# 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

## ERRATUM 4

## REPLACEMENT PAGES

## FOR DIAGNOSTIC ADVISORY COMMITTEE MEETING 2

| | |
|---|---|
| **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]<br>Jonathan Shepherd, Principal Research Fellow[1]<br>Emma Maund, Research Fellow[1]<br>Bogdan Grigore, Associate Research Fellow[2]<br>Lois Woods, Senior Research Assistant t[1]<br>Joanne Lord, Professorial Research Fellow in Health Economics[1]<br>David Alexander Scott, Principal Research Fellow[1]<br>Chris Hyde, Professor of Public Health and Clinical Epidemiology[2] |
| **Correspondence to** | Dr Jonathan Shepherd<br>Principal Research Fellow<br>Southampton Health Technology Assessments Centre (SHTAC)<br>University of Southampton<br>Alpha House<br>Enterprise Road, Southampton Science Park<br>Southampton, SO16 7NS, UK.<br>Email: |
| **Date completed** | 19th June 2024 |

**Table 1 Overview of included studies (people over 16 years)**

| Study ID | Country. No. centres | Study design | Intervention. Setting | Comparator/reference standard. Setting | Study population. No. patients | Outcome measures |
|---|---|---|---|---|---|---|
| **Novel devices compared to home RP** | | | | | | |
| Devani et al (2021)[19] | UK (London). Single centre | Prospective, single cohort | AcuPebble SA100 automated diagnosis. Home-based, (Unattended) (Patients were not trained on the use of the device under evaluation) | Cardiovascular respiratory polygraphy (CR-PG) (Embletta MPR Sleep System). with manual scoring Home-based (Unattended) | People with suspected OSA referred for examination. N=182 enrolled N=150 analysed | Diagnostic test accuracy; Accuracy in event classification, including central versus obstructive apnoeas Diagnostic test agreement; Diagnostic test failure rates; Patient acceptability and usability; Healthcare resources used and costs. |
| Alsaif et al (2023)[17] SOSAT trial | UK (Scottish highlands and inner-London). Two centres | Prospective, randomised, single cohort blinded pilot study. | Sunrise (MMs) with autoscoring). Home-based. (Unattended) | RP (ApneaLink Air device) with manual scoring. Home-based. (Unattended) | People with suspected OSA undergoing investigation. [redacted] | Time to treatment decision (days); Diagnostic test accuracy; Diagnostic test agreement (AHI; treatment decisions); Diagnostic test failure rates. [redacted] |
| Storey et al (2022)[28] | UK [redacted] Single Centre | Prospective randomised study | WatchPAT ONE Home-based (Unattended) | RP (NOX T3) Home-based (Unattended) | Patients referred by Sleep, ENT, Insomnia, Dental or Respiratory consultants N=600 enrolled (300 randomised to WatchPAT ONE and 300 to NOX T3 | Mean patient time (including travel time) to receive and return equipment; Number of appointments not attended by patients for intervention versus comparator; cost per appointment (equipment, room staff, postage); mean staff time taken per appointment (excluding analysis) |
| Mueller et al (2022)[29] | Germany. Single centre | Prospective randomised study | WatchPAT 300 with manual scoring (based on manual editing with software) Home-based. (Unattended) | RP (Miniscreen plus device) with manual scoring. Home-based. (Unattended) | People with suspected OSA needing home sleep testing. N=61 enrolled N=56 analysed | OSA diagnosis rates; OSA severity classification; Diagnostic test failure rates; Time spent in supine sleep position; Number of repeat sleep studies; Perceived quality of sleep and test related discomfort. |
| **Novel devices compared to PSG** | | | | | | |

| Study ID | Country. No. centres | Study design | Intervention. Setting | Comparator/reference standard. Setting | Study population. No. patients | Outcome measures |
|---|---|---|---|---|---|---|
| Sanchez Gomez et al (2024) [20] | Spain. Single centre | Prospective, single cohort | AcuPebble SA100 automated diagnosis. Sleep laboratory-based. | PSG (Philips Sleepware G3 version 2.8.78) with manual scoring. Sleep laboratory-based (Attended) | Patients referred for assessment of potential OSA N=80 enrolled N=63 analysed | Diagnostic test accuracy; Diagnostic test failure rates; OSA severity classification. |
| Martinot et al (2017)[21] | Belgium. Single centre | Prospective, single cohort | Brizzy (MMs) with manual scoring. Sleep laboratory-based. | Routine PSG (SomnoscreenPlus) with manual scoring. Sleep laboratory-based. (Unattended) | People with suspected OSA referred for laboratory sleep test (with moderate to high pre-test probability) N=100 enrolled N=92 analysed (inc. 13 healthy volunteers) | Diagnostic test accuracy; [b] Diagnostic test agreement; Diagnostic test failure rates OSA severity classification. |
| Massie et al (2018)[22] | Belgium. Single centre | Prospective, single cohort | NightOwl (reusable version) with autoscoring Sleep laboratory-based. | PSG (device not stated) with a combination of manual and automated scoring Sleep laboratory-based. | Patients who underwent a diagnostic in-hospital PSG in the sleep laboratory N=101 enrolled N= 101 analysed | Diagnostic test accuracy; Diagnostic test agreement; OSA severity classification. |
| Massie et al (2022)[23] | USA and Belgium. Four centres (3 in USA, 1 in Belgium) | Prospective, single cohort. | NightOwl with autoscoring. Sleep laboratory based. (Attended) | Routine PSG (Alice 6 PSG (European centres) or Cadwell Easy PSG (USA centres)). Sleep-laboratory based (Attended)<br><br>Each PSG manually scored independently by local centre & by a separate expert centre (reference standard) | People with suspected OSA scheduled for in-lab PSG. N=261 enrolled N=261 analysed | Diagnostic test accuracy;[a] Diagnostic test agreement;[a] Minimum required REM sleep time; OSA diagnosis rates. |

### 4.4 Results of critical appraisal of study methodology

In this section we summarise the results of our critical appraisal of all the studies included in this systematic review (i.e. for the 2-16 years age group and for the 16 years and older group). Further detail on our critical appraisal judgements are presented in Appendix 5.

We applied the QUADAS-2 tool[13] to each of the included studies to assess the risk of bias and the applicability of the study to the decision problem. The QUADAS-2 tool appraises the likelihood of bias arising from: the selection of participants; the conduct and interpretation of the index test and the reference standard; the flow of participants through a study and the timing of the index test and reference standard. It also assesses the applicability of the participants selected and the index test and reference standard to the review's research question. **Table 2** shows the results of our critical appraisal and **Figure 1** presents the results graphically, for all studies included in this systematic review (i.e. both the 'people over 16 years' and children and young people aged 2 to 16 years' sub-groups).

The majority of studies were judged to be at low risk of bias overall, but in five studies a high risk of bias judgement was made in one of the four bias domains. Patient selection was judged to be at high risk of bias in the study by Pillar et al., 2020. The study selectively recruited heart-failure patients but it is not clear if this resulted in inappropriate exclusions. The intentional bias towards selecting patients with congestive heart failure may, therefore, have introduced other unintentional biases.[31] A high risk of bias was judged in the conduct or interpretation of the index test in four studies (Kelly et al., 2022; Martinot et al., 2017; Martinot et al., 2022 and Pepin et al., 2020) )[21 26 27 35] all of which used post-hoc analyses to optimise diagnostic cut-off points, potentially over-estimating novel device diagnostic accuracy.

Regarding applicability to the decision problem, most studies were judged as low concern for the patient selection and the reference standard domains. However in many studies it was unclear whether the conduct, or interpretation of the index test was relevant to the decision problem. This judgement was made for all studies where the novel testing device was used in a sleep laboratory (concomitant to PSG testing), rather than its intended setting (i.e. the patient's home). Two studies were also rated unclear for this domain although they were conducted in a home setting. Alsaif et al (2023) did not report on the thresholds used in their study and Storey et al., 2022 did not report details of the conduct and interpretation of the index test. For four studies, the judgements were of high concern – in Kelly et al., 2022, Martinot et al., 2022 and Pepin et al., 2020 all used post-hoc analyses to optimise diagnostic cut-off points, while in Martinot et al., 2015 diagnostic accuracy results for against PSG or

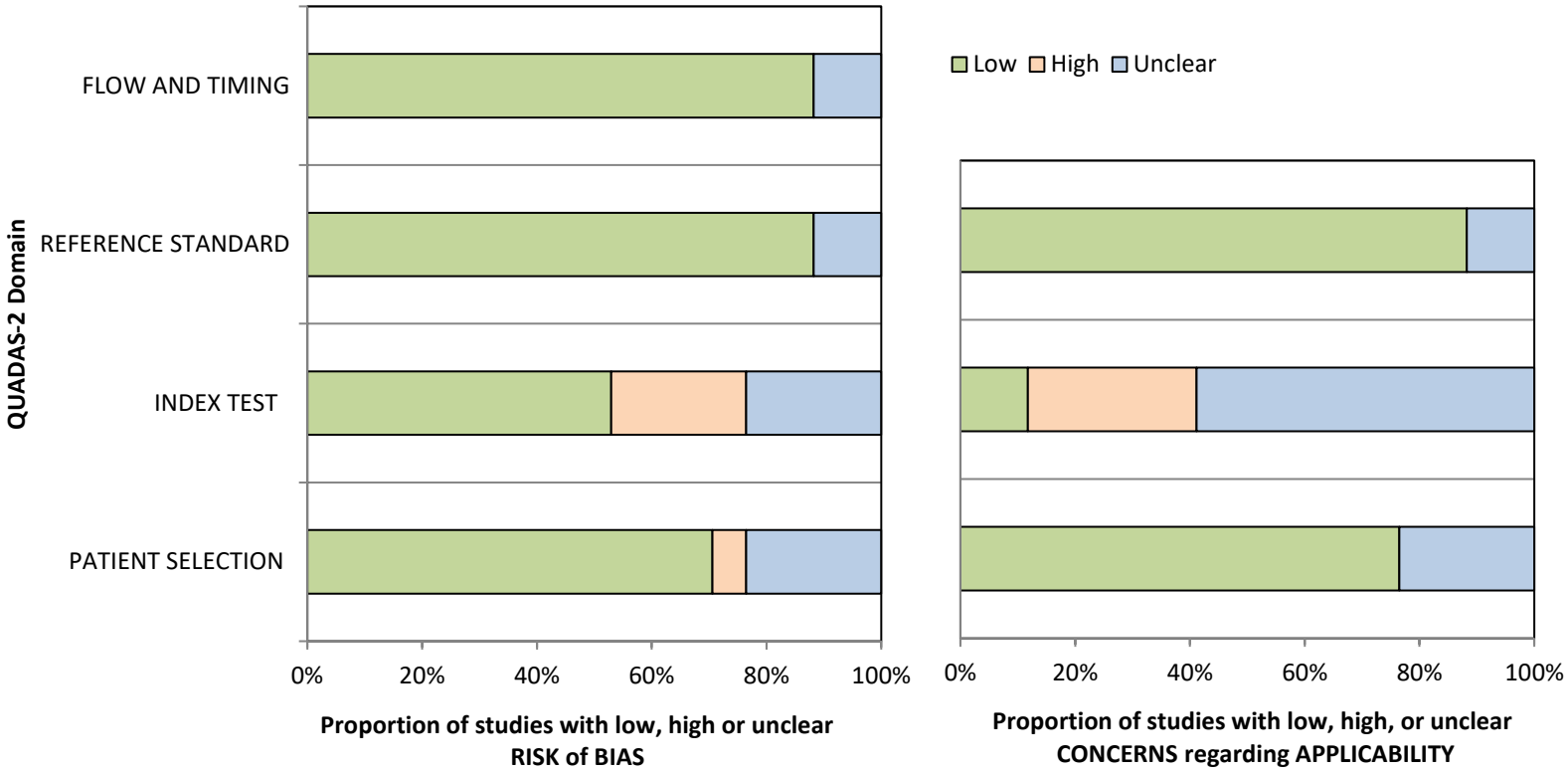
any other reference standard were not reported and the study was conducted in a sleep laboratory rather than the home setting.

**Table 2 Overview of QUADAS-2 assessments for all studies**

| Study | RISK OF BIAS | | | | APPLICABILITY CONCERNS | | |
|---|---|---|---|---|---|---|---|
| | PATIENT SELECTION | INDEX TEST | REFERENCE STANDARD | FLOW AND TIMING | PATIENT SELECTION | INDEX TEST | REFERENCE STANDARD |
| **AcuPebble SA100** | | | | | | | |
| Devani 2021 | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ | ☺ |
| Sanchez Gomez 2024 | ☺ | ☺ | ☺ | ☺ | ☺ | ? | ☺ |
| NCT04031950 (child) | ☺ | ☺ | ☺ | ☺ | ☺ | ? | ☺ |
| **Brizzy** | | | | | | | |
| Martinot 2017 | ☺ | ☹ | ☺ | ☺ | ☺ | ☹ | ☺ |
| Martinot 2015 (child) | ☺ | ? | ☺ | ☺ | ☺ | ☹ | ? |
| **NightOwl** | | | | | | | |
| Massie 2018 | ? | ☺ | ☺ | ☺ | ? | ? | ☺ |
| Massie 2022 | ☺ | ☺ | ☺ | ☺ | ☺ | ? | ☺ |
| Van Pee 2022 | ☺ | ☺ | ☺ | ☺ | ☺ | ? | ☺ |
| Lyne 2023 | ? | ☺ | ☺ | ☺ | ☺ | ? | ☺ |
| **Sunrise** | | | | | | | |
| Pepin 2020 | ☺ | ☹ | ☺ | ☺ | ☺ | ☹ | ☺ |
| Kelly 2022 | ☺ | ☹ | ☺ | ☺ | ☺ | ☹ | ☺ |
| Alsaif 2023 | ? | ? | ☺ | ? | ? | ? | ☺ |
| Martinot 2022 (child) | ☺ | ☹ | ☺ | ☺ | ☺ | ☹ | ☺ |
| **WatchPAT 300/ONE** | | | | | | | |
| Mueller 2022 | ☺ | ? | ? | ☺ | ☺ | ☺ | ☺ |
| Storey 2022 | ☺ | ? | ? | ? | ? | ? | ? |
| **Supporting evidence (WatchPAT 200U)** | | | | | | | |
| Pillar 2020 | ☹ | ☺ | ☺ | ☺ | ? | ? | ☺ |
| Tauman 2020 | ? | ☺ | ☺ | ☺ | ☺ | ? | ☺ |

☺Low Risk ☹High Risk ? Unclear Risk

**Figure 1 Proportion of studies with low, unclear or high risk of bias and proportion of studies with low, unclear or high concerns regarding applicability (all included studies)**



has a time horizon of 12 months to capture any delays to the start of treatment (should treatment be offered). A lifetime Markov model is used to estimate the longer-term impacts associated with the performance of the devices. It models the risks of cardiovascular events and RTAs for people with OSAHS and includes death from other causes for the total cohort.

In the base case analysis, all six novel devices are estimated to be less costly than respiratory polygraphy, but they are also associated with a small estimated reduction in QALYs. For AcuPebble and Sunrise compared to respiratory polygraphy, the reduction in QALYs is considered cost-effective compared to the reduction in costs (i.e. INMB > £0 at the £20,000 and £30,000 per QALY thresholds). Compared to oximetry, all novel devices have a positive INMB at both £20,000 and £30,000 per QALY thresholds.

However, it is important to recognise the high level of uncertainty over the cost-effectiveness results. This is apparent from the probabilistic and scenario analyses. In the probabilistic base case analysis, there are wide and overlapping confidence ranges for the incremental costs and QALYs for each novel device compared with oximetry, which is more pronounced for the comparisons with respiratory polygraphy. For example, the incremental costs for WatchPAT 300 compared with respiratory polygraphy range from -£298 to £235 and the incremental QALYs range from -0.040 to 0.033. This uncertainty is reflected in wide confidence ranges around the INMBs for the novel devices, for example, comparing Sunrise to respiratory polygraphy, the INMB at £20,000 per QALY gained ranges from -£238 to £572.

These results are sensitive to a number of assumptions, as the scenario analyses indicate, including the data source used to estimate the diagnostic performance and failure rates associated with respiratory polygraphy, the proportion of people diagnosed with mild OSAHS who are treated with CPAP, alternative parameterisation of the decision tree (using 4x4 contingency table data), and the impacts associated with false positives. See section **Error! Reference source not found.** below for a discussion of the key sources of uncertainty and their effect on cost-effectiveness results. Moreover the data used in the base case analysis to inform the accuracy estimates for novel devices are all derived from a clinical setting, with three based on post-hoc optimisation of thresholds, which is likely to overestimate the accuracy of the devices.

## 6.2 Strengths and limitations of the assessment

### 6.2.1 Strengths

The cost-effectiveness model is adapted from one that was used to inform recent NICE guidance on the diagnosis and management of obstructive sleep apnoea/ hypopnoea syndrome and obesity hypoventilation syndrome in over 16s (NG202).[52] The NG202 economic model was developed in consultation with the Guideline Committee, and was itself adapted from the model developed to inform the NICE appraisal on CPAP treatment for people with OSAHS (TA139).[54] We believe that the attention that versions of this model have received, by experts in the field during development and in consultation processes is a strength. We updated parameter values from those used in the TA139 and NG202 models where we could identify more recent, relevant data of better quality. The choice of data for the model parameters was informed by our systematic review of clinical and diagnostic assessments of the novel devices, and economic evidence on cost-effectiveness, resource use and costs and health-related quality of life. We also conducted targeted reviews for other key model parameters. Throughout our adaptation of the model, we consulted with experts, especially on the validity of base case and scenario analysis assumptions

### 6.2.2 Limitations

The cost-effectiveness analysis is limited by the availability and quality of data for many of the model components. This included:

- Limited diagnostic accuracy data for novel devices evaluated in the home, rather than the clinic – data from a home setting were only available for AcuPebble and Sunrise, and as the AcuPebble study (Devani et al) did not use PSG as the reference standard, and the Sunrise study (Kelly 2022) was very small, neither were used in our base case.
- Lack of evidence on current versions of devices, e.g. WatchPAT 300 and ONE, although the manufacturer has reported that these versions of the WatchPAT device produce identical signals and use the same algorithm as the previous 200U version.
- Inconsistency in reference standards used across devices – in Devani et al (2021) the reference standard was home RP, in Kelly et al (2022) it was home PSG.
- Post-hoc optimisation of diagnostic thresholds within accuracy studies, such as for Brizzy in Martinot et al (2017), and for Sunrise in Pepin et al (2020) and Kelly et al (2022).
- Lack of data on health-related quality of life (utility) associated with different severities of OSAHS.

| | | |
|---|---|---|
| **Signalling question 3:**<br>Did patients receive the same reference standard? | Yes | *"Routine laboratory-based PSG was recorded with a Dream Medatec device, Brussels, Belgium*" (p.568) |
| **Signalling question 4:**<br>Were all patients included in the analysis? | Yes | Paper does not suggest that any data were missing for any reason. |
| **Judgment:**<br>Could the patient flow have introduced bias? | **RISK:** LOW | No comment. |
| [a] with caveat that the device may be contraindicated in certain patient populations<br>[b] with caveat that if the index test was automatically scored by the software only, it could be considered independent of the results of the reference standard<br>[c] with caveat that 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. | | |

**Martinot et al., 2017[21]**

| **Device:** Brizzy<br>**Secondary papers:** none | | |
|---|---|---|
| **DOMAIN 1: PATIENT SELECTION** | **Assessment** | **Comments** |
| **A. Risk of Bias** | | |
| **Signalling question 1:**<br>Was a consecutive or random sample of patients enrolled? | Yes | *"The patients eligible to participate were consecutive subjects"* (p.568). Thirteen participants with "no specific sleep complaints were recruited by word of mouth" (p. 568), but the reviewers do not expect that this would have introduced bias. |
| **Signalling question 2:**<br>Was a case-control design avoided? | Yes | The computed distance from the MM probes was transmitted to the PSG when PSG was conducted so both measures done on the same patient. |
| **Signalling question 3:**<br>Did the study avoid inappropriate exclusions?[a] | Yes | Consecutive patients were consenting adults "18 years and older with symptoms suggestive of sleep-disordered breathing (SDB) undergoing a single PSG." |
| **Judgment:**<br>Could the selection of patients have introduced bias? | **RISK:** LOW | No comment |
| **B. Concerns regarding applicability** | | |
| **Judgment:**<br>Is there concern that the included patients do not match the review question? | **CONCERN:** LOW | No comment |
| **DOMAIN 2: INDEX TEST(S)** | **Assessment** | **Comments** |
| **A. Risk of Bias** | | |

| | | |
|---|---|---|
| **Signalling question 1:**<br>Were the index test results interpreted without knowledge of the results of the reference standard?[b] | Yes | *"Scoring for MM was performed by two blinded independent readers who had been trained to read MM tracings, while a different experienced reader analysed the standard PSG"* (p.568). |
| **Signalling question 2:**<br>If a threshold was used, was it pre-specified?[c] | No | Post-hoc optimisation was done to select the diagnostic cut-offs for the Sunrise device.<br><br>*"The outcome variable related to the diagnostic of the disease was based on a sensitivity/specificity analysis of MM device with the two different polysomnographic pre-specified cut-off values of RDI recommended in ICSD-3 (PSG-RDI ≥ 5 and ≥15/h TST). OSAS severity was evaluated from AHI, with <5, 5–15, 15–30 and >30/h TST representing the four severity categories*." (p.569)<br><br>*"ROC curve analyses were performed to evaluate the ability of MM-RDI to detect PSG-defined OSAS at three pre-specified selected cut-off points (Fig. 3). The characteristics as well as the best cut-off point of these three classifications are given in Table 2"* |
| **Judgment:**<br>Could the conduct or interpretation of the index test have introduced bias? | **RISK:** HIGH | High risk of bias due to post-hoc optimisation to select the diagnostic cut-offs for the Sunrise device. |
| **B. Concerns regarding applicability** | | |
| **Judgment:**<br>Is there concern that the index test, its conduct, or interpretation differ from the review question? | **CONCERN:** HIGH | Index test carried out in a lab not a home setting.<br>Post-hoc optimisation to select the diagnostic cut-offs for the Sunrise device. |
| **DOMAIN 3: REFERENCE STANDARD** | **Assessment** | **Comments** |
| **A. Risk of Bias** | | |
| **Signalling question 1:**<br>Is the reference standard likely to correctly classify the target condition? | Yes | Study used "standardized in-laboratory PSG for the diagnosis of OSAS (ICSD-3, International Classification of Sleep Disorders, Third Edition)" and "a different experienced reader analysed the standard PSG, after de-identification of records." (p.568) |
| **Signalling question 2:**<br>Were the reference standard results interpreted without knowledge of the results of the index test? | Unclear | Does not explicitly state that the PSG reader interpreted the PSG without knowledge of the index test results, but the PSG was analysed by a different reader. States PSG analyses were conducted after 'de-identification of records' but unclear what aspect of identification was removed (i.e. MM results or patient name/hospital number). |