all hypotheses for one instrument, rather than judging convergent validity per included study.

In **chapter four**, we describe the development of a new PROM for patients with sleep apnea, measuring apnea-specific quality of life: the patient-reported apnea questionnaire (PRAQ). The choice to make a new PROM was made in the working group, which consisted of a patient, a pulmonologist, and two researchers on this project. This working grouped judged the four PROMs identified in chapter two on their face validity and suitability for clinical practice. One of the PROMs was not publically available or retrievable via the developer (the OSAPOSI) so this PROM was excluded from the deliberation. The SAQLI required an interviewer for administering the PROM because there were many, and relatively complicated questions, which the working group did not deem feasible for clinical practice. The QSQ and MOSAS questionnaire we did not consider ideal because some aspects of quality of life potentially important for clinical practice were missing, and because some of the items contained unclear phrasing. Therefore, we decided to develop a new PROM, which also gave us the opportunity to pay specific attention to its usefulness in clinical practice. Because the mentioned PROMs were developed with patient input, the working group did consider them suitable to serve as a basis for the PRAQ.

Patients and healthcare professionals were intensively involved in the development of the PRAQ via membership of the development team, online surveys and focus groups, as well as two rounds of cognitive validation to check the understanding and the phrasing of the items. They helped us select topics and items from the existing PROMs, and add or adjust topics or items, based on what was considered relevant for clinical practice. This resulted in a first version of the PRAQ, consisting of 43 items and 10 preliminary domains. Patients indicated that PRAQ was comprehensive and its length acceptable.

In **chapter five** we describe the further development and validation of the first version of the PRAQ, in the context of how we aim for the PRAQ to be used: patients complete the PRAQ before their OSA-consultation, where the results can be discussed with a healthcare professional; and the aggregate outcomes of groups of patients can then be studied by making use of an optimal subselection of the completed items. We conducted a study in which 180 patients with suspected OSA completed the preliminary version of the PRAQ. The collected data was used to 1) perform the final item selection for individual use of the PRAQ in clinical practice, 2) create the domains for (aggregate) outcome measurement, and 3) assess the measurement properties internal consistency, test-retest reliability, convergent validity and responsiveness. We selected 40 items and 10 domains for the final version of the PRAQ for use in daily clinical practice. A subselection of 33 items in 5 domains was selected for optimal outcome measurement with the PRAQ. The results for the outcome measurement domains were: Cronbach's α 0.88–0.95, intraclass correlation coefficient (ICC) 0.81–0.88, and > 75% of hypotheses correct for convergent validity and responsiveness. This lead us to the conclusion that the PRAQ shows good measurement properties in patients with (suspected) OSA.

In this chapter we also describe the development of the PRAQ-report, an overview of an individual patient's PRAQ results in which colored 'smileys' show which domains are more and less problematic for that patient.

In **chapter six** we present the outcomes of an explorative pilot study for the impact of the PRAQ on clinical practice, for which the PRAQ was implemented in three Dutch sleep centers. We had several expectations with regard to how the PRAQ could impact clinical practice. First, we expected that completing the PROM would educate and empower patients suspected of having OSA by making them more aware of which of their symptoms and problems in daily life can be related to OSA. Interviews with patients showed that this expectation seems to be met: completing the PRAQ has the potential to teach patients about the potential impact of OSA on their lives, and through discussing the questions of the PRAQ with their family it also provided more insight into their own situation. Second, we expected that the PRAQ would shift focus during a consultation from the medical facts towards the experiences and problems of patients with OSA and result in more 'patient-centered' care. However, the interviews with healthcare professionals showed that this was not met for a majority of healthcare professionals. Though most professionals did discuss the PRAQ with patients in some way during their consultations, this was usually briefly, and some mentioned they only did it for the sake of the study. Most professionals let the PRAQ play only a minor role during their consultations.

We also conducted a complimentary electronic health record study, before and after implementation of the PRAQ, in which we studied: treatment choice; referrals to other healthcare professionals; and patients' compliance with treatment. We had expected that these aspects of care could change based on the changes in the consultations. However, we did not find any differences, in line with our qualitative findings that there was generally not

much change in the consultations. The availability of individual OSA-related quality of life data did therefore not seem to impact care during this study.

In **chapter seven**, we reflect on why the PRAQ did not show the impact on clinical practice that we had expected. The following factors that potentially influenced the lack of impact are discussed: the OSA healthcare professionals and the context in which they work, the patients, the nature of OSA, technology, the PRAQ itself, and efforts of the research team to improve implementation. One of our conclusions is that even though care for patients with OSA is undergoing a shift towards more personalized, patient-centered care, the healthcare professionals in this study did for the most part not yet take this approach. This is likely in part due to limited time for consultations. In the future, the further shift to personalized care may lead to a greater interest of sleep centers and healthcare professionals in the use of the PRAQ. Since the effectiveness of the PRAQ in clinical practice has not been shown in this thesis, it will be up to sleep centers to decide whether they *consider* it to be potentially useful enough to warrant the effort and money of implementation and continuous measurement. They will also have to consider if they are willing and able to provide the context in which the PRAQ can be employed in a more useful way for individual patients, by for example planning more time for an intake consultation. Currently, one Dutch sleep center is using individual PRAQ data in its clinical practice, and has shown interest in also using this data on an aggregated level.

We then discuss the challenges related to using PROMs in clinical practice. We argue for collecting PROM data for two goals at once: improving care at the level of an individual patient, and measuring quality with aggregated PROM data. However, the question via which platform this data must be collected to be able to use for both purposes is still a challenge. In addition, patient burden should also be taken into account, especially those with multiple chronic diseases who in the future may be overloaded with requests to complete PROMs. The interdisciplinary use of generic item banks, such as the Patient Reported Outcomes Measurement Information System (PROMIS), may be a solution.

Lastly, we raise the question whether using PROM data on its own as a single intervention to bring about patient-centered care is the optimal approach. PROM information may be more useful when it is part of a larger intervention to increase patient-centeredness of care. This may involve a restructuring of the care process and/or an intervention that focuses on more than only using PROM data. In future evaluations of the impact of PROM data on clinical practice, academia may be involved though participatory research in which the health care organization and the involved professionals share ownership of the research with the researchers.

Samenvatting

Het doel van dit proefschrift was om meer inzicht te geven in hoe individuele PROM-resultaten een mogelijk effect bereiken wanneer ze gebruikt worden in de dagelijkse klinische praktijk. Dit werd onderzocht in het kader van de zorg voor patiënten met obstructieve slaapapneu (OSA), een aandoening waarbij de adem van een patiënt tijdens de slaap frequent stokt. Dit veroorzaakt een verscheidenheid aan symptomen en functionele problemen. We hadden de volgende specifieke doelen:

- Nagaan of een PROM van voldoende kwaliteit beschikbaar is voor patiënten met OSA, die OSA-gerelateerde kwaliteit van leven meet;
- Als er geen bestaande PROM is die de OSA-gerelateerde kwaliteit van leven meet, en van voldoende kwaliteit is: het ontwikkelen van een nieuwe PROM voor patiënten met OSA, specifiek voor gebruik in de klinische praktijk, met het doel om geschikt te zijn voor gebruik bij zowel een individuele patiënt en geaggregeerd niveau;
- Het ontwikkelen van een 'patiëntvriendelijke' manier om de resultaten van de PROM te presenteren, zodat ze makkelijk geïnterpreteerd kunnen worden;
- Het beoordelen van de validiteit, betrouwbaarheid en responsiviteit van de PROM in een Nederlandse setting;
- Het bestuderen van de impact van individuele resultaten van de PROM op de zorg voor patiënten met OSA, en te onderzoeken waarom deze impact wel of niet wordt gevonden.

Hieronder vatten we de hoofdstukken van dit proefschrift samen, waarin deze doelen aan bod zullen komen.

Hoofdstuk twee is een systematische review over de kwaliteit van bestaande PROMs die gevalideerd zijn bij patiënten met OSA. We identificeerden 22 PROMs in de literatuur: acht die OSA-gerelateerde kwaliteit van leven meten, acht die OSA-gerelateerde symptomen meten (bijvoorbeeld slaperigheid), en zes PROMs die generieke kwaliteit van leven meten. De resultaten lieten zien dat de meeste meeteigenschappen van de geïdentificeerde PROMs niet, of niet adequaat, onderzocht zijn. Veel van de geïdentificeerde PROMs waren ontwikkeld zonder de betrokkenheid van patiënten, en zonder beoordeling van inhoudsvaliditeit voordat de PROM werd getest bij patiënten met OSA. Positieve uitzonderingen daarop zijn vier PROMs die de OSA-gerelateerde kwaliteit van leven meten en die ontwikkeld zijn met input van patiënten: de *Sleep Apnea Quality of Life index* (SAQLI), de *Maugeri Obstructive Sleep Apnea Syndrome* (MOSAS) vragenlijst, de *Quebec Sleep Questionnaire* (QSQ) en de *Obstructive Sleep Apnea Patient-Oriented Severity Index* (OSAPOSI). Voor geen van deze PROMs was er voldoende bewijs om hun kwaliteit volledig te beoordelen, dus ze zouden allemaal verder gevalideerd moeten worden.

We hebben overwogen welke deze PROMs we zouden aanbevelen voor onderzoek en voor gebruik in de klinische praktijk. De SAQLI was de PROM met het meeste positieve bewijs voor de meeteigenschappen. Omdat deze PROM afgenomen wordt door middel van een interview, hebben we hem niet aanbevolen voor gebruik in de klinische praktijk. Hij is mogelijk

wel te gebruiken in onderzoek. De QSQ, MOSAS vragenlijst en OSAPOSI zijn mogelijk geschikt voor gebruik in onderzoek en de klinische praktijk, mits men hun resultaten voorzichtig interpreteert. Deze PROMs waren echter niet ontwikkeld met de klinische praktijk in het achterhoofd en hun *face validity* werd niet bestudeerd in deze review.

In **hoofdstuk drie** doen we aanbevelingen voor problemen die we tegenkwamen bij de systematische review van hoofdstuk twee, wat betreft de interpretatie van convergente validiteit. Convergente validiteit is het type validiteit dat wordt beoordeeld door middel van 'hypothesetesten': in dit geval het bepalen of de scores van het onderzochte instrument correleren met andere instrumenten die gerelateerde constructen meten in de mate die men zou verwachten. Auteurs van systematische reviews over meeteigenschappen van PROMs kunnen validatieartikelen aantreffen waarin correlaties tussen instrumenten worden gepresenteerd zonder dat er hypothesen zijn opgesteld waaraan ze moeten worden getoetst. We stellen voor dat de reviewers in deze gevallen hun eigen hypothesen opstellen om de mate van convergente validiteit te bepalen voor de PROM in kwestie. Maar het opstellen van hypothesen en het interpreteren van uitkomsten is niet altijd eenvoudig. We hebben de volgende aanbevelingen gedaan voor auteurs van reviews: neem een actieve rol aan bij het beoordelen van de geschiktheid van de vergelijkingsinstrumenten; wees transparant over welke hypothesen zijn opgesteld, de onderliggende aannames waarop ze zijn gebaseerd, en of de auteurs van het validatieartikel ze zelf hebben opgesteld, of dat de reviewer dit gedaan heeft; reflecteer op hypothesen die niet juist blijken te zijn, zeker als de convergente validiteit van de PROM niet voldoende blijkt op basis van deze hypothesen; en tel de resultaten van alle hypothesen voor één instrument op, in plaats van de convergente validiteit per geïncludeerd onderzoek te beoordelen.

In **hoofdstuk vier** beschrijven we de ontwikkeling van een nieuwe PROM voor patiënten met slaapapneu, die apneuspecifieke kwaliteit van leven meet: de *Patient-Reported Apnea Questionnaire* (PRAQ). De keuze om een nieuwe PROM te ontwikkelen is gemaakt met een werkgroep bestaande uit een patiënt, een longarts, en twee onderzoekers op dit project. Deze werkgroep heeft de vier beste PROMs uit de review van hoofdstuk twee beoordeeld op *face validity* en op de vraag of ze geschikt lijken voor de klinische praktijk. Een van die PROMs was niet publiekelijk beschikbaar, noch opvraagbaar via de ontwikkelaar (de OSAPOSI), dus deze PROM is daarbij niet meegenomen. De SAQLI vond de werkgroep niet haalbaar voor de klinische praktijk, omdat deze afgenomen moet worden door een interviewer en omdat hij veel en relatief gecompliceerde vragen bevat. De QSQ en de MOSAS vragenlijst beschouwde de werkgroep niet als ideaal omdat sommige aspecten van OSA-specifieke kwaliteit van leven die mogelijk belangrijk zijn voor de klinische praktijk ontbraken. Daarnaast waren sommige items in deze vragenlijsten onduidelijk geformuleerd. De werkgroep heeft daarom besloten om een nieuwe PROM te ontwikkelen. Dit gaf ons ook de mogelijkheid om bij de ontwikkeling specifiek aandacht te besteden aan het nut van deze nieuwe vragenlijst voor de klinische

praktijk. Omdat de hiervoor genoemde PROMs ontwikkeld waren met behulp van patiëntinput achtte de werkgroep deze wel geschikt als uitgangspunt van the PRAQ.

Bij de ontwikkeling van de PRAQ waren zowel patiënten als zorgprofessionals intensief betrokken. Ze namen deel aan het ontwikkelteam, aan online enquêtes en focusgroepen, en aan twee ronden van cognitieve validatie om de formulering en interpretatie van de vragen te controleren. Ze speelden een belangrijke rol bij het selecteren van onderwerpen en items uit de bestaande PROMs, en kregen de mogelijkheid om onderwerpen of items toe te voegen of aan te passen op basis van wat relevant werd geacht voor de klinische praktijk. Dit resulteerde in een eerste versie van de PRAQ, bestaande uit 43 items en 10 voorlopige domeinen. De geconsulteerde patiënten gaven aan dat ze de PRAQ alomvattend vonden en dat de lengte van de vragenlijst acceptabel was.

In **hoofdstuk vijf** beschrijven we de doorontwikkeling en validatie van de voorlopige versie van de PRAQ, in de context van hoe we willen dat de PRAQ wordt gebruikt: patiënten vullen de PRAQ in vóór hun OSA-consult, zodat de resultaten kunnen worden besproken met hun zorgverlener; en de geaggregeerde uitkomsten van de op deze manier verzamelde PRAQ-data kunnen gebruikt worden voor uitkomstmetingen door gebruik te maken van een (door ons geselecteerde) optimale subselectie van de volledige PRAQ. We hebben een onderzoek uitgevoerd waarin 180 patiënten met verdenking op OSA de voorlopige versie van de PRAQ hebben ingevuld. De verzamelde gegevens werden gebruikt om 1) de definitieve selectie van *items* uit te voeren voor individueel gebruik van de PRAQ in de klinische praktijk, 2) de *domeinen* voor (geaggregeerde) uitkomstmeting te maken, en 3) het bestuderen van de meeteigenschappen interne consistentie, test-hertestbetrouwbaarheid, convergente validiteit en responsiviteit. We selecteerden 40 items en 10 domeinen voor de definitieve versie van de PRAQ voor gebruik in de dagelijkse klinische praktijk. Een subselectie van 33 items in 5 domeinen werd geselecteerd voor optimale uitkomstmetingen met de PRAQ. De waarden van de meeteigenschappen van de domeinen voor uitkomstmetingen waren: Cronbach's α 0.88-0.95, *intraclass correlation coefficient* (ICC) 0.81-0.88, en >75% van de hypothesen correct voor convergente validiteit en responsiviteit. Dit bracht ons tot de conclusie dat de PRAQ goede meeteigenschappen laat zien bij patiënten met verdenking op OSA.

In dit hoofdstuk staat ook de ontwikkeling van de PRAQ-rapportage beschreven: een overzicht van de PRAQ-resultaten van een individuele patiënt, waarbij gekleurde 'smileys' laten zien welke domeinen voor die patiënt meer of minder problematisch zijn.

In **hoofdstuk zes** presenteren we de uitkomsten van een exploratieve pilotstudie naar de impact van de PRAQ op de klinische praktijk. Hiervoor werd de PRAQ geïmplementeerd in drie Nederlandse slaapcentra. We hadden een aantal verwachtingen met betrekking tot hoe de PRAQ de klinische praktijk zou kunnen beïnvloeden. Ten eerste verwachtten we dat het invullen van de PROM de positie van patiënten binnen het consult zou versterken en hen meer bewust zou kunnen maken van welke van hun symptomen en problemen in het dagelijks leven verband

kunnen houden met OSA. Interviews met patiënten lieten zien dat aan deze verwachting lijkt te zijn voldaan: het invullen van de PRAQ heeft de potentie om patiënten dingen te leren over de mogelijke impact van OSA op hun leven, en het bespreken van de vragen van de PRAQ met hun familie gaf hen ook meer inzicht in hun situatie. Ten tweede verwachtten we dat de PRAQ tijdens een consult de focus zou kunnen verleggen van medische feiten naar ervaringen en problemen van patiënten met OSA. Dit kan resulteren in meer 'patiëntgerichte' zorg. De interviews met zorgprofessionals toonden echter aan dat dit bij het merendeel van hen niet was gebeurd. Hoewel de meeste zorgprofessionals de PRAQ wel op enige manier bespraken met patiënten was dit meest kort, en een aantal professionals gaf aan dat ze het vooral deden voor de studie. De meeste zorgverleners lieten de PRAQ slechts een minimal rol spelen tijdens hun consulten.

We voerden ook een aanvullende studie uit met behulp van patiëntdossiers, voor en na de implementatie van de PRAQ, waarin we hebben onderzocht: behandelingskeuze; verwijzingen naar andere beroepsbeoefenaren in de gezondheidszorg; en therapietrouw van de patiënt. We hadden verwacht dat deze aspecten van zorg zouden kunnen veranderen omdat het gebruik van de PRAQ tot veranderingen in het consult zou leiden. We vonden echter geen verschillen, in lijn met onze kwalitatieve bevindingen dat er weinig verandering was in de consulten. De beschikbaarheid van informatie over individuele OSA-gerelateerde kwaliteit van leven leek daarom niet van invloed te zijn op de zorg in deze studie.

In **hoofdstuk zeven** reflecteren we op de vraag waarom de PRAQ niet de impact op de klinische praktijk liet zien die we hadden verwacht. De volgende factoren die van invloed kunnen zijn op dit gebrek aan impact worden besproken: de OSA-zorgprofessionals en de context waarin zij werken, de patiënten, de aard van OSA, de gebruikte technologie, de PRAQ zelf en de inspanningen van het onderzoeksteam om de implementatie te verbeteren. Een van onze conclusies is dat, hoewel er een overgang gaande is naar meer gepersonaliseerde, patiëntgerichte zorg voor patiënten met OSA, de meeste zorgprofessionals in deze studie deze benadering nog niet hebben gevolgd. Dit heeft waarschijnlijk onder andere te maken met de beperkte tijd voor consulten. Mogelijk zal de PRAQ in de toekomst wel van nut zijn in slaapcentra en voor zorgprofessionals als er een verschuiving plaatsvindt naar een meer gepersonaliseerde zorg. Aangezien de effectiviteit van de PRAQ in de klinische praktijk in dit proefschrift niet is aangetoond, is het aan slaapcentra om te beslissen of zij de PRAQ potentieel nuttig genoeg vinden om de inspanning en het geld van implementatie en continue meting te rechtvaardigen. Ze zullen ook moeten overwegen of ze bereid en in staat zijn om de context te bieden waarin het gebruik van de PRAQ voor individuele patiënten zinvol is, door bijvoorbeeld meer tijd te reserveren voor een intakegesprek. Er is momenteel één Nederlands slaapcentrum dat werkt met individuele PRAQ-gegevens in de klinische praktijk, en dat belangstelling heeft om deze gegevens ook op geaggregeerd niveau te gebruiken.

Vervolgens bespreken we de uitdagingen voor het inzetten van PROMs in de klinische praktijk. Daarbij pleiten we onder andere voor het in één keer verzamelen van PROM-data

voor twee doelen: het verbeteren van de zorg op het niveau van een individuele patiënt, en het meten van kwaliteit met geaggregeerde PROM-data. De vraag via welk digitaal systeem deze data verzameld moet worden om het voor beide doelen goed te kunnen gebruiken is echter nog een uitdaging. Daarnaast moet er ook rekening gehouden worden met patiënten, vooral die met meerdere chronische ziektes die in de toekomst misschien overvraagd gaan worden met verzoeken om PROMs in te vullen. Het interdisciplinair gebruik van generieke item banks, zoals bijvoorbeeld het *Patient Reported Outcomes Measurement Information System* (PROMIS) kan hier mogelijk een oplossing bieden.

Als laatste stellen we de vraag of het afnemen van een PROM op zichzelf de optimale aanpak is als een enkele interventie om patiëntgerichte zorg tot stand te brengen. PROM-informatie kan nuttiger zijn wanneer deze deel uitmaakt van een meer uitgebreide interventie om de patiëntgerichtheid van de zorg te verbeteren. Dit kan een herstructurering van het zorgproces zijn en/of een interventie die zich richt op meer dan alleen het gebruik van PROM-gegevens. Wij zien het als een goede aanpak om in toekomstige evaluaties van de impact van PROM-gegevens op de klinische praktijk de academische wereld te betrekken door middel van participerend onderzoek. Hierbij zijn de zorgorganisatie en de betrokken zorgprofessionals samen met de onderzoekers eigenaar van het onderzoek.

DATA MANAGEMENT

For each study of this PhD involving participant data, the research protocol was submitted to the local Medical Ethics Committee CMO Arnhem-Nijmegen. All studies were officially declared exempt from ethical approval for human subjects research. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

The survey and questionnaire data described in chapters 4 and 5 were collected via a secure version of Limesurvey, and data collected in chapter 6 via a secure platform of VitalHealth Software. The contact details of the patients that completed the survey in chapter 4 have been discarded. All PROM, survey, and electronic health record data from chapters 5 and 6 were collected by the involved healthcare providers, and no identifying patient information was shared with the researchers. Identifying information of the participants of focus groups and interviews held for thesis were stored separately from the data, in a secured folder to which only the main researcher and the quality officer had access. The identifying information was deleted after finishing the respective studies. Recordings of the focus groups and interviews were deleted, only the (anonymised) transcripts/summaries are saved.

Until chapter 6 of this PhD has been published, the raw and processed data and accompanying files (descriptive files, syntaxes, etc.) of the projects of this thesis will be stored in a folder on the department server of IQ healthcare which is accessible only by the main researchers of this project. Thereafter, the data will stored on the secured IQ healthcare archive server in a folder called "PROMs bij slaapapneu" for 10 years, which is accessible only by the secretary of IQ healthcare. Since the participants of the studies in this PhD did not give informed consent for sharing their data publically, requests for data can be made via receptie.iqh@radboudumc.nl. A suitable way to share the data will then be sought.

DANKWOORD

Mijn dank gaat allereerst uit naar mijn promotieteam. Philip, je begon als mijn copromotor en werd uiteindelijk eerste promotor, en erg verdiend. Jouw vertrouwen, advies, en onze goede inhoudelijke discussies hebben mijn promotietraject verreikt. Maroeska, jouw kritische blik en commentaar zetten mij regelmatig flink aan het werk – en ik zou het niet anders hebben gewild. De persoonlijke openheid waarmee je dit combineerde, bijvoorbeeld bij onze EBS heidagen, waardeer en bewonder ik. Ik zie ernaar uit om als postdoc met jullie beiden samen te blijven werken. Gert, jouw bijdrage was meer vanaf de zijlijn maar het was altijd fijn om een nieuwe, frisse blik te krijgen op het onderzoek en mijn artikelen.

Marijke IJff en Bernard Hol, patiënt en arts in mijn onderzoeksteam, ik wil ook jullie hartelijk bedanken voor jullie belangrijke rol bij de totstandkoming van dit proefschrift. Jullie hebben vanaf dag één jullie enthousiasme uitgesproken over het idee van een PROM bij slaapapneu, en bij elke stap van het onderzoek meegedacht en waar nodig meegeholpen. Op de momenten dat ik worstelde met de zin en onzin van de PRAQ stonden jullie altijd klaar om jullie visie te geven en mij weer nieuwe energie te geven voor het project. Daarnaast was de samenwerking niet alleen nuttig, maar ook gezellig. Dank hiervoor!

Leden van de manuscriptcommissie, Prof. Judith Prins, Prof. Henri Marres en Dr. Caroline Terwee, mijn dank aan jullie voor het beoordelen van mijn proefschrift.

Masha, het was erg leuk om jou als stagiaire en interviewmaatje aan mijn zijde te hebben tijdens de laatste studie van mijn promotietraject. Bedankt voor je hulp, en succes met je verdere studie!

Ook wil ik graag de zorgverleners bedanken die hebben bijgedragen aan de studies van dit promotietraject. Winnie Fok en Annemieke Houweling van het slaapcentrum van het Albert Schweizerziekenhuis, dank jullie wel voor de coördinatie en hulp bij de validatiestudie van de PRAQ. Dr. Yvonne Berk en Josina Claassen, Janet Dijkstra en Hans Varenbrink, Dr. Michiel Eijsvogel, Simone Bos en Loes Wansing, en anderen die bij het laatste onderzoek van deze promotie betrokken zijn geweest, ik wil jullie allen hartelijk danken voor jullie steun bij de implementatie en evaluatie van de PRAQ in jullie slaapcentra.

Leden van (voormalig) Celsus, academie voor betaalbare zorg, wat was het leuk om bij dit team te horen. Ondanks dat ik qua onderwerp van mijn promotieonderzoek bij de inhoud niet helemaal aansloot, voelde ik me echt 'thuis' bij Celsus en heb ik er veel geleerd. Joost, Floris, Niek, Wieteke en Florien, ik heb veel gehad aan het laagdrempelig sparren over onze respectievelijke onderzoeken en artikelen, en genoten van de gezelligheid op onze kamers. Speciale shout-out aan Wieteke en de grote pret die wij konden hebben over Lync, en aan

Florien voor de verkleedpartij bij de door ons georganiseerde IQ kerstlunch! Dank ook dat jullie tijdens de verdediging van mijn proefschrift aan mijn zijde staan als paranimfen.

Huidige en voormalige collega's van Evidence-Based Surgery, of we nu een 'team', 'groep' of 'netwerk' zijn, ik heb het altijd leuk en interessant gevonden om jullie en jullie onderzoeken te leren kennen. Ik heb het als bijzonder ervaren om regelmatig samen te komen met een groep PhDs uit verschillende hoeken van het ziekenhuis. De EBS-lunches en vooral ook de leerzame en leuke heidagen hebben mijn promotietraject verrijkt. Machteld, daarnaast nog speciale dank aan jou, voor de gezellige en inspirerende lunches/koffiemomenten, jouw positieve blik en enorme vertrouwen.

Collega-promovendi van IQ healthcare over de jaren, en anderen op de benedenverdieping, dank jullie wel voor de gezelligheid en de positieve en ontspannen sfeer in 'de kelder'. Een goede omgeving om een proefschrift te schrijven.

Mam en pap, dank voor jullie steun en vertrouwen! Ik kan bij jullie altijd mijn verhaal kwijt en jullie denken altijd met mij mee, wat erg fijn is.

Ten slotte dank aan Ayla, mijn viervoetige huisgenootje, voor haar enorme enthousiasme als ik na een dag hard werken aan mijn proefschrift weer thuiskwam.

ABOUT THE AUTHOR

Inger Abma was born on the 31st of May 1987, in Nijmegen, The Netherlands. She obtained her grammar school degree at the Stedelijk Gymnasium Nijmegen (2005) and then started the Bachelor program of Biomedical Sciences (BMS) at Nijmegen University (now Radboud University). She obtained the Bachelor's degree in 2008 and then took a year off studying to work: helping at practicals of several BMS courses, and working as customer service representative at OHRA.

She continued her studies in 2009 with the BMS Master program major 'Health Technology Assessment' and two minors: Epidemiology, and Health Policy Innovation and Management - the latter consisted of courses at Maastricht University. Her Master's program also included the 'consultancy' track, for which she did an internship at the Quality Institute for Healthcare CBO (Utrecht). This resulted in an advisory report on how patients can be involved in the design and evaluation of clinical pathways in the hospital. Her final internship for the Master's degree was at Imperial College London (2011). Here she stayed for an additional two years to work on a research project about barriers to home haemodialysis. During this period she realised how much she liked doing scientific research, and started looking for a PhD position in her home country.

In 2014 she ended up back Nijmegen to do her PhD research at IQ healthcare, Radboudumc, the results of which are presented in this thesis. At the Radboudumc, she joined two interesting research groups/teams: the Celsus Academy for Sustainable Healthcare, and the Evidence-Based Surgery team. Furthermore, she became part of the quality commission of IQ healthcare, and functioned as internal auditor. During her years as PhD candidate she wrote reports for several other projects, most notably on the international implementation pilot for EuroFIT, a lifestyle program for European football fans. IQ healthcare is where Inger now continues doing research in the area of quality of healthcare as a post-doc researcher.

Scientific publications (*published in this thesis)

Abma IL, Rovers M, IJff M, Hol B, Westert GP, Van der Wees PJ. Instrument completion and validation of the Patient-Reported Apnea Questionnaire (PRAQ). *Health and Quality of Life Outcomes* 2018 16:158. DOI 10.1186/s12955-018-0988-6.*

Abma IL, Rovers M, IJff M, Hol B, Westert GP, Van der Wees PJ. The development of a patient-reported outcome measure for patients with obstructive sleep apnea: the Patient-Reported Apnea Questionnaire (PRAQ). *Journal of Patient-Reported Outcomes* 2017, 1:14. DOI 10.1186/s41687-017-0021-6.*

Abma IL, Rovers M, Van der Wees PJ. Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes. *BMC Res Notes* 2016, 9:226, Doi:10.1186/s13104-016-2034-2.*

Abma IL, van der Wees PJ, Veer V, Westert GP, Rovers M. Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): A systematic review. *Sleep Med Rev* 2015*, 28*, 14-27, doi:10.1016/j.smrv.2015.07.006.*

Abma IL, Jayanti A, Bayer S, Mitra S, Barlow J. Perceptions and experiences of financial incentives: a qualitative study of dialysis care in England. *BMJ Open* 2014 12;4(2):e004249.

Jayanti A, Wearden AJ, Morris J, Brenchley P, Abma I, Bayer S, Barlow J, Mitra S. Barriers to successful implementation of care in home haemodialysis (BASIC-HHD):1. Study design, methods and rationale. *BMC Nephrol* 2013;14:197.

Dutch publications

Abma IL, Ten Hove K, Ranke S, Rovers M, Adang E, Jeurissen P,Van der Wees PJ. Doelmatige innovatie in de zorg. (Sustainable innovation in healthcare.) A publication of Celsus Academy of Sustainable healthcare. Nijmegen, 2016. DOI: 10.13140/RG.2.2.14338.99522.

Van der Wees, PJ, Abma IL, Vajda, I, Verbiest-Hoppenbrouwer, M. Patiënt-gerapporteerde uitkomstmetingen: Hoe zetten we de patiënt echt centraal? (Patient-reported outcome measurements: How can we really achieve patient-centeredness?) Nijmegen, 2016. DOI: 10.1007/s12468-016-0029-6.

Institute for Health Sciences
Radboudumc

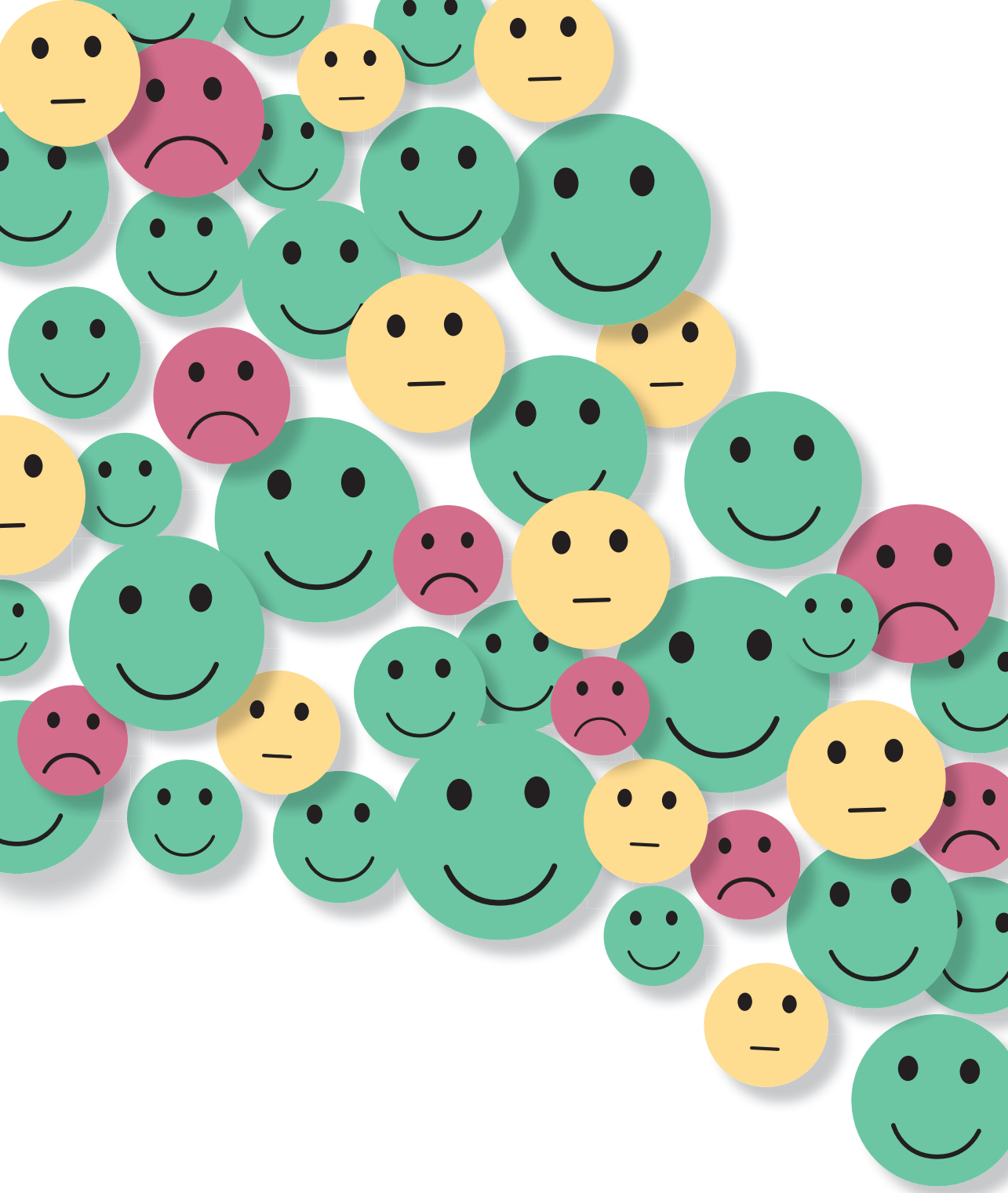
PHD PORTFOLIO

Name PhD candidate: I.L. Abma	**PhD period: 01-02-2014 - 01-01-2019**
Department: IQ healthcare	**Promotor(s): Prof. P.J. van der Wees. Prof. M. Rovers, Prof. G. Westert**
Graduate School: Radboud Institute for Health Sciences	

	Year(s)	ECTS
TRAINING ACTIVITIES		
a) Courses & workshops		
- NCEBP Introduction Course for PhD students	2014	1.5
- Introduction to Endnote (Radboudumc)	2014	0.1
- Qualitative research methods in health care (CaRe, Radboudumc)	2014	1.0
- Summer school: Implementation Science in Health Care (Radboudumc)	2014	2.0
- Clinimetrics: Assessing Measurement Properties of Health Measurement Instruments (VUmc)	2015	1.0
- Winter Academy 'Economy and Policy' (Celsusacademie en Rijksacademie voor Financiën, Economie en Bedrijfsvoering)	2015	0.7
- Scientific integrity (Radboud University)	2015	0.7
- BROK certificate for good clinical practice	2015	1.0
- Training 'writing clear reports'	2015	0.1
- Education in a Nutshell (Radboud University)	2016	1.0
- Value-Based Health Care Delivery Intensive Seminar, lectures only (Harvard Business School)	2016	0.2
- Career management for PhD students (Radboud University)	2017	0.7
- Analyzing Qualitative data (Radboud University)	2018	1.5
- Qualitative interviewing (Evers Research)	2018	0.7
b) Symposia, congresses & seminars		
- ISOQOL symposium "Cliëntenparticipatie bij onderzoek over de zorgl?"	2014	0.2
- National PROMs summit, London	2014	0.3
- IQ healthcare conference	2014, 2016, 2018	0.6
- Seminar Value-Based Healthcare (VitalHealth)	2015	0.2
- Celsus 'knowledge at the table' session, theme oncology & network care	2015	0.2
- Celsus preconference meeting 'Ethics and sustainable healthcare'	2015	0.2
- Celsus conference	2014-2018[a]	1.0
- National spring meeting ApneuVereniging	2016	0.2
- PHD-retreat	2015[b]	0.7
- ICHOM conference	2016	0.3
- ISPOR conference	2017[a]	1.0
- Symposium for Professionals (ApneuVereniging)	2018	0.2
TEACHING ACTIVITIES		
c) Lecturing		
- Course 'Health Outcomes Measurement'	2017, 2018	1.6
- Working groups for course 'Evidence-Based Practice'	2017, 2018	0.4
- Working groups for course 'Interview techniques'	2017	0.1
- Working groups for course 'Evidence-Based Guidelines'	2017	0.2
d) Other		
- Reviewer for scientific publications	2015-2018	4.0
- Supervisor of internship	2017	1.0
TOTAL		**24.6 ECTS**

a. poster presentation
b. 3-slide oral presentation

Radboud University
Radboudumc university medical center



Institute for Health Sciences
Radboudumc