ORIGINAL ARTICLE

Minimum important difference of the Epworth Sleepiness Scale in obstructive sleep apnoea: estimation from three randomised controlled trials

Sarah Crook,[1] Noriane A Sievi,[2] Konrad E Bloch,[2,3] John R Stradling,[4] Anja Frei,[1] Milo A Puhan,[1] Malcolm Kohler[2,3]

[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[2]Department of Pneumology, University Hospital of Zurich, Zurich, Switzerland
[3]Center of Interdisciplinary Sleep Research, University of Zurich, Zurich, Switzerland
[4]Oxford Centre for Respiratory Medicine and Oxford NIHR Biomedical Research Centre, Churchill Campus, Oxford University, Oxford, UK

Correspondence to
Dr Malcolm Kohler, Department of Pulmonology, University Hospital Zurich, Zurich 8091, Switzerland;
malcolm.kohler@usz.ch

Received 19 April 2018
Revised 21 June 2018
Accepted 23 July 2018
Published Online First
12 August 2018

To cite: Crook S, Sievi NA, Bloch KE, *et al*. *Thorax* 2019;**74**:390–396.

ABSTRACT

Background The Epworth Sleepiness Scale (ESS) is a widely used tool for assessing sleepiness in patients with obstructive sleep apnoea (OSA). We aimed to estimate the minimal important difference (MID) in patients with OSA.

Methods We used individual data from three randomised controlled trials (RCTs) in patients with OSA where the preintervention to postintervention change in ESS was used as a primary outcome. We used anchor-based linear regression and responder analysis approaches to estimate the MID. For anchors, we used the change in domains of the Functional Outcomes of Sleep Questionnaire and 36-Item Short Form Health Survey. We also used the distribution-based approaches Cohen's effect size, SE of measurement and empirical rule effect size to support the anchor-based estimates. The final MID was determined by triangulating all estimates to a single MID.

Findings A total of 639 patients with OSA were included in our analyses across the three RCTs with a median (IQR) baseline ESS score of 10 (6–13). The median (IQR) ESS change score overall was −2 (−5 to 1). The anchor-based estimates of the MID were between −1.74 and −4.21 points and estimates from the responder analysis were between −1 and −3 points. Distribution-based estimates were smaller, ranging from −1.46 to −2.36.

Interpretation We propose an MID for the ESS of 2 points in patients with OSA with a disease severity from mild to severe. This estimate provides the means to plan trials and interpret the clinical relevance of changes in ESS.

Trial registration number Provent, NCT01332175; autoCPAP trial, NCT00280800; MOSAIC,ISRCTN (3416388).

Key messages

What is the key question?

- We aimed to provide an estimate of the Epworth Sleepiness Scale (ESS) that can be used in practice to determine an important change in sleepiness in patients with a wide spectrum of obstructive sleep apnoea (OSA) severity and undergoing different interventions.

What is the bottom line?

- The presented manuscript, based on individual data from three different randomised controlled trials, suggests the use of a single minimal important difference of 2 points for the ESS in patients with OSA.

Why read on?

- It facilitates the calculation of required sample sizes in the planning of future trials, as well as providing the means to interpret the clinical relevance of changes in daytime sleepiness following an intervention in both research and clinical practice.

INTRODUCTION

Excessive daytime sleepiness is a key symptom of obstructive sleep apnoea (OSA) and can greatly impact patient's everyday life and well-being.[1 2] The high clinical relevance means that sleepiness is frequently assessed in intervention studies as a primary outcome. One method for assessing sleepiness is the patient-reported Epworth Sleepiness Scale (ESS).[3 4] The ESS is a simple, 8-item self-reported questionnaire where patients answer questions based on how likely they are to doze off or fall asleep during sedentary activities. It is an attractive alternative to objective methods for assessing sleepiness, such as the multiple sleep latency test which is conducted in a sleep laboratory and takes a full day to complete,[5] since it is more practical and cost-efficient to complete.

Investigations into the measurement properties of the ESS have shown it to have good test–retest reliability,[6] internal consistency[7] and moderate construct validity.[8 9] However, only one study has investigated the minimal important difference (MID) of the ESS in patients with OSA.[10] This study estimated the MID as falling between −2 and −3 points in patients with OSA undergoing CPAP for 3 months. While this provides some insight into the ESS MID, a range of estimates can be difficult to use in practice where a single estimate is required. Since there is no evidence in favour of either MID estimates, the final choice of MID to use is unclear. Furthermore, the MIDs were estimated by relating the change in ESS to a global rating of change in sleepiness questionnaire, with the score based on how the respondents feel their sleepiness has improved from baseline. This may be susceptible to recall bias[11] and can be influenced when patients are

not blinded to the treatment.[12] Although specific to sleepiness, it relies on a single question and answer and is not a validated instrument with an existing MID in OSA. When establishing an MID, it is also important to repeat the estimation in different populations of a disease because the MID can vary depending on the setting and population characteristics.[13]

Although the ESS is commonly used both in clinical practice and in randomised controlled trials (RCTs), the interpretation of changes in sleepiness remains challenging without a well-established MID. We aimed to provide an estimate of the minimum change in sleepiness measured by the ESS that is clinically important and that can be used to guide practice and design of future trials, in patients with a wide spectrum of OSA severity.

METHODS

Participants

We used individual data from three separate RCTs of patients with OSA. RCT 1 (The Multi-centre Obstructive Sleep Apnoea Interventional Coardiovascular trial (MOSAIC)) investigated CPAP therapy in minimally symptomatic patients with OSA in the UK and Canada, with a follow-up of 6 months.[14]Patients were aged between 45 and 75 years and had an oxygen desaturation index (ODI)>7.5/hour, but did not have daytime symptoms considered by patient and physician to be sufficient to require CPAP therapy. RCT 2 (autoCPAP) was an equivalence trial comparing the efficacy of continuous automatic mask pressure adjustment to conventional fixed mask pressure for reducing sleepiness in moderate to severe patients with OSA in Switzerland, with follow-up assessments at 3, 12 and 24 months.[15]Patients were aged 18 to 75 years, had an ESS score ≥8 points, an apnoea/hypopnoea index (AHI) ≥10/hour and had completed a 2 to 4 week CPAP adaption period. RCT 3 (Provent) investigated the efficacy of Provent, an expiratory nasal resistance valve, for preventing the reoccurrence of OSA following withdrawal of CPAP therapy in patients with moderate to severe disease in the UK and Switzerland, with a follow-up of 2 weeks.[16]Patients were aged 20 to 75 years, had an ODI>10/hour and had received CPAP treatment for at least 12 months prior to the trial with an average compliance of ≥4 hours per night.

Epworth Sleepiness Scale

All three RCTs assessed the change in ESS as a primary outcome. A total score was calculated for the ESS by summing the scores of the eight items, ranging from 1 to 24 points with higher values reflecting a higher level of sleepiness. The change score was calculated as the difference between the total ESS score at baseline and follow-up. For the purpose of our analyses, in the autoCPAP study the primary change in ESS was calculated as the difference between scores at baseline and 3 months as the observed ESS change score was similar between the different time points and so that we could retain the largest sample size possible.

Other outcomes

Other outcomes assessed at the same time points were considered as potential anchor variables if they were already validated for use in patients with OSA and had an MID. Outcomes assessed in autoCPAP and MOSAIC were the EuroQol 5 Dimensions Questionnaire (EQ-5D)[17] and 36-Item Short Form Health Survey (SF-36),[18] which both assess general health-related quality of life (HRQoL). The five dimensions of the EQ-5D are summarised to a utility index score ranging from −0.208 (worst possible health) to 1.000 (best possible health) and patients also rate their health on a Visual Analogue Scale on a scale of 0 (the worst health you can imagine) to 100 (the best health you can imagine). The SF-36 is composed of eight domains and two summary components on a scale of 0 to 100, with higher scores indicating better health. We used an MID of 5 points for the SF-36, based on previous estimates in COPD and rheumatoid arthritis.[19–21] In autoCPAP and Provent, we used the Functional Outcomes of Sleep Questionnaire[22] (FOSQ), which is specific to sleep disorders and assesses the impact of sleepiness on activities of everyday life. The FOSQ total score has a scale of 5 to 20 points with an MID of 0.75 and 5 domains measured on a scale of 1–4 with an average MID of 0.3 points, with higher scores indicating a smaller effect of sleepiness.[23]The intimacy domain was not routinely assessed in the RCTs and therefore was not analysed.

Statistical analysis

We used a combination of anchor-based and distribution-based approaches to calculate MID estimates. Where possible, we pooled data from the RCTs. For the anchor-based analyses, data were pooled where studies assessed the same anchor outcomes resulting in two datasets: (1) autoCPAP +MOSAIC (n=574) and (2) autoCPAP +Provent (n=267)). As the distribution-based analyses use only the ESS score, these were conducted in all three RCTs pooled together as well as in each RCT individually.

The primary anchor-based method used linear regression with the change in ESS score as the dependent variable and change in anchor variable as the independent variable. The resulting coefficient for the anchor was multiplied by the MID of the anchor and added to the intercept coefficient to determine the change in ESS that is mathematically equivalent to an important change in the anchor.[24] Domains were used as anchors if they had a meaningful Pearson correlation coefficient (≥0.3)[13] with the change score of the ESS. In autoCPAP (treatment and placebo groups), we used the anchor variables most strongly correlated with the ESS to define patients as responders or non-responders based on whether they improved by more than the anchor MID or not. We calculated receiver operator characteristic (ROC) curves from logistic regression models with the response classification as the dependent variable and ESS change score as the independent variable. The optimal cut-point (MID) was determined as the ESS change score with the highest sensitivity (true positives) and specificity (false negatives) for classification of the anchor response, with equal weighting for sensitivity and specificity.[25]

We used three distribution-based methods to support the anchor-based approaches: Cohen's effect size (0.5*SDΔ), empirical rule effect size (0.08*6*SDΔ) and SE of measurement (SEM) (SDbaseline*sqrt[1-intraclass correlation coefficient (ICC)]). For the SEM, we calculated the ICC from a random-effects model of two ESS assessments taken at diagnosis and study inclusion in the Provent study, for patients with less than 20 days between the two assessments.

We used recommended methods for triangulating all estimates to a single final MID estimate.[26] A consensus was reached between the investigators by judging the importance of each estimate based on several criteria: the quality of the anchor MID, responsiveness of the anchor, the statistical relationship and similarity of content between the anchor and ESS, the size and characteristics of the population we estimated the MID in, and the statistical method we used. Analyses were conducted using Stata for Mac (version 14.1; StataCorp, College Station, Texas, USA), except for the ROC curves analyses, which were conducted using R (version 3.4.1; www.r-project.org).

Table 1 Baseline patient characteristics for each RCT combination

Characteristic	Pooled RCT combination	
	1. (N=574) MOSAIC (n=372) +autoCPAP (n=202)	2. (N=267) Provent (n=65) +autoCPAP (n=202)
Age, years	58.0 (51.0–63.0)	59.0 (49.9–64.2)
Male sex, n (%)	470 (82)	225 (84)
BMI, kg/m^2	32.0 (28.9–36.0)	32.7 (29.4–36.9)
Neck circumference (cm)	43.0* (40.5–45.0)	43.5† (42.0–46.0)
Waist circumference (cm)	107.0* (100.5–115.0)	114.5‡ (106.0–124.0)
AHI>4%, events/hour	49.5‡ (31.0–69.0)	Provent: 1.4§ (0.7–3.1); autoCPAP: 49.5 (31.0–69.0)
ODI>3%, events/hour	21.1 (12.4–40.0)	Provent: 2.5§ (0.9–4.5); autoCPAP: 42.8 (29.9–63.0)

Data are presented as median (IQR) unless otherwise stated.
*Only MOSAIC.
†Only Provent.
‡Only autoCPAP.
§on CPAP.
AHI, apnoea/hypopnoea index; BMI, body mass index; ODI, oxygen desaturation index; RCT, randomised controlled trial.

RESULTS

Five hundred and seventy-four patients were included in the first pooled RCT combination (MOSAIC (n=372) plus autoCPAP (n=202)) and 267 in the second (Provent (n=65) plus autoCPAP). Baseline characteristics of each RCT combination are shown in table 1. ESS score improved by 2.4 (SD 4.1) points in RCT combination 1 and by 3.9 (SD 4.6) points in RCT combination 2 (table 2). Baseline and follow-up change scores for potential anchor instruments are presented in table 2.

Pearson correlation coefficients between the change in ESS score and change in potential anchor instruments are shown in table 3. The SF-36 energy/vitality and physical component domains met the methodological criterion for use as an anchor (correlation strength ≥0.3) in study combination 1. In study combination 2, the FOSQ general productivity, activity level and vigilance domains, and total score were all negatively correlated <−0.5, whereas the social outcome domain was less strongly correlated (r=−0.31). MID estimates based on these anchors were lower when estimated by the SF-36 (−1.74 and −2.66) compared with the FOSQ (−3.03 to −4.21) (table 4).

The SF-36 energy/vitality domain, FOSQ total score, activity level domain and vigilance domain were used to define responders/non-responders. Based on the anchor MIDs (SF-36=5,[20 21 27] FOSQ total score=0.75, FOSQ domains=0.3,[23] 402 out of 574, and 156, 139 and 136 out of 267 patients were classified as responders in the SF-36 energy/vitality, FOSQ total, activity level and vigilance, respectively. Results of the ROC curve analyses are displayed in table 4, where the MID estimates were higher based on the FOSQ (≥2 points) than the SF-36 (1 point). Graphical displays of the ROC curves with area under the ROC curve (AUC), sensitivity and specificity for each anchor can be seen in figure 1. MID estimates were between 1 and 3 points. A

Table 2 Summary of ESS scores and potential anchor instruments

Instrument	RCT combination			
	1. autoCPAP+MOSAIC		2. autoCPAP+Provent	
	0–3 months (autoCPAP) 0–6 months (MOSAIC) n=574		0–3 months (autoCPAP) 0–2 weeks (Provent) n=267	
	Baseline	Δ (Mean, SD)	Baseline	Δ (Mean, SD)
ESS	10 (6–13)	−2.4 (4.1)	12* (9–15)	−3.9 (4.6)
EQ-5D index	0.89 (0.76–1.0)	0.02 (0.15)	–	–
EQ-5D VAS	70 (60–80)	3.5 (14.5)	–	–
FOSQ total	–	–	16.5 (14.6–18.2)	1.7 (2.7)
FOSQ general productivity	–	–	3.5 (3.1–3.9)	0.3 (0.5)
FOSQ social outcome	–	–	3.5 (3–4)	0.3 (0.8)
FOSQ activity level	–	–	3.2 (2.7–3.6)	0.4 (0.6)
FOSQ vigilance	–	–	3.0 (2.6–3.4)	0.4 (0.7)
SF-36 physical functioning	80 (60–90)	2.5 (15.4)	–	–
SF-36 role physical	75 (25–100)	12.3 (36.2)	–	–
SF-36 bodily pain	75 (44.4–100)	7.7 (30.8)	–	–
SF-36 general health perception	62 (45–77)	4.2 (16.4)	–	–
SF-36 energy/vitality	45 (30–65)	11.8 (19.7)	–	–
SF-36 social functioning	87.5 (62.5–100)	5.8 (20.6)	–	–
SF-36 role emotional	100 (33.3–100)	8.3 (35.2)	–	–
SF-36 mental health	76 (60–88)	4.4 (14.6)	–	–
SF-36 change in health	50 (50–50)	−6.9 (29.4)	–	–
SF-36 physical component	56.4 (41.0–84.0)	2.9 (9.4)	–	–
SF-36 mental component	49.4 (39.6–56.0)	3.8 (9.0)	–	–

Data are present as median (IQR) unless otherwise stated.
*In the Provent study, baseline ESS score was measured prior to CPAP withdrawal.
EQ-5D, EuroQol 5 Dimension Questionnaire; ESS, Epworth Sleepiness Scale; FOSQ, Functional Outcomes of Sleep Questionnaire; RCT, randomised controlled trial; SF-36, 36-Item Short Form Health Survey; VAS, Visual Analogue Scale.

Table 3 Pearson correlation coefficients between ESS Δ and potential anchor Δ scores

	RCT combination	
	1. autoCPAP+MOSAIC	2. autoCPAP+Provent
	0–3 months (autoCPAP) 0–6 months (MOSAIC) n=574	0–3 months (autoCPAP) 0–2 weeks (Provent) n=267
FOSQ total	–	**−0.56**
FOSQ general productivity	–	**−0.50**
FOSQ social outcome	–	**−0.31**
FOSQ activity level	–	**−0.57**
FOSQ vigilance	–	**−0.60**
SF-36 physical functioning	−0.15	–
SF-36 role physical	−0.25	–
SF-36 bodily pain	−0.22	–
SF-36 general health perception	−0.29	–
SF-36 energy/vitality	**−0.43**	–
SF-36 social functioning	−0.22	–
SF-36 role emotional	−0.14	–
SF-36 mental health	−0.18	–
SF-36 change in health	0.29	–
SF-36 physical component	**−0.30**	–
SF-36 mental component	−0.23	–
EQ-5D index	−0.07	–
EQ-5D VAS	−0.05	–

Values in bold indicate correlations sufficiently strong enough to use in the anchor-based methods.
EQ-5D, EuroQol 5 Dimension Questionnaire; ESS, Epworth Sleepiness Scale; FOSQ, Functional Outcomes of Sleep Questionnaire; RCT, randomised controlled trial; SF-36, 36-Item Short Form Health Survey ; VAS, Visual Analogue Scale.

cut-point of 2 points in the ESS was identified as the ESS change score that best discriminated between responders and non-responders based on the FOSQ total score MID. This estimate had the highest AUC out of the four domains analysed (0.82).

Only 13 patients from the Provent RCT provided data sufficient for calculating the ICC (≤20 days between tests), giving an ICC of 0.65. Therefore, we identified two previous studies that calculated the ICCs as 0.78 and 0.81 in translated versions of the ESS in OSA[28 29] and triangulated these with our ICC to a single ICC of 0.75 to use in our calculation of the SEM. All distribution estimates are shown in table 5. The estimates were mostly consistent across studies with estimates around two points. Overall, the distribution-based estimates were lower than the anchor-based estimates. The estimates calculated for each follow-up time point of the autoCPAP trial show minor increases of between 0.02 and 0.05 as the length of follow-up increases (table 5).

DISCUSSION

This is the first study to provide a single estimate of the MID for the ESS in OSA and our results show that the MID of the ESS is 2 points in this patient population. This MID was calculated using a combination of anchor-based and distribution-based approaches in a broad population of patients with OSA from three RCTs. Anchor-based estimates were slightly higher than distribution-based estimates.

Although change score correlations were too low to use the EQ-5D as an anchor, we observed moderate correlations between the ESS and FOSQ (r≥0.5 except for the social outcome domain) and slightly weaker correlations with the SF-36 (≥0.3), where only two domains were considered as anchors. This was expected as the FOSQ assesses the impact of sleepiness specifically on daily activities, whereas the SF-36 and EQ-5D measure general HRQoL. While the anchor-based estimates differed in the two populations, the two domains of the anchors used that best reflect the content of the ESS were the FOSQ energy/vitality and SF-36 activity domains. The estimates using these domains were on the smaller end within each population at −3.45 (95% CI −4.19 to −2.72) and −1.7 (95% CI −2.18 to −1.30), respectively. The distribution-based estimates were on average 1.9 and 1 point lower than the anchor-based estimates, with the estimates from the pooled data of all three RCTs being higher than the single RCTs. This is because a greater variability in change scores is represented, which may better reflect the true MID.[13] Examination of the distribution-based estimates revealed no evidence for different MIDs dependent on the follow-up time. Provent (2 weeks) and MOSAIC (6 months) gave almost identical estimates, and in autoCPAP estimates only increased negligibly as the length of follow-up increased (table 5). There was also no clear evidence for different MIDs between studies which included patients with different OSA severities; however, this should be explored in further studies with a combination of anchor and distribution methods used.

The triangulation of all estimates weighted according to our criteria led to a final MID of 2 points. In the anchor-based

Table 4 MID estimates and 95% CIs for the ESS based on anchors with correlations≥0.3

		RCT combination	
		1. autoCPAP+MOSAIC	2. autoCPAP+Provent
Anchor instrument	No of patients exceeding the anchor MID (%)	0–3 months (autoCPAP) 0–6 months (MOSAIC) n=572	0–3 months (autoCPAP) 0–2 weeks (Provent) n=264
FOSQ total	157 (59)	–	−3.03 (−3.71 to −2.35)
FOSQ general productivity	99 (38)	–	−4.21 (−5.05 to −3.37)
FOSQ social outcome	108 (41)	–	−4.00 (−4.78 to −3.21)
FOSQ activity level	137 (51)	–	−3.62 (−4.40 to −2.84)
FOSQ vigilance	134 (51)	–	−3.45 (−4.19 to −2.72)
SF-36 energy/vitality	398 (70)	−1.74 (−2.18 to −1.30)	–
SF-36 physical component	204 (37)	−2.66 (−3.19 to −2.13)	–

ESS, Epworth Sleepiness Scale; FOSQ, Functional Outcomes of Sleep Questionnaire; MID, minimal important difference; RCT, randomised controlled trial; SF-36, 36-Item Short Form Health Survey.

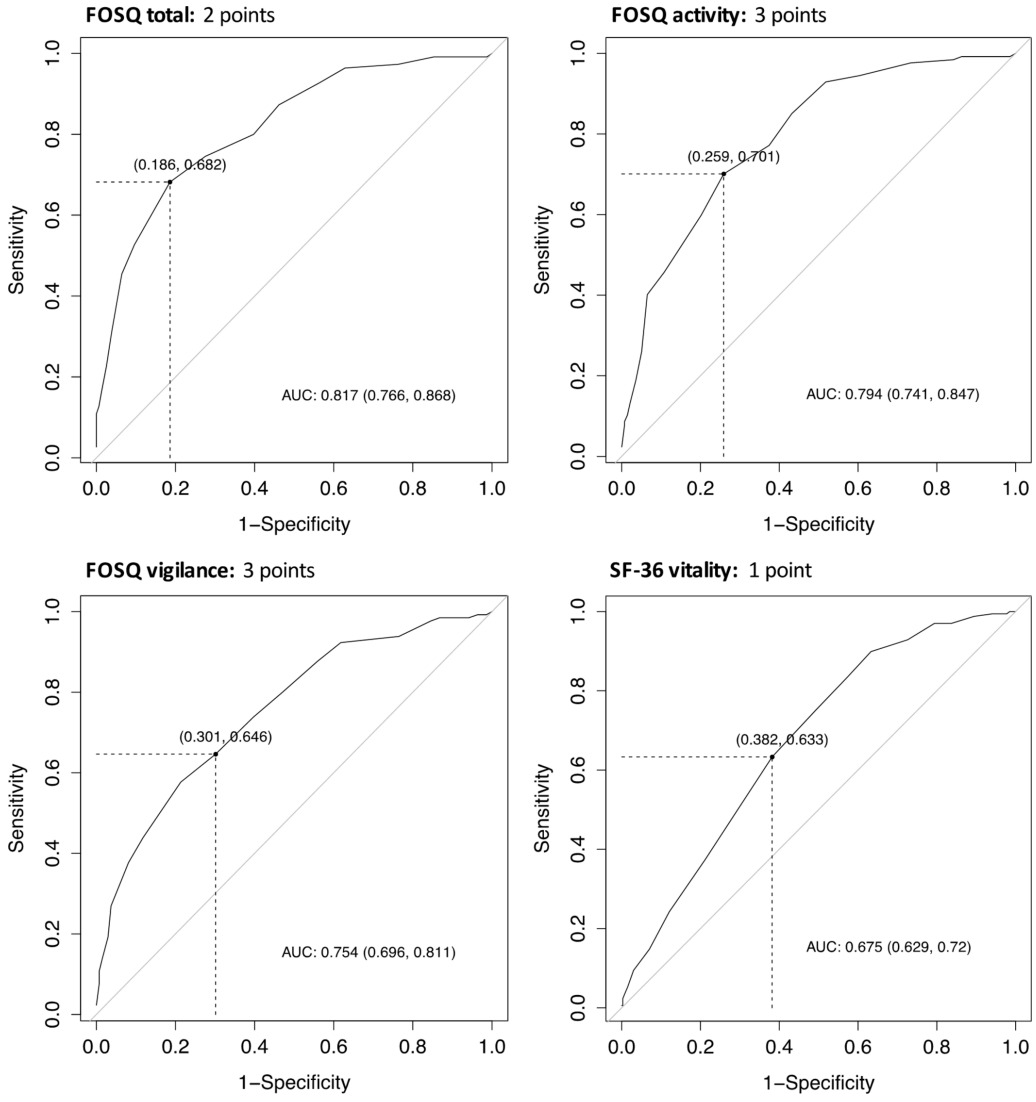

Figure 1 Receiver operating characteristic (ROC) curves identifying the change in ESS that best classifies responders and non-responders based on a change of more than or equal to anchor MID. FOSQ total score and domains are based on data from autoCPAP and Provent, and SF-36 vitality is based on autoCPAP and MOSAIC, using the 3-month follow-up in autoCPAP. Each plot shows the sensitivity and specificity, and the AUC (95% CI). AUC, area under the ROC curve; ESS, Epworth Sleepiness Scale; FOSQ, Functional Outcomes of Sleep Questionnaire; SF-36, 36-Item Short Form Health Survey.

estimates, we gave less weight to the FOSQ estimates because the MID used is not well-established since it calculated in a single study using only one distribution approach.[23] On the other hand, the SF-36 MID we used comes from several studies where several methods have been used to calculate it. Although the MID has not been established in patients with OSA, one study estimated the MID to be between 1 and 5 points in patients with rheumatoid arthritis,[21] who often experience symptoms of daytime sleepiness and may have a similar MID to patients with OSA.[30] We also gave more weight to study combination 1 (autoCPAP and MOSAIC) as the total sample size was much larger (n=574 vs n=267) and covers a wider range of patient severities. The distribution of change scores was also more evenly spread than study combination 2 (autoCPAP and Provent), where fewer patients deteriorated. In study combination 1, the ESS was generally more responsive relative to the FOSQ, with fewer patients improving more than the FOSQ domain MID compared with the ESS. This could lead to an overestimation of the ESS MID since a larger improvement in ESS will be seen for a 1 MID improvement in the FOSQ domain. This is also true for the SF-36 physical component score and reflects the anchors which gave larger MID estimates. Based on an ESS MID of 2 points, 72% of patients improved in study combination 1 and 55% of patients improved in study combination 2. Anchor-based

Table 5 Distribution-based estimates of the MID

Distribution method	Provent	autoCPAP	MOSAIC*	All
SE of measurement	1.89	1.77	2.14	2.36
Cohen's effect size	1.53	3 months: 1.97 12 months: 1.99 24 months: 2.04	1.52	2.08†
Empirical rule effect size	1.47	3 months: 1.89 12 months: 1.91 24 months: 1.96	1.46	1.99†

*Calculated only in the treatment group due to the lack of effect in the placebo group.
†Based on 3-month follow-up for autoCPAP.
MID, minimal important difference.

methods are generally considered to be superior to distribution-based methods as distribution methods are influenced by the data they are estimated in, and can underestimate the MID if based on a single study where strict inclusion criteria limit the population. In our data, the distribution estimates are less likely to be biased by these limitations as we pooled data from three RCTs with very different inclusion criteria. Therefore, we considered these estimates to provide valuable information about the MID and considered them equally in the triangulation. Overall, after consideration of all of these criteria, we came to a consensus that 2 points would be the appropriate estimate as the evidence for a higher MID from the FOSQ was not strong enough to increase the estimate to 3 points.

Our estimate of 2 points is similar to a previous study where an important improvement was estimated to be a change of between −2 and −3 points.[10] As this study was unable to determine a single value, interpretation of change remained complicated by an uncertainty around which MID to use. If an MID of 3 points was prespecified as the MID in the design of an RCT, an overall group improvement of 2.5 points would be identified as not being clinically significant, whereas our results indicate that a change of 2 points is sufficient and that 3 points is too large. Using an MID that is larger than the 'true' MID in a sample size calculation for an RCT would also result in a smaller sample size requirement that would then lack the power to see a statistically and clinically significant effect. One study that investigated the ESS in narcolepsy suggested a 25% change from baseline to be the threshold for improvement.[31] Applied to the three RCTs in our study, this would equate to a change of 2 points in Provent and MOSAIC, which are in agreement with our MID estimate, and 3.3 points in autoCPAP, which is higher than our estimate. This difference may reflect the larger change in ESS score seen in this study of narcolepsy (mean decrease of 8 points) compared with OSA, and 25% is likely to be too high for patients with OSA.

Improving the assessment of sleepiness in patients with OSA was recently identified as a key area that future research should prioritise.[32] Since the ESS is an important instrument for assessing sleepiness, it is especially important to determine its MID. The lack of a well-established MID means that, up until now, researchers using the ESS as an outcome had to make an a priori assumption about the change in ESS needed to determine the clinical significance of an intervention or rely on statistical significance. This results in an inconsistency of criteria used across studies, which can influence the final conclusion of the results and limits comparability. For example, out of the three RCTs used in this study, one assumed a change of 2 points to be important,[15] and two did not report a prespecified criterion and relied on statistical significance of the treatment effect.[14 16] Our robust calculation of the MID means that clinicians and researchers can now confidently use an MID that is based on strong evidence. The MID of 2 points can be used in the preparation of clinical trials to calculate the sample size required and determine the treatment effect needed to see an improvement. Clinicians can also use this MID in the management of patients over time to identify worsening of the disease or to assess response to treatment.

A major strength of this study is that we were able to include a wide spectrum of patients with OSA with different disease severities in several countries. Across the intervention and control groups in the three RCTs, patients underwent several different treatments including standard CPAP, autoadjusting CPAP, expiratory nasal resistance valve therapy after withdrawal from CPAP (actual and placebo) or standard care (without CPAP). This means that our results can be generalised to a wide range of patients and intervention study designs. We were able to pool data from the three RCTs, which increases the variability of between-patient responses. This leads to stronger correlations between the change scores of the ESS and anchors as well as a higher ESS baseline and change score SDs, which in turn increases the validity of our results.[33] The pooled data and increased variability also avoids underestimation of the MID in the distribution approaches. Another strength of our study is that we used variety of methodological approaches to produce a large number of estimates (27) on which to base our findings. We followed the US Food and Drug Administration guidance for patient-reported outcome development, where they recommend basing an important difference on several approaches, including anchor-based approaches using a responder definition (ie, anchor MID) and distribution-based approaches to provide supporting information.[12] A limitation of our study is that we were not able to pool all three RCTs into one dataset for the anchor-based approaches due to the different instruments assessed in each RCT. Although we were still able to combine two studies for each instrument, we may have seen stronger correlations with three different populations combined. Another limitation is that we relied on measures of HRQoL as anchors and were not able to use an anchor that reflects sleepiness specifically, such as an objective measure of daytime sleepiness, which could only be used as anchors if there was an established MID for these measures. Using anchors that measure the same (or similar) construct increases the validity of the MID estimate and reduces the chance of misleading results. However, HRQoL is known to be associated with the degree of sleepiness in OSA, and in our data the SF-36 and FOSQ met the requirements for being an anchor (having an appreciable association and being interpretable).[24] Where possible, future studies should look at measures directly relating to sleepiness as anchors. While the ESS has been found to have good reliability when assessed 7 days apart,[6] a recent study identified that patients with OSA had a high within-patient variability in scores when not undergoing an intervention.[34] Although all patient reported outcomes will have some measurement error, it is necessary to know whether change is due to measurement error or due to natural fluctuation in disease status or intervention.

CONCLUSION

In conclusion, we suggest to use an MID of 2 points for the ESS in patients with OSA. This MID is applicable to patients with a broad spectrum of OSA severity and to different study designs. Our estimate will facilitate the planning of clinical trials and provides the means to interpret the clinical relevance of changes in daytime sleepiness following interventions in patients with OSA.

Contributors SC, NAS and MK had full access to all data in the study. SC, NAS, KEB, JRS, AF, MAP and MK take responsibility for the integrity of the data and the accuracy of the data analysis. MAP and MK contributed to study design and obtained funding for the study. KEB, JRS and MK contributed to data collection. All authors contributed to analysis and interpretation of data. SC, AF, MAP and MK contributed to writing of the report. NAS, KEB, JRS, AF, MAP and MK contributed to critical revision.

Funding This work was supported by the Swiss National Science Foundation (32003B_162534) and an unrestricted grant by Bayer, Germany.

Competing interests MK and JRS declare advisory board fees from Bayer. Other authors have no completing interests to declare.

Patient consent Not required.

Ethics approval Research ethics committees in Zurich and Oxford.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Mulgrew AT, Ryan CF, Fleetham JA, *et al*. The impact of obstructive sleep apnea and daytime sleepiness on work limitation. *Sleep Med* 2007;9:42–53.
2 Lacasse Y, Godbout C, Sériès F. Health-related quality of life in obstructive sleep apnoea. *Eur Respir J* 2002;19:499–503.
3 Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 1991;14:540–5.
4 Bloch KE, Schoch OD, Zhang JN, *et al*. German version of the Epworth Sleepiness Scale. *Respiration* 1999;66:440–7.
5 Littner MR, Kushida C, Wise M, *et al*. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. *Sleep* 2005;28:113–21.
6 van der Heide A, van Schie MK, Lammers GJ, *et al*. Comparing treatment effect measurements in narcolepsy: the sustained attention to response task, epworth sleepiness scale and maintenance of wakefulness test. *Sleep* 2015;38:1051–8.
7 Hagell P, Broman JE. Measurement properties and hierarchical item structure of the Epworth Sleepiness Scale in Parkinson's disease. *J Sleep Res* 2007;16:102–9.
8 Chervin RD, Aldrich MS, Pickett R, *et al*. Comparison of the results of the epworth sleepiness scale and the multiple sleep latency test. *J Psychosom Res* 1997;42:145–55.
9 Kendzerska TB, Smith PM, Brignardello-Petersen R, *et al*. Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review. *Sleep Med Rev* 2014;18:321–31.
10 Patel S, Kon SSC, Nolan CM, *et al*. The epworth sleepiness scale: minimum clinically important difference in obstructive sleep apnea. *Am J Respir Crit Care Med* 2018;197:961–3.
11 Schwarz N, Sudman S. *Autobiographical memory and the validity of retrospective reports*. New York: Springer, 1994.
12 US Department of Health and Human Services Food and Drug Administration. Guidance for industry: patient- reported outcome measures: use in medical product development to support labeling claims. 2009 https://www.fda.gov/%20downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf%20Date
13 Revicki D, Hays RD, Cella D, *et al*. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102–9.
14 Craig SE, Kohler M, Nicoll D, *et al*. Continuous positive airway pressure improves sleepiness but not calculated vascular risk in patients with minimally symptomatic obstructive sleep apnoea: the MOSAIC randomised controlled trial. *Thorax* 2012;67:1090–6.
15 Bloch KE, Huber F, Furian M, *et al*. Autoadjusted versus fixed CPAP for obstructive sleep apnoea: a multicentre, randomised equivalence trial. *Thorax* 2018;73:174–84.
16 Rossi VA, Winter B, Rahman NM, *et al*. The effects of Provent on moderate to severe obstructive sleep apnoea during continuous positive airway pressure therapy withdrawal: a randomised controlled trial. *Thorax* 2013;68:854–9.
17 EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199–208.
18 Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;30:473–83.
19 Stewart AL, Greenfield S, Hays RD, *et al*. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *JAMA* 1989;262:907–13.
20 Wyrwich KW, Fihn SD, Tierney WM, *et al*. Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease: an expert consensus panel report. *J Gen Intern Med* 2003;18:196–202.
21 Kosinski M, Zhao SZ, Dedhiya S, *et al*. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis Rheum* 2000;43:1478–87.
22 Weaver TE, Laizner AM, Evans LK, *et al*. An instrument to measure functional status outcomes for disorders of excessive sleepiness. *Sleep* 1997;20:835–43.
23 Billings ME, Rosen CL, Auckley D, *et al*. Psychometric performance and responsiveness of the functional outcomes of sleep questionnaire and sleep apnea quality of life instrument in a randomized trial: the HomePAP study. *Sleep* 2014;37:2017–24.
24 Guyatt GH, Osoba D, Wu AW, *et al*. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–83.
25 McLeod LD, Coon CD, Martin SA, *et al*. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11:163–9.
26 Revicki DA, Erickson PA, Sloan JA, *et al*. Interpreting and reporting results based on patient-reported outcomes. *Value Health* 2007;10(Suppl 2):S116–24.
27 Swigris JJ, Brown KK, Behr J, *et al*. The SF-36 and SGRQ: validity and first look at minimum important differences in IPF. *Respir Med* 2010;104:296–304.
28 Izci B, Ardic S, Firat H, *et al*. Reliability and validity studies of the Turkish version of the Epworth Sleepiness Scale. *Sleep Breath* 2008;12:161–8.
29 Cho YW, Lee JH, Son HK, *et al*. The reliability and validity of the Korean version of the Epworth sleepiness scale. *Sleep Breath* 2011;15:377–84.
30 Omachi TA. Measures of sleep in rheumatologic diseases: Epworth Sleepiness Scale (ESS), Functional Outcome of Sleep Questionnaire (FOSQ), Insomnia Severity Index (ISI), and Pittsburgh Sleep Quality Index (PSQI). *Arthritis Care Res* 2011;63:S287–96.
31 Scrima L, Emsellem HA, Becker PM, *et al*. Identifying clinically important difference on the Epworth Sleepiness Scale: results from a narcolepsy clinical trial of JZP-110. *Sleep Med* 2017;38:108–12.
32 McNicholas WT, Bassetti CL, Ferini-Strambi L, *et al*. Challenges in obstructive sleep apnoea. *Lancet Respir Med* 2018;6:170–2.
33 Puhan MA, Chandra D, Mosenifar Z, *et al*. The minimal important difference of exercise tests in severe COPD. *Eur Respir J* 2011;37:784–90.
34 Campbell AJ, Neill AM, Scott DAR. Clinical reproducibility of the epworth sleepiness scale for patients with suspected sleep apnea. *J Clin Sleep Med* 2018;14:791–5.