Diagnostic Assessment Report commissioned by the NIHR on behalf of the National Institute for Health and Care Excellence

Novel home-testing devices for diagnosing obstructive sleep apnoea/hypopnoea syndrome - a systematic review and economic evaluation

ADDENDUM

Produced by	Southampton Health Technology Assessments Centre (SHTAC), University of Southampton, SO17 1BJ, UK[1] and Exeter Test Group, University of Exeter, EX1 2LU, UK [2]
Authors	Jaime Peters, Senior Research Fellow[2] Jonathan Shepherd, Principal Research Fellow[1] Emma Maund, Research Fellow[1] Bogdan Grigore, Associate Research Fellow[2] Lois Woods, Senior Research Assistant t[1] Joanne Lord, Professorial Research Fellow in Health Economics[1] David Alexander Scott, Principal Research Fellow[1] Chris Hyde, Professor of Public Health and Clinical Epidemiology[2]
Correspondence to	Dr Jonathan Shepherd Principal Research Fellow Southampton Health Technology Assessments Centre (SHTAC) University of Southampton Alpha House Enterprise Road, Southampton Science Park Southampton, SO16 7NS, UK. Email: jps@southampton.ac.uk
Date completed	19TH June 2024, updated 8th October 2024

1. Background

This is an addendum to the external assessment group (EAG) report produced by SHTAC and the Exeter Test Group for the NICE diagnostic assessment DG70. This addendum was originally produced in response to comments made by consultees on the draft NICE guidance published in 2024. It has been updated in October 2024 with information provided by the company (Sunrise) in response to Diagnostic Appraisal Committee (DAC) meeting 2 held in June 2024.

Stakeholder comments on the first draft NICE guidance for this topic noted a study of the Sunrise device in adults by Martinot et al 2022 had erroneously been classified as a secondary publication of an existing included study (Pepin et al 2020) in the EAG's systematic review. The company pointed out that the Martinot et al 2022 publication relates to a completely separate study, and its findings should therefore be included in the synthesis of study findings for consideration by the DAC. Below we present a narrative review of the study and its results, and the critical appraisal using the QUADAS2 instrument.

In this updated addendum, the text of the original addendum is in black while text describing new information provided by the company following DAC meeting 2 is in green.

2. Study design and characteristics

The primary focus of the publication was to explore the approach of near boundary labelling (NBL). The authors postulated that the risk of AHI-based severity mis-classification due to inter-human PSG rating could be reduced when considering borderline zones around the traditional fixed AHI thresholds. They applied the NBL approach to a clinical study aiming to validate a machine learning–based algorithm for mandibular movement signals (Sunrise, Namur, Belgium). Additional information provided by the company, subsequent to DAC meeting 2, clarified that the main objective of the study was to assess the diagnostic accuracy of Sunrise at the conventional AHI thresholds of 5, 15, and 30 events per hour using the PSG-AHI as the reference standard. Two methods were used: the NBL approach described above, and the standard diagnostic and severity classification rules set by the AASM (source: Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf).

Since the issue of NBL is not central to the scope of this diagnostic assessment, and for brevity, we focus below on the diagnostic performance of Sunrise in terms of sensitivity, specificity and other metrics. These estimates are presented without the use of NBL, for comparability with the results of other studies in our systematic review.

The study included 289 participants. Additional information provided by Sunrise, in consultee comments on the NICE draft guidance, stated that participants were aged 18 years and older and eligible for an in-laboratory sleep test for suspected OSAHS. The participants were initially referred due to a history of excessive daytime sleepiness, loud snoring, and/or witnessed apnoea.

[redacted]. Further information on the study population provided by Sunrise following DAC meeting 2 stated that: [redacted] (source: Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf).

The participants underwent an in-laboratory PSG (Somnoscreen; channels included EEG (Fz-A+, Cz-A+, Pz-A+), right and left EOG, submental EMG, tibial EMG, chest and abdominal wall motion by respiratory inductance plethysmography, nasal and oral airflow measured via a pressure transducer and a thermistor, respectively, and O_2 saturation assessed by a digital oximeter displaying pulse waveform) coupled with simultaneous MM recordings using the Sunrise device. Sleep technicians, who had undergone basic training on using the mobile app, placing the sensor, and performing basic troubleshooting, assisted patients with the setup of the MJM recording (source: DA70 Request for information Sunrise from EAG 110924 Sunrise [AIC].pdf). The PSG data were then manually scored by two experienced and blinded investigators. The collected MM data were automatically analysed by a machine learning algorithm developed by Sunrise.

3. Study results

The study reports that, based on the conventional rules for severity grading, the participants were categorized into non-OSA (n = 14; 4.8%), mild (n = 109; 37.7%), moderate (n = 113; 39.1%), and severe OSA (n = 53; 18.4%).

Table 1 below is a confusion matrix showing the distribution of participants classified across severity groupings by Sunrise and PSG.

Table 1 distribution of PSG-AHI scores within four conventional severity levels for PSG scoring and sunrise classification (NB. EAG converted proportions presented in study publication Figure 1 to numbers of patients)

	OSA Severity Sunrise				
OSA Severity PSG	**Normal**	**Mild**	**Moderate**	**Severe**	**Total**
Normal	12	2	0	0	14
Mild	2	100	7	0	109
Moderate	0	13	97	3	113
Severe	0	1	9	43	53
Total	14	116	113	46	289

Table 2 below gives diagnostic accuracy estimates based on the figures given in table 1 above. This is based on the threshold for test positivity incorporating mild, moderate and severe groupings combined.

Table 2 Diagnostic accuracy based on figures from above table (true positive = mild, moderate or severe OSA; true negative = not mild, moderate or severe OSA)

	Reference standard positive	**Reference standard negative**	**Total**
Index test positive	273	2	275
Index test negative	2	12	14
Total	275	14	289
Accuracy	98.62% (95% CI 96.49% to 99.62%)		
Diagnosis	**Value**		**95% CI**
Clinical sensitivity	99.27%		97.40% to 99.91%
Clinical specificity	85.71%		57.19% to 98.22%
PPV	99.27%		97.42% to 99.80%
NPV	85.71%		59.73% to 96.04%
Positive likelihood ratio [sensitivity/(1-specificity)]	6.95		1.93 to 25.07
Negative likelihood ratio [(1-sensitivity)/specificity]	0.01		0.00 to 0.03
Disease prevalence	95.16%		92.01% to 97.33%

Abbreviations: CI confidence interval; h, hour; NPV, negative predictive value; PPV, positive predictive value

Table 3 and Table 4 below, gives the diagnostic accuracy estimates of the Sunrise device at AHI threshold ≥ 15 events/hour and AHI threshold ≥ 30 events/hour, respectively

Table 3 Diagnostic accuracy of Sunrise to detect OSA at the AHI threshold ≥ 15 events/h

	Reference standard positive	Reference standard negative	Total
Index test positive	[redacted]	[redacted]	[redacted]
Index test negative	[redacted]	[redacted]	[redacted]
Total	[redacted]	[redacted]	[redacted]
Accuracy	[redacted]		
Diagnosis	**Value**		**95% CI**
Clinical sensitivity	[redacted]		[redacted]
Clinical specificity	[redacted]		[redacted]
PPV	[redacted]		[redacted]
NPV	[redacted]		[redacted]
Positive likelihood ratio [sensitivity/(1-specificity)]	[redacted]		[redacted]
Negative likelihood ratio [(1-sensitivity)/specificity]	[redacted]		[redacted]
Disease prevalence	[redacted]		[redacted]

Source: pages 8 and 9 and Table 4 in Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf
Abbreviations: CI confidence interval; h, hour; NPV, negative predictive value; PPV, positive predictive value
[a]Balanced accuracy – a metric that measures the average of sensitivity and specificity, providing a more balanced view of model performance when dealing with imbalanced datasets

Table 4 Diagnostic accuracy of Sunrise to detect OSA at the AHI threshold ≥ 30 events/h

	Reference standard positive	Reference standard negative	Total
Index test positive	[redacted]	[redacted]	[redacted]
Index test negative	[redacted]	[redacted]	[redacted]
Total	[redacted]	[redacted]	[redacted]
Accuracy	[redacted]		
Diagnosis	**Value**		**95% CI**
Clinical sensitivity	[redacted]		[redacted]
Clinical specificity	[redacted]		[redacted]
PPV	[redacted]		[redacted]
NPV	[redacted]		[redacted]
Positive likelihood ratio [sensitivity/(1-specificity)]	[redacted]		[redacted]
Negative likelihood ratio [(1-sensitivity)/specificity]	[redacted]		[redacted]
Disease prevalence	[redacted]		[redacted]

Source: pages 9 and 10 and Table 5 in Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf
Abbreviations: h, hour; NPV, negative predictive value; PPV, positive predictive value
[a]Balanced accuracy – a metric that measures the average of sensitivity and specificity, providing a more balanced view of model performance when dealing with imbalanced datasets

Following DAC meeting 2, NICE requested the company to provide accuracy estimates from the Martinot et al., 2022 study data using the ORDI thresholds values established in Pepin et al., 2020. The rationale behind this request was that the company had previously applied ORDI thresholds established in Pepin et al., 2020 to study data from Kelly et al., 2022. In response, the company provided [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

The EAG echoes the company's note of caution in the interpretation of the results: because the Martinot et al. 2022 publication focuses on AHI, applying ORDI threshold values identified in Pépin et al. 2020 study to the Martinot et al. 2022 study data may not provide relevant or meaningful insights (source: DA70 Request for information Sunrise 190924 Sunrise [AIC].docx)

In a second EAG addendum (October 2024) we compare the results from applying Pepin's thresholds to Martinot et al (2022) to the results when Pepin's thresholds are applied to the Kelly et al (2022) study data. We also estimate cost-effectiveness based on the results.

4. Critical appraisal

Appendix 1 gives the EAG's critical appraisal of the study. Additional information provided by the company following DAC meeting 2 allows us to make a more informed critical appraisal of study validity than previously. We now judged the study to be at low risk of bias for all domains.

References

Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf

DA70 Request for information Sunrise from EAG 110924 Sunrise [AIC].pdf

DA70 Request for information Sunrise 190924 Sunrise [AIC].docx

Martinot JB PJ, Malhotra A, Le-Dong N. Near-boundary Double-labelling Based Classification: The New Standard When Evaluating Performances of New Sleep Apnoea Diagnosis Solution Against Polysomnography? Sleep 2022;45(10) doi: https://doi.org/10.1093/sleep/zsac188

Pepin JL, Letesson C, Le-Dong NN, et al. Assessment of Mandibular Movement Monitoring With Machine Learning Analysis for the Diagnosis of Obstructive Sleep Apnea. *JAMA Network Open* 2020;3(1):e1919657.

Appendix 1. DAP70: QUADAS- 2 Risk of bias and applicability study assessments

Study - First Author: **Jean-Benoit Martinot** **Martinoot et al (2022a)**	**Year:2022**	**Rayyan No: 566581088**
DOMAIN 1: PATIENT SELECTION	**Assessment (delete as appropriate)**	**Comments**
A. Risk of Bias		

Signalling question 1: Was a consecutive or random sample of patients enrolled?	Yes	"Consecutive participants presenting with obstructive sleep apnea (OSA) suspicion"
Signalling question 2: Was a case-control design avoided?	Yes	"Consecutive participants presenting with obstructive sleep apnea (OSA) suspicion"
Signalling question 3: Did the study avoid inappropriate exclusions? *(Note: Remember that the device may be contraindicated in certain patient populations)*	Yes	Sunrise confirmed the participants were 18 years and older and eligible for an in-laboratory sleep test for suspected OSAHS. They were initially referred due to a history of excessive daytime sleepiness, loud snoring, and/or witnessed apnoea. Sunrise confirmed there were no exclusions.
Judgment: Could the selection of patients have introduced bias?	**RISK:** LOW	Consecutive participants presenting with obstructive sleep apnea (OSA) suspicion. There were no exclusions.
B. Concerns regarding applicability		
Judgment: Is there concern that the included patients do not match the review question?	**CONCERN:** LOW	"Consecutive participants presenting with obstructive sleep apnea (OSA) suspicion" Participants were eligible for an in-laboratory sleep test for suspected OSAHS. They were initially referred due to a history of excessive daytime sleepiness, loud snoring, and/or witnessed apnoea. There were no exclusions.
DOMAIN 2: INDEX TEST(S)	**Assessment (delete as appropriate)**	**Comments**

A. Risk of Bias		
Signalling question 1: Were the index test results interpreted without knowledge of the results of the reference standard? *(Note: Consider whether the index test was automatically scored by the software only, and could therefore be considered independent of the results of the reference standard)*	Yes	Data were automatically analysed
Signalling question 2: If a threshold was used, was it pre-specified? *(Note: for AHI and ODI, the following thresholds are standard (NICE scope, EAG protocol): Mild OSAHS: 5 or more to less than 15 events per hour; Moderate OSAHS: 15 or more to less than 30 events per hour; Severe OSAHS: 30 or more events per. If these specific thresholds are used but NOT prespecified we will not consider this an increase risk of bias)*	Yes	Conventional rules for severity grading based on the AHI
Judgment: Could the conduct or interpretation of the index test have introduced bias?	**RISK:** LOW	No comment
B. Concerns regarding applicability		
Judgment: Is there concern that the index test, its conduct, or interpretation differ from the review question?	**CONCERN:** UNCLEAR	No comment
DOMAIN 3: REFERENCE STANDARD		
A. Risk of Bias		
Signalling question 1: Is the reference standard likely to correctly classify the target condition?	Yes	In Laboratory PSG
Signalling question 2:	Yes	“The PSG data were then manually scored by two

Were the reference standard results interpreted without knowledge of the results of the index test?		experienced and blinded investigators”
Judgment: Could the reference standard, its conduct, or its interpretation have introduced bias?	**RISK:** LOW	
B. Concerns regarding applicability		
Judgment: Is there concern that the target condition as defined by the reference standard does not match the review question?	**CONCERN:** LOW	
DOMAIN 4: FLOW AND TIMING		
A. Risk of Bias		
Signalling question 1: Was there an appropriate interval between index test(s) and reference standard?	Yes	Simultaneous testing of Sunrise and PSG
Signalling question 2: Did all patients receive a reference standard?	Yes	(“Based on the conventional rules for severity grading, the participants could be categorized into non-OSA (n = 14; 4.8%), mild (n = 109; 37.7%), moderate (n = 113; 39.1%), and severe OSA (n = 53; 18.4%). Corresponding proportions of the seven categories in the NBL classification are presented in Table 1” – if you add the number of participants in each category the total is 289, which is the total sample of enrolled participants).
Signalling question 3:	Yes	In Laboratory PSG

Did patients receive the same reference standard?		
Signalling question 4: Were all patients included in the analysis?	Yes	"Based on the conventional rules for severity grading, the participants could be categorized into non-OSA (n = 14; 4.8%), mild (n = 109; 37.7%), moderate (n = 113; 39.1%), and severe OSA (n = 53; 18.4%)." – The sum of patients across the above categories is 289, which matches is number of enrolled participants.
Judgment: Could the patient flow have introduced bias?	**RISK:** LOW	

Diagnostic Assessment Report commissioned by the NIHR on behalf of the National Institute for Health and Care Excellence

Novel home-testing devices for diagnosing obstructive sleep apnoea/hypopnoea syndrome - a systematic review and economic evaluation

Second EAG addendum

Produced by	Southampton Health Technology Assessments Centre (SHTAC), University of Southampton, SO17 1BJ, UK[1] and Exeter Test Group, University of Exeter, EX1 2LU, UK [2]
Authors	Jaime Peters, Senior Research Fellow[2] Jonathan Shepherd, Principal Research Fellow[1] Emma Maund, Research Fellow[1] Bogdan Grigore, Associate Research Fellow[2] Lois Woods, Senior Research Assistant t[1] Joanne Lord, Professorial Research Fellow in Health Economics[1] David Alexander Scott, Principal Research Fellow[1] Chris Hyde, Professor of Public Health and Clinical Epidemiology[2]
Correspondence to	Dr Jonathan Shepherd Principal Research Fellow Southampton Health Technology Assessments Centre (SHTAC) University of Southampton Alpha House Enterprise Road, Southampton Science Park Southampton, SO16 7NS, UK. Email: jps@southampton.ac.uk
Date completed	8th October 2024

1. **Introduction**

This is an addendum to the external assessment report (EAR) report produced by SHTAC and the Exeter Test Group (the external assessment group, EAG) for the NICE diagnostic assessment DG70. This addendum been produced by the EAG in response to stakeholder comments on the draft NICE guidance published on July 17th 2024. We refer to this as the *'second EAG addendum'*.

In this addendum we provide a summary of additional information provided in response to the July 2024 draft guidance by two of the companies with devices relevant to the decision problem – Sunrise (Sunrise) and Nomics (Brizzy).

A previous addendum produced by the EAG for the second NICE diagnostic advisory committee meeting, held on 19th June 2024, has been updated in response to stakeholder comments on the July 2024 NICE draft guidance. In that addendum we report a summary and critical appraisal of the study by Martinot et al (2022) which investigated the Sunrise device. Since June 2024 the company (Sunrise) have provided further information on the Martinot et al (2022) study in response to a request from the EAG via NICE. The previous addendum (which we now refer to as the *'updated addendum (October 2024)'* has been updated with this further information. Please refer to the updated addendum for more details of our summary and critique of Martinot et al (2022).

2. Sunrise (Sunrise) - additional information

2.1 Application of diagnostic thresholds from Pepin et al (2020) to Kelly et al (2022) and Martinot et al (2022)

The company (Sunrise) previously provided diagnostic accuracy estimates for the Kelly et al (2022) study based on the post hoc optimised diagnostic thresholds first reported by Pepin et al (2020). These estimates were discussed at the second diagnostic advisory committee meeting for this topic on 19th June 2024.

Following a request from NICE (19th September 2024) Sunrise provided diagnostic accuracy estimates using the diagnostic thresholds reported by Pepin et al (2020) applied to the study sample from Martinot et al (2022). (DA70 Request for information Sunrise 190924 Sunrise [AIC].docx). Thus, it is now possible to compare the diagnostic accuracy of Sunrise at the optimised thresholds from Pepin et al (2020) to the accuracy based on these thresholds in two separate study samples (Kelly et al, 2022 and Martinot et al (2022)).

Table 1 below shows the diagnostic accuracy estimates for Kelly et al and Martinot et al from the company's retrospective application of Pepin's thresholds. Table 1 also shows the diagnostic accuracy estimates for the three studies at the thresholds reported in their respective study publications (NB. for Pepin and Kelly these are the same accuracy estimates as previously reported in Table 9 of the EAR; for Martinot these estimates are the same as those reported in the updated addendum (October 2024) and these have not been previously available to the DAC).

In Pepin's study the diagnostic accuracy of Sunrise (referenced to hospital sleep-laboratory PSG) was high, with sensitivity and specificity generally above 90% for the post hoc optimised thresholds of Sunrise RDI 7.63 and 12.65 (according to PSG conventional thresholds PSG-RDI ≥5 and PSG-RDI ≥15, respectively).

When the Sunrise RDI 7.63 threshold was retrospectively applied to Kelly's study data there was a slight increase in sensitivity (from 91% to 96%) but a notable decrease in specificity (from 94% to 60%) compared to Pepin et al. Of note, the Sunrise RDI threshold of 12.65 was the optimal threshold independently derived from post-hoc analyses in both Pepin et al and Kelly et al's studies. As would be expected, the diagnostic accuracy estimates reported by Kelly et al and the estimates reported by the company when applying the Pepin Sunrise threshold of RDI 12.65 are identical.

When Pepin's thresholds were applied to the Martinot et al study data, the differences in diagnostic accuracy estimates between the two studies were [redacted] compared to the differences between Pepin et al and Kelly et al (above).

- At the Sunrise RDI 7.63 threshold, sensitivity [redacted] from 91% (95% CI 89% to 92%) to [redacted], whilst specificity [redacted] from 94% (95% CI 91% to 97%) to [redacted].
- At the Sunrise RDI 12.65 threshold, sensitivity [redacted] from 92% (95% CI 90% to 94%) to [redacted], whilst specificity [redacted] from 84% (95% CI 81% to 87%) to [redacted].

2.2 Interpretation

The inclusion of Martinot et al (2022) provides greater insights into this issue and more certainty than was the case previously when thresholds were applied to just the Kelly et al study. The key observation from this current exercise is that where Pepin's thresholds are applied to all studies the diagnostic accuracy estimates from Pepin et al (2020) and Martinot et al (2022) are [redacted].

Martinot et al (2022) has a substantially larger sample size available for analysis compared to Kelly et al (n=289 versus n=31, respectively), and is more on a par with the Pepin sample size of n=376.
[redacted] The favourable EAG critical appraisal of Martinot et al (2022) (reported in the updated addendum) also provides reassurance in these findings.

However, there are some limitations to consider, particularly clinical heterogeneity across the studies. For example, the location of testing was the hospital sleep laboratory in both the Pepin and Martinot studies, whereas in Kelly's study testing was located in the patient's home. PSG was the reference standard in all three studies and our assumption is that there would be no difference in diagnostic performance. Notably, Sunrise caution against applying the ORDI threshold values identified in Pépin et al. to the AHI data from Martinot et al 2022 as it "may not provide relevant or meaningful insights." (page 2, "DA70 Request for information Sunrise 190924 Sunrise [AIC].docx"

Finally, the EAG would like to emphasise, as we have done previously, that neither the Kelly et al (2022) nor the Martinot et al (2022) studies were originally designed as diagnostic threshold validation studies. Kelly et al performed a post hoc optimisation of diagnostic thresholds based on their study sample (N=40 enrolled; N=31 analysed). In Martinot et al (2022), there was no mention of optimising thresholds from the study data, nor mention of applying (validating) previously optimised thresholds. Rather, the aim of Martinot et al (2022) was to assess the diagnostic accuracy of Sunrise using the Apnoea-Hypopnea Index (AHI) at the conventional thresholds of 5, 15, and 30 events per hour. The company's rationale for using the AHI rather than the ORDI (as done by Pepin and Kelly) is based on the assertion that the AHI is typically used by clinicians to diagnose OSAHS and categorise severity. It is evident, therefore, that diversity exists between these three studies in terms of their aims and objectives and their findings. This diversity somewhat limits the ability to make comparisons between the studies.

Table 1 Sunrise OSAHS diagnostic accuracy estimates for Pepin et al, Kelly et al and Martinot et al.

Author (year), **Novel device**	**No. pts**	**Cut-offs** Novel device, Reference standard	**Sensitivity % (95% CI)**	**Specificity % (95% CI)**	**PPV % (95% CI)**	**NPV % (95% CI)**	**Accuracy % (95% CI)**
Pepin et al (2020), Sunrise	376	Sunrise-RDI 7.63 PSG-RDI ≥5	91 (89 to 92)	94 (91 to 97)	99 (99 to 99)	59 (55 to 63)	92 (90 to 94)[c]
		Sunrise-RDI 12.65 PSG-RDI ≥15	92 (90 to 94)	84 (81 to 87)	89 (88 to 91)	88 (85 to 91)	88 (86 to 90)[c]
Kelly et al (2022), Sunrise	31	MM-ORDI 9.53 PSG-ORDI >5	88 (69 to 97)	100 (54 to 100)	100 (85 to 100)	89 (NR)	94 (NR)
		MM-ORDI 12.65 PSG-ORDI >15	100 (79 to 100)	75 (45 to 92)	80 (NR)	100 (NR)	88 (NR)
		MM-ORDI 24.81 PSG-ORDI >30	79 (NR)	96 (NR)	95 (NR)	82 (NR)	87 (NR)
Kelly et al (2022), Sunrise *Pepin thresholds*	31	Sunrise-RDI 7.63 PSG-ORDI ≥5	96 (NR)	60 (NR)	(NR)	(NR)	(NR)
		Sunrise-RDI 12.65 PSG-ORDI ≥15	100 (NR)	75 (NR)	(NR)	(NR)	(NR)
Martinot et al (2022) Sunrise	289	Sunrise AHI ≥5 PSG AHI ≥5	99 (98 to 100)	86 (69 to 99)	99 (98 to 100)	86 (69 to 100)	92 (84 to 98)
		Sunrise AHI ≥15 PSG AHI ≥15	92 (88 to 95)	94 (91 to 97)	96 (93 to 98)	89 (84 to 93)	93 (90 to 95)
		Sunrise AHI ≥30 PSG AHI ≥30	81 (72 to 90)	99 (97 to 100)	93 (87 to 98)	96 (94 to 98)	90 (85 to 94)
Martinot et al (2022)	289	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Author (year), **Novel device**	**No.** **pts**	**Cut-offs** Novel device, Reference standard	**Sensitivity %** **(95% CI)**	**Specificity %** **(95% CI)**	**PPV %** **(95% CI)**	**NPV %** **(95% CI)**	**Accuracy %** **(95% CI)**
Sunrise *Pepin thresholds*		[redacted] [redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

AHI Apnoea-hypopnoea index; NPV Negative predictive value, NR Not reported; ODI Oxygen Desaturation Index; ORDI Obstructive Respiratory Disturbance Index; PPV positive predictive value; Pts patients; PSG Polysomnography, RDI Respiratory disturbance index;

2.2 Cost effectiveness based on external application of diagnostic thresholds from Pepin et al (2020)

We ran the model using different data sources for Sunrise compared to oximetry and respiratory polygraphy (using accuracy estimates from the NG202 meta-analysis). The results are shown in Table 2.

Table 2 Deterministic cost-effectiveness results for Sunrise compared to respiratory polygraphy and oximetry using different data sources for Sunrise accuracy estimates

Setting of evaluation	**Sunrise data**			
Data	**Pepin et al**	**Kelly et al**	**Martinot et al**	**Martinot et al**
Cut-off used	**Pepin**	**Pepin**	**Pepin**	**Martinot**
Compared to respiratory polygraphy				
Incremental cost	-£40	£167	[redacted]	-£73
Incremental QALYs	0.022	0.048	[redacted]	0.018
ICER (£ per QALY gained)	Dominant	£3,506	[redacted]	Dominant
INMB at £20,000 per QALY gained	£480	£787	[redacted]	£428
INMB at £30,000 per QALY gained	£700	£1,265	[redacted]	£606
Compared to oximetry				
Incremental cost	£664	£871	[redacted]	£631
Incremental QALYs	0.096	0.122	[redacted]	0.092
ICER (£ per QALY gained)	£6,897	£7,140	[redacted]	£6,855
INMB at £20,000 per QALY gained	£1,261	£1,569	[redacted]	£1,210
INMB at £30,000 per QALY gained	£2,223	£2,788	[redacted]	£2,130
Abbreviations: ICER incremental cost-effectiveness ratio; INMB incremental net monetary benefit; QALYs quality-adjusted life-years				

For all data sources and at all cut-offs used, Sunrise is estimated to be more cost-effective than respiratory polygraphy and oximetry. When using Kelly et al. accuracy data at Pepin cut-offs, the model estimates that Sunrise is more expensive, but also more effective than respiratory polygraphy. This is mainly due to the estimate of 1 for sensitivity at the high cut-off, meaning that everyone with moderate or severe OSA is correctly identified as such, and so the costs and benefits of treatment in these people are fully captured. When using accuracy data from Martinot et al 2022 at the cut-offs from Pepin et al, Sunrise is estimated to be slightly more expensive than respiratory polygraphy, but more effective.

[redacted]

[redacted] Where they are misdiagnosed as not having OSA, there are no benefits for these individuals. Where they are misdiagnosed as having moderate or severe OSA, they are incurring higher treatment costs.

When using the accuracy data as reported in Martinot et al 2022, Sunrise is estimated to dominate respiratory polygraphy. This is due to all estimates of sensitivity and specificity (at the high and low cut-offs) being estimated as higher than that for home RP.

We undertook probabilistic analyses to demonstrate the greater uncertainty in estimates from the Kelly et al study, since only 31 participants contribute to the estimates of sensitivity and specificity. There appears to be slightly more variation in the probabilistic analysis based on estimates from Kelly et al than from Pepin et al or Martinot et al, but this is not particularly marked, see Figure 1 and Figure 2.

In the probabilistic analysis, we observe more variation in the inputs for Pepin et al and Martinot et al than for Kelly et al. It is, however, the specificity estimates where there is more variation in Kelly et al than in Pepin et al or Martinot et al. Estimates of specificity only affect the lower half of the decision tree, where diagnosis is made of those who are truly mild or truly have no OSA. Differences in this part of the model generally have smaller impacts on the total costs and QALYs than any variation in the sensitivity estimates. Variation in estimates of sensitivity, which affects what happens to be people with moderate or severe OSA, may be diluted by the fact that people with mild or severe OSA who are misdiagnosed as having no OSA, will receive another test and then be more likely to be correctly diagnosed. We believe this and the fact that estimates of sensitivity and specificity are limited (0,1) explains why we are not seeing much variance, even though Kelly et al has 31 participants, and Pepin et al has almost 400 participants.

[redacted]

Figure 1 Cost-effectiveness plane for Sunrise compared to oximetry using different data sources at the thresholds used by Pepin et al

[redacted]

Figure 2 Cost-effectiveness plane for Sunrise compared to oximetry using different data sources at the thresholds used by Pepin et al

3. Brizzy (Nomics) – additional information

3.1 The JawRhin1 study

3.1.1 Study aims and methods

The manufacturer of the Brizzy device, Nomics, submitted details of a study in progress, the ‘Sleep Respiratory Disorders in Patients With Moderate to Severe Persistent Rhinitis (JawRhin1) study’ (ClinicalTrials.gov ID NCT04012216), evaluating the diagnostic performance of the Brizzy device in detecting sleep disordered breathing in patients with moderate to severe persistent rhinitis. The company states that the results of this study “provides important additional information” in response to recommendation 1.4 in the July 2024 draft NICE guidance that “more research is needed on using Brizzy home-testing device to diagnose and assess the severity of OSAHS in people 16 years and over before it can be used in the NHS”.

The company provided:

- A brief narrative textual summary of the study, covering its aims, methods and some findings. (Nomics_comment_SEP24.docx).
- A study report containing selected study findings illustrated in scatter plots, ROC curves and tables (Rapport Jawrhin.pdf). The study report provides no accompanying narrative summary or discussion of any of the data presented.
- [redacted]. The EAG notes that the study record on clinicaltrials.gov has additional information on the trial methods and design, but results are not yet publicly available. The study is due for completion by 31st July 2025.
- The study report also contains a series of meta-analysis forest plots showing the pooled diagnostic accuracy estimates of Brizzy from the JawRhin1 study and a study labelled as ‘Martinot JB., et al’. We believe this is the same study as Martinot et al (2017) included in our systematic review. (Martinot et al. 2017 assessed the performance of Brizzy against the reference standard sleep laboratory PSG). The results of the meta-analyses are presented based on two diagnostic accuracy thresholds established by post hoc optimisation in Martinot et al. (2017), approximating to mild OSAHS (RDI PSG ≥5 =RDI Sunrise JAWAK >5.9) and moderate-to-severe OSAHS (RDI PSG ≥15 =RDI Sunrise JAWAK >13.5).

The aim of the study is “to demonstrate that the measurement of respiratory effort assessed by mandibular movements during sleep is a useful measure for the screening of sleep disordered breathing in patients with moderate to severe persistent rhinitis”. The study also compared the results from the automated analysis of the JAWAC signal alone (i.e. Brizzy) versus standard practice using simulated respiratory polygraphy. Simulation involved examining the channels of the PSG that would have been available in a respiratory polygraphy (nasal flow, thoracic and abdominal belts, oximetry, accelerometer).

3.1.2 Relevance of JawRhin1 to the decision problem

The EAG assessed the relevance of the study by applying the inclusion/exclusion criteria from our systematic review of diagnostic test accuracy. In the EAG’s judgement this study does not meet the inclusion criteria and therefore will not be included as relevant evidence for Brizzy in the systematic review and the economic evaluation. The primary reason for its exclusion is because the study population is clinically different from the population in the systematic review. As stated in the review protocol “The relevant population for this assessment is people presenting with signs and symptoms suggestive of OSAHS, considered suitable for home-testing”. To be enrolled in the JawRhin1 study patients had to present with persistent, moderate-to-severe rhinitis. However, there is no explicit statement that they had to present with signs and symptoms suggestive of OSAHS. We excluded other studies (not necessarily of the Brizzy device) in rhinitis patients when screening studies for inclusion in the systematic review.

The EAG’s understanding is that rhinitis is a risk factor for OSAHS and that some people with rhinitis also have OSAHS, but some don’t. For example, a meta-analysis of 44 studies containing 6086 participants (Cao et al, 2018) found that for adults, the prevalence of allergic rhinitis was 22.8% (95% CI, 15.0–30.6) in people with sleep disordered breathing and 35.2% (95% CI, 25.6–44.7) in people with OSA.

A secondary reason for excluding the JawRhin1 study is because it investigated the performance of Brizzy in the screening of sleep disordered breathing, rather than diagnosis. Screening and diagnostic testing in health are done for different purposes, but the focus in this NICE DA is on solely on diagnostic testing in people suspected with OSAHS.

The company believes this study helps to address the DAC’s concerns about the lack of evidence on the diagnostic performance of Brizzy in adults. The committee were concerned that diagnostic performance of Brizzy in the Martinot 2017 study was based on post hoc optimisation of cut-offs. There were no other relevant studies of Brizzy included in the review, and hence no independent validation of these cut-offs in separate sample of patients for comparison.

3.1.3 Diagnostic threshold validation

Whilst the JawRhin study is outside the scope of this diagnostic assessment, in the EAG's opinion it *partly* addresses the committee's concerns:

- The study report gives diagnostic accuracy results for Brizzy (PSG reference standard) using the optimised cut offs from the Martinot 2017 study (RDI_JAWAK at >5.9; >13.5 and >32.5 per hour).
- The company don't explicitly describe JawRhin as a validation study, but we note that the objective was "to determine a mandibular movement respiratory disturbance index (MM-RDI) threshold associated with a polysomnography respiratory disturbance index (PSG-RDI) ≥ 15 / h in a population of patients with moderate-to-severe persistent rhinitis"). Without further elaboration this statement is open to interpretation, but the fact that optimised RDI cut offs from a previous study (i.e. Brizzy) are applied to this study suggests some intent for external validation, albeit in a different condition.

The meta-analyses forest plots provided in the study report allow the diagnostic accuracy estimates from JawRhin to be compared with those of Martinot 2017 at the same thresholds. They also show the results when data from the two studies are pooled statistically. However, caution is needed in the interpretation because they are clinically distinct patient groups (i.e. Martinot et al 2017 suspected OSA; JawRhin1 moderate to severe rhinitis). The company acknowledges this: [redacted]. The EAG notes an additional uncertainty due to lack of detail on the meta-analysis methods used, including whether a bivariate approach was taken to assess the correlation between sensitivity and specificity. We note, however, that meta-analyses estimates are given according to a common effect and a random effects model, thus illustrating the degree of variation in precision.

3.1.4 Simulated respiratory polygraphy

Regarding the comparison of Brizzy with the reference standard of simulated respiratory, there may be issues of validity and generalisability arising from the simulation itself. Further information on the procedures are needed to arrive at a judgement on its appropriateness.

As far as the EAG can tell, none of the results currently available are based on this comparison (results seem to be against PSG reference standard).

3.1.5 EAG conclusion

We are mindful that in the absence of further relevant evidence for Brizzy the DAC may wish to consider the currently available results of the JawRhin study in their deliberations. Clearly there are limitations because of the differences in the study populations, but also due to the lack of clarity in the reporting of study methods and results, as we have stated above. The EAG has not been able to perform an informed QUADAS2 critical appraisal of the study, so there the risk of bias is currently uncertain. If the DAC considers this study as potentially informative then further information from the company, when available, will aid our understanding of the study's aims and objectives, its methods, and interpretation of its results.

3.2 Updated price

When the cost of £44 was assumed for Brizzy and the RP accuracy data were taken from the NG202 meta-analysis, Brizzy dominated RP. Assuming the reduced costs for Brizzy, leads to slightly improved estimates of cost-effectiveness, i.e. the estimated INMB have increased, see Table 3.

Table 3 INMB for Brizzy compared to oximetry and RP for differing costs of Brizzy (where home RP and oximetry accuracy estimates are from NG202 meta-analysis)

	INMB at £20,000		**INMB at £30,000**	
Brizzy device cost	**v oximetry**	**v RP**	**v oximetry**	**v RP**
£44 (base case)[a]	£1,119	£337	£1,934	£410
£39.02	£1,124	£342	£1,939	£415
£35.10	£1,128	£346	£1,943	£419

[a] as shown in Table 40 of DAP70 EAG report

INMB, incremental net monetary benefit; RP, respiratory polygraphy

References

Cao Y, Wu S, Zhang L, Yang Y, Cao S, Li Q. Association of allergic rhinitis with obstructive sleep apnea: A meta-analysis. Medicine (Baltimore). 2018 Dec;97(51):e13783. doi: 10.1097/MD.0000000000013783. PMID: 30572534; PMCID: PMC6319794.
Clinical Validation of SunriseEvaluation v0.1 [AIC].pdf

DA70 Request for information Sunrise 190924 Sunrise [AIC].docx

Kelly JL, Ben Messaoud R, Joyeux-Faure M, et al. Diagnosis of Sleep Apnoea Using a Mandibular Monitor and Machine Learning Analysis: One-Night Agreement Compared to in-Home Polysomnography. Frontiers in Neuroscience 2022;16 doi: 10.3389/fnins.2022.726880

Martinot J-B, Borel J-C, Cuthbert V, et al. Mandibular position and movements: Suitability for diagnosis of sleep apnoea. Respirology 2017;22(3):567-74. doi: https://doi.org/10.1111/resp.12929

Martinot JB PJ, Malhotra A, Le-Dong N. Near-boundary Double-labelling Based Classification: The New Standard When Evaluating Performances of New Sleep Apnoea Diagnosis Solution Against Polysomnography? Sleep 2022;45(10) doi: https://doi.org/10.1093/sleep/zsac188

Nomics_comment_SEP24.docx

Pepin JL, Letesson C, Le-Dong NN, et al. Assessment of Mandibular Movement Monitoring With Machine Learning Analysis for the Diagnosis of Obstructive Sleep Apnea. *JAMA Network Open* 2020;3(1):e1919657.

Nomics Rapport Jawrhin.pdf