HEALTHTECH ASSESSMENT PROGRAMME

Home-testing devices for diagnosing obstructive sleep apnoea hypopnoea syndrome

Draft guidance – Comments

Committee date: 22 October 2024

THEME: Recommendation for the Sunrise test

Comment number	Name and organisation	Section number	Comment	NICE Response
1	Consultee 2 Acurable	Not specified	Context: The guideline recommendation has revised the earlier decision not to endorse the Sunrise device. This change is based on Sunrise manufacturer's retrospective analysis, which evaluated the device's performance using the dataset from Kelly's study along with the post-hoc thresholds from Pepin's study. Despite a decline in performance when using these non-optimal thresholds, the device's effectiveness was found to be comparable to that of another system, WatchPat, which had been recommended. Initially, the device was not endorsed because the committee correctly determined that the evidence was highly biased, given that the thresholds were determined post-hoc and varied between the two studies. However, the justification provided by the manufacturers for the new evaluation of the evidence is scientifically weak and not conclusive enough to be incorporated into the health economics model. Furthermore, recommending this technology as part of the guideline could pose significant risks to patients that cannot be quantified with the existing data. These risks are unique to the Sunrise device and do not apply to any other of the devices that have been recommended. Specifically, it is crucial to consider that the Sunrise technology is based on Machine Learning (ML), which significantly influences the strength of the evidence supporting the recommendation. All other recommended technologies are based on conventional signal processing (i.e algorithms are based on physiological and clinical knowledge. They are not data driven).	Thank you for your comments which the committee considered. During consultation, the company provided further details of a study with a larger sample size (Martinot 2022). The external assessment group (EAG) reviewed the information and further analysis of the study results provided by the company. Please see EAG updated addendum 1 and addendum 2 for further details.

Comment number	Name and organisation	Section number	Comment	NICE Response
			Hence, the reasons listed below, supporting the case of the evidence not being strong enough to change the recommendation on the basis of the re-interpretation of the data, are directly related to the fact that Sunrise's technology relies on ML, distinguishing it from the other recommended technologies. ● A major challenge with machine learning (ML) is generalization and overfitting. Overfitting occurs when a model learns the details and noise (anything in the signal not relevant to the problem of interest) in the training data to such an extent that it performs poorly on new data. The only way to validate that overfitting is not occurring is by using a sufficiently large testing database. All patients included in Pepin's paper should still be considered training data since the thresholds were chosen after analyzing the ROC curve. Therefore, the number of testing data points, as presented in Kelly's paper, is only 31. This number is not large enough to determine whether overfitting has occurred. It is important to note that this issue affects only the Sunrise device. The other devices on the recommended list are based on conventional signal processing, which relies on general theories and does not "learn" from data in the same manner as ML. ● The fact that the optimal cut-off value in Kelly (2022), 9.53 events per hour, is so far from that reported in Pépin (2020), 7.63 events per hour, clearly demonstrates that the method is highly sensitive to the patient population and not necessarily generalisable. ● With such a small validation dataset of only 31 patients, it is impossible to determine whether the training data used to develop the models for the device was adequate to avoid the well-recognized issue of biased predictions. In machine learning, inadequate or "noisy" training data can inadvertently render a model ineffective or inaccurate when applied to the intended population. Furthermore, even if the models did not suffer from this issue—which is impossible to ascertain given the limited data in Kelly's paper—the performance	The committee considered this further evidence and agreed this provided further reassurance about the diagnostic accuracy of the Sunrise test (as described in section 3.3 to 3.4 of the guidance).

Comment number	Name and organisation	Section number	Comment	NICE Response
			significantly deteriorates when the thresholds from Pepin's study are applied. This introduces even greater doubts about the validity of the model. ● The authors of Pépin (2020) and Kelly (2022) do not state how many features their machine learning model employs to make a diagnosis; however, Kelly (2022) states "Input features consisted of a combination of axes of the accelerometer/gyroscope, processing modes (filter with several frequency bands, moving average) and statistical functions.", implying 12 or more features. To ensure a validation sample appropriately describes the feature space for a machine learning method, it is necessary to have many times more subjects than features (e.g. https://www.nature.com/articles/s41592-018-0019-x). With possibly only 3 times as many subjects as features, this study cannot possibly have sufficient samples of patient variability to draw any significant conclusions about its generalisability to future patients. ● The fact that the training data was not representative of the UK population and relied on an unclear, non-UK referral process for patient recruitment is concerning, especially when combined with the reliance on performance results from testing on only 31 individuals. To illustrate, a study by DeGrave et al. (2020) on machine learning (ML) classification of chest radiographs for COVID-19 detection found that their model's performance deteriorated when tested on a dataset collected from a different source. A detailed investigation revealed that the model was taking 'shortcuts' by exploiting unrelated features present only in the initial dataset. The heterogeneity of data sources or a strong understanding of the source is crucial in mitigating these effects for ML algorithms. It is therefore worrisome that the training was not conducted with a UK population or following UK patient referral pathways, and there is already a noticeable degradation in performance in the testing sample, which also does not represent the UK population. This issue is specific to ML-based algorithms. ● A further concern arises from the supplementary information in Pepin's study,	

Comment number	Name and organisation	Section number	Comment	NICE Response
			which describes the development of the algorithm. The authors seem to state that they used 30-second epochs as training data. If this is the case, this would be scientifically wrong because the device produces an output at 10-second epochs or less—the duration necessary for diagnosis since events are 10 seconds long so it cannot be any longer. While the performance might seem acceptable, the reliance of machine learning approaches on non-physiological/clinical knowledge poses the significant risk that this is an artifact and in fact the algorithms (which have been trained on 30s epochs as opposed to 10s) might not be functioning correctly. This issue could easily be overlooked due to the small size of the validation database; with only 31 patients, it is insufficient to ensure robust validation. ● One further concern is the fact that the authors claimed that there was a previous database of 100 patients. However they don't use the thresholds in that database when they evaluate the algorithm in Pepin and instead they choose the optimum one in the ROC curve. Why is that? And why is that again different in Kelly? A validation study for a marketed device must be done with a "frozen algorithm" (i.e. not an algorithm that has got programmable parameters). ● Which threshold is in the marketed device? Has it changed with respect to the publication they are now using in the evidence (ie. Supporting the recommendation). If that threshold has changed that publication cannot be used as evidence. ● Even disregarding the limitations of the very small sample size in Kelly (2022), guidelines on the clinical use of Machine Learning methods (e.g. NHS Guidance on AI and ML -https://www.england.nhs.uk/long-read/artificial-intelligence-ai-and-machine-learning/) state that external and prospective studies are necessary for the validation of Machine Learning methods before clinical use. Kelly (2022) is undoubtedly not a prospective validation study.	

Comment number	Name and organisation	Section number	Comment	NICE Response
			In addition to the ML related reasons as to why the re-interpretation of the evidence does not make it strong enough there is another additional reasons to the ones already identified by the EAG which is also listed below: ● In the Sunrise system, the range for the moderate region changes from the clinically accepted [5,15] (i.e. 10 units) to [7.63, 12.65] (i.e. 5 units). This is not a one to one mapping of indexes, which makes interpretation impossible for clinicians and defeats the point of having indexes. The Sunrise device does not provide the conventional channels, and without conventional mapping of the indexes that the clinicians can follow would make any attempt for interpretation impossible.	
2	Consultee 3	Not specified	A colleague passed this consultation to me for a second opinion on their scientific evaluation of the validation of the Sunrise device and its subsequent recommendation by NICE. In their expert opinion, the criteria followed by NICE for the recommendation completely ignored the fact that this is an ML-based device. Such devices need to be treated differently from non-ML devices due to their inherent risks. While I am not an expert on OSA, I specialise in ML and medical devices and lead a world-renowned centre in this field. I think NICE should reconsider its recommendation when it comes to the Sunrise device. Recommending for wide use a device that uses machine learning (ML) algorithms validated on such a small sample size (~30 people since the paper with a bigger sample size was effectively used for training), poses very significant risks, which are inherent to ML (i.e. they don't apply to devices that are based on conventional signal processing): Limited Generalisability: A sample of 30 individuals is unlikely to represent the broader population accurately. Variations, amongst others in demographics, and usage scenarios might not be captured, leading to an algorithm that performs poorly outside the test group. Related to this, the population this was developed on is in general very different demographically to that in the UK	Thank you for your comments which the committee considered. Please see the responses to comment 1.

Comment number	Name and organisation	Section number	Comment	NICE Response
			which significantly increases the probability of this risk for the UK population. Overfitting: In ML there's always a very significant risk of the algorithm overfitting, where it learns the specific details and noise of the training data rather than general patterns. This can result in a model that works well on a small validation set but poorly in real-world applications. In this context the diagnostic threshold has been fitted to the data to justify the validity of the post-hoc determination to NICE. But in ML this is not enough justification for final performance data when the validation data set is small (as in this case). Bias and Variability: A small sample size can introduce significant bias, especially if the sample is not randomly selected or is homogeneous. This can skew the algorithm's performance, favouring certain groups while disadvantaging others. See the comment in point 1 about the demographics of the population in the country where this was tested being very different to the UK, so the “consecutive” criteria in recruitment does not minimise this risk because the referred population is very different to start with. Validation and Reliability: It is well known by anyone working in ML that the reliability and validity of the model are questionable when validated on a small sample. Reproducibility issues might arise, where subsequent tests on larger or different datasets yield inconsistent results. Hence there is a significant risk recommending this for wider use in the UK population. User Trust and Safety: NICE recommending a ML device based on insufficient evidence can erode public trust in NICE, all new technologies, innovation, and ML as a whole if such a device later proves ineffective. This could significantly hinder progress in improving healthcare. In my opinion, in order to mitigate these risks, as with any other ML based device, prior to recommendation, evidence has to be collected involving a much larger blind validation datasets in a UK population, with a trial fully powered in advance, and conducting extensive testing in varied real-world conditions before commercial deployment. This is the only way of assuring that the ML algorithms are reliable, generalisable, and safe for all intended users.	

Comment number	Name and organisation	Section number	Comment	NICE Response
			One additional comment, not directly related to the above, is that while becoming familiar with the device, I checked the UK distributor's (Sefam) website. I noticed that they recommend the device for screening only, not for diagnostics, which is interesting given that there is no reimbursement pathway for screening in the UK. While this could be a potential use case scenario, reducing the risks through proper implementation would incur additional costs. Moreover, what would the pathway for this be? It does not exist within the NHS, meaning it falls outside the assumptions of the health economics model. However, recommending specific population screening could be something for NICE to consider in the future. On a different note, since NICE asks for comments on equality, the distributor's website mentions that the device is not suitable for individuals unwilling to shave. That rules out a very significant percentage of the UK population.	
3	Consultee 4	3.3 The accuracy evidence for Sunrise is acceptable for decision making	This section is confusing. Upon reading 3.3, the conclusion of a reader is that, with the committees concerns being justified, the Sunrise device should not in fact be used for diagnosis. It is not explained why, therefore, the previous recommendation changed?	Thank you for your comments which the committee considered. The committee considered the analysis of the Kelly (2022) study using the cut-off values established in the Pepin (2020) study as acceptable for decision making. The committee has now reviewed further evidence on the

Comment number	Name and organisation	Section number	Comment	NICE Response
				Sunrise test which was provided by the company during this consultation. Please see EAG updated addendum 1 and addendum 2 for further details. The committee agreed that this provided further reassurance about the accuracy of the Sunrise test. Section 3.3 in the guidance has been updated.
4	Consultee 4	3.3 The accuracy evidence for Sunrise is acceptable for decision making	Are the cut-off values from Pepin 2020, used in Kelly 2022 as described, the same cut-off values which are used in this product in real world use? The only relevant cut offs, and consequent accuracy estimates, should surely be the ones that are used in the NHS or general clinical use? I have read the comments from the previous consultation to understand this better but think the recommendation and report would benefit from further explanation as to how the optimal cut offs in these papers relate to the "final version" used in the NHS eg in Scotland.	The company confirmed that the Sunrise device reports both AHI (using conventional cut-off values of 5, 15, 30 events per hour) and obstructive respiratory disturbance index (ORDI) (using cut-off values of 7.63 and 12.65 events per hour established in the Pépin study) when used in clinical

Comment number	Name and organisation	Section number	Comment	NICE Response
				practice (as described in section 3.4 of the guidance).
5	Consultee 4	3.3 The accuracy evidence for Sunrise is acceptable for decision making	If the concern is justified, why has the additional data resulted in a change of the recommendation? This does not appear to be a logical conclusion to the statement.	Thank you for your comments which the committee considered. Please see the response to comment 3.
6	Consultee 5 Sunrise	3.3	Regarding the statement, “But when asked in the first committee meeting why it had characterised the difference in cut-off values as small, and to what extent this may have affected test accuracy estimates, the company was unable to justify this comment,” could you please specify, “… was unable to justify this comment, as the Sunrise representative in charge of these aspects of the development was unable to attend the meeting.” Regarding the statement, “During consultation on the first draft of the guidance, the company provided accuracy estimates from the Kelly (2022) data set. These were produced using cut-off values that had been established in the	Thank you for your comments which the committee considered. The external assessment group (EAG) explained that Kelly (2022) was not intended as a prospective validation study. The approach taken is not the same

Comment number	Name and organisation	Section number	Comment	NICE Response
			previous Pepin study. The EAG highlighted that these were applied retrospectively, that is, after the study had been completed, but that it did consider this data informative," could you please specify that accuracy estimates for all possible cut-off values are provided in Figure 3 of the Kelly publication, including therefore the cut-off values established in the previous Pepin study. Indeed, we would like to reiterate and emphasize the following points: • In Kelly et al. 2022, the diagnostic accuracy was evaluated by the authors for all possible cut-off values of the Sunrise-ORDI scale at conventional PSG cut-off values of 5 and 15 events/hour, providing an unbiased and extensive evaluation of diagnostic accuracy across the entire Sunrise-ORDI measurement scale for the study. The results of this evaluation were presented in the publication (Figure 3) for utmost transparency and are available for everyone to review. The objective of this method was under no circumstances to increase or bias the diagnostic accuracy of Sunrise. Instead, these results have allowed reviewers and editors of the leading journal to recognize the diagnostic accuracy of Sunrise. • Figure 3 of the Kelly et al. 2022 publication, available for everyone to review, provides therefore the Sunrise diagnostic accuracy at the pre-established cut-off values identified in Pepin et al. 2020. This responded to the recommendation of the diagnostics advisory committee to apply the Pepin et al. 2020 cut-off values to another data set. The results unequivocally confirmed that there is no uncertainty about the high Sunrise diagnostic accuracy and the correct identification of the Sunrise cut-off values in Pepin et al. 2020. The term "retrospectively" seems irrelevant in this context. While the cut-off values from the Pepin et al. 2020 study were identified before the Kelly et al. 2022 study began, the Kelly et al. 2022 publication assessed independently and	as prospectively applying the cut-offs from an existing study to a purposively designed validation study. The application of cut-off values from previous studies to Kelly's sample is a retrospective/post hoc exercise, but the EAG did consider this data informative. The EAG has provided a more informed critical appraisal for Martinot (2022) study with the additional information. Please see the EAG updated addendum 1 for further detail.

Comment number	Name and organisation	Section number	Comment	NICE Response
			in full transparency the Sunrise diagnostic accuracy at all possible cut-off values, including therefore the cut-off values established in the Pepin et al. 2020 study. Pointing out the diagnostic accuracy at these pre-established cut-off values from the figure, rather than from the optimized cut-off values of the Kelly study data set, did not introduce any bias. This could be clarified as follows: “During consultation on the first draft of the guidance, the company highlighted that accuracy estimates from the Kelly (2022) data set were provided for all cut-off values in the publication, including the cut-off values established in the previous Pepin study. The EAG did consider this data informative.” Considering the precisions above, could you please specify “initially” in the second sentence of the section: “But the EAG initially judged both studies to be at high risk of bias for interpreting the index test, because they reported accuracy data using test cut-off values that were not predefined.” Regarding the statement, “Accuracy estimates from the Kelly (2022) data used the cut-off value for OSAHS set in the Pepin (2020) study (7.63 events per hour) rather than the optimised value set in the Kelly study (9.53 events per hour),” could you please specify that this concerns the conventional PSG cut-off value of 5 events per hour. This could be clarified as follows: “For the conventional PSG cut-off value of 5 events per hour, accuracy estimates from the Kelly (2022) data used the cut-off value for OSAHS set in the Pepin (2020) study (7.63 events per hour) rather than the optimised value set in the Kelly study (9.53 events per hour).”	

<table>
<tr><th>Comment number</th><th>Name and organisation</th><th>Section number</th><th>Comment</th><th>NICE Response</th></tr>
<tr><td></td><td></td><td></td><td>Regarding the statement, “But the EAG commented that limited detail was available for this study, so it was not possible to do a critical appraisal,” we hope that the additional information provided in the comments below allows for a full critical appraisal of the risk of bias and that the overall judgment is now clear.

We are happy to assist in facilitating communication with the authors to ensure the committee receives all necessary details.</td><td></td></tr>
<tr><td>7</td><td>Consultee 5 Sunrise</td><td>3.17</td><td>Regarding the statement, “The committee discussed that disposable devices would have an environmental cost,” it is certainly worth mentioning that the Sunrise device, although designed for single use, includes a prepaid envelope for returning it by post. This allows for the recycling of the device's components, thereby mitigating its environmental impact.</td><td>Thank you for your comments which the committee considered.</td></tr>
<tr><td>8</td><td>Consultee 5 Sunrise</td><td>EAG addendum page 2</td><td>Regarding the details of the socio-demographic and health characteristics of participants, please find additional information provided by the authors below:

• This clinical validation includes 289 patients (18 years and older) eligible for an in-laboratory sleep test for suspected OSAHS. The patients were initially referred due to a history of excessive daytime sleepiness, loud snoring, and/or witnessed apnoea. They were consecutively included when referred for a single overnight in-laboratory PSG.
• The characteristics of the 289 included patients (144 men and 145 women) are described in the table below as mean, SD, median, and 5th-95th percentiles:

<table>
<tr><td>Parameters</td><td>Mean ± SD</td><td>Median</td><td>5th – 95th percentiles</td></tr>
<tr><td>Age (years)</td><td>47.55 ± 13.07</td><td>47.45</td><td>25.63 – 69.79</td></tr>
<tr><td>BMI (kg/m2)</td><td>32.41 ± 8.06</td><td>31.05</td><td>21.12 – 46.89</td></tr>
<tr><td>Neck circ. (cm)</td><td>40.00 ± 3.73</td><td>40.00</td><td>34.00 – 47.00</td></tr>
</table></td><td>Thank you for your comments which the committee considered.

The external assessment group (EAG) has updated the EAG addendum. Please see the EAG updated addendum 1 for further detail.</td></tr>
</table>

Comment number	Name and organisation	Section number	Comment	NICE Response
			We are happy to assist in facilitating communication with the authors to ensure the committee receives all necessary details.	
9	Consultee 5 Sunrise	EAG addendum Appendix 1	Regarding the row "Signalling question 3: Did the study avoid inappropriate exclusions?" on page 5, the assessment should be "Yes." The sample consisted of 289 consecutive patients presenting with suspected OSAHS, and results were provided for all 289 patients. Therefore, there were no exclusions. This information is present in the publication. This assessment is coherent with the "Yes" assessment provided for the following publications for example: • Martinot et al., 2015 (child): "Consecutive children and no exclusion criteria given so presume there were no exclusions." • Martinot et al., 2017: "Consecutive patients were consenting adults '18 years and older with symptoms suggestive of sleep-disordered breathing (SDB) undergoing a single PSG.'" • Pepin et al., 2020: "Consecutive adults with suspected OSA and no exclusion criteria given so presume there were no exclusions. Furthermore, 'The final data set included all 376 patients recruited' (p4)." Additional information provided by the authors as mentioned above: • This clinical validation includes 289 patients (18 years and older) eligible for an in-laboratory sleep test for suspected OSAHS. The patients were initially referred due to a history of excessive daytime sleepiness, loud snoring, and/or witnessed apnoea. They were consecutively included when referred for a single overnight in-laboratory PSG.	Thank you for your comments which the committee considered. The external assessment group (EAG) has provided a more informed critical appraisal for Martinot (2022) study with the additional information. Consequently, all but one of the domains of the QUADAS-2 instrument are judged as low risk of bias. Please see the EAG addenda for further detail. The committee considered the additional information and agreed this provided further reassurance about the

<table>
<tr><th>Comment number</th><th>Name and organisation</th><th>Section number</th><th>Comment</th><th>NICE Response</th></tr>
<tr><td></td><td></td><td></td><td>

The characteristics of the 289 included patients (144 men and 145 women) are described in the table below as mean, SD, median, and 5th-95th percentiles:

<table>
<tr><td>Parameters</td><td>Mean ± SD</td><td>Median</td><td>5th – 95th percentiles</td></tr>
<tr><td>Age (years)</td><td>47.55 ± 13.07</td><td>47.45</td><td>25.63 – 69.79</td></tr>
<tr><td>BMI (kg/m2)</td><td>32.41 ± 8.06</td><td>31.05</td><td>21.12 – 46.89</td></tr>
<tr><td>Neck circ. (cm)</td><td>40.00 ± 3.73</td><td>40.00</td><td>34.00 – 47.00</td></tr>
</table>
Based on the above information, the assessments for the rows “Judgment: Could the selection of patients have introduced bias?” and “Judgment: Is there concern that the included patients do not match the review questions?” on page 5 should be “Low.”

Regarding the row “Signalling question 2: If a threshold was used, was it pre-specified?” on page 6, the assessment should be “Yes.”

Indeed:

The publication states “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 EAG assessment template includes the following caveat: “for AHI and ODI, the following thresholds are standard (as per NICE scope, and 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 hour. If these specific thresholds are used but NOT prespecified, this is not considered an increased risk of bias.”

</td><td>accuracy of the Sunrise device.</td></tr>
</table>

Comment number	Name and organisation	Section number	Comment	NICE Response
			• We previously confirmed that the evaluation of Sunrise diagnostic accuracy with the AHI was performed using the conventional PSG cut-off values, like for other home-testing devices. This assessment is coherent with the "Yes" assessment provided for the following publications for example: • Devani et al., 2021: "'Using the current AASM diagnostic criteria.'(p4)." • Mueller et al., 2022: "Standard thresholds were used 'Mild OSA was defined as an AHI of 5 to <15, moderate OSA was defined as an AHI of 15 to <30, and severe OSA was defined as an AHI ≥ 30.'" • Lyne et al., 2023: "Used standard thresholds for OSA severities (no reference cited for these) 'Secondary outcomes included the agreement between NOM and NOR and PSG with respect to diagnostic classification of OSA across four categories: no-OSA group (AHI < 5 events per hour), mild OSA (AHI 5–14 events per hour), moderate OSA (AHI 15–29 events per hour), and severe OSA ≥ 30 events per hour.' (p.1430)." Following the comment above, the assessment of the row "Judgment: Could the conduct or interpretation of the index test have introduced bias?" on page 6 should be "Low." The assessment of the row "Judgment: Is there concern that the index test, its conduct, or interpretation differ from the review question?" on page 6 may remain "Unclear" because the index test was conducted in a sleep laboratory and not in a home setting. We are happy to assist in facilitating communication with the authors to ensure the committee receives all necessary details.	

Comment number	Name and organisation	Section number	Comment	NICE Response
10	Consultee 5 Sunrise	EAG addendum page 4	We hope that the additional information provided in the comments above and below allows for a full critical appraisal of the risk of bias and that the overall judgment is now clear. We are happy to assist in facilitating communication with the authors to ensure the committee receives all necessary details.	Thank you for your comments. The committee has considered the additional information.

THEME: Recommendation for the Brizzy test

Comment number	Name and organisation	Section number	Comment	NICE Response
11	Consultee 6 Nomics	1.2	Recommendation 1.2 highlights the need for the choice of device to be tailored to the patient. Since the first diagnostics consultation, Nomics has received interest from NHS clinicians who want to use Brizzy to conduct clinical evaluations or service evaluations, which will generate data to address the gaps that are common to all the devices evaluated and Brizzy in particular. If the committee maintains the current evaluation of Brizzy in “further research needed” over “conditionally recommended”, these evaluations might no longer happen, reducing the change of UK-based evidence generation and thus stifling innovation. This evaluation could be done without risk to the quality of care afforded to patients. Indeed, as noted in the guidance, current practices already allow for the use of either ventilatory polygraphy or oximetry. And, as noted in Table 1	Thank you for your comments which the committee considered. The committee considered the amount and quality of evidence available and factored this into its decision making. Overall, the committee

Comment number	Name and organisation	Section number	Comment	NICE Response
			"Device specifications" of the guidance, Brizzy has the option to use oximetry (and thus provide ODI). Although Nomics is a small company, it has nevertheless generated a very large number of studies on the mandibular movements, and it remains committed to robust evidence generation. This will be significantly impaired if the current draft recommendation remains unchanged. The UK, and England specifically, will be essential in generating new evidence for the use of novel devices for diagnostic purposes, as other countries are not yet ready to recommend novel device for this purpose (see comment 2 for more details).	considered that there was not sufficient evidence to recommend use of Brizzy (as described in section 3.5 and 3.6 of the guidance).
12	Consultee 6 Nomics	1.4	Recommendation 1.4 does not consider all of the relevant evidence, bearing in mind the clinical context in which the current devices are being investigated. All international guidelines on the diagnosis of sleep apnoea specify that only a ventilatory polygraph– also known as Type-3 devices in the AASM classification– can be used for the diagnosis of sleep apnoea. Similarly, in NG202, NICE recommend home respiratory polygraphy for the diagnosis of sleep apnoea and only considers oximetry if access to polygraphy is limited. The novel devices being evaluated, including Brizzy, are not Type-3 devices nor home respiratory polygraphy as defined in NG202. The current guidance is, therefore, taking a step in extending the type of device that can be used for diagnostic purposes. As the reviewed guidance is taking a new approach to the diagnostic of sleep apnoea, it explains why, although the JAWAC sensor at the core of Brizzy has been included in 23 clinical studies (all identified in the Request for Information submitted to NICE), most of these studies were not strictly within the scope of this assessment and, therefore, not considered by the EAG. Instead, most studies looked at Brizzy as a screening tool in accordance with the available guidelines.	Thank you for your comments which the committee considered. The external assessment group (EAG) explained that the reason not all Brizzy studies were included in the systematic review is, because they were not fully relevant to the scope of this assessment. Thus, they did not meet all of the inclusion criteria. In the case of the study by Rotty et al

Comment number	Name and organisation	Section number	Comment	NICE Response
			One of those studies was led by Rothy et al. in 2017 and used the pre-specified cut-off of 13.5/h that was identified in the Martinot et al. 2017 paper. This study was conducted at home and provided an ambulatory failure rate, addressing a further evidence gap identified by the by the EAG and the committee. This study was designed with Brizzy as a screening device, but this was consistent with the available guidelines, and still provide important additional details. In summary, it is unfair to single out Brizzy for a 'more research needed' recommendation when the evidence base for all of the technologies is limited in quality and quantity.	(2017), it did not meet the inclusion criteria as the title of the publication indicated the study was screening for OSAHS rather than diagnosing OSAHS. The committee considered the evidence for each device and concluded that there was not enough evidence to recommend use of Brizzy at this time.
13	Consultee 6 Nomics	1.4	Recommendation 1.4 is not a suitable basis for guidance to the NHS and is not a fair summary of the clinical and cost effectiveness evidence, because the committee's decision to change the recommendation on Brizzy (from 'use' to 'more research needed') is disproportionate in response to the EAG's revised quality assessment of one domain of one study. It is also inconsistent with recent guidance on other technologies with similar levels of uncertainty in the evidence base, where NICE has frequently made recommendations for use with prospective evidence generation in line with section 6.4.12 of the Health Technology Evaluations Manual. Recent examples include: tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer (DG58); the use of artificial intelligence (AI)-derived software to help clinical decision-making in stroke (DG57); and devices for remote monitoring of Parkinson's disease (DG51).	Thank you for your comments which the committee considered. The committee considered the amount and quality of evidence available. The committee reiterated its opinion that accuracy estimates should be

Comment number	Name and organisation	Section number	Comment	NICE Response
				generated from a different data set to that used to set test cut-off values. They agreed there is not enough evidence to reduce the uncertainty about the accuracy of the Brizzy device to identify and assess severity of obstructive sleep apnoea hypopnoea syndrome.
14	Consultee 6 Nomics	1.4	Not all of the relevant evidence has been considered. The main issue that the committee has raised regarding the inclusion of Brizzy is the revised assessment by the EAG of the index test domain of the QUADAS-2 tool of the Martinot et al. 2017 study. The committee considered (section 3.4) that "accuracy estimates should be generated from a different data set to that used to set test cut-off values". As mentioned in comment #2, the lack of evidence for Brizzy as a diagnostic tool is mostly the result of the framing the current guidelines rather than a lack of research. At the time of writing, several studies comparing the JAWAC sensor at the core of Brizzy with a PSG are ongoing. [redacted]	Thank you for your comments which the committee considered. The external assessment group (EAG) assessed the new evidence provided by Nomics. Please see EAG's addendum 2 for further detail.

Comment number	Name and organisation	Section number	Comment	NICE Response
			[redacted] The new evidence will specifically address the methodological weakness of the Martinot et al. 2016 (Respirology) study on which the EAG's updated critical appraisal changed the committee's recommendations. This is because the evidence will provide more data comparing polysomnography and Brizzy at the specified cut-off of 13.5/h. The new evidence will include a meta-analysis of at least three studies (Martinot 2017 and two new studies) to calculate the specificity and sensitivity data for the 13.5/h cut-off. We expect this to be of sufficient quality for decision-making. The reasons Nomics cannot provide NICE with the data on the 7th of August are manifold: - Some comments on Brizzy were raised at the last meeting, which means we had to respond quickly – and during the challenging summer season when partners are often less available. - Most of the data are part of a recent study, and patient recruitment has finished recently. Therefore, more time is needed before submitting evidence of the highest standard that NICE would expect. [redacted] Nomics it is a small company with few employees and limited resources. Nomics only entered the UK market after the start of the current guidance	

Comment number	Name and organisation	Section number	Comment	NICE Response
			development and was approach by NICE to be part of this guidance. Therefore, this guidance development process has been challenging for Nomics, especially with the often short turnaround time to submit response. For example, Nomics does not directly employ a health economist or someone with a background capable on working on the EAG model, meaning independent consultants were needed. These considerations also explain why it is challenging for Nomics to reply promptly to the committee, especially when some points are raised late in the process.	
15	Consultee 6 Nomics	3.3/3.4	Sections 3.3 and 3.4 of the second draft guidance differentiate Brizzy and Sunrise based on the diagnostic accuracy studies, leading to different recommendations. However, both devices are based on the study of mandibular movement, although they differ technically and in the way they are used. This is reflected in the clinical development studies where, for example, the studies selected by the EAG to assess Sunrise include many references to Brizzy or JAWAC (the sensor used in Brizzy). Kelly et al. (2022) cites nine studies on Brizzy/JAWAC, while Pepin et al. (2020) cite four. Despite the technical differences between the two devices (e.g., Sunrise has one sensor point, Brizzy has two, Sunrise uses AI; Nomics uses mathematical algorithms), both shows, despite similar uncertainties in the evidence, similar sufficient promise for the home diagnosis of sleep apnoea to be recommended to be used with prospective evidence generation. This would be consistent with NICE's guidance on transperineal biopsy for diagnosing prostate cancer (DG54), which states: "There are technical differences between them, but they all work in a similar way using the same biopsy technique. So, these devices are recommended as options for diagnosing prostate cancer."	Thank you for your comments which the committee considered. The committee noted the similarities and differences between the Sunrise and Brizzy devices. Experts on the committee explained that the use of mandibular movement as a measure is not standardised so evidence must be reviewed on each device. The committee reiterated its opinion that the current

Comment number	Name and organisation	Section number	Comment	NICE Response
				evidence for Brizzy is not enough to reduce the uncertainty around its accuracy to identify and assess severity of obstructive sleep apnoea hypopnoea syndrome. More research is needed on Brizzy to address the concern.
16	Consultee 6 Nomics	3.4	As stated elsewhere, the committee has decided not to recommend Brizzy based on the revised assessment by the EAG of the index test domain of the QUADAS-2 tool of the Martinot et al. 2017 study, as “accuracy estimates should be generated from a different data set to that used to set test cut-off values.” While Nomics understands the committee's point and is determined to provide new evidence in a timely manner, the committee's decision to no longer recommend Brizzy is severe in light of the methodological limitation that was raised. This is because Martinot et al. 2017 was designed to follow the guidance of the AASM task force on the assessment of the Sleep Apnoea test device (Collop et al. 2009), which requires the authors to provide a post-hoc analysis of the results to generate an ideal cut-off point. The AASM guidance was also used for studies of other devices in the evaluation. In addition, the data presented by Martinot et al. (2017) includes an idea of the expected variation of true sensitivity and specificity through the confidence interval for each value. At the lowest bound, the sensitivity and specificity at the 13.5/h cut-off remain high at 79.83% and 83.18%, respectively. Nomics did not	Thank you for your comments which the committee considered. The committee reiterated its opinion that accuracy estimates should be generated from a different data set to that used to set test cut-off values. The current evidence for Brizzy is not enough to reduce the uncertainty around its accuracy to identify and assess severity

Comment number	Name and organisation	Section number	Comment	NICE Response
			raised those points during the last committee meeting as it was not asked to respond to the committee's concern on this point. [redacted]	of obstructive sleep apnoea hypopnoea syndrome. More research is needed on Brizzy to address the concern.
17	Consultee 6 Nomics	1.4	Recommendation 1.4 is not a sound basis for guidance to the NHS because it would deprive patients of the option of one of the only few devices in the evaluation, which does not require that the patients have internet access or a smartphone to be used according to Table 1 "Device specifications" of the guidance. The importance of this is articulated in the committee's equality consideration in section 3.13 of the draft guidance.	Thank you for your comments which the committee considered.
18	Consultee 6 Nomics	Table 1 Device specifications	When the current Diagnostics Assessment Programme started, Nomics had not yet entered the UK market, nor was its British subsidiary, Nomicscare UK Ltd, been created. Therefore, at that time we could only provide the cost of our test in other markets and in euros. This was converted and used by the EAG. We now have a clear pricing structure for Brizzy. As a reminder, Nomics provides the test/analysis with all cost included (devices consignment, consumable, software, CDS tool, support) rather than the devices outright. Prices for Brizzy as of September 2024, VAT EXCLUSIVE - For 100 tests, £ 39.02 per test - For 200 tests, £ 36.93 per test	Thank you for your comments which the committee considered. The external assessment group (EAG) did further cost-effective analyses based on the updated pricing structure. Please see EAG's addendum 2

Comment number	Name and organisation	Section number	Comment	NICE Response
			- For 400 tests, £ 36.17 per test - For 600 tests, £ 35.10 per test	for further detail. The committee has considered the additional information for Brizzy.
19	Consultee 6 Nomics	1. Recommendation, point 1.4	Please find, attached, additional data regarding the question raised by the committee. The provided data comes from the JawRhin1 study (ClinicalTrials.gov ID NCT04012216). This study was conducted on a particularly difficult population for home-testing devices as patients with rhinitis suffers from an increased resistance in the upper airways which can lead to increased respiratory and to events associated with micro-arousals (such as Respiratory Effort-Related Arousals (RERAs), but also hypopneas with micro-arousals. This primary outcome of this study was 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. However, the researchers also wanted to identify the added value of jaw movements compared to standard practice (i.e. respiratory polygraphy). To do so, they compared the results from the automated analysis of the JAWAC signal alone (i.e. Brizzy) vs a simulated respiratory polygraphy. To simulate the respiratory polygraphy and recreate an REI, the researchers used the channels of the polysomnography that would have been available in a respiratory polygraphy (nasal flow, thoracic and abdominal belts, oximetry, accelerometer). From there, they calculate an REI that includes the apnoea as well as the hypopnea followed by a 3% O_2 desaturation. The researched teams then compared the performance of the REI form the JAWAC alone (Brizzy) vs the simulated respiratory polygraphy.	Thank you for your comments which the committee considered. The external assessment group (EAG) assessed the relevance of the study. Please see EAG's addendum 2 for further detail. The committee has considered the additional information from Brizzy.

Comment number	Name and organisation	Section number	Comment	NICE Response
			This data provides important additional information. It compares Brizzy with the gold standard (PSG) at the pre-specified cut-off of 13.5/h, but also with the current practice (RP) which is what the committee is looking to understand. Moreover, the results provide significant insights. While the specificity and sensitivity in this cohort are lower than in Martinot et al (2016), this can be explainable by the choice of population (patient with moderate/severe rhinitis vs general population). More importantly, when comparing the results of the Brizzy vs the simulated polygraphy, it is evident that Brizzy provide a much higher accuracy in that population. The methodology has some limitations. Indeed, Brizzy was not compared directly to a respiratory polygraphy, but to a simulated one. However, the research team believes that this provided the simulated RP with an increased advantage. Indeed, not only are the signal used to simulate the RP directly extracted from the gold standard, but the scoring for the simulated RP was done manually by experts in Sleep Medicine. Conversely, Brizzy events were scored automatically by its algorithms. Please see https://clinicaltrials.gov/study/NCT04012216 for the published protocol.	

THEME: Comments on the guidance and recommendations

Comment number	Name and organisation	Section number	Comment	NICE Response
20	Consultee 1 ARTP (Association for Respiratory Technology & Physiology)	2	These devices use new technologies that clinicians may be unfamiliar with. Consideration must given to the support required to allow clinicians to interpret the data obtained and to determine what is a high quality study and what is an unacceptably study.	Thank you for your comments which the committee considered.
21	Consultee 1 ARTP (Association for Respiratory Technology & Physiology)	3.1 Impact of using home-testing devices for people with suspected OSAHS	This is a very important point. Estimates of capacity within Sleep Clinics for initiating treatment(s) are based on numbers of patients tested. If large volumes of unplanned referrals are received for treatment(s) this has the potential to shift the bottleneck from diagnosis to treatment, which will not ultimately reduce patient wait times.	Thank you for your comments which the committee considered.
22	Consultee 4	Not specified	A really interesting read and shows promising options for patients undergoing sleep testing going forward. Thank you to NICE and the teams that have put this together for taking comments from the public into consideration.	Thank you for your comments.
23	Consultee 4	1	Overall, it's exciting to see NICE consider and review novel technologies which have the potential to really help the NHS recover from the current long waiting times experienced.	Thank you for your comments.

Comment number	Name and organisation	Section number	Comment	NICE Response
24	Consultee 4	3.1 Impact of using home-testing devices for people with suspected OSAHS	All of these devices sound like they could be much easier to use than traditional systems and look like they would be less annoying for patients. It's great to see the committee highlight this.	Thank you for your comments.
25	Consultee 4	3.7 Home-testing devices may reduce healthcare resource use, but the extent is uncertain	My experience was that the wait can be much more than 6 weeks, closer to 6 months	Thank you for your comments which the committee considered. The 6-week estimate reflects the situation cited by clinical experts consulted for this assessment. The draft guidance section 3.8 notes expert opinion that 'waiting lists are still growing' and that 'services are under considerable pressure'.
26	Consultee 4	3.14 Some home-testing devices are cost effective for diagnosing OSAHS in people 16 years and over	It is not clear what the cut-off values used in the commercially available product are, and these should be the ones used to calculate cost-effectiveness data.	Thank you for your comments which the committee considered. The company confirmed that the Sunrise device reports both Apnoea-hypopnoea index (AHI) (using

Comment number	Name and organisation	Section number	Comment	NICE Response
				conventional cut-off values of 5, 15, 30 events per hour) and Obstructive respiratory disturbance index (ORDI) (using cut-off values of 7.63 and 12.65 events per hour established in the Pépin study) when used in clinical practice (as described in section 3.4 of the guidance).

THEME: Requests for clarification and minor corrections

Comment number	Name and organisation	Section number	Comment	NICE Response
27	Consultee 4	2.3 Care pathway and clinical need	I suggest changing this phrasing to “NICE currently recommends”, or similar, as it otherwise sounds like this sentence is in contrast to the novel device recommendations made here.	Thank you for your comments which the committee considered. The text has been amended to: “NICE currently recommends.”
28	Consultee 4	3.5 Using test accuracy data from	Should this say "the disposable NightOwl device that is available in the UK"? The way this is currently written suggests this is not currently available, which	Thank you for your comments which the

Comment number	Name and organisation	Section number	Comment	NICE Response
		previous and similar versions was acceptable for NightOwl and the WatchPAT devices	is at odds with the external report scoping, which says devices included "are commercially available in England".	committee considered. At the time of writing the guidance he NightOwl device is not currently available in the UK market because it does not yet have regulatory approval.
29	Consultee 4	3.8 Evidence in children and young people under 16 years is limited	How does body size alone affect accuracy of devices? It would be interesting to have further information on this, or a reference to where more information can be read in the external report.	Thank you for your comments which the committee considered. The sentence referred to cites clinical expert opinion that the main difference between adults and children is body size, not that this is the only factor. The external assessment group (EAG) briefly summarised diagnostic criteria for children and implications for device accuracy in EAG's report sections 1.3.7 and 5.11.1. See also

Comment number	Name and organisation	Section number	Comment	NICE Response
				the British Thoracic Society guideline (reference 9 in the EAG's report) for further detail.
30	Consultee 4	3.11 Diagnostic accuracy in people with brown or black skin	This conclusion does not seem to follow from the comments made above in 3.10, relating to an independent report about pulse oximeters (which "found extensive evidence of poorer performance of pulse oximeters for patients with darker skin tones"). Would it not be safer to advise further research before potentially recommending light-based technologies if there is concern?	Thank you for your comments which the committee considered. Clinical experts explained that when diagnosing obstructive sleep apnoea/hypopnoea syndrome (OSAHS) using blood oxygen levels, they focus on relative changes from the person's baseline, and any impact of skin tone on device performance is unlikely to affect accuracy. They also noted that diagnosis considers symptoms and the impact of sleepiness, not just device outputs. The committee concluded

Comment number	Name and organisation	Section number	Comment	NICE Response
				that light-based home-testing devices are suitable for people with brown or black skin but suggested more research would be beneficial (as described in section 3.13 of the guidance).
31	Consultee 5 Sunrise	1.6	Regarding the statement, “More research is needed on how accurately the home-testing devices diagnose and assess the severity of OSAHS in people with black or brown skin,” it should be mentioned that this concern specifically applies to light-based measurement (either integrated or through a third-party pulse oximeter) devices, as explained in sections 3.10 and 3.11 of the draft guidance. This could be clarified as follows: “More research is needed on how accurately the home-testing devices using light-based measurement (either integrated or through a third-party pulse oximeter) diagnose and assess the severity of OSAHS in people with black or brown skin.”	Thank you for your comments which the committee considered. The committee agreed that they were satisfied that the home-testing devices that use light-based technologies for assessment are appropriate to use for people with brown or black skin (see section 3.13 of the guidance). However, they noted there would be a large advantage to using the home-testing

Comment number	Name and organisation	Section number	Comment	NICE Response
				devices if they improved detection of OSAHS for people with brown and black skin compared with currently used tests. So, they agreed it would be beneficial to have data showing how accurate any of the devices are for people with brown or black skin to see if any could be recommended over others.
32	Consultee 5 Sunrise	3.10 3.11	The final part of the sentence, “which may provide useful additional outputs (see section 3.9)” in “AcuPebble SA100 and Brizzy also have the option to use a third-party pulse oximeter alongside the home-testing device, which may provide useful additional outputs (see section 3.9),” is not pertinent to this section and should be removed. Regarding the statement, “But, it agreed it would be beneficial to have data showing how accurate the devices are for people with brown or black skin, to understand if any of the devices should be recommended over others,” it should be mentioned that this concern specifically applies to light-based measurement devices (either integrated or through a third-party pulse oximeter), as explained in sections 3.10 and 3.11 of the draft guidance.	Thank you for your comments which the committee considered. Please see response to comment 31. The text has been amended to: “AcuPebble SA100 and Brizzy also have the option to use a third-party pulse oximeter alongside

Comment number	Name and organisation	Section number	Comment	NICE Response
			This could be clarified as follows: “But, it agreed it would be beneficial to have data showing how accurate the devices using light-based measurement (either integrated or through a third-party pulse oximeter) are for people with brown or black skin, to understand if any of the devices should be recommended over others.” This could also be clarified in the title corresponding to the two sections and the first sentence of section 3.10 as follows: “Diagnostic accuracy of devices using light-based measurement in people with brown or black skin” and “The committee considered how well the devices using light-based measurement (either integrated or through a third-party pulse oximeter) work for people with brown or black skin.”	the home-testing device.”
33	Consultee 5 Sunrise	EAG erratum Table 1	Part of the table appears to be missing. For the portions of the table that are presented, no changes from the previous version of this table have been identified.	Thank you for your comments. EAG erratum contains replacement pages and as such the portion of the table in question is only included as it features on page 308.
34	Consultee 5 Sunrise	EAG erratum (Critical appraisal of studies included in the systematic review of clinical effectiveness)	There appear to be some typographical errors in the comments of the following rows of the table, where the name ‘Sunrise’ should be replaced by ‘Brizzy’. Row 2 on page 309 (“Signalling question 2: If a threshold was used, was it pre-specified?”) “Post-hoc optimisation was done to select the diagnostic cut-offs for the Sunrise device” should be replaced by “Post-hoc optimisation was done to select the diagnostic cut-offs for the Brizzy device.”	Thank you for your comments. The EAG has changed all mentions of “Sunrise” to “Brizzy” on page 309 in EAG Erratum.

Comment number	Name and organisation	Section number	Comment	NICE Response
			Row 3 on page 309 (“Judgment: Could the conduct or interpretation of the index test have introduced bias?”) “High risk of bias due to post-hoc optimisation to select the diagnostic cut-offs for the Sunrise device” should be replaced by “High risk of bias due to post-hoc optimisation to select the diagnostic cut-offs for the Brizzy device.” Row 5 on page 309 (“Judgment: Is there concern that the index test, its conduct, or interpretation differ from the review question?”) “Post-hoc optimisation to select the diagnostic cut-offs for the Sunrise device” should be replaced by “Post-hoc optimisation to select the diagnostic cut-offs for the Brizzy device.”	
35	Consultee 6 Nomics	EAR Erratum 4	On page 309 of erratum 4 of the EAG’s report, the device studied is repeatedly referred to as “Sunrise”. This should be corrected to read “Brizzy”.	Thank you for your comments. The EAG has changed all mentions of “Sunrise” to “Brizzy” on page 309 in EAG Erratum.