Engineering Science

DPhil Thesis

Enhancing the clinical utility of minimally obtrusive technologies for the diagnosis of sleep disorders



Frederik Massie

Kellogg College

University of Oxford

Supervised by Professor Jeroen Bergmann

Michaelmas Term 2025

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Keywords: Home Sleep Apnea Testing, Peripheral Arterial Tonometry, Photoplethysmography, Clinical Validation, Central Sleep Apnea Detection, REM OSA Phenotyping, Minimally Obtrusive Sleep Staging, Context-Aware Autoscoring

Copyright 2025 by Frederik Massie

Summary

Sleep apnea is a prevalent and underdiagnosed sleep disorder characterised by breathing interruptions during sleep. It adversely impacts quality of life and causes or aggravates severe health conditions, including cardiovascular disease and depression. The gold standard diagnostic method, in-hospital polysomnography (PSG), is costly, inconvenient, and commonly subject to long waiting times. As an alternative, home sleep apnea testing (HSAT) allows simplified testing in the home environment. HSAT based on peripheral arterial tonometry (PAT HSAT) is one of the most widely used methods. It uses an optical sensor placed on the fingertip, from which fluctuations in peripheral arterial tone, heart rate, and blood oxygen saturation are inferred. Characteristic changes in these physiological signals are used to automatically detect breathing interruptions.

Despite its widespread use, PAT HSAT has considerable limitations affecting its clinical utility. This thesis addresses four key limitations of PAT HSAT.

Limitation 1: The lack of standardisation in HSAT clinical validation studies has led to selective reporting and difficulties in reproducing or comparing results. We first demonstrated, through a meta-analysis, that the historic overreliance on correlation coefficients as a substitute for diagnostic accuracy can lead to misleading conclusions about the clinical utility of HSAT devices. We subsequently proposed a more robust set of validation endpoints and introduced methods that explicitly account for the inter-scorer variability of the reference PSG.

Limitation 2*:* PAT HSAT cannot differentiate between central sleep apnea and obstructive sleep apnea without adding extra sensors to the body. Differentiating these types of sleep apnea is clinically important because therapeutic approaches vary greatly. To this end, we developed and clinically evaluated a novel method that differentiates between these two types using only PAT HSAT data.

Limitation 3: PAT HSAT is consistently less accurate than PSG for sleep staging. This reduces the accuracy of sleep stage-related diagnostic outputs such as the identification of rapid eye movement-related obstructive sleep apnea (REM OSA), a phenotype characterised by more severe apnea during REM sleep. We first compared the performance of PAT HSAT-based REM OSA detection against PSG and proposed a multi-night testing protocol to improve performance. Secondly, we complemented PAT HSAT with a miniaturised biopotential device to substantially enhance sleep staging performance.

Limitation 4: Current PAT HSAT autoscoring algorithms analyse each apnea event in isolation, limiting diagnostic accuracy. We developed a context-aware autoscoring algorithm that considers repetitive and characteristic patterns spanning multiple apnea episodes. This algorithm significantly improved the accuracy of PAT HSAT apnea event autoscoring.

By systematically addressing these limitations, this thesis contributes to improved accuracy, expanded capabilities, and greater overall clinical utility of PAT HSAT.

Table of contents

Summary **3**

Table of contents **5**

Chapter 1: Clinical context and research aims **9**

Clinical context 9

Overview of research aims 16

Conclusion of research aims and chapter structure overview 22

Chapter 2: Improving the performance evaluation of home sleep apnea tests **25**

Research Aim 1.1: Addressing the prevalent misuse of correlation coefficients to support diagnostic performance 25

Reiteration of research aim and introductory context 25

Abstract 26

Introduction 27

Methods 30

Results 32

Discussion 36

Research Aim 1.2: Proposing robust endpoint parameters with consideration for the inter-rater-disagreement of PSG 38

Reiteration of research aim and introductory context 38

Abstract 39

Introduction 40

Methods 42

Results 51

Discussion....63

Conclusion 64

Chapter 3: Enabling PAT HSAT to differentiate between CSA and OSA....67

Research Aim 2.1: Investigating a PPG-based approach for differentiating obstructive and central sleep apnea....67

Reiteration of research aim and introductory context 67

Abstract....68

Introduction 69

Methods 73

Results 84

Discussion....90

Conclusion 94

Chapter 4: Enhancing the ability of PAT HSAT to measure sleep accurately 96

Research Aim 3.1: Evaluating and improving REM OSA phenotyping....96

Reiteration of research aim and introductory context 96

Abstract....97

Introduction 98

Methods 100

Results 106

Discussion....108

Conclusion....114

Research Aim 3.2: Enhancing PAT HSAT sleep staging by incorporating a single-lead biopotential measurement....116

Reiteration of research aim and introductory context....116

Abstract....117

Introduction....118

Methods....121

Results....129

Discussion....139

Conclusion....141

Chapter 5: Rendering PAT HSAT autoscoring context-aware....144

Research Aim 4.1: Investigating context-aware PAT HSAT autoscoring....144

Reiteration of research aim and introductory context....144

Abstract....144

Introduction....146

Methods....148

Results....161

Discussion....168

Conclusion....173

Chapter 6: Discussion and conclusion....175

Limitation 1: The absence of clinical validation reporting standards for HSAT leads to selective reporting and difficulties in reproducing or comparing results....175

Limitation 2: PAT HSAT does not differentiate between CSA and OSA....176

Limitation 3: PAT HSAT does not measure sleep accurately....178

Limitation 4: The limited analysis scope of PAT HSAT autoscoring algorithms restricts performance....179

Future research....181

Conclusion....184

Note on affiliations....185

List of abbreviations....186

References....190

Acknowledgements....211

Chapter 1: Clinical context and research aims

Clinical context

A brief introduction to sleep apnea, its pathophysiology, and epidemiological consequences

Sleep apnea is a disorder characterised by repeated episodes of reduced or ceased breathing during sleep. Complete cessations and reductions of breathing lasting more than 10 seconds are called apnea and hypopnea, respectively, and are both referred to as respiratory events[1].

The primary index determining sleep apnea severity is called the apnea-hypopnea index (AHI), which is defined as the total number of respiratory events divided by the total sleep time (TST). A related measure, the respiratory disturbance index (RDI), additionally includes respiratory effort-related arousals (RERAs) in the event count, which are brief arousals caused by increased breathing effort that do not meet criteria for apnea or hypopnea. The magnitude of the AHI defines the sleep apnea severity category. An AHI below 5 is considered nonpathological. An AHI between 5 and 15 is regarded as mild sleep apnea. An AHI between 15 and 30 is considered moderate sleep apnea, and an AHI above 30 is considered severe sleep apnea[1].

Two main types of sleep apnea exist: obstructive sleep apnea (OSA) and central sleep apnea (CSA). In OSA, regular respiratory effort (drive to breathe) remains intact, and the breathing interruptions are due to a physical obstruction[1]. These obstructions result from a partial or complete collapse of the upper airway caused by the relaxation of the muscles controlling the soft palate and tongue[2]. CSA is characterised by a lack of respiratory effort. In this case, the brain fails to transmit signals to the respiratory muscles. When both types of events occur consecutively, they are referred to as mixed apnea[3].

Almost 1 billion people globally are affected by mild-to-severe sleep apnea[4], with 425 million adults aged between 30 and 69 having moderate-to-severe sleep apnea[5]. In the United States, it has been reported that 25% to 30% of men and 9% to 17% of women meet the criteria for sleep apnea[6]. As sleep apnea is a chronic condition, leaving it untreated can substantially impair quality of life and worsen long-term health outcomes. The condition is linked to a wide range of comorbidities, including depression, cardiovascular disease, obesity, diabetes, and (drug-resistant) hypertension[7].

The degree of sleep apnea underdiagnosis is estimated to be above 80% worldwide[8]. Nevertheless, treatment is cost-effective[3,9,10]. As such, it is crucial to detect the disease early to limit its negative impact on both the individual and society. Standard therapeutic approaches for sleep apnea include lifestyle modifications, positional therapy, oral appliances, continuous positive airway pressure (CPAP), and surgical interventions. CPAP therapy is generally considered the default therapy for moderate to severe OSA. It delivers continuous air pressure through a nasal or facial mask, keeping the upper airway open during sleep and effectively preventing its collapse[11]. Despite CPAP's proven efficacy in reducing AHI and improving symptoms, patient adherence remains a challenge, influenced by mask discomfort, pressure intolerance, and inconvenience[12]. Alternative or complementary approaches, such as oral appliances, which reposition the mandible forward, or upper airway surgeries, might be employed when CPAP adherence is problematic or insufficient[13].

Overview of diagnostic equipment deployed in clinical practice

The American Academy of Sleep Medicine (AASM) is the world's leading institution for issuing clinical guidelines for evaluating, managing, and treating sleep apnea. The AASM considers three technologies technically adequate for diagnosing sleep apnea: polysomnography (PSG), home sleep testing based on polygraphy (PG), and home sleep testing based on peripheral arterial tonometry (PAT HSAT, or P-HSAT)[1].

Polysomnography (PSG)

PSG is the gold standard for analysing sleep physiology and pathophysiology. It comprises an extensive set-up in a clinical environment where the patient is monitored for one night[14]. PSG includes a wide range of sensors and measurements, including electrocardiography (ECG), airflow, respiratory effort, arterial blood oxygen saturation (SpO_2), electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and body position[1]. Typically, a trained sleep technician manually analyses the presence of sleep apnea, which can take over two hours per patient[14]. The method can be used to identify all forms of sleep apnea[1].

Full-night supervised laboratory PSG has major drawbacks. First, it is costly and requires a complicated setup in a hospital environment, which is inconvenient for patients. Second, visual analysis of the recordings by a sleep expert is time-consuming and can take more than two hours per patient[15,16]. In addition, hospitals have limited capacity, and hence waiting times for PSG can exceed one year[17]. The limited time window to track a person's sleep, usually a single night, also fails to portray important trends in sleeping behaviour and to correlate results to external factors like environmental or behavioural parameters and medication[18].

Polygraphy (PG)

PG is a home sleep apnea testing setup that typically consists of all sensor modalities of PSG except for the biopotential sensors (EEG, EOG, and EMG) which are used to determine sleep stages[1]. Like PSG, the analysis is generally performed by a trained technician. Since no biopotential electrodes are present, this procedure is less expensive than PSG and is more easily conducted in the home environment.

The absence of biopotential electrodes implies that sleep stages cannot be scored. Instead, when computing parameters that require the total sleep time in the denominator, like the AHI, the total

recording duration is usually taken instead of the total sleep time. This introduces substantial inaccuracies in these parameters[19].

Home sleep testing based on peripheral arterial tonometry (PAT HSAT)

The final category of technically adequate diagnostics is the PAT HSAT (also abbreviated as P-HSAT). This technology uses a fingertip-mounted photoplethysmography (PPG) probe[1] (see Figure 1) and automated respiratory event detection algorithms.

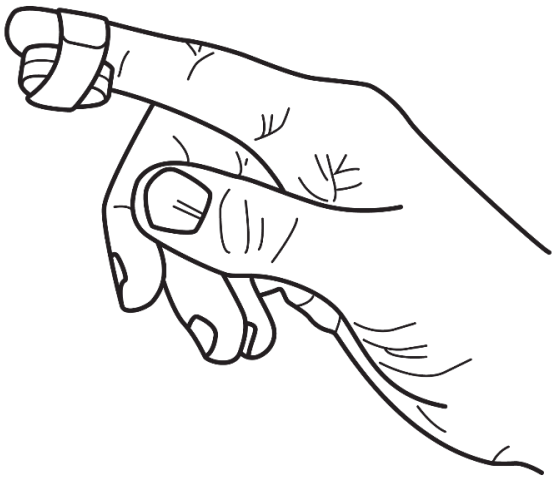

Figure 1: An illustration of a PAT HSAT.

PPG uses simple optical technology to detect blood volume changes in the tissue's microvascular bed (see Figure 2). It consists of a light-emitting diode (LED) that illuminates the local tissue and a photodiode (PD) that generates an electrical signal called the PPG signal, proportionate to its incident light intensity[20]. As such, the PD signal amplitude fluctuates when the composition of absorbing or scattering particles in the tissue illuminated by the LED fluctuates. When the illuminated tissue contains more (arterial) blood, an increased portion of the light emitted by the LED will be scattered or absorbed before it reaches the PD, resulting in a lower PPG signal amplitude[20].

PPG measurements are used to derive SpO_2, pulse rate (PR), and changes in peripheral arterial tone (PAT). Respiratory events (apnea and hypopnea) can be inferred from changes in these PPG-derived measurements[21].

Peripheral arterial tone refers to the tone of the peripheral arterial smooth muscle tissue (located in the fingertips). When the muscle tone of peripheral arteries increases, the arteries' diameter decreases, resulting in reduced perfusion and, thus, a decrease in blood volume displaced between diastole and systole in the peripheral tissue. The latter manifests as a drop in the PPG signal's amplitude swing between systole and diastole[21].

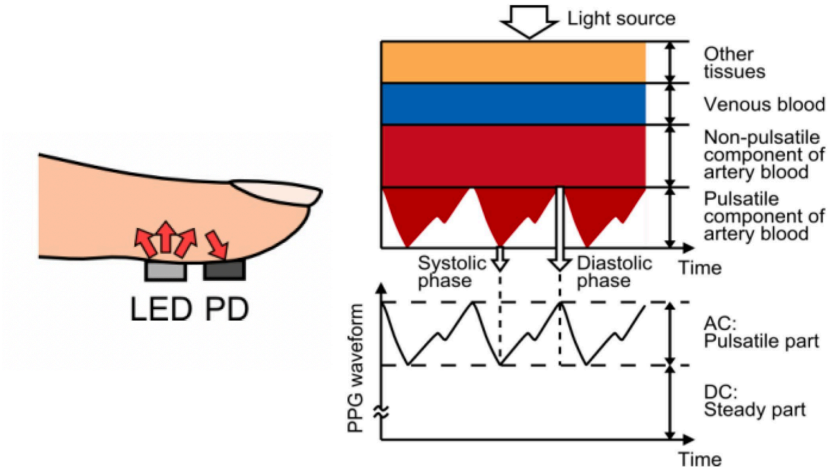

Figure 2: An overview of the working principle of photoplethysmography (PPG). An increase in finger tissue perfusion by arterial blood during systole results in increased light absorption and scattering by this arterial blood, resulting in less light reaching the photodiode (PD) of the sensor. The Light intensity picked up by the PD is referred to as the PPG signal. Adapted from Dzedzickis et al., 2020[22].

The PPG-derived signal trending such pulsatile blood volume reductions is called the PAT signal. As such, an *increase* in peripheral arterial (smooth muscle) tone is reflected in the PAT signal as an amplitude *decrease*[21]. To avoid ambiguity, note that the acronym 'PAT' is used in two related ways. When used on its own, PAT refers to the peripheral arterial tone. When used as part of PAT HSAT, PAT denotes peripheral arterial tonometry, the method/technology underlying HSAT systems that analyse the peripheral arterial tone signal.

During a respiratory event, as airflow is reduced, the oxygen supply to the lungs decreases from baseline, resulting in distinct SpO_2 drops called desaturations. Near the end of a respiratory event, the sympathetic nervous system activity surges to arouse the patient so they can resume normal breathing. This sympathetic activation results in a surge in pulse rate and a release of norepinephrine in the bloodstream, which binds to alpha-adrenergic receptors innervating the arterial smooth muscle tissue in the finger. Such binding of norepinephrine causes sudden increases in peripheral arterial tone (vasoconstrictions), picked up by a drop in the PAT signal[20,21,23]. PAT HSAT analyses this concurrence of desaturations, PAT signal drops and pulse rate surges[21]. See Figure 3 for an illustration of this principle.

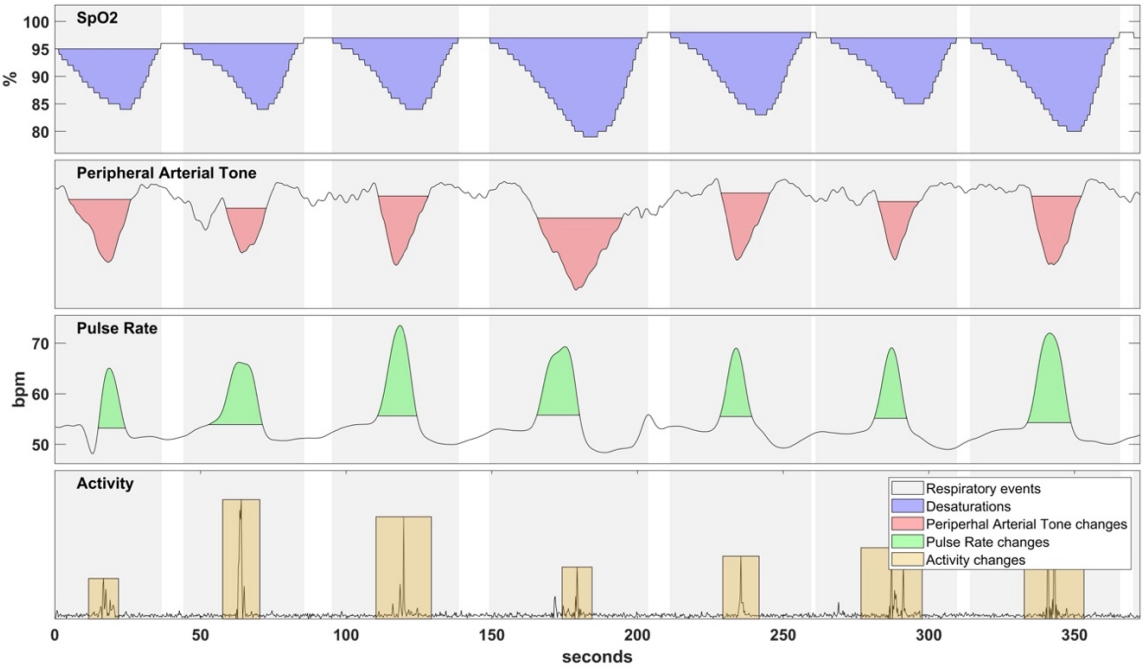

Figure 3: Display of the Repetitive SpO_2 desaturations, PAT amplitude decreases (reflecting vasoconstrictions), pulse rate surges, and activity surges (due to sudden movement after autonomic arousal) that characterise a sequence of respiratory events. Each signal-type event is indicated in a different colour. The grey bars mark the area around which a respiratory event is occurring.

The rapidly expanding clinical deployment of PAT HSAT

The COVID-19 pandemic reshaped the clinical practice of sleep apnea diagnosis. During outbreaks, sleep labs were forced to restrict in-lab PSG, which accelerated the shift toward HSAT[24]. In particular, PAT HSAT has gained substantial traction because it is cost-effective and simple for patients to self-apply as it relies on compact and low-complexity hardware[25]. It is now one of the most widely used diagnostic platforms for sleep apnea[26,27] and is considered to hold immense promise for further transforming the field[28]. Nevertheless, it still has several important limitations that constrain its clinical and scientific utility.

The research described in this thesis focuses on exploring novel approaches to reduce or remove these limitations. The following section elaborates on these limitations as well as the concrete research aims that were formulated to address them.

Overview of research aims

As introduced earlier, PAT HSAT is the least obtrusive HSAT technology considered technically adequate by the AASM and has become a highly active area of scientific and clinical research[1,28].

However, this technology and the scientific reporting on its performance suffer from considerable limitations. As such, this thesis reports on how addressing these limitations could improve the clinical utility of PAT HSAT.

The following sections introduce four substantial limitations of PAT HSAT. These limitations were selected based on the criterion that their resolution would directly enhance the technology's clinical utility for the diagnosis of sleep-related breathing disorders.

Limitation 1: The absence of clinical validation reporting standards for HSAT leads to selective reporting and difficulties in reproducing or comparing results

Introductory context

As HSAT in general, and PAT HSAT in particular, are increasingly deployed in clinical practice and more devices reach the stage of clinical validation, there is a growing need to standardise reporting for performance validation studies[24,26,27]. Such efforts will enable researchers, healthcare providers, regulators, device manufacturers, and other stakeholders to efficiently and reproducibly design and assess HSAT performance validation studies. Indeed, Collop et al. already highlighted in 2011 that criteria need to be established to evaluate the appropriateness of HSAT devices[29]. Nevertheless, to date, no consensus or attempts for standardisation exist on identifying endpoint parameters that promote the robust and transparent evaluation of the performance of HSAT devices. As such, multiple validation studies have supported the performance of HSAT based on selective or

inadequately conditioned endpoint parameters, of which the Pearson correlation appears to be most frequently misused[30]. Therefore, before addressing PAT HSAT limitations of a more technological nature, we endeavoured to enhance the quality of its validation reporting. This permitted a more robust investigation of the clinical merit of the PAT HSAT innovations described in this thesis. It prompted the formulation of two research aims related to this limitation.

Research Aim 1.1: Addressing the prevalent misuse of correlation coefficients to support diagnostic performance

The clinical performance of HSAT can be described by its (diagnostic) accuracy (and related metrics), defined as the percentage agreement with PSG on the sleep apnea severity category (normal, mild, moderate, and severe). Rather than reporting on accuracy and related metrics, there has been a firm reliance in the literature on reporting correlation coefficients between the AHI of HSAT and PSG to support claims of diagnostic performance. This is concerning, as it has been previously described that correlation coefficients are inadequate to evaluate equivalence between 2 parameters[31].

As such, we systematically investigated the magnitude of the discrepancies between correlation coefficients and diagnostic accuracy reported in or retrievable from HSAT validation studies to compile a conclusion that may inform future efforts for robust endpoint parameter selection. This effort could improve researchers' ability to evaluate the clinical utility of a particular HSAT device and to compare results across different studies.

Research Aim 1.2: Proposing robust endpoint parameters with consideration for the inter-rater-disagreement of PSG

As mentioned, the AHI obtained from PSG is traditionally used as the ground-truth reference to which the PAT HSAT AHI is compared. The PSG-derived AHI is inherently an imperfect ground truth since significant inter-rater disagreement exists[32]. Indeed, the apnea events are visually scored from PSG data by a sleep technician, and the rules of scoring such events leave room for the technician's interpretation. This implies that an AHI derived by PSG that is close to any of the sleep apnea severity category boundaries (5, 15, and 30) should be treated with caution. For example, an AHI of 15.1 would qualify as moderate sleep apnea, whereas an AHI of 14.9 would qualify as mild sleep apnea, which could have different therapeutic implications[33]. However, this difference in AHI is much smaller than the typical inter-rater variability of the AHI[32]. Considering this clinically relevant imperfection of PSG as a gold standard reference technology, we proposed novel endpoint reporting methodologies that adequately acknowledge and handle this imperfection, based on data from a clinical study.

Limitation 2: PAT HSAT does not differentiate between CSA and OSA

Introductory context

Both obstructive and central events are characterised by airflow reduction. However, central events are accompanied by a cessation of respiratory effort as traditionally inferred from the abdominal and thoracic respiratory effort belts of PSG, which monitor respiratory-related fluctuations in the abdominal and thoracic circumference[34]. Figure 4 compares the respiratory effort during central and obstructive apnea.

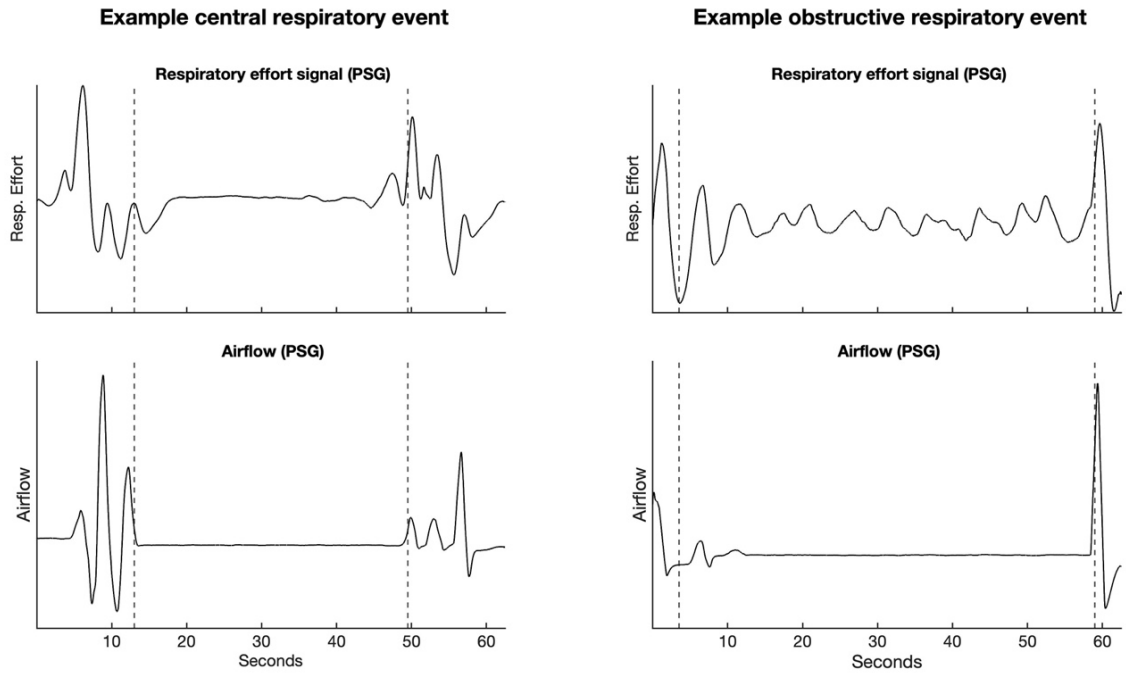

Figure 4: Comparison of a typical central and obstructive apnea event. Both events are accompanied by cessation of airflow. Still, a central event is characterised by a lack of respiratory effort, reflected in a flat region of the respiratory effort belt.

Since PAT HSAT relies solely on a finger PPG probe for respiratory event detection, it lacks airflow and respiratory-effort channels that would allow direct labelling of respiratory event types. As such, PAT HSAT (without adding additional sensors) cannot differentiate between the two types of sleep apnea.

Research Aim 2.1: Investigating a PPG-based approach for differentiating obstructive and central sleep apnea

Since OSA and CSA are two types of sleep apnea with very distinct aetiologies and therapy options, their differentiation is of high clinical relevance[35]. As such, we developed a method to detect CSA using only finger PPG data from PAT HSAT and clinically evaluated the proposed method by comparing it to simultaneous in-lab PSG.

Limitation 3: PAT HSAT does not measure sleep accurately

Introductory context

The gold standard method for staging sleep, as performed by PSG, is based on biopotential (EEG, EOG, and EMG) analysis[1]. Repeated efforts have been made to use PAT HSAT's PPG signal for sleep staging; however, the accuracy has consistently been inferior to that of PSG[36–39]. Accurately identifying sleep stages and, in addition, identifying the tendency for respiratory events to cluster within certain sleep stages (especially REM sleep) has relevant clinical implications[40,41]. The REM OSA phenotype is an important subcategorisation of the sleep apnea condition, characterised by the tendency for respiratory events to increase in frequency or severity during REM sleep[42]. Furthermore, as the most important index for sleep apnea severity determination is the AHI, the accuracy of the diagnosis depends on an accurate determination of the total sleep time[43], which is the denominator of the AHI. As such, it is important to explore to what degree the sleep stage determination of PAT HSAT in general, and REM OSA phenotyping in particular, can be enhanced. Two research aims were formulated to this end.

Research Aim 3.1: Evaluating and improving REM OSA phenotyping

To date, no assessment of REM OSA phenotyping performance has been conducted for any type of sleep testing technology. Therefore, our first aim was to compare the sensitivity and specificity of PAT HSAT-based and PSG-based REM OSA phenotyping. Secondly, we assessed the potential of a multi-night PAT HSAT testing protocol to increase the performance of REM OSA phenotype detection.

Research Aim 3.2: Enhancing PAT HSAT sleep staging by incorporating a single-lead biopotential measurement

With the aim of improving the sleep stage determination of PAT HSAT, we introduced a single-lead biopotential unit into the test setup that captures a superposition of EEG, EOG, and EMG signals. PSG determines sleep stages using multiple biopotential leads, most of which are positioned within the patient's hair, and this configuration can substantially inconvenience patients[1]. A major factor driving the widespread clinical adoption of PAT HSAT is its minimal obtrusiveness compared to alternative technologies[28]. Therefore, we only considered attempts to enhance PAT HSAT's sleep staging capability that preserve the minimally obtrusive character of the original method.

Single-lead biopotential devices can be designed in a highly miniaturised form factor and positioned without placing electrodes in the hair. We hypothesised that, with careful consideration of electrode placement, the device could meaningfully improve sleep stage classification accuracy of PAT HSAT systems. To test this hypothesis, we developed sleep scoring algorithms tailored to this single-lead biopotential setup and evaluated the performance of the extended PAT HSAT system in a clinical study.

Limitation 4: The limited analysis scope of PAT HSAT autoscoring algorithms restricts performance

Introductory context

PAT HSAT infers respiratory events from the co-occurrence of oxygen desaturations, pulse-rate surges, and reductions in PAT signal amplitude[1]. The main limitation compared with PSG is the absence of airflow and EEG channels[25], which requires respiratory events to be inferred indirectly from autonomic and oximetric signals. This inevitably introduces some discordance relative to PSG.

In both PSG and PAT HSAT, events are scored using only the physiological features present within a short, isolated event window. Information in the surrounding signal, before or after the event, is not considered[44]. Yet respiratory events often occur in repetitive sequences with similar morphological characteristics[45,46] (Figure 3), suggesting that contextual information could help distinguish true respiratory events from artefacts.

Research Aim 4.1: Investigating context-aware PAT HSAT autoscoring

We therefore investigated whether an autoscoring approach that incorporates the context surrounding each respiratory event could improve PAT HSAT performance and narrow the performance gap to PSG.

Conclusion of research aims and chapter structure overview

This thesis reports on the investigation into how the clinical utility of PAT HSAT can be enhanced by addressing a series of limitations of the current state-of-the-art of this technology. Limitation 1 relates to the general lack of clinical validation reporting standards for HSAT, leading to concerns about selective performance reporting and reproducibility. As such, it was important to address these structural scientific reporting issues first before addressing the limitations of a more technical nature (Limitations 2 to 4). From each identified limitation, one or two distinct research aims were formulated. The results of the research related to each individual aim were published as a separate peer-reviewed journal paper. Table 1 provides an overview of the sequential nature in which each of the following chapters of this thesis focuses on the findings of the research related to one limitation and its associated research aim(s).

Chapter	PAT HSAT Limitation	Research Aim	Publication
2	**Limitation 1**: The absence of clinical validation reporting standards for HSAT leads to selective reporting and difficulties in reproducing or comparing results	**Research Aim 1.1**: Addressing the prevalent misuse of correlation coefficients to support diagnostic performance	Massie F, Van Pee B, Bergmann J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. **J Clin Sleep Med**. 2022 Mar 1;18(3):871-876. doi: 10.5664/jcsm.9744.
		Research Aim 1.2: Proposing robust endpoint parameters with consideration for the inter-rater-disagreement of PSG	Van Pee B, Massie F, Vits S, Dreesen P, Klerkx S, Bijwadia J, Verbraecken J, Bergmann J. A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry. **Sleep**. 2022 May 12;45(5):zsac028. doi: 10.1093/sleep/zsac028.
3	**Limitation 2**: PAT HSAT does not differentiate between CSA and OSA	**Research Aim 2.1**: Investigating a PPG-based approach for differentiating obstructive and central sleep apnea	Massie F, Vits S, Khachatryan A, Van Pee B, Verbraecken J, Bergmann J. Central Sleep Apnea Detection by Means of Finger Photoplethysmography. **IEEE J Transl Eng Health Med**. 2023 Jan 12;11:126-136. doi: 10.1109/JTEHM.2023.3236393.
4	**Limitation 3**: PAT HSAT does not measure sleep accurately	**Research Aim 3.1**: Evaluating and improving REM OSA phenotyping	Massie F, Van Pee B, Vits S, Verbraecken J, Bergmann J. Phenotyping REM OSA by means of peripheral arterial tone-

			based home sleep apnea testing and polysomnography: A critical assessment of the sensitivity and specificity of both methods. **J Sleep Res**. 2022 Apr;31(2):e13481. doi: 10.1111/jsr.13481.
		Research Aim 3.2: Enhancing PAT HSAT sleep staging by incorporating a single-lead biopotential measurement	Massie F, Vits S, Verbraecken J, Bergmann J. The evaluation of a novel single-lead biopotential device for home sleep testing. **Sleep**. 2025 Apr 11;48(4):zsae248. doi: 10.1093/sleep/zsae248.
5	**Limitation 4**: The limited analysis scope of PAT HSAT autoscoring algorithms restricts performance	**Research Aim 4.1**: Investigating context-aware PAT HSAT autoscoring	Massie F, Vits S, Verbraecken J, Bergmann J. Context-aware analysis enhances autoscoring accuracy of home sleep apnea testing. **J Clin Sleep Med**. 2025 May 1;21(5):789-804. doi: 10.5664/jcsm.11534.

Table 1: Overview of chapter structure, the relation to the identified PAT HSAT limitations, and the associated research aims.

Chapter 2: Improving the performance evaluation of home sleep apnea tests

Research Aim 1.1: Addressing the prevalent misuse of correlation coefficients to support diagnostic performance

Published as: Massie F, Van Pee B, Bergmann J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. *Journal of Clinical Sleep Medicine*. 2022 Mar 1;18(3):871-876. doi: 10.5664/jcsm.9744.

Reiteration of research aim and introductory context

As discussed in Chapter 1, the clinical performance of HSAT can be described by its (diagnostic) accuracy (and related metrics), defined as the percentage agreement with the sleep apnea severity category (normal, mild, moderate, and severe) determined by the polysomnography serving as ground truth. Rather than reporting on accuracy and related metrics, there has been a firm reliance in the literature on reporting correlation coefficients between the AHI of HSAT and PSG to support claims of diagnostic performance[30]. In the manuscript introduced in this subchapter, we first demonstrated theoretically why correlation coefficients are inherently unsuitable for supporting claims about the diagnostic performance of any device that outputs a continuous index related to disease severity. We subsequently systematically investigated the magnitude of the discrepancies between correlation coefficients and diagnostic accuracy reported in or retrievable from HSAT validation studies. The analysis confirmed the issue with the overreliance on correlation parameters. It reemphasised the need for increased scientific rigour in selecting primary performance endpoints to support HSATs' clinical performance claims.

Abstract

Study Objectives

The clinical performance of HSAT can be described by its (diagnostic) accuracy, defined as the percentage agreement with the obstructive sleep apnea severity category (normal, mild, moderate, and severe) based on polysomnography. Rather than reporting on accuracy, there has been a strong reliance in the literature on reporting correlation coefficients between the apnea-hypopnea index of HSATs and polysomnography to support claims of diagnostic performance. This is surprising, as it has been well described that correlation coefficients are inadequate to evaluate equivalence between 2 parameters. The aim of this study was to systematically investigate the magnitude of the discrepancies between correlation coefficients and diagnostic accuracy reported in or retrievable from HSAT validation studies.

Methods

We compared the discrepancy between accuracy and apnea-hypopnea index correlation coefficients of all validation papers that met the inclusion criteria. A total of 20 papers were retained.

Results

The weighted average apnea-hypopnea index correlation across all 20 papers was 0.82, and the weighted average accuracy was 0.61, highlighting a discrepancy of 0.21 and an overall misdiagnosis rate of 39%.

Conclusions

The results of our study confirm the need for increased scientific rigor in selecting primary performance endpoints to support clinical performance claims of HSATs.

Introduction

The American Academy of Sleep Medicine clinical guideline for the evaluation, management, and long-term care of obstructive sleep apnea in adults[43] defines the severity of OSA by whether the AHI is lower than 5 (normal), between 5 and 15 (mild OSA), between 15 and 30 (moderate OSA), or above 30 events per hour of sleep (severe OSA).

The performance of an HSAT can be evaluated by determining the percentage agreement of the OSA severity (normal, mild, moderate, or severe) estimated by HSAT with that determined by concurrently administered gold-standard PSG. This percentage agreement is commonly referred to as the (diagnostic) accuracy or the (diagnostic) concordance.

Yalamanchali et al.[47] performed a comprehensive meta-analysis of the clinical performance of the most frequently studied HSAT by systematically reviewing all papers that reported on the AHI correlation between the HSAT and PSG. From the apparently very strong correlation coefficients reported in the included papers, the authors concluded that the HSAT represents a viable alternative to PSG for confirmation of clinically suspected sleep apnea.

However, it has been well described in domains outside of sleep medicine that correlation coefficients such as Pearson's and Spearman's correlations are inadequate in describing the degree of one-on-one correspondence of 2 diagnostic endpoint parameters[31]. Almost 2 decades ago, Flemons et al.[48,49] pointed out, based on an example comparing the RDI of HSAT and the AHI of

PSG, how (Pearson) correlation coefficients are only able to indicate whether 2 measurements are (linearly) related, and not whether they have a similar magnitude.

Indeed, the Pearson correlation coefficient attains the maximum value of 1 upon a perfect linear relationship between 2 endpoint parameters, but it does not penalise a constant offset or scaling factor between the parameters. Worse in this context is the Spearman correlation coefficient, as it attains the maximum value of 1 when there is a perfect monotonously increasing relationship between the 2 parameters, without penalising for the nonlinearity of such a relationship. Correlation coefficients are also heavily influenced by extreme datapoints, such as very high AHIs.

These issues can be illustrated by the examples presented in Figure 5. All 4 panels show how AHI correlation coefficients can be arbitrarily high for devices with little to no clinical utility (as illustrated by diagnostic accuracies of ≤ 0.5). The top left and right panels, respectively, show how the Pearson correlation does not take scaling mismatches and constant offsets between the AHIs of HSAT and PSG into account. The bottom left panel shows how low AHIs (< 5 events/h) and extreme AHIs (> 30 events/h) skew Pearson correlations to misleadingly large values, while the intrinsic diagnostic accuracy is very low. The bottom left panel has the same diagnostic accuracy of 0.5 as the bottom right panel but has fewer extreme AHI values in the sample pool, highlighting the collapse of the Pearson correlation in the absence of extremity-introduced bias. Importantly, extremely low and high AHIs, while excessively influencing correlations, have the least relevance in the performance evaluation of an HSAT. Indeed, it is comparatively straightforward to detect a near-complete absence of respiratory events (e.g., by a completely flat oxygen saturation trace) or an extreme prevalence of respiratory events (e.g., by the presence of extremely frequent oxygen desaturations). The patients most challenging to assess with any HSAT device are those with AHIs between 5 and 30 events/h, a range of particular importance for therapeutic decisions.

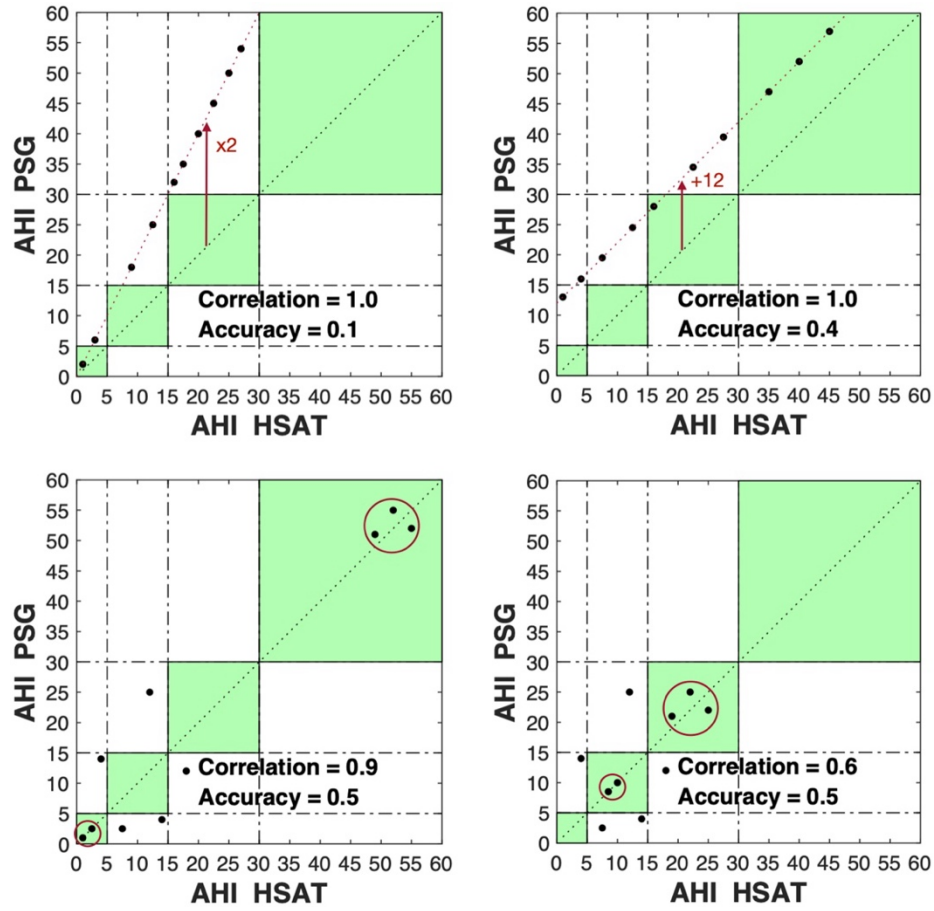

Figure 5: Sample scatterplots highlighting issues of relying on correlation coefficients to describe equivalence between 2 parameters. The green-shaded squares mark the areas in which PSG and HSAT agree on the OSA severity category. Top left: AHIs differ by a constant scaling factor (AHI PSG = 2 × AHI HSAT). Top right: AHIs differ by a constant offset (AHI PSG = AHI HSAT + 12). The differences between AHI PSG and AHI HSAT do not impact the correlation coefficient but strongly impact the accuracy. Bottom: Extreme values disproportionally influence correlation coefficients. Moving extreme values (dots within red circles) more to the centre of the spectrum, without altering the accuracy, has a strong effect on the correlation coefficient. This is observed by comparing the bottom left (correlation = 0.9) and bottom right (correlation = 0.6) panels.

A recent study by Ioachimescu et al.[50], in which the HSAT device described by Yalamanchali et al.[47] was evaluated against PSG on 500 participants, reported a very apparent discrepancy between the study's AHI Pearson correlation of 0.80 and diagnostic accuracy of 0.53. These findings drove a starkly different conclusion than the one by Yalamanchali et al.[47] Indeed, the authors stated that because of the large diagnostic misclassification rate, patients without OSA or mild disease assessed by the HSAT should undergo repeat in-laboratory PSG, as such, limiting the device's utility to

inclusion screening. These conflicting conclusions highlight the peril of reliance on correlation coefficients as opposed to diagnostic accuracy as a primary clinical performance endpoint of HSAT devices.

Prompted by the scientific community's surprise[51] about these findings, we systematically compared the discrepancy between diagnostic accuracy and AHI correlation coefficients of all published validation studies of the HSAT devices that met the search criteria.

Methods

To allow for cross-trial, weighted-average aggregation of accuracy and correlation parameters, we limited the inclusion of HSAT types to the single most frequently validated HSAT device. The WatchPAT device (Itamar Medical, Caesarea, Israel) was found to be the most frequently included in clinical validation studies (~ 20% of all HSAT validation studies report on this device).

As such, we searched PubMed on April 3, 2021, for the search term "'watchpat' or 'watch-pat' or 'watch pat.'"

Any paper that met at least 1 of the exclusion criteria was discarded from further analysis. The following exclusion criteria were used:

1. The paper was not available in the English language.
2. The paper did not report on the performance validation of the HSAT device.
3. The HSAT was not concurrently administered with PSG.
4. No diagnostic accuracy could be retrieved from the paper.
5. No AHI or RDI correlation coefficient could be retrieved from the paper.

For all included papers, the correlation between the (automated) AHI estimate of HSAT and PSG was retrieved. We additionally registered the type of the reported correlation coefficient (Spearman, Pearson, or intraclass correlation coefficient). For all included papers, we retrieved the accuracy of the HSAT. For papers for which the accuracy was not explicitly reported, or for which it could not be precisely calculated from the confusion matrix, we estimated the accuracy based on the datapoints of the scatter plots or the Bland-Altman plots, or inverse-calculated the accuracy from the disease prevalence, sensitivity, and specificity values. Such scatterplots are equivalent to those presented in Figure 5. We counted the number of dots falling in the green rectangles and divided this number by the total number of dots to arrive at the accuracy value. Any discrepancy between the estimated participant size from the scatter plots or Bland-Altman plots and the reported participant size was also identified.

As accuracy estimates could be impacted by incomplete or incorrect visual retrieval of the datapoints, we performed a subgroup analysis in which papers that did not allow exact diagnostic accuracy retrieval were excluded.

For papers in which only an RDI correlation and accuracy could be retrieved, we used the RDI instead of the AHI. Finally, the scoring rules used to score the AHI by PSG were noted for each paper. We performed a subgroup analysis on papers that reported on the 2012 AASM respiratory event scoring rules[1].

We then plotted the diagnostic accuracy and the AHI correlation coefficient for each included paper, ranked by decreasing sample size. Finally, we calculated the overall (sample-size weighted average) accuracy and the AHI correlation coefficient of the total participant pool.

Results

As illustrated in Figure 6, the search term resulted in the identification of 86 papers, of which 20 were retained after application of the exclusion criteria. Fifty out of 86 identified papers were not considered HSAT validation studies, rendering this the main reason for exclusion. The results of the systematic analysis are summarised in Figure 7 and Table 2.

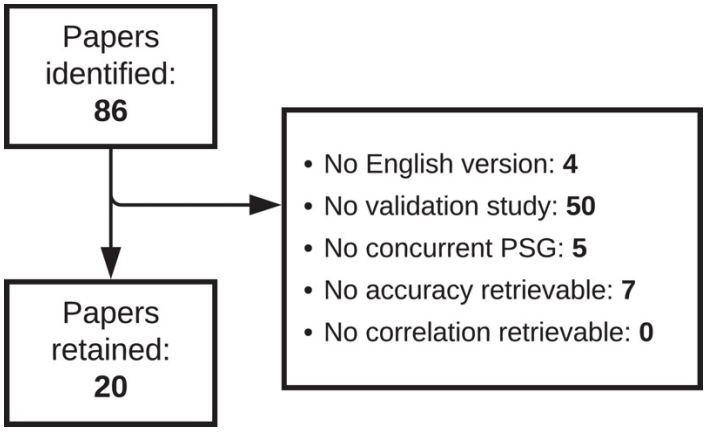

Figure 6: Overview of identified and retained papers and the reasons for any exclusions.

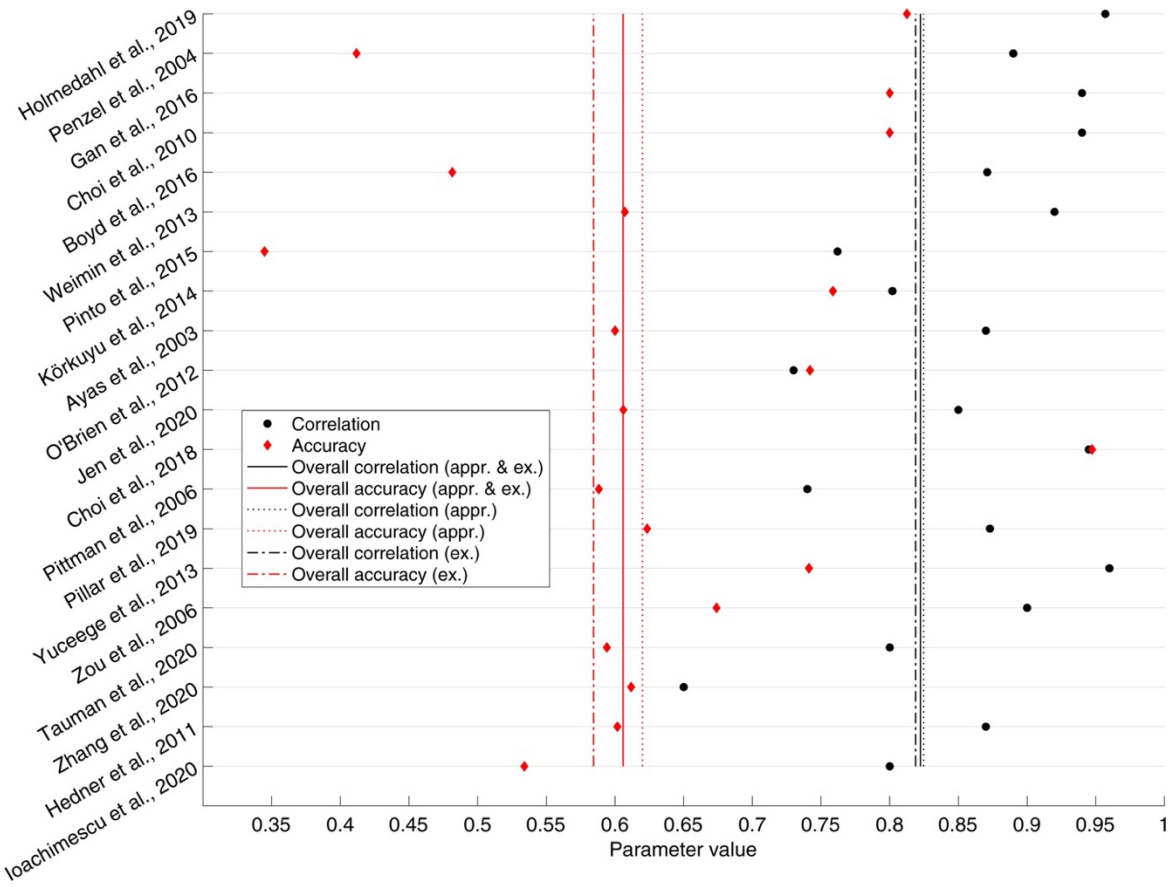

Figure 7: Comparison of accuracy and correlations for each included paper. Comparison of accuracy and AHI correlations (horizontal axis) as retrieved from each included paper (vertical axis). Papers are ranked by increasing sample size (largest at the bottom). Overall accuracy and correlation were calculated from a weighted average of all individual trials, with the weights equal to the trial population size. appr. = accuracy was approximated from the graphs or inverse-calculated; ex. = accuracy was exactly inferred.

First author and date	AHI or RDI	Correlation (type)	Accuracy (a or e)	Difference	N (reported)	Scoring rules
Holmedahl et al., 2019	AHI	0.96 (ICC)	0.81 (a)	0.15	16	O
Penzel et al., 2004	RDI	0.89 (Pearson)	0.41 (a)	0.48	17	O
Gan et al., 2016	AHI	0.94 (Pearson)	0.80 (a)	0.14	20	A
Choi et al., 2010	AHI	0.94 (Pearson)	0.80 (e)	0.14	25	O

Boyd et al., 2016	AHI	0.87 (Pearson)	0.48 (a)	0.39	27 (28)	O
Weimin et al., 2013	AHI	0.92 (Pearson)	0.61 (e)	0.31	28	O
Pinto et al., 2015	AHI	0.76 (Spearman)	0.35 (a)	0.42	29 (30)	O
Körkuyu et al., 2014	AHI	0.80 (Spearman)	0.76 (a)	0.04	29 (30)	O
Ayas et al., 2003	AHI	0.87 (Pearson)	0.60 (e)	0.27	30	O
O'Brien et al., 2012	AHI	0.73 (Pearson)	0.74 (e)	-0.01	31	O
Jen et al., 2020	AHI	0.85 (Pearson)	0.61 (a)	0.24	33	O
Choi et al., 2018	AHI	0.95 (Spearman)	0.95 (e)	0.00	38	A
Pittman et al., 2006	AHI	0.74 (ICC)	0.59 (a)	0.15	68 (70)	O
Pillar et al., 2020	AHI	0.87 (Pearson)	0.62 (a)	0.25	77 (84)	A
Yuceege et al., 2013	AHI	0.96 (Pearson)	0.74 (a)	0.22	85	O
Zou et al., 2006	AHI	0.90 (Pearson)	0.67 (a)	0.23	92	O
Tauman et al., 2020	AHI	0.80 (Pearson)	0.59 (a)	0.21	101	A
Zhang et al., 2020	AHI	0.65 (Spearman)	0.61 (a)	0.04	170	A
Hedner et al., 2011	RDI	0.87 (ICC)	0.60 (a)	0.27	236	O
Ioachimescu et al., 2020	AHI	0.80 (Pearson)	0.53 (e)	0.27	500	A and B
Overall (A or B)	AHI	0.79	0.59 (a or e)	0.20	906	A or B
Overall (O)	AHI or RDI	0.87	0.63	0.24	746	O

Overall (a)	AHI or RDI	0.83	0.62 (a)	0.21	1000	A, B or O
Overall (e)	AHI or RDI	0.82	0.58 (e)	0.24	652	A, B or O
Overall	AHI or RDI	0.82	0.61 (a or e)	0.21	1652	A, B or O

Table 2: Tabulation of accuracy and correlations for each included paper. Papers are ranked by increasing sample size. Papers exclusively reporting on the RDI were highlighted as such in column 2. For 'appr.' papers, the participant number (n) is displayed based on the paper's graphs, with the reported number between parentheses when it was different from the sample size determined from the graphs or metrics used to estimate the accuracy. 'Overall (A or B)' and 'Overall (O),' respectively, contain the sample-size weighted average of all studies scored by 'A or B' rules or 'O' rules. 'Overall (appr.)' and 'Overall (ex.),' respectively, contain the sample-size weighted average of all studies with an 'appr.' or 'ex.' accuracy. A = AASM 2012 scoring rules with 1A rule for hypopnea; appr. = accuracy was approximated from the graphs or inverse-calculated; B = AASM 2012 scoring rules with 1B rule for hypopnea; ex. = accuracy was exactly inferred; ICC = intraclass correlation coefficient; O = scoring rule other than any AASM 2012 rule; RDI = respiratory disturbance index.

The weighted average AHI correlation across all 20 papers was 0.82, and the weighted average accuracy was 0.61, highlighting a discrepancy of 0.21. The lowest observed accuracy was 0.35 (correlation 0.76) and the highest observed accuracy was 0.95 (correlation 0.95), which illustrates a large degree of performance variation between studies. The lowest observed AHI correlation was 0.65 (accuracy 0.61), and the highest observed AHI correlation was 0.96 (accuracy 0.74).

The weighted average AHI correlation and accuracy across all papers for which the accuracy could exactly be retrieved were, respectively, 0.82 and 0.58. For 6 out of 20 papers, the accuracy was exactly retrievable. Six out of 20 studies reported on the AASM 2012 scoring rules using the 1A rule for hypopnea requiring 3% desaturations[1]. Only 1 study reported (partially) on the AASM 2012 scoring rules using the 1B rule for hypopnea requiring 4% desaturations. Comparing the studies that scored using any of the AASM 2012 rules with the papers that applied an older scoring rule revealed a difference of 0.04 (from 0.59 to 0.63) for accuracy and 0.08 (0.79 to 0.87) for correlation, where the highest number was found for the older rules.

Discussion

Similar to the results reported by Ioachimescu et al.[50], our meta-analysis determined a discrepancy of 21% between the diagnostic accuracy of 61% and correlation coefficient of 82%, derived from a total participant pool size of 1,652. When only considering papers reporting on the 2012 AASM scoring rules, a misdiagnosis rate of 41% was found. For only 6 out of the 20 included papers, the diagnostic accuracy could directly be retrieved or exactly calculated, revealing a large underreporting of this important endpoint parameter.

These observations confirm the urgent need for researchers and device manufacturers to increase scientific rigor and transparency in the presentation of HSAT devices' clinical performance by complementing their reporting with diagnostic accuracy, in a move away from the overreliance on correlations. This study can be viewed as a precursor to much-needed future work on the determination of a statistically robust set of parameters to characterise the performance of HSATs in general, and the agreement between HSATs and PSG in particular.

This meta-analysis had several limitations. First, the accuracy had to be estimated from alternative data, such as scatter plots and Bland-Altman plots, for 14 out of 20 papers, due to a lack of explicit reporting of the parameter. Second, only 6 papers reported on the 2012 AASM scoring rules, and only 1 paper reported on the 2012 1B rule for the scoring of hypopnea. Finally, there was heterogeneity in participant population characteristics. Most papers (11/20) reported on a participant population with a high suspicion of OSA, while other papers discussed specific participant populations, such as those with cardiovascular or pulmonary comorbidities. Nevertheless, the overall participant pool closely reflects the population in which HSAT is typically used in clinical practice.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests.
Publication status:	*Published*
Publication details:	Massie F, Van Pee B, Bergmann J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. J Clin Sleep Med. 2022 Mar 1;18(3):871-876. doi: 10.5664/jcsm.9744.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed data acquisition and publication design and writeup. B Van Pee provided publication editing. J Bergmann provided supervisory guidance and publication editing.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Research Aim 1.2: Proposing robust endpoint parameters with consideration for the inter-rater-disagreement of PSG

Published as: Van Pee B, Massie F, Vits S, Dreesen P, Klerkx S, Bijwadia J, Verbraecken J, Bergmann J. A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry. *SLEEP*. 2022 May 12;45(5):zsac028. doi: 10.1093/sleep/zsac028.

Reiteration of research aim and introductory context

Our findings related to Research Aim 1.1 confirmed the need for researchers and device manufacturers to increase scientific rigor and transparency in the presentation of HSAT devices' clinical performance. Furthermore, as explained in Chapter 1, the AHI obtained from PSG is traditionally used as the ground-truth reference to which the PAT HSAT AHI is compared. The PSG-derived AHI is inherently an imperfect ground truth since significant inter-rater disagreement exists[32]. Indeed, the apnea events are visually scored from the PSG data by a sleep technician, and the rules of scoring such events leave room for the technician's interpretation. Considering this imperfection of PSG as a gold standard reference technology, we proposed novel endpoint reporting methodologies that adequately acknowledge and handle this imperfection. In the paper introduced in the following paragraphs, we report on the extensive clinical validation of a PAT HSAT with the aim of serving as a 'case in point' to introduce a novel concept of handling the imperfections of the PSG reference technology. This concept was called 'near-boundary double-labelling (NBL)'. In addition, the paper proposes a selection of robust endpoint parameters that, contrary to an

overreliance on correlation parameters, allow for transparent and clinically relevant reporting on HSAT performance.[a]

Abstract

Study objectives

This paper reports on the multicentric validation of a novel FDA-cleared PAT HSAT with the aim of proposing a set of robust endpoint parameters that specifically address the inter-rater variability of PSG scoring.

Methods

One hundred sixty-seven participants suspected of having OSA were included in a multicentric cohort. All patients underwent simultaneous PSG and PAT HSAT, and all PSG data were independently double scored using both the Recommended 1A rule for hypopnea, requiring a 3% desaturation or arousal (3% rule), and the Acceptable 1B rule for hypopnea, requiring a 4% desaturation (4% rule). The double-scoring of PSG enabled a comparison of the agreement between

[a]Several authors adopted or expressed the endorsement of this novel reporting method in their publications[28,152–157]. Ioachimescu[152] published a journal article to corroborate the clinical and scientific relevance of the NBL method. It concluded that this unique approach of double-labelling sleep studies allows for a more in-depth analysis of the agreement between PAT HSAT and PSG while considering the PSG inter-rater agreement when dealing with spurious categorisation variance generated by reference AHIs close to the traditional thresholds or boundaries. Martinot et al.[153] published an article titled '*Near-boundary double-labelling-based classification: the new standard when evaluating performances of new sleep apnea diagnostic solutions against polysomnography?*' The authors concluded that applying the NBL approach could help make better diagnostic and therapeutic decisions in patients presenting PSG results within near-boundary intervals. They further state that using the NBL procedure in clinical research would allow for better standardisation, transparency, and reproducibility in PSG data analysis.

PAT HSAT, and PSG to the inter-rater agreement of PSG. Clinical endpoint parameters were selected to evaluate the device's ability to determine the OSA severity category. Finally, a correction for near-boundary apnea-hypopnea index values was proposed to adequately handle the inter-rater variability of the PSG benchmark.

Results

For both the 3% and the 4% rules, most endpoint parameters showed substantial agreement with PSG. The 4-way OSA severity categorisation accuracy of PAT HSAT was nevertheless lower than the inter-rater agreement of PSG (70% vs 77% for the 3% rule and 78% vs 81% for the 4% rule).

Conclusions

This paper reported on a multitude of robust endpoint parameters while also benchmarking clinical performances against double-scored PSG. To address the inter-rater variability of PSG, the concept of near-boundary double labelling was introduced. This study demonstrated promising agreement between PAT HSAT and PSG.

Introduction

The COVID-19 pandemic has reshaped how OSA diagnosis is being performed. Sleep labs face restrictions on the number of patients they can admit to their facilities during the outbreaks, which resulted in an accelerated shift from in-lab PSG to home sleep apnea testing devices (HSATs)[24]. PAT HSAT was especially well positioned to accommodate this shift as it is cost-effective to produce and can be deployed in a compact form, driving efficiencies in logistics and ecological footprint[26,27,66].

A brief history of peripheral arterial tonometry

In 1937, Hertzman published a paper titled 'Photoelectric plethysmography of the Fingers and Toes in Man'[67], which would later be credited as the founding paper for the research into PPG[68]. PPG operates based on optical technology to detect pulsatile blood volume changes in the tissue, from which blood oxygen level estimates can also be derived (i.e., pulse oximetry). Hertzman observed how changes in the tone of the peripheral arterial smooth muscle tissue, also referred to as peripheral arterial tone and itself triggered by changes in sympathetic tone, were observable in the pulsatile blood volume changes as registered by the PPG. In a follow-up paper from 1942, Hertzman et al.[69] described the occurrence of such periodic changes in PAT in a snoring individual, in what we may today speculate to have been a patient with sleep apnea.

In the early 1970s, Lugaresi et al.[70] further complemented Hertzman's work in his reporting of the simultaneous occurrence of respiratory disturbances with an increase in PAT, an acceleration in pulse rate, and the presence of a cortical arousal. This observation closely resembles the AASM's definition of the Peripheral Arterial Tonometry HSAT technique [1].

At the time of this publication there are two HSATs cleared by the U.S. Food and Drug Administration (FDA) in the PAT category: WatchPAT (Itamar Medical, Israel)[50] and NightOwl (ResMed, San Diego, USA)[21], the latter of which is the device studied in this manuscript (study device). Both devices make use of signal conditioning methods to derive a sensitive PAT measurement from the PPG but differ in the mechanisms used to obtain such a measurement. WatchPAT uses mostly hardware implementations: an approximately isosbestic PPG wavelength provided through a third optical emitter helps compensate for fluctuations in the PPG driven solely by blood oxygen changes. A cuff-like pneumo-optic probe applies an approximately uniform and constant pressure, with claims of improving the signal-to-noise ratio of the PPG as well as preventing venous pooling from affecting the PPG[71]. The study device comprises a wrap-around sensor probe

which is the size of a fingertip and does not fully envelop the finger. Instead of a uniform pressure probe, it relies on signal processing techniques to compensate for varying levels of blood oxygen saturation and the effects of venous pooling, as well as to obtain a more linear measurement of PAT. Relying on signal processing instead of physical components allowed for further miniaturisation of the PAT technology.

Study objective

The aim of this paper was to report on the multicentric validation of a novel home sleep apnea test based on PAT with a particular emphasis on rethinking robust clinical endpoint design and evaluation.

Methods

Participants

One hundred sixty-seven participants suspected of having OSA were consecutively included in a cohort across four different centres of which one was located in Belgium (Ziekenhuis Oost Limburg, ZOL, Genk) and three in the United States (where all centres were part of the United Health Systems Group in Miami, FL). All participants were scheduled for one overnight in-lab PSG. Participants were asked for informed consent. The USA branch of the study was approved by Aspire Institutional Review Board (IRB), part of the WIRB-Copernicus Group. The European branch of the study was approved by the Ethics Committee of ZOL. Underaged or cognitively impaired participants were excluded from participation in the study. For the European centre, recruitment took place between July 2018 and September 2018. For the USA-based centres, recruitment took place between December 2019 and January 2020. For all participants, gender, age, and Body-Mass-Index (BMI)

were recorded. For the USA branch of the study, participants completed the FDA's *self-completion questionnaire for ethnicity and race*.

Protocol and devices

A graphical representation of the study setup is provided in Figure 8. Routine PSG was performed in all study participants. Qualified lab technicians at each participating study centre were responsible for setting up the equipment and capturing PSG data. During the setup of PSG, the PAT HSAT (NightOwl) was attached to the middle finger of the hand to which the pulse oximeter of PSG was applied. All PSG data were double scored by two independent centres, which were blinded from one another's analysis.

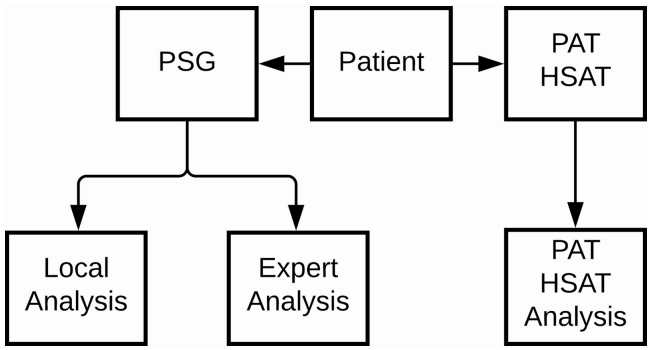

Figure 8: Diagram of the data acquisition setup. This diagram depicts that PSG and PAT HSAT were administered concurrently for each study participant. PSG was analysed by two independent scoring centres (Expert and Local), and PAT HSAT was analysed automatically.

Polysomnography

For the European centre, the Alice 6 PSG (Philips Respironics, USA) was used, whereas a Cadwell Easy PSG (Cadwell, USA) was applied in the USA centres.

PSG was scored by two independent scoring centres. The first scoring was performed by the team of sleep technicians of the centre where the patient was admitted (further referred to as the 'Local

Analysis'). A second independent scoring was performed by scorers of Cerebra Medical (CM, Canada), which provides computer-aided sleep scoring services to support PSG scoring for clinical centres and clinical trials. The studies were first analysed by their proprietary Michele Sleep Scoring System (MSSS) and were subsequently complemented with complete manual rescoring by an expert technologist. All expert scorers of CM had received Registered Polysomnographic Technologist certification through the Board of Registered Polysomnographic Technologists.

Malhotra et al.[32] confirmed in a multicentric trial that the MSSS, complemented with manual editing by an expert scorer, is more robust than the results of a single expert scorer. Because of these conclusions, CM's analysis served as the expert benchmark to which the Local Analysis and PAT HSAT's analysis were compared. The analysis by CM is therefore referred to as the 'Expert Analysis.'

All PSG data were scored according to the 2012 AASM scoring rules[1]. Data were first scored using the Recommended rule 1A for the scoring of hypopnea (3% rule), requiring a 3% desaturation or an arousal for the scoring of hypopnea. An alternative scoring of all PSGs was also performed using the Acceptable rule 1B (4% rule) for the scoring of hypopnea by discarding all hypopneas that did not coincide with an oxygen desaturation of at least 4%.

Data Analysis

General

Statistical analysis was performed using MATLAB (version 2019a, MathWorks, USA). For all endpoint parameters, 95% confidence intervals were computed. For proportion-based endpoints, confidence intervals were computed by approximating the distribution of error of a binomially distributed observation with a normal distribution. Significance levels were set at a p-value of 0.05. The PAT HSAT outcome was compared to the Expert Analysis for both the 3% rule and the 4% rule

by using the PAT HSAT AHI (pAHI) scored by its 3% rule and 4% rule scoring variant. Significant differences between two proportions were assessed by means of a *two-proportion z-test.*

Data synchronisation

PSG and PAT HSAT data were algorithmically synchronised by matching the instantaneous heart rate traces derived from the ECG trace of PSG and the pulse rate trace of PAT HSAT. Data epochs that were of insufficient quality to be interpreted by the sleep technician or PAT HSAT were rejected from both PSG and PAT HSAT traces. This resulted in a median rejection rate of 12% of data epochs per recording.

Data adequacy: Technical failure of PSG and missing PSG data or annotations

The AASM defines HSAT as technically adequate if at least 4 h of analysable signal can be obtained[33]. In our study, the same cutoff criterion of technical adequacy was used, and all technically inadequate recordings were excluded from further analysis. When the PSG recording was technically inadequate, for instance, when one of the channels could not be interpreted by the technicians, the participant was excluded from analysis. Similarly, when PSG data or any annotations of the two scoring centres were missing due to administrative errors, the participant was removed from further analysis.

Participants with missing patient characteristics, such as age and gender data, were omitted from the analysis of population demographic statistics.

Performance endpoints

The clinical performance of HSAT can be described by its (diagnostic) accuracy, defined as the percentage agreement with polysomnography of the obstructive sleep apnea severity category (normal, mild, moderate, and severe)[43].

Secondary performance endpoints, which characterise the device's bias and variance in estimating the AHI, may provide additional insights as to the device's propensity to over- or underestimate the OSA severity. In light of the considerations mentioned above, a list of primary and secondary endpoints is proposed.

Primary endpoints

OSA severity categorisation accuracy (4-way categorisation accuracy)

The 4-way categorisation accuracy expresses the percentage of agreement between the OSA severity determined by HSAT and the OSA severity determined by PSG. Its main advantage is its straightforward interpretation. Its main disadvantage is its lack of insight into whether the categorisation performance of HSAT exceeds the agreement that can be obtained by random guessing (chance level). Consider an extreme example where 90% of the study participants have mild OSA. In such a case, it is trivial for HSAT to obtain a 90% categorisation accuracy by outputting mild OSA 100% of the time without performing any meaningful inference. The 4-way accuracy of 90% would misleadingly suggest that this test is effective.

Cohen's Kappa (κ)

To address the main limitation of categorisation accuracies, Cohen's Kappa[72] is an alternative agreement metric that takes into account the chance level. Cohen's Kappa is formulated as follows:

$$\kappa = \frac{categorization\ accuracy - chance\ level}{1 - chance\ level}$$

The downside of this metric is its less straightforward interpretation. Applying this formula to the previous example, Cohen's Kappa corresponding to the 90% categorisation accuracy would be 0.

Confusion matrix, sensitivity (Se), specificity (Sp), negative predictive value (NPV), positive predictive value (PPV), and cutoff agreement (Acc)

Confusion matrices and their derived parameters provide additional granularity beyond categorisation accuracy and Cohen's Kappa, as they reveal whether HSAT tends to over- or underestimate specific OSA severity categories.

Secondary endpoints

Bland-Altman analysis

In order to describe the bias and variance of AHI estimates, a Bland-Altman analysis can be performed, which sets out the average AHI of the reference and comparator against their difference[73]. The standard Bland-Altman analysis is sensitive to extreme values, typically occurring at higher AHIs. Therefore, we propose to complement the standard Bland-Altman analysis with a sub-analysis in which only reference AHIs smaller than 30 are retained. For non-normally distributed differences, the limits of agreement (LoA) were determined as the 97.5th and 2.5th percentiles of the differences. For normally distributed differences, the LoA were determined as the mean ±1.96 times the standard deviation of the residuals. We also performed a Bland-Altman analysis for the (estimated) TST.

ICC(A,1)

The degree of absolute agreement between two AHI estimates (and other parameters such as TST) can be described by the intraclass correlation coefficient of the type of *two-way fixed model with single measures of absolute agreement* (ICC(A,1))[74].

In a context where absolute agreement rather than merely a linear relationship is important, the ICC(A,1) is a more robust and targeted parameter than the commonly used Pearson or Spearman correlation coefficient. The Pearson correlation coefficient attains the maximum value of 1 upon a perfect linear relationship between the two raters' observations, but it does not penalise a constant offset or a scaling factor between them. For example, if the AHI determined by HSAT would be consistently equal to twice the AHI of PSG increased by 10 events per hour, a perfect linear relationship would exist, and the Pearson correlation coefficient would attain the maximum value of 1. Nevertheless, such HSAT would have impaired clinical utility. Worse in this context is the Spearman correlation coefficient, as it attains the maximum value of 1 when there is a perfect monotonously increasing relationship between the two variables without penalising for non-linearity of such a relationship[30]. When absolute agreement needs to be assessed, these coefficients provide misleadingly high values for HSAT and should be avoided. The ICC(A,1) penalises both issues and attains the maximum value of 1 only upon a perfect match (i.e., absolute agreement) between the raters' observations. Nevertheless, the ICC(A,1), like most other correlation coefficients, is heavily influenced by outliers. Therefore, to assess this influence, we included an additional ICC(A,1), which was calculated on only those participants for whom the Expert Analysis' AHI was less than 30. Confidence intervals for the ICC(A,1) were calculated as described by McGraw et al.[74].

Endpoint assessment

No consensus exists on what endpoint parameter values are required to conclude that an HSAT performs adequately. To avoid defining arbitrary performance targets, we benchmarked the HSAT endpoint parameters against the scorer-to-scorer parameters obtained by comparing the Local PSG Analysis with the Expert PSG Analysis. For each endpoint parameter, we then assessed whether the HSAT–Expert Analysis value was significantly lower than the corresponding PSG inter-scorer value.

Handling AHIs close to OSA severity category boundaries

Significant inter-rater disagreement on key diagnostic parameters, such as the AHI, exists[32]. This implies that an AHI derived by PSG that is close to any of the OSA severity category boundaries (5, 15, and 30) should be treated with caution. For example, an AHI of 15.1 would qualify as moderate, whereas an AHI of 14.9 would qualify as mild, which could have different therapeutic implications. However, this difference in AHI is much smaller than the typical inter-rater variability of the AHI. A dataset in which a significant proportion of AHIs are close to the OSA severity boundaries (near-boundary AHIs) could provide an overly pessimistic assessment of HSAT performance. Therefore, we complemented any endpoint analysis based on AHI cutoffs with an alternative endpoint parameter calculation that corrects for near-boundary AHIs. Concretely, we allocated two possible OSA severity categories to near-boundary Expert Analysis' AHIs. This process was called near-boundary double-labelling (NBL). For example, an Expert Analysis' AHI of 14.9 would receive the label of *mild OR moderate OSA* rather than just *mild OSA*. As a result, if an HSAT detects *moderate OSA*, this scoring should be considered in agreement with the Expert Analysis. For endpoint parameters that are evaluated at a single AHI cutoff, the same NBL principle can be applied. For example, if the agreement at AHI cutoff 15 is evaluated and if the Expert Analysis' AHI is very close to 15, the ground truth OSA severity category is similarly likely to be either mild or moderate and is as such to be considered in default agreement with HSAT's or Local Analysis' AHI categorisation at cutoff 15.

When implementing boundary corrections, it is important to establish adequate ranges for near-boundary-zones (NBZ), for which Expert Analysis' AHIs falling within should receive NBL. We determined the NBZ from analysing the double-scored PSG data obtained in this study.

As the first step, we estimated the probability that a second independent scoring of the PSG data would end up in a different OSA severity category from the Expert Analysis (OSA severity

disagreement probability). For each AHI value ranging from 0 to 40 (the reference AHI), evaluated at increments of 0.2 events per hour, we gathered the observed AHI differences of the two PSG scorings (,i.e. the Expert and Local scorings) for which one of the two scorer's AHI was within a range of 5 events per hour of that reference AHI. We included those AHI differences in what was named the nearby sample set for that particular reference AHI. We could then fit a normal distribution onto the nearby sample set of each reference AHI by taking the reference AHI as the mean of the distribution with a standard deviation equal to that found within the nearby sample set of AHI differences. From the resulting AHI disagreement probability distribution, we could straightforwardly calculate the OSA severity disagreement probability for each reference AHI. The ±5 events/h window used to define the nearby sample set was deliberately not narrowed further. A smaller window would have required a larger dataset to maintain a smooth OSA severity disagreement probability curve, as quantified by the number of slope sign changes of that curve. Normality of the nearby sample set was evaluated by using the Anderson-Darling test[75]. We defined the NBZ as those AHI ranges for which there is a probability of one out of three (33%) OSA severity disagreement. The rationale for selecting a 33% disagreement-probability cutoff is twofold. First, if the cutoff were set at 50%, meaning that AHIs are double-labelled only when the probability of falling into a different OSA severity category is at least even, then almost no AHIs would qualify. Only AHIs lying exactly on the boundary values (5.0, 15.0, and 30.0) would reach this probability, resulting in practically no double-labelling. Second, a cutoff that is too low would have the opposite effect. For example, with a 15% cutoff in this dataset, almost all AHIs except those in the very severe range would fall within the near-boundary zone and would therefore be double-labelled. A cutoff of 33% provides a balanced compromise between these extremes and can be finetuned by individual practitioners' preference.

Sample size determination

Statistical power was determined by postulating that a 10% decrease in OSA severity categorisation accuracy should be detected as statistically significant with an alpha level of 0.05 and a power of 0.8. Assuming a minimum 4-way categorisation accuracy parameter value of 0.75 for the Local Analysis, a minimum sample size of 165 participants was found.

Results

Participant inclusion and population statistics

Figure 9 provides a flowchart highlighting the number of recruited participants as well as administrative and technical failures, including the reason for failure. Out of the 228 participants who gave informed consent, concurrent PSG and PAT HSAT data were successfully acquired for 180 participants. For respectively 4 and 7 participants, there was an issue with the airflow or SpO_2 channel of PSG, rendering scoring impossible. For 9 participants, PSG was only single-scored, or the link between the PAT HSAT, and PSG could no longer be retrieved. Three participants received a defective PAT HSAT sensor that did not acquire any data. Two participants detached the PAT HSAT sensor early in the study. For 7 participants who gave informed consent, eventually, no PAT HSAT data acquisition was started. For 16 participants in the European branch of the study, the PAT HSAT data acquisition app experienced stability issues, resulting in a loss of data. For 13 out of the 180 successful inclusions, the PAT HSAT recordings were not considered technically adequate since less than 4 h of interpretable data could be acquired. As a result, technically adequate

data were acquired for 167 participants. Of these 167 technically adequate inclusions, 74 were performed in the United States and 93 in Europe.

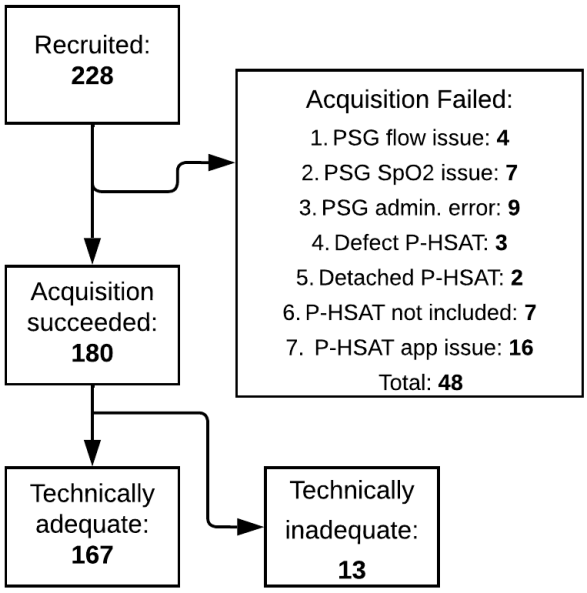

Figure 9: Flowchart of participant recruitment. This flowchart depicts how many participants gave informed consent (228), for how many participants PSG and PAT HSAT data acquisition was successful (180), and for how many participants the PAT HSAT was technically adequate (> 4 h of interpretable data). Additionally, the flowchart summarises the reasons for data acquisition failure.

As elaborated in Table 3, participants were predominantly male (63%), of middle age (mean 56 years, standard deviation (SD) 15), and overweight (mean BMI 30.7 and SD 6.3). The mean AHI was 32.7 (SD 26.8). Twenty-two participants had no OSA, 37 participants had mild OSA, 33 participants had moderate OSA, and 75 participants had severe OSA. In the USA branch of the study, 26% of participants identified as black, 74% as white, and 55% as Hispanic, Latino, or Spanish.

	Mean	SD	Min	Max
Age	56	15	21	84
BMI	30.7	6.3	18.2	53.8
AHI	32.7	26.8	0.0	114.9
	N (%)			
No OSA	22 (13%)			
Mild OSA	37 (22%)			
Moderate OSA	33 (20%)			
Severe OSA	75 (45%)			
Black	43 (26%)			
White	124 (74%)			
Hispanic, Latino, or Spanish	92 (55%)			

Table 3: Demographic and clinical characteristics of participants in the dataset. N = number of participants; SD = standard deviation.

Near-boundary determination

Figure 10 displays the results of the near-boundary AHI determination, as described in the methods section. The figure displays the OSA severity disagreement probability for each AHI. The red zones highlight those AHI values for which this probability is larger than 33% (i.e., the near-boundary zones). As such, these near-boundary zones can be summarised as displayed in Table 4. The values presented in this table were used to calculate primary endpoint parameters using near-boundary double-labelling. For AHIs ranging from 0 to 12 and from 33 to 40, the normal distribution assumption of the nearby sample set was not met according to the Anderson-Darling test. A violation of this assumption might render the error probability estimates less accurate in those regions.

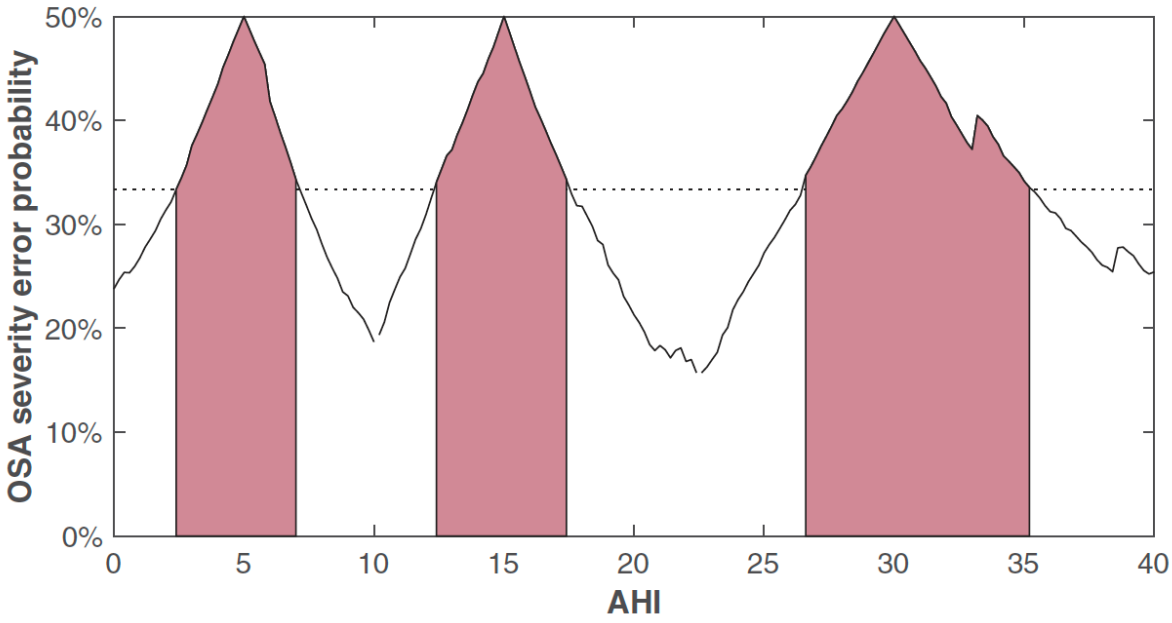

Figure 10: OSA severity error probability. This figure displays for each AHI the probability that a second PSG scoring would result in an AHI with a different OSA severity category. The red zones highlight those AHI values for which this probability is larger than 33% (dotted line).

OSA severity categories using NBL	**AHI Range**	**N (%)**
Normal	0.0 ≤ AHI < 2.4	12 (7.2%)
Normal OR Mild (NBZ)	2.4 ≤ AHI < 7.0	21 (12.6%)
Mild	7.0 ≤ AHI < 12.4	17 (10.2%)
Mild OR Moderate (NBZ)	12.4 ≤ AHI < 17.4	15 (9.0%)
Moderate	17.4 ≤ AHI < 26.6	21 (12.6%)
Moderate OR Severe (NBZ)	26.6 ≤ AHI < 35.2	14 (8.4%)
Severe	35.2 ≤ AHI	67 (40.0%)

Table 4: OSA severity categories when applying near-boundary double-labelling. The near-boundary zone is defined as the zone for which the probability that a second scoring of the AHI would fall in a different OSA severity category exceeds 33%. NBZ are highlighted in italics. N = number of participants within each category (according to Expert Analysis); NBL = near-boundary double-labelling; NBZ = near-boundary zones.

Primary endpoint parameters

Table 5 displays all primary endpoint parameters calculated for PAT HSAT as well as the Local PSG Analysis, for both the 3% rule and 4% rule, and for both the regular endpoint parameter calculation and the alternative calculation with NBL. Whenever PAT HSAT, or Local Analysis statistically significantly outperformed one another, the outperforming endpoint parameter is highlighted with an asterisk.

Endpoint parameter		**NBL**	**3% Rule**		**4% Rule**	
			PAT HSAT	Local PSG	PAT HSAT	Local PSG
4-way categorisation accuracy		*off*	0.70 (0.63 – 0.77)	0.77 (0.70 – 0.83)	0.78 (0.72 – 0.84)	0.81 (0.76 – 0.87)
		on	0.81 (0.75 – 0.87)	0.88 (0.83 – 0.93)*	0.89 (0.84 – 0.93)	0.89 (0.84 – 0.93)
4-way Cohen's Kappa		*off*	0.56 (0.46 – 0.66)	0.66 (0.57 – 0.76)*	0.70 (0.62 – 0.79)	0.75 (0.67 – 0.83)
		on	0.72 (0.64 – 0.81)	0.83 (0.76 – 0.90)*	0.85 (0.78 – 0.91)	0.85 (0.78 – 0.91)
Cutoff AHI 5	Se	*off*	0.93 (0.89 – 0.97)	0.95 (0.93 – 0.99)	0.95 (0.91 – 0.99)	0.97 (0.94 – 1.00)
		on	0.97 (0.94 – 1.00)	0.99 (0.98 – 1.00)	0.97 (0.94 – 1.00)	0.98 (0.96 – 1.00)
	Sp	*off*	0.72 (0.54 – 0.91)	0.72 (0.54 – 0.91)	0.80 (0.68 – 0.92)	0.84 (0.73 – 0.95)
		on	0.85 (0.71 – 0.95)	0.88 (0.74 – 1.00)	1.00 (0.96 – 1.00)*	0.88 (0.79 – 0.98)
	NPV	*off*	0.62 (0.42 – 0.80)	0.73 (0.54 – 0.91)	0.85 (0.75 – 0.96)	0.90 (0.81 – 0.99)
		on	0.85 (0.71 – 0.99)	0.95 (0.87 – 1.00)	0.90 (0.81 – 0.99)	0.95 (0.88 – 1.00)
	PPV	*off*	0.96 (0.92 – 0.99)	0.96 (0.93 – 0.99)	0.93 (0.88 – 0.97)	0.94 (0.90 – 0.98)
		on	0.97 (0.94 – 1.00)	0.98 (0.96 – 1.00)	1.00 (0.97 – 1.00)	0.99 (0.93 – 1.00)
	Acc	*off*	0.90 (0.86 – 0.95)	0.93 (0.89 – 0.97)	0.91 (0.87 – 0.95)	0.93 (0.90 – 0.97)
		on	0.95 (0.92 – 0.98)	0.98 (0.95 – 1.00)	0.98 (0.95 – 1.00)	0.96 (0.93 – 0.99)
	Kappa	*off*	0.60 (0.43 – 0.79)	0.69 (0.52 – 0.86)	0.76 (0.65 – 0.88)	0.83 (0.73 – 0.93)
		on	0.82 (0.70 – 0.94)	0.90 (0.80 – 1.00)*	0.93 (0.87 – 1.00)	0.89 (0.81 – 0.97)
Cutoff AHI 15	Se	*off*	0.91 (0.85 – 0.96)	0.94 (0.89 – 0.98)	0.94 (0.90 – 0.99)	0.94 (0.90 – 0.99)
		on	0.94 (0.89 – 0.98)	0.95 (0.92 – 0.99)	0.96 (0.92 – 1.00)	0.97 (0.93 – 1.00)
	Sp	*off*	0.76 (0.65 – 0.87)	0.85 (0.76 – 0.94)	0.91 (0.85 – 0.97)	0.86 (0.78 – 0.94)
		on	0.83 (0.73 – 0.93)	0.90 (0.82 – 0.98)	0.95 (0.90 – 1.00)	0.90 (0.83 – 0.96)
	NPV	*off*	0.82 (0.72 – 0.92)	0.87 (0.79 – 0.96)	0.93 (0.88 – 0.99)	0.93 (0.87 – 0.99)
		on	0.87 (0.79 – 0.96)	0.91 (0.84 – 0.99)	0.95 (0.90 – 1.00)	0.96 (0.91 – 1.00)
	PPV	*off*	0.88 (0.81 – 0.94)	0.92 (0.87 – 0.97)	0.92 (0.87 – 0.98)	0.89 (0.82 – 0.95)
		on	0.91 (0.86 – 0.96)	0.95 (0.90 – 0.99)	0.96 (0.92 – 1.00)	0.92 (0.86 – 0.97)
	Acc	*off*	0.86 (0.80 – 0.91)	0.90 (0.86 – 0.95)	0.93 (0.89 – 0.97)	0.90 (0.86 – 0.95)
		on	0.90 (0.85 – 0.94)	0.93 (0.90 – 0.97)	0.95 (0.92 – 0.99)	0.93 (0.90 – 0.97)
	Kappa	*off*	0.68 (0.56 – 0.80)	0.79 (0.69 – 0.89)*	0.86 (0.78 – 0.93)	0.81 (0.72 – 0.90)
		on	0.77 (0.67 – 0.88)	0.85 (0.77 – 0.94)*	0.90 (0.84 – 0.97)	0.87 (0.79 – 0.94)

Cutoff AHI 30	Se	*off*	0.91 (0.84 – 0.97)	0.91 (0.84 – 0.97)	0.86 (0.77 – 0.95)	0.95 (0.89 – 1.00)
		on	0.93 (0.88 – 0.99)	0.95 (0.90 – 1.00)	0.91 (0.83 – 0.98)	1.00 (0.96 – 1.00)*
	Sp	*off*	0.91 (0.86 – 0.97)	0.92 (0.87 – 0.98)	0.96 (0.93 – 1.00)	0.95 (0.90 – 0.99)
		on	0.94 (0.88 – 0.99)	0.96 (0.92 – 1.00)	0.97 (0.93 – 1.00)	0.95 (0.91 – 0.99)
	NPV	*off*	0.92 (0.87 – 0.98)	0.92 (0.87 – 0.98)	0.93 (0.88 – 0.98)	0.97 (0.94 – 1.00)
		on	0.95 (0.90 – 0.99)	0.96 (0.92 – 1.00)	0.96 (0.92 – 0.99)	1.00 (0.97 – 1.00)
	PPV	*off*	0.90 (0.83 – 0.96)	0.91 (0.84 – 0.97)	0.92 (0.85 – 1.00)	0.90 (0.82 – 0.98)
		on	0.92 (0.86 – 0.98)	0.95 (0.90 – 1.00)	0.92 (0.85 – 1.00)	0.90 (0.82 – 0.98)
	Acc	*off*	0.91 (0.87 – 0.95)	0.92 (0.87 – 0.96)	0.93 (0.89 – 0.97)	0.91 (0.91 – 0.98)
		on	0.93 (0.90 – 0.97)	0.95 (0.92 – 0.99)	0.95 (0.91 – 0.98)	0.96 (0.94 – 0.99)
	Kappa	*off*	0.82 (0.73 – 0.91)	0.83 (0.75 – 0.92)	0.84 (0.75 – 0.93)	0.88 (0.81 – 0.96)
		on	0.87 (0.79 – 0.94)	0.90 (0.83 – 0.97)	0.88 (0.80 – 0.95)	0.92 (0.86 – 0.98)

Table 5: Endpoint parameter results summary table. This table displays all endpoint parameter values and their confidence intervals (between brackets) for PAT HSAT and Local PSG, and for the 3% rule and the 4% rule, with and without NBL. An asterisk signifies a significant outperformance of the endpoint parameter (within the same scoring rule selection). Acc = accuracy; Kappa = Cohen's Kappa; NBL = near-boundary double-labelling; NPV = negative predictive value; PPV = positive predictive value; Se = sensitivity; Sp = specificity.

The 4-class accuracy using NBL, as well as the Cohen's Kappa (with and without NBL), was significantly lower for PAT HSAT compared to the Local Analysis for the 3% rule. The specificity at AHI cutoff 5 after NBL was significantly higher for PAT HSAT for the 4% rule. The Cohen's Kappa for the same cutoff after NBL was significantly lower for PAT HSAT for the 3% rule. Another significant underperformance of PAT HSAT was found for the AHI cutoff of 15 for the calculation of Cohen's Kappa, both with and without NBL. Finally, the sensitivity of the Local Analysis at AHI cutoff 30 calculated with NBL was significantly higher than the PAT HSAT's for the 4% rule. No other statistically significant differences were found.

For 84% of all reported endpoint parameter values, the value for the 4% rule was higher than or equal to that of the 3% rule. The impact of near-boundary AHIs on performance calculations becomes apparent from the performance increase reported in all endpoint parameters when applying NBL. Table 6 displays the confusion matrices for the OSA severity of PAT HSAT compared to the Expert Analysis, as well as for the Local Analysis compared to the Expert Analysis. The confusion matrices were generated for the 3% rule and the 4% rule as well as for OSA severity categorisation

with and without application of NBL, resulting in four different confusion matrices. Inspection of the confusion matrices shows that almost all misclassifications occur between adjacent severity categories, with very limited cases where normal OSA is classified as severe or vice versa.

		PAT HSAT I Local PSG – 3% Rule				**PAT HSAT** I Local PSG – 4% Rule			
		No	Mild	Mod.	Severe	No	Mild	Mod.	Severe
Expert PSG I NBL off	No	**16** I 16	**4** I 5	**2** I 1	**0** I 0	**35** I 37	**9** I 3	**0** I 4	**0** I 0
	Mild	**9** I 5	**16** I 24	**12** I 8	**0** I 0	**5** I 3	**21** I 23	**7** I 7	**0** I 0
	Mod.	**1** I 0	**8** I 6	**16** I 20	**8** I 7	**1** I 1	**3** I 4	**26** I 23	**4** I 6
	Severe	**0** I 1	**1** I 0	**6** I 6	**68** I 68	**0** I 0	**1** I 0	**7** I 3	**48** I 53
Expert PSG I NBL on	No	**22** I 21	**2** I 2	**2** I 1	**0** I 0	**37** I 39	**0** I 1	**0** I 4	**0** I 0
	Mild	**3** I 0	**21** I 29	**8** I 5	**0** I 0	**3** I 1	**31** I 27	**4** I 4	**0** I 0
	Mod.	**1** I 0	**5** I 4	**22** I 26	**6** I 4	**1** I 1	**2** I 2	**32** I 29	**4** I 6
	Severe	**0** I 1	**1** I 0	**4** I 3	**70** I 71	**0** I 0	**1** I 0	**4** I 0	**48** I 53

Table 6: OSA severity category confusion matrix. This table displays the confusion matrix comparing the OSA severity category as determined by the Expert PSG Analysis to respectively PAT HSAT (bold font) and Local PSG Analysis (normal font) for the 3% rule (L) and the 4% rule (R), with NBL (Bottom) and without NBL (Top). Mod. = moderate sleep apnea; NBL = near-boundary double-labelling.

Secondary endpoint parameters

AHI

Figure 11, Figure 12, and Table 7 show the results of a Bland-Altman and correlation analysis for both scoring rules. A significantly higher ICC(A,1) was found for PAT HSAT compared to the Local PSG for the 4% rule. Removing AHIs larger than 30 significantly reduced the ICC(A,1) for both PAT HSAT and Local Analysis, which highlights the influence of extreme values on this parameter. The width of the limits of agreement significantly reduced after removing AHIs larger than 30.

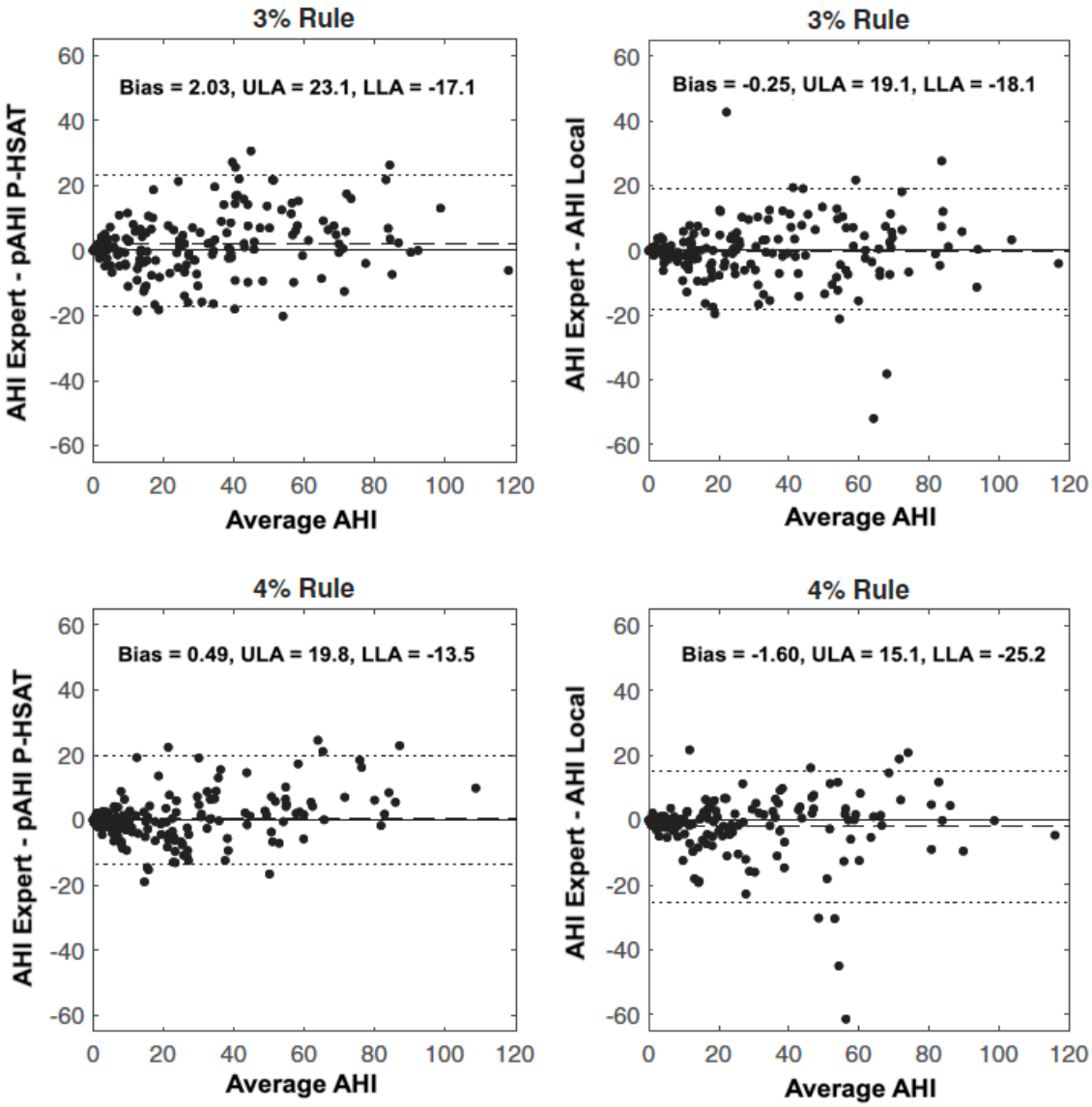

Figure 11: Bland-Altman plots of AHI determined by PAT HSAT (L) and Local Analysis (R) versus Expert Analysis for the 3% rule (Top) and the 4% rule (Bottom). The dotted line represents the mean Limits of Agreement. The dashed line represents the mean difference or bias. Average AHI = average of AHI Expert and pAHI PAT HSAT (L) or AHI Expert and AHI Local (R); LLA = lower limit of agreement; pAHI = AHI estimated by peripheral arterial tonometry; ULA = upper limit of agreement.

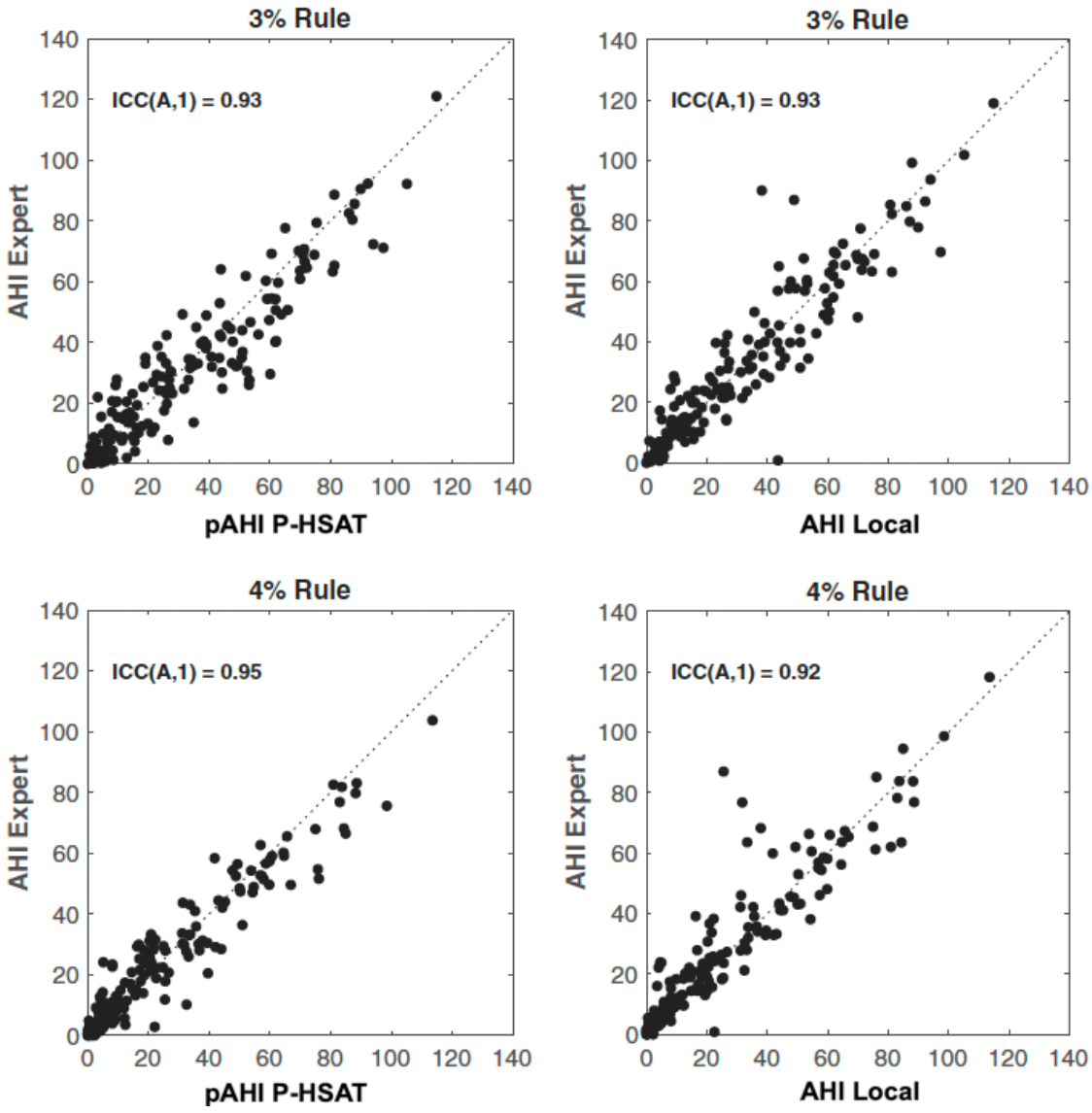

Figure 12: Scatter plots of AHI determined by PAT HSAT (L) and Local Analysis (R) versus Expert Analysis for the 3% rule (Top) and the 4% rule (Bottom). The dotted line represents the points for which the y-axis values equal the x-axis values of the graph (identity line). ICC(A,1) = Intraclass correlation coefficient of type (A,1); pAHI = AHI estimated by peripheral arterial tonometry.

		3% Rule		4% Rule	
		PAT HSAT	Local PSG	PAT HSAT	Local PSG
All AHIs	Bias	2.03	-0.35	0.49	-1.60
	ULA	23.1	19.1	19.8	15.1
	LLA	-17.1	-18.1	-13.5	-25.2
	ICC(A,1)	0.93 (0.90 – 0.95)	0.93 (0.91 – 0.95)	0.95 (0.94 – 0.96)*	0.93 (0.90 – 0.95)
AHIs < 30	Bias	-1.29	-1.81	-1.77	-2.40
	ULA	10.9	8.6	8.5	6.6
	LLA	-17.0	-16.8	-14.1	-19.1
	ICC(A,1)	0.74 (0.63 – 0.82)	0.79 (0.69 – 0.86)	0.81 (0.72 – 0.87)*	0.68 (0.55 – 0.77)

Table 7: Summary of Bland-Altman and correlation analysis of AHI. This table displays the bias, ULA, LLA, and ICC(A,1) from the Bland-Altman and correlation analysis comparing the pAHI of PAT HSAT and the AHI of the Local PSG Analysis to the Expert PSG Analysis. When confidence intervals were calculated, they are displayed between brackets. An asterisk signifies a significant outperformance of the endpoint parameter (within the same scoring rule selection). These parameters are calculated for the 3% rule and the 4% rule and for all AHIs as well as only AHIs < 30 (as determined by Expert PSG Analysis). AHI = apnea-hypopnea index; ICC(A,1) = intraclass correlation coefficient of type (A,1); LLA = lower limits of agreement; pAHI = AHI estimated by peripheral arterial tonometry; ULA = upper limits of agreement.

TST

Figure 13 shows the Bland-Altman and correlation analysis for the TST estimate. A significantly lower correlation was found for PAT HSAT, as well as a wider distance between the limits of agreement.

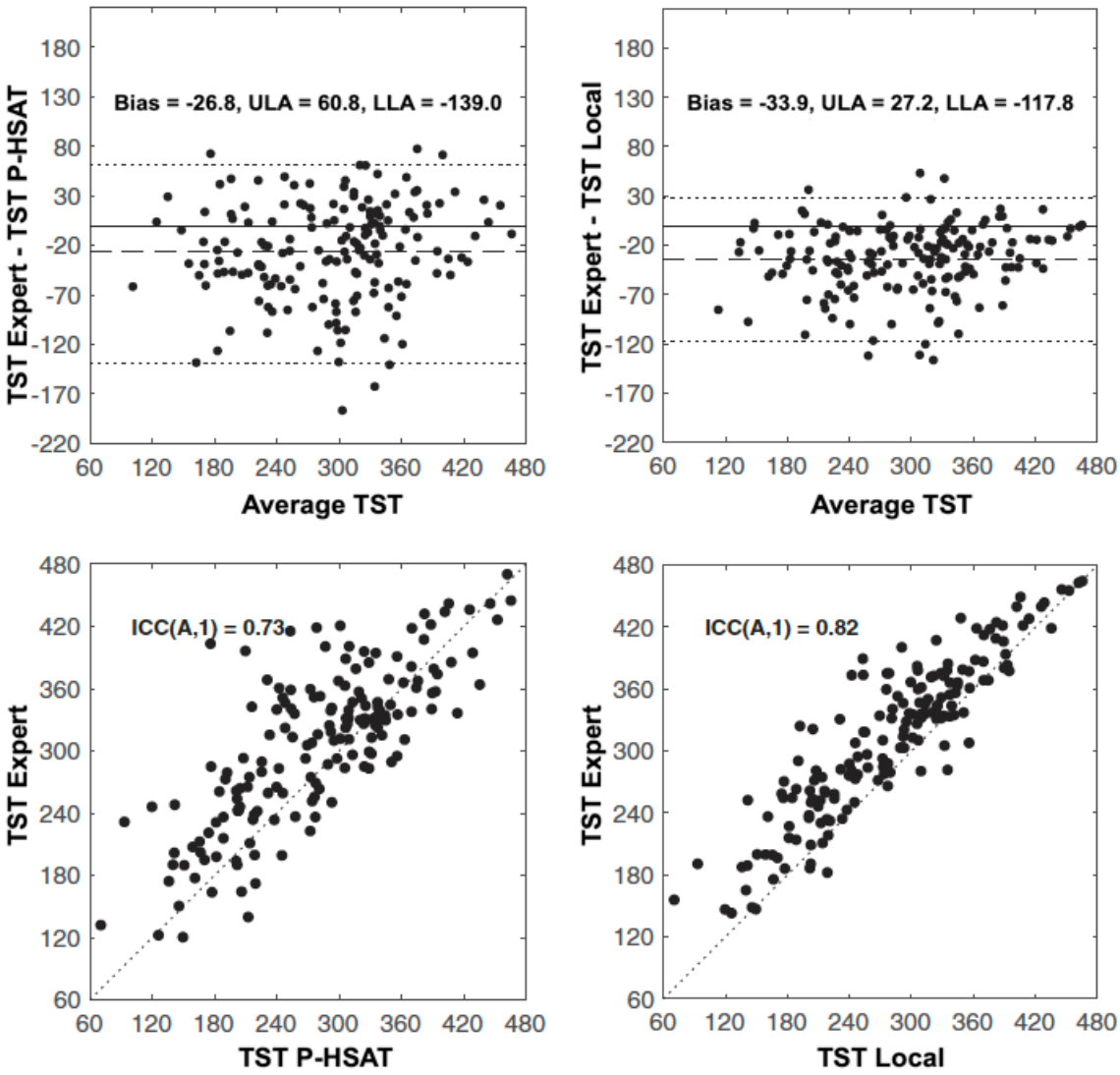

Figure 13: Bland-Altman(Top) and Scatter (Bottom) plots of TST determined by PAT HSAT (L) and Local Analysis (R) versus Expert Analysis. Top: The dotted line represents the mean Limits of Agreement. The dashed line represents the mean difference or bias. Bottom: The dotted line represents the points for which the y-axis values equal the x-axis values of the graph (identity line). Average TST = average of TST Expert and TST PAT HSAT (L) or TST Expert and TST Local (R); ICC(A,1) = Intraclass correlation coefficient of type (A,1); LLA = lower limit of agreement; ULA = upper limit of agreement.

Discussion

Endpoint analysis highlights and discussion

For both the 3% and the 4% rule, most primary endpoint parameters showed substantial agreement with PSG. The disparity between PAT HSAT and the Local Analysis' performance was the smallest for the 4% rule. This is unsurprising since the 4% rule differs from the 3% rule in that the latter does not consider arousals. The scoring of cortical arousals suffers from significant inter-scorer variability[76], trickling through to a larger inter-scorer variability for hypopnea scoring. The NPV at AHI cutoff 5 of 62% for PAT HSAT and 73% for the Local PSG Analysis (using the 3% rule) highlights a tendency of both to underscore OSA. However, only in one instance (and only for the 3% rule) was a patient classified as normal by PAT HSAT and as having moderate OSA by the Expert Analysis. In no instance was a patient classified as normal by PAT HSAT as having severe OSA by the Expert Analysis.

A significantly lower ICC(A,1) as well as significantly wider LoA interval was found for the total sleep time estimate of PAT HSAT compared to the Local PSG Analysis. This confirms the previously reported underperformance of PAT HSAT compared to PSG[21,36] for estimating sleep time. This is unsurprising as PAT HSAT makes use of actigraphy, which is merely an approximation of true sleep time as estimated by EEG. Therefore, PAT HSAT cannot be considered a valid substitute for PSG as it pertains to the assessment of sleep (stages).

Finally, the strong increase in most endpoint parameter values when applying NBL further illustrates the significant impact of near-boundary AHIs. These findings warrant further discussion on whether patients with an AHI in NBZ would benefit from further diagnostic evaluation to increase confidence in therapy decision-making.

Strengths and limitations of the study

The first strength of this study is its adequately powered multicentric design, which incorporates centres located in Europe and the United States. The second strength is its unique approach in double labelling of PSG so as to allow for the comparison of the agreement of PAT HSAT with PSG to the inter-rater agreement of PSG. The third strength lies in its critical discussion of clinical endpoint parameters. The fourth strength of the study is its unique approach to dealing with spurious endpoint parameter variability caused by reference AHIs close to OSA severity boundaries.

A limitation of the study is the lack of assessment of the impact of multi-night testing on the endpoint parameters, as PAT HSAT is typically administered for multiple nights. Inter-night variability has been shown to be a significant contributor to diagnostic errors[18].

Conclusion

This multicentric validation study of the PAT HSAT was designed to robustly assess whether the device can adequately navigate the patient through the diagnostic pathway, i.e., whether it can adequately determine the OSA severity.

For both the 3% and the 4% rules, most endpoint parameters showed substantial agreement with PSG when compared with the inter-rater variability of PSG. The 4-way categorisation accuracy of PAT HSAT was nevertheless lower than the inter-rater agreement of PSG (70% vs 77% for the 3% rule and 78% vs 81% for the 4% rule).

The unique cornerstones of the study design are the double-labelling of PSG so as to establish a performance target for HSAT, adequate treatment of AHIs close to the OSA severity category boundaries, and the avoidance of reliance on potentially misleading clinical endpoint parameters such as Pearson and Spearman correlation coefficients[30]. Replication of these design cornerstones

increases transparency of clinical endpoints and can enable more generalisation and standardisation of future HSAT validation studies.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry.
Publication status:	*Published*
Publication details:	Van Pee B, Massie F, Vits S, Dreesen P, Klerkx S, Bijwadia J, Verbraecken J, Bergmann J. A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry. Sleep. 2022 May 12;45(5):zsac028. doi: 10.1093/sleep/zsac028.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed PAT HSAT algorithm development and evaluation, clinical study design and writeup, and publication design and writeup. B Van Pee and S vits performed clinical study data acquisition, PAT HSAT algorithm evaluation, and publication editing. P Dreesen, S Klerkx, and J Bijwadia supported clinical study data acquisition. J Verbraecken provided publication editing. J Bergmann provided supervisory guidance and publication editing. F Massie, Van Pee B, and S Vits are shared first authors.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Chapter 3: Enabling PAT HSAT to differentiate between CSA and OSA

Research Aim 2.1: Investigating a PPG-based approach for differentiating obstructive and central sleep apnea

Published as: Massie F, Vits S, Khachatryan A, Van Pee B, Verbraecken J, Bergmann J. Central Sleep Apnea Detection by Means of Finger Photoplethysmography. *IEEE Journal of Translational Engineering in Health and Medicine*. 2023 Jan 12;11:126-136. doi: 10.1109/JTEHM.2023.3236393.

Reiteration of research aim and introductory context

As introduced in Chapter 1, both obstructive and central events are characterised by airflow reduction. However, central events are accompanied by a cessation of respiratory effort as traditionally inferred from the abdominal and thoracic respiratory effort belts of PSG, which monitor respiratory-related fluctuations in the abdominal and thoracic circumference[34]. PAT HSAT, when using only a finger PPG probe for respiratory event detection, does not have airflow and respiratory effort channels that can be used for respiratory event-type labelling. As such, PAT HSAT (without adding additional sensors) cannot differentiate between the two types of sleep apnea. Since OSA and CSA are two types of sleep apnea with very distinct aetiologies and therapy options, such differentiation is of high clinical relevance[35]. The article included in this chapter addresses this limitation. It describes the first end-to-end physiological data processing pipeline for central sleep apnea detection based solely on finger PPG, for which adequate clinical performance could be demonstrated using representative clinical data.

Abstract

Study objectives

Obstructive sleep apnea and central sleep apnea are two types of sleep apnea with different aetiologies and treatment options. PAT HSAT has become one of the most widely deployed outpatient sleep apnea diagnostic methods. Being able to differentiate between CSA and OSA based solely on photoplethysmography data would further increase PAT HSAT's clinical utility. The present work proposes a method to detect CSA using finger PPG data and evaluates the proposed method against simultaneous in-lab PSG.

Methods

For 266 patients with suspected sleep apnea, concurrent in-lab PSG and PPG data were acquired. The respiratory information embedded in the PPG data were extracted and used to train an *ensemble of trees* classifiers that predicts the central or obstructive nature of each respiratory event. The classifier performance was evaluated using patient-wise leave-one-out cross-validation, where an expert analysis of PSG served as ground truth. A second, independent analysis of PSG was also evaluated against the ground truth to allow benchmarking of the PPG-based method.

Results

The method achieved a sensitivity of 81%, a specificity of 99%, a positive predictive value of 90%, and a negative predictive value of 98% at the central apnea-hypopnea index cutoff of 10 events per hour of sleep.

Conclusions

The present study aimed to evaluate a method for detecting CSA in sleep apnea patients using only PPG data. Such a method could be used to flag patients with possible CSA and thereby support more optimal therapy decision-making.

Introduction

Sleep apnea is a common sleep disorder affecting millions of people worldwide[5,77]. An apnea event is defined as a cessation of airflow during sleep lasting 10 seconds or more, whereas a hypopnea event is characterised by an airflow reduction rather than a full cessation[78,79]. A patient's sleep apnea severity can be expressed by their AHI, which is simply the number of apnea and hypopnea events per hour of sleep[34,80,81].

Sleep apnea can be obstructive (OSA) or central (CSA) in nature. OSA is the more common form of sleep apnea, and some of its symptoms and consequences are fatigue, daytime sleepiness, cardiac arrhythmia, and systemic hypertension[79]. During OSA, breathing effort continues, but the upper airway is mechanically obstructed, resulting in interruptions of airflow. CSA, on the other hand, is characterised by a lack of neural drive to breathe during sleep[82,83]. Both types of respiratory events can coexist in one patient. A respiratory event can also be of mixed origin, characterised by a lack of respiratory drive followed by obstructive breathing. Typically, CSA is the primary diagnosis when at least 50% of respiratory events are scored as central in origin. CSA can be characterised by the cAHI (central AHI), which is calculated as the number of central respiratory events per hour of sleep.

Whereas the symptoms of CSA are often similar to those of OSA, the choice of therapy depends on the type of sleep apnea. Treatment of OSA includes lifestyle measures (for instance, weight loss), mandibular advancement devices, surgical procedures, and, most commonly, CPAP[84,85]. Comparably, a wide variety of therapies for CSA exist, of which some are distinctly different from

therapies for OSA. These methods include drug intervention, oxygen therapy, nocturnal mechanical ventilation (via nasal mask or tracheostomy and tracheal tube), and diaphragm pacing[35,79]. As such, for clinicians to make optimal therapy decisions, it is important to differentiate between CSA and OSA in sleep apnea patients.

The gold standard for diagnosing sleep apnea is PSG[86,87]. It utilises electroencephalography, electro-oculography, electromyography, electrocardiography, and pulse oximetry, as well as airflow and respiratory effort to assess underlying causes of sleep disturbances[88]. However, due to the inconvenience of performing an in-lab PSG and in part due to the global COVID pandemic[24], PAT HSAT has rapidly gained popularity and currently comprises one of the most widely deployed types of HSAT[26,27]. PAT HSAT allows for minimally obtrusive multi-night testing and is available in a fully disposable format[50]. This study makes use of the FDA-cleared NightOwl PAT HSAT (study device)[21]. It comprises a finger probe of the size of a fingertip and a cloud-based analysis software (Figure 14).

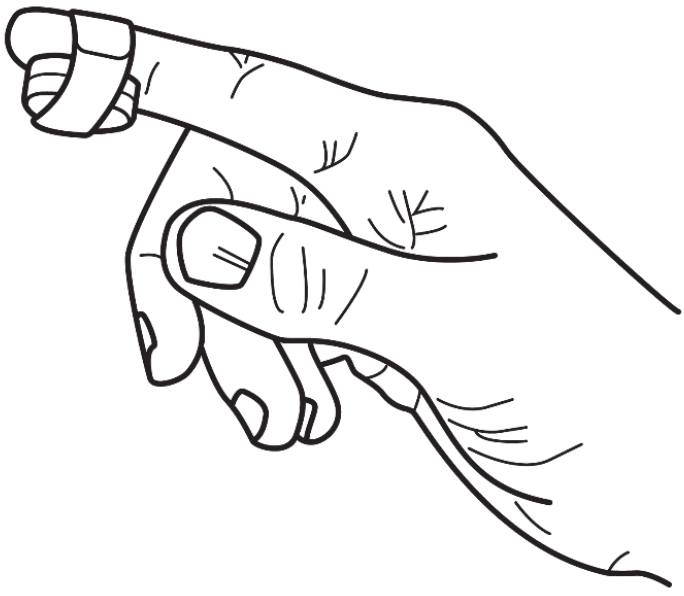

Figure 14: Illustration of the study device.

Compared to PSG, PAT HSATs have a major drawback, namely, a reduced-channel design. Indeed, a PAT-based HSAT obtains most of its modalities from finger PPG, from which it derives SpO_2, PR, and PAT. Inherently, this means that PAT HSATs do not have access to information sources such as airflow reduction and cortical arousals and must instead rely on alternative sources of information to infer the presence of respiratory events[89].

One more challenge for PAT HSAT is the differentiation between the types of respiratory events. Both types of events are characterised by airflow reduction. However, central respiratory events are accompanied by a cessation of respiratory effort as traditionally inferred from the abdominal and thoracic respiratory effort belts of PSG (which monitor respiratory-related fluctuations in the abdominal and thoracic circumference)[34]. Figure 15 shows a comparison of the respiratory effort during a central and obstructive respiratory event.

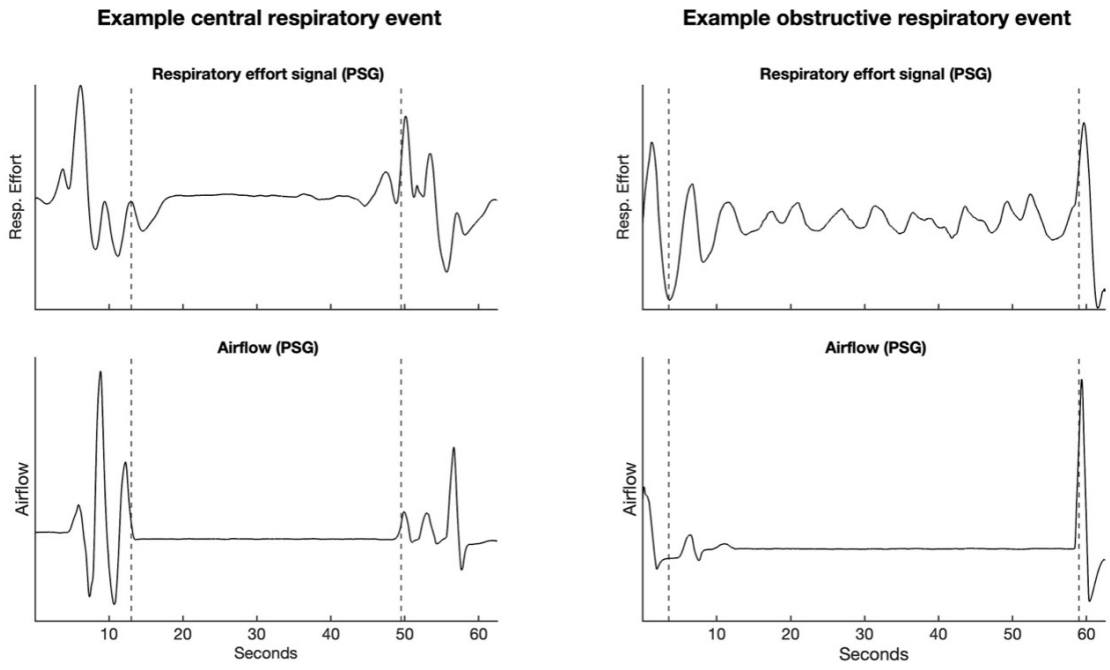

Figure 15: Comparison of a central and obstructive respiratory event. Both events are accompanied by cessation of airflow, but a central event is characterised by a lack of respiratory effort, which is reflected in a flat region of the respiratory effort belt signal.

The study device relies solely on a finger PPG probe for respiratory event detection and therefore lacks airflow and respiratory effort channels that would allow direct identification of the respiratory event type. The present study aims to evaluate a method for detecting CSA using only the PPG data acquired from the finger by the device. A validated PPG-based CSA detection method could be used to flag patients with suspected CSA and facilitate referral for an in-lab PSG to determine CSA severity and subtype.

The key idea behind the method is to extract and analyse respiratory-effort-related information from the finger PPG data, thereby deriving a signal that resembles PSG-based respiratory effort as closely as possible. This extracted breathing signal is then used to extract features to train and validate an *ensemble of trees* classifier that attempts to learn how to differentiate between the two types of respiratory events.

To our knowledge, the present work is the first evaluation of an end-to-end approach for CSA detection based solely on finger PPG data with patient-wise performance evaluation.

Related works

Recently, Pillar et al. published a validation study of their novel CSA detection method using the WatchPAT PAT HSAT[64] with extra sensors placed on the chest. In addition to finger PPG, the system comprises an optional snoring and body position (SBP) sensor positioned under the sternal notch. The SBP sensor is used to derive upper chest wall respiratory movements that are diminished during central respiratory events. According to Pillar et al.[64], the PPG's systolic upstroke shows more variability during obstructive respiratory events compared to central ones. The method uses both the upstroke variability and the respiratory effort data acquired by the SBP sensor to identify central events.

Lazazzera et al. also published a method to detect and classify respiratory events into apnea and hypopnea and their origin (central or obstructive) using PPG and SpO_2 signals[90]. Their method uses PPG features, SpO_2 features, pulse rate time domain features, and pulse rate frequency domain features to classify respiratory events. However, they do not report on any patient-wise performance analysis, nor do they report on the cAHI estimation accuracy. As such, their method's potential for clinical deployment remains unclear.

Methods

PAT HSAT and its principle of operation

PAT HSAT makes use of a fingertip-mounted PPG probe[21]. PPG operates on simple optical technology to detect blood volume changes in the tissue's microvascular bed[20,91]. PPG measurements are used to derive SpO_2, PR, and changes in PAT, which are then used to detect respiratory events[21]. Peripheral arterial tone refers to the tone of the peripheral arterial smooth muscle tissue. When the muscle tone of peripheral arteries increases, the arteries' diameter decreases, resulting in a reduction of perfusion and thus a decrease in pulsatile blood volume in the peripheral tissue. The latter is picked up as a drop in the PPG signal swing between systole and diastole. The PPG-derived signal, trending such pulsatile blood volume reductions, is referred to as the PAT signal. A PAT HSAT analyses the concurrence of drops in SpO_2, surges in pulse rate, and increases in peripheral arterial tone. As airflow is reduced, the oxygen supply to the lungs decreases from baseline, resulting in distinct SpO_2 drops. Near the end of a respiratory event, the sympathetic nervous system activity surges to arouse the patient, resulting in resumption of ventilation, a surge in pulse rate, and a release of norepinephrine in the bloodstream. Consequently, norepinephrine binds to alpha-adrenergic receptors innervating the arterial smooth muscle tissue in the finger, causing sudden vasoconstrictions, and corresponding increases in PAT, picked up by the PPG sensor[21,92].

The same PPG sensor of the study device used to detect respiratory events is used in this study to extract respiratory effort-related information from which we attempt to differentiate between central and obstructive respiratory events.

The dataset

The dataset used in this study comprises 266 patients with suspicion of sleep apnea who were prospectively recruited across four different sleep clinic centres of which three were located in the USA (where all centres were part of the United Health Systems Group in Miami, Florida) and one in Belgium (Ziekenhuis Oost Limburg, ZOL, Genk). The first patient was recruited on the 4th of July 2017, and recruitment was concluded on the 25th of January 2020. The USA branch of the study was approved by Aspire IRB, part of the WIRB-Copernicus Group. The Belgian branch of the study was approved by the Ethics Committee of ZOL.

All patients were scheduled for one overnight in-lab PSG. The patient cohort was not stratified for CSA prevalence to retain the normal cohort prevalence of CSA. Non-adult and/or cognitively impaired patients were excluded. As the cohort matches the intended target population of the study device, this leads to the most representative prevalence-dependent performance endpoints such as NPV and PPV.

Qualified lab technicians at each participating study centre were responsible for setting up the equipment and capturing PSG data. During the setup of PSG, the PAT HSAT was attached to the middle finger of the hand to which the pulse oximeter of PSG was applied.

All PSG data were double scored by two independent centres, which were blinded from one another's analysis. The first scoring was performed by the team of sleep technicians of the centre where the patient was admitted (Local Analysis). Another independent scoring was performed by scorers of Cerebra Medical (CM, Canada). The studies were first analysed by their Michele Sleep

Scoring System (MSSS) and were subsequently complemented with complete manual rescoring by an expert technologist. Malhotra et al.[32] confirmed in a multicentric trial that the MSSS, complemented with manual editing by an expert scorer, is more robust than the results of a single scorer. Because of this conclusion, CM's analysis served as the expert benchmark (Expert Analysis) to which the Local Analysis and our PPG-based method were compared. All PSG data were scored according to the 2012 AASM scoring rules using the Recommended 1A (3%) rule for hypopnea scoring[1].

Working principles of the CSA detection method

Respiratory event extraction from raw PPG

The PAT HSAT uses PPG and PPG-derived signals such as PR and SpO_2 to detect respiratory events. At this point, no distinction has been made between obstructive and central events. The labelling of respiratory events as central or obstructive for the CSA detection model training is performed using PSG annotations.

Respiratory signal extraction from raw PPG

The PPG signal measured at the fingertip inherently contains respiratory information because the blood flow to body extremities is affected by alterations in thoracic pressure throughout the respiratory cycle[93]. Therefore, the PPG signal amplitude oscillates in synchrony with the respiratory cycle. This amplitude modulation can be isolated to retain a signal representing respiratory effort. Figure 16 shows the PPG signal along with the respiratory modulations present in the signal. Our method utilises the presence or absence of these breathing modulations in the finger PPG to infer the type of respiratory event. The respiratory modulation was extracted from the raw PPG signal using the steps described in this section. The raw PPG signal was high-pass-filtered at 0.15 Hz to

remove the slowly varying baseline modulations. The breathing frequency at rest is usually 0.2 Hz or higher (the normal breathing frequency range for an adult is 0.2-0.33 Hz) and, therefore, the breathing modulations remain substantially unaffected by such filtering. Subsequently, the peak envelope of the filtered PPG signal was calculated, which was further normalised by subtracting a 6-second moving average of this envelope and dividing it by a 30-second moving interquartile range operation. The 6-second window was chosen to remove signal baseline fluctuations while retaining breathing modulations, whereas the 30-second window was chosen to normalise any steady-state changes in the signal amplitude. The resulting signal, the PPG-derived respiratory effort (PPGDR) signal, was used for further analysis. Figure 17 compares a segment of the PPGDR to the corresponding respiratory effort segment from the thoracic effort belt of PSG.

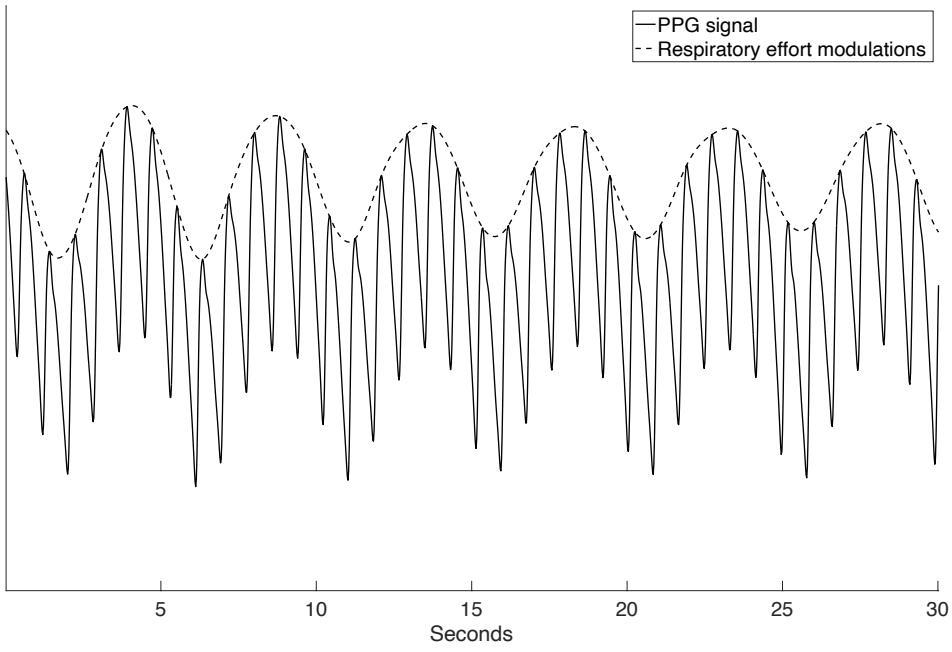

Figure 16: Example segment of PPG signal with the envelope reflecting respiratory effort modulations.

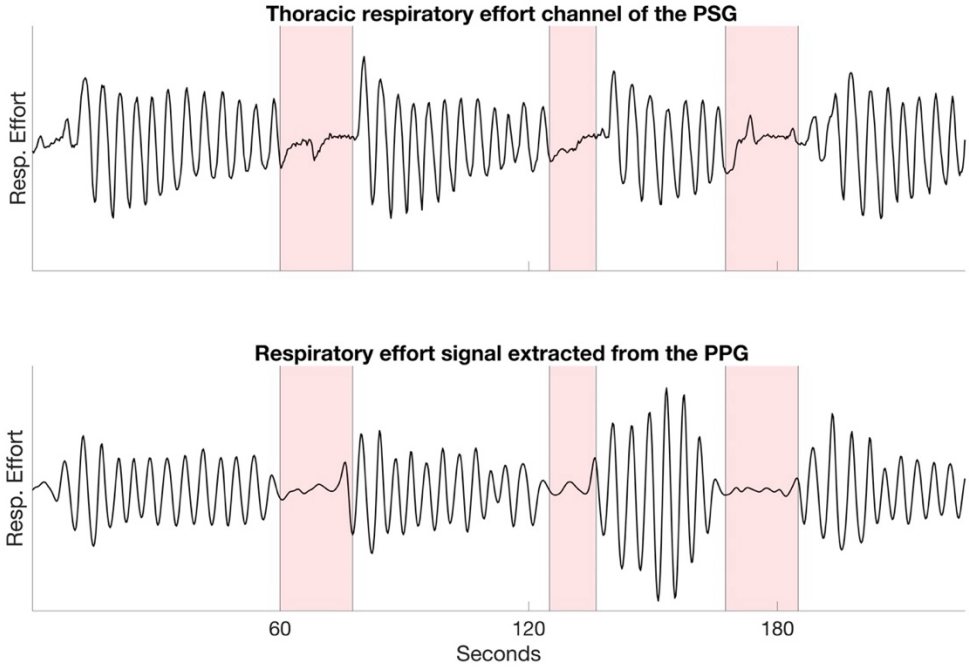

Figure 17: Comparison of the respiratory effort signal extracted from the finger PPG with the respiratory effort channel of PSG. Zones highlighted in pink indicate flat respiratory effort zones.

Feature calculation

After PPGDR signal extraction, features were extracted from the PPGDR using a sliding-window operation. Two odd-numbered window sizes were used: 11 and 21 seconds. The window shifted over the PPGDR signal with a stride of 1 second, and the feature values calculated for each window were ascribed to the one-second interval at the centre of the window.

The features computed for each window were the range, the interquartile range, the variance, and the number of PPGDR local maxima (peaks). The range of the signal was additionally computed for window sizes of 5 and 7 seconds.

For local maxima-related features, local maxima detection was performed on the PPGDR (minimum peak-to-peak distance was set to 250 milliseconds), where the peak-to-peak interval represents the duration of a breathing cycle. A lack of local maxima suggests a lack of breathing effort.

In the next step, features were ascribed to each respiratory event detected by the PAT HSAT. Successful feature design must account for the variability in duration and location of the flat zones in the PPGDR during a respiratory event. As such, for each respiratory event detected by the PAT HSAT, the moving-window features for which the centre-second overlapped with the respiratory event period were selected. From these selected features, the minimum value, the average of the three lowest values, and the 10th and 25th percentiles of the feature values were calculated.

For local maxima-related features, the aggregation operations comprised computing the minimum value and the numbers and proportions of windows with zero, one, and two peaks. The resulting aggregated feature sets for each PAT HSAT respiratory event were used to support subsequent steps.

Labelling respiratory events

For model training purposes, each respiratory event inferred by the PAT HSAT was labelled using the Expert Analysis's event type labels (central vs. obstructive). Since for model training, we only wanted to retain respiratory events that were identified with a comparatively high level of certainty, respiratory events identified by the Expert Analysis that were not also identified by Local Analysis were removed. The respiratory events for which there was consensus by the Expert and Local Analysis on their presence (but not necessarily on their type) were used to label PAT HSAT events (Ground Truth Events). The PAT HSAT events that did not overlap with or were within ten seconds after any Ground Truth Event were removed from the training data. This removal of Expert Analysis or PAT HSAT events was only performed for the training sets and not for model validation. Each Expert Analysis' annotation was used at most once for PAT HSAT event labelling.

Figure 18 illustrates the labelling procedure by means of examples. If a PAT HSAT event had a uniquely overlapping Expert Analysis annotation, the PAT HSAT event was labelled according to the corresponding Expert Analysis's event type. If there was more than one annotated event

overlapping with a PAT HSAT event, the PAT event was labelled as central if any of the overlapping Expert Analysis' annotations were central events; otherwise, the respiratory event was labelled as obstructive. Since the PAT HSAT respiratory event locations may be slightly shifted in relation to Expert Analysis' annotations, any preceding Expert Analysis' annotations of up to ten seconds were matched to a PAT HSAT event if there was no other annotation overlapping with the PAT HSAT event. Each Expert Analysis' annotation was used at most once for PAT HSAT event labelling. Out of a total of 44,420 PAT HSAT events, 26,208 had a corresponding Expert Analysis annotation and were labelled according to the above-described procedure. Out of the 26,208 labelled PAT HSAT events, 2,881 were central (or mixed) and 23,327 were obstructive.

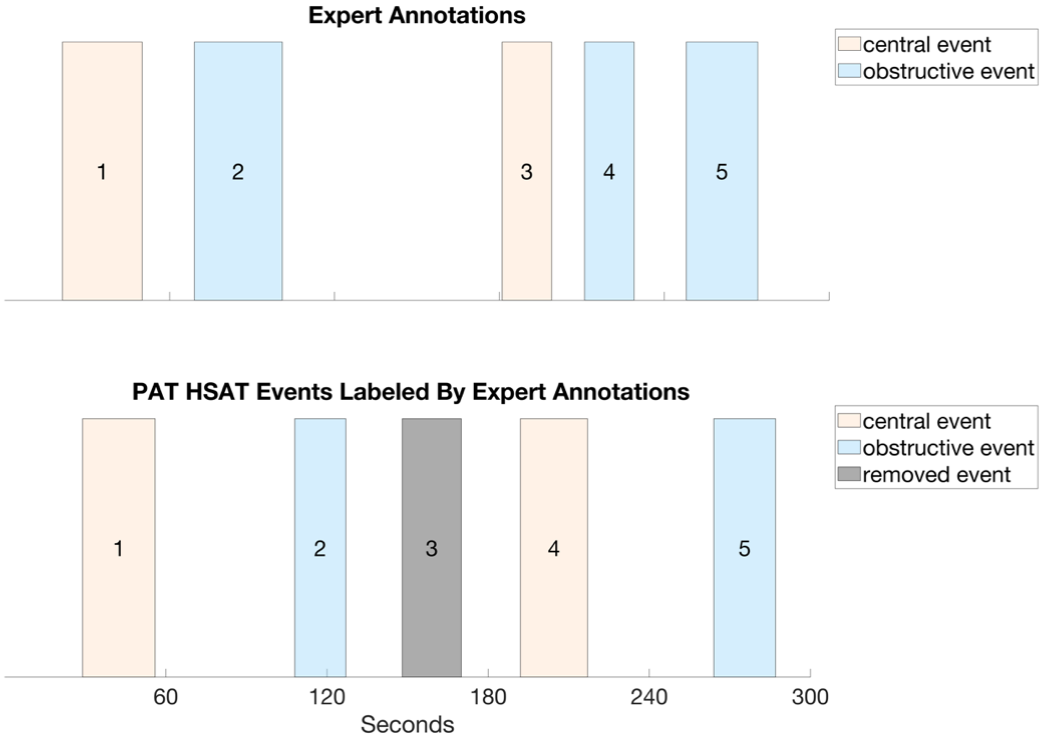

Figure 18: The procedure used to label PAT HSAT events. PAT HSAT event 1 is labelled as central because it has a uniquely overlapping central Expert annotation. PAT HSAT event 2 is labelled as obstructive because it has an Expert obstructive annotation preceding it by not more than 10 seconds. PAT HSAT event 3 is removed because it cannot be labelled using the Expert annotations. This event will not be used for training. PAT HSAT event 4 has two overlapping annotations, but since at least one of them is central, it is labelled as central. PAT HSAT event 5 is labelled as obstructive because it has a single overlapping obstructive annotation.

Since an airflow channel is necessary to mark the exact start and end of a respiratory event, a PAT HSAT has an imperfect delimitation of respiratory events. As such, it is not readily possible to infer via a PAT HSAT whether an episode without breathing effort followed by an episode with breathing effort corresponds to a central or a mixed event. For that reason, the Expert Analysis labels were binarised into *central* and *obstructive*, considering mixed events as *central*. The same approach is used for the Local Analysis labels. Therefore, the reported cAHI of our method is an estimate of the sum of central and mixed respiratory events per hour of sleep.

Signal rejection

Because we use only PPG to extract the respiratory signal, the quality of our predictions is heavily dependent on the underlying PPG data quality. Strong vasoconstrictions reflected in sudden drops in PPG pulse amplitude frequently contain low signal-to-noise ratio episodes and do not get sufficiently cancelled out by normalisation operations. This reduces PPGDR quality. We recognise vasoconstrictions by identifying prominent troughs in a filtered version of the PPG signal. The filtering is performed to remove baseline shifts and shifts in perfusion that might otherwise impact trough detection. Figure 19 shows an example of such a vasoconstriction-affected zone. PPGDR extracted from such vasoconstriction-affected zones was rejected. That is, if an event had a vasoconstriction-affected zone, features from that zone were not used for the compilation of the aggregated feature set for this event.

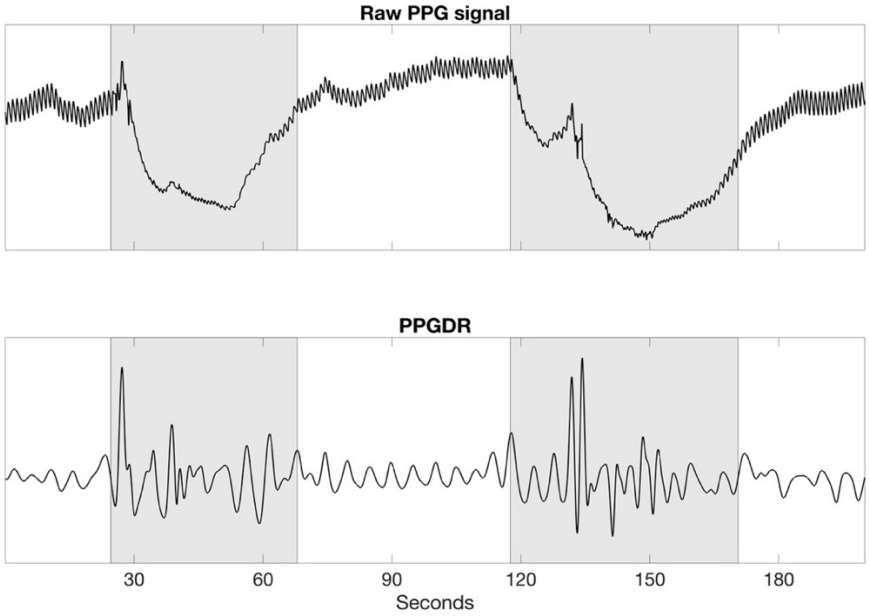

Figure 19: PPG signal and PPGDR of two vasoconstriction-affected zones (visualised by grey areas). The PPGDR extracted from these zones is of lower quality, apparent from the comparatively more erratic swings in the PPGDR signal and is therefore rejected. PPGDR = PPG-derived respiratory effort.

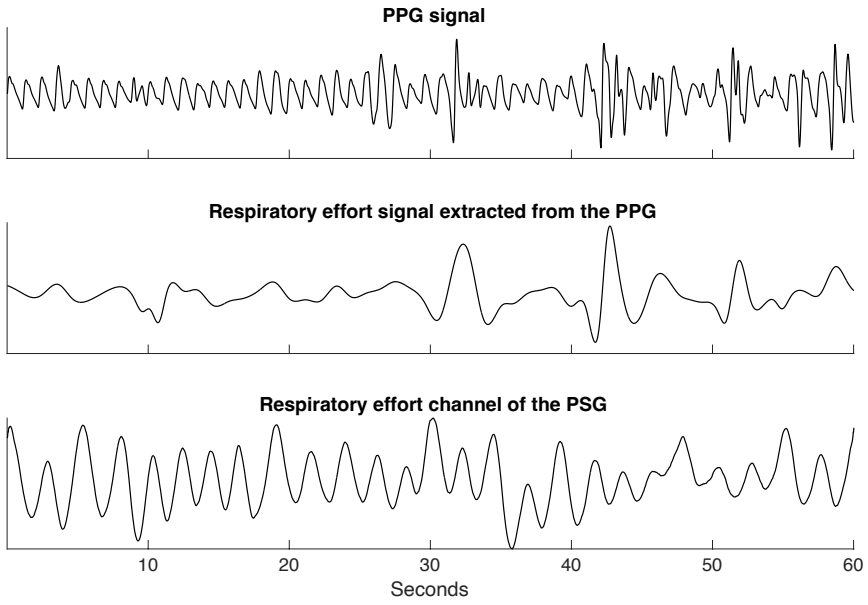

Figure 20: Example of a PPG signal with artefacts (top), PPGDR extracted from the PPG signal (middle), and the corresponding respiratory effort channel of PSG (bottom). The PPGDR extracted from this zone with artefacts does not reflect the true respiratory modulations. PPGDR = PPG-derived respiratory effort.

Event rejection

Another issue with PPGDR signal quality arises when motion or other artifacts impact the PPG quality or when severe cardiac arrhythmia are present. When the peak-to-peak intervals of the PPG are irregular, either due to artifacts or severe cardiac arrhythmia, the breathing modulations present in the PPG get overpowered by such irregularities and PPGDR extracted from these regions would show spurious oscillations unrelated to the true respiratory modulations.

To make sure that the PPGDR can be utilised for respiratory event type prediction, irregular PPG peak-to-peak interval zones were detected, and any respiratory events that had more than 20% of irregular PPG peak-to-peak intervals were rejected. Peak-to-peak intervals were labelled as irregular when the absolute value of the difference of two neighbouring peak-to-peak intervals was higher than 20% of the average of those neighbouring intervals. Figure 20 shows a segment of a rejected PPG signal and the corresponding PPGDR compared against the respiratory effort channel of the PSG. The proportion of rejected respiratory events for each recording was called the Event Rejection Proportion (ERP).

Patient rejection

Predictions made for patients that have a substantial number of rejected events would be unreliable due to the small proportion of retained events used for cAHI estimation. Therefore, patients with an ERP higher than 0.3 were rejected and removed from further analysis.

cAHI prediction and evaluation

Unlike for training, for evaluation purposes, all 44,420 PAT HSAT events were retained regardless of the existence of a corresponding PSG annotation. To maximise utilisation of the strongly imbalanced dataset caused by the low prevalence of CSA, training and evaluation were performed

via patient-wise leave-one-out cross-validation. For each patient, an *ensemble of trees* classifier (ensemble method: bagging, false positive cost: 2.5, number of trees: 30, max. number of splits: 150) was trained on the labelled data of the remaining patients. The model was insensitive to a wide range of tree and split numbers, alleviating the need for extensive hyperparameter optimisation. The false positive cost parameter of the classifier was used to tune the event-wise sensitivity and specificity of the predictions, used in this case to prevent over-detection of CSA. Using this classifier, predictions were made on the non-rejected PAT HSAT events of the considered patient. The predicted cAHI was calculated as the number of central events as predicted by the model, divided by the PAT HSAT TST estimate. Since predictions were made only with non-rejected events, the predicted cAHI was extrapolated by dividing it by one minus the ERP to account for the total number of respiratory events (rejected and non-rejected) of the patient.

Performance of the method was evaluated for cAHI cutoffs of 5, 10, and 15. The evaluation was performed on a patient-wise basis by flagging a patient's sleep apnea as central if the cAHI was above the given cutoff value.

Statistical analysis

General

Statistical analysis was performed using MATLAB (version 2020a, MathWorks, USA). Patient demographic information was obtained during the clinical study and is reported in the results section. For each of the cAHI cutoffs, cohort-wise sensitivity, specificity, positive and negative Likelihood Ratios (LR+, LR-), NPV, PPV, accuracy, and Cohen's Kappa values were computed, and a Receiver Operating Characteristic (ROC) and Precision Recall curve was generated. The ROC and Precision Recall curves were generated by obtaining classifier probability scores for each respiratory event and varying the probability cutoff threshold for binary classification. For all sensitivity, specificity,

LRs, PPV, NPV, accuracy, and Cohen's Kappa endpoint parameters, 95% confidence intervals were computed. The classifier confusion matrix was constructed for four cAHI categories: [cAHI < 5], [5 ≤ cAHI < 10], [10 ≤ cAHI < 15], and cAHI ≥ 15. Statistical significance was defined as $p \leq 0.05$.

Where possible, the same analysis was repeated to compare the Local Analysis to the Expert Analysis. To estimate the importance of each feature, we summed changes in the risk score due to tree splits on every feature and divided the sum by the number of branch nodes.

Endpoint analysis

Since there are no standardised performance targets established for cAHI prediction, the Local Analysis's performance against the Expert Analysis serves as a performance benchmark. In line with Pillar et al.[64], cAHI cutoffs of 10, and 15 were used for performance analysis. AHI cutoffs of 5 and 30 are also standard in the literature; therefore, performance analysis was also reported for the cAHI cutoff of 5. Because only six patients had a cAHI ≥ 30, this cutoff was excluded from the analysis, as the sample size was too small to support generalisable conclusions.

The endpoints used for performance assessment were the sensitivity, specificity, LRs, NPV, PPV, accuracy, and Cohen's Kappa of the PPG-based method at the examined cAHI cutoff values.

Results

Patient rejection

After removing patients with an ERP above the 0.3 cutoff from further analysis, 245 of the original 266 patients were retained. All statistical analyses and evaluations were performed on this final set of accepted patients.

Demographic information

Patients were predominantly male (60%), of middle age (mean 53.9 years, SD 13.7), and overweight (mean BMI 29.9 kg/m2 and SD 5.9). The mean AHI was 31.1 (SD 24.8). More demographic information can be found in Table 8.

	Mean	SD	Min	Max
Age	53.9	13.7	21.0	84.0
AHI	31.1	24.8	0.0	117.3
cAHI	3.2	7.5	0.0	45.5
BMI	29.9	5.9	18.2	53.8

Table 8: Demographic and clinical characteristics of patients in the dataset. SD = standard deviation.

To investigate rejection bias related to skin pigmentation, which is a known source of PPG signal quality deterioration[94], the patient population was divided into two skin pigmentation categories. Category 1 consists of 186 patients with light skin pigmentation (Fitzpatrick Scale below III), while category 2 consists of 79 patients with darker skin pigmentation (Fitzpatrick Scale of III and above). For one patient, the skin pigmentation information was not registered. The rejection ratios for categories 1 and 2 were 8.1% and 7.6%, respectively. As such, no clear rejection bias related to skin tone was found. 24 patients had no clinical sleep apnea, 56 patients had mild sleep apnea, 64 patients had moderate sleep apnea, and 101 patients had severe sleep apnea. 213 patients had a cAHI lower than 5, 11 patients had a cAHI between 5 and 10, 10 patients had a cAHI between 10 and 15, 5 patients had a cAHI between 15 and 30, and 6 patients had cAHI $\geq$ 30.

Cohort-wise performance

Table 9 shows the sensitivity, specificity, LRs, NPV, PPV, accuracy, and Cohen's Kappa for three cAHI cutoffs for the PPG-based method against the Expert Analysis. The same is displayed for the Local Analysis against the Expert Analysis. For cAHI cutoff 10, the method achieved a sensitivity of 81%, a specificity of 99%, an LR+ of 90.7, an LR- of 0.19, a PPV of 90%, an NPV of 98%, an accuracy of 97.6%, and a Cohen's Kappa of 0.84. Figure 21 and Figure 22, respectively, show the ROC and the Precision Recall curves for each of the cAHI cutoffs. The largest ROC area under the curve (AUC) of our method, 0.98, was found at a cAHI cutoff of 15. The largest Precision Recall AUC of 0.80 was found at a cAHI cutoff of 5.

		cAHI ≥ 5	**cAHI ≥ 10**	**cAHI ≥ 15**
Sensitivity (%)	**PPG – Expert**	**78.1 (63.8 – 92.4)**	**81.0 (64.2 – 97.8)**	**72.7 (46.4 – 99.0)**
	Local – Expert	71.9 (56.3 – 87.5)	57.1 (35.9 – 78.3)	72.7 (46.4 – 99.0)
Specificity (%)	**PPG – Expert**	**95.3 (92.5 – 98.1)**	**99.1 (97.9 – 100.0)**	**97.8 (95.9 – 99.7)**
	Local – Expert	95.3 (92.5 – 98.1)	99.1 (97.9 – 100.0)	98.3 (96.6 – 100.0)
LR+	**PPG – Expert**	**16.6 (8.8 – 31.3)**	**90.7 (22.5 – 365.9)**	**34.0 (13.3 – 87.1)**
	Local – Expert	15.3 (8.1 – 29.1)	64.0 (15.3 – 267.1)	42.5 (15.1 – 120.0)
LR-	**PPG – Expert**	**0.23 (0.12 – 0.44)**	**0.19 (0.08 – 0.46)**	**0.28 (0.11 – 0.73)**
	Local – Expert	0.30 (0.17 – 0.51)	0.43 (0.26 – 0.71)	0.28 (0.11 – 0.73)
PPV (%)	**PPG – Expert**	**71.4 (56.4 – 86.4)**	**89.5 (75.7 – 100.0)**	**61.5 (35.0 – 88.0)**
	Local – Expert	69.7 (54.0 – 85.4)	85.7 (67.4 – 100.0)	66.7 (40.0 – 93.4)

NPV (%)	**PPG – Expert**	**96.7 (94.3 – 99.1)**	**98.2 (96.5 – 99.9)**	**98.7 (97.2 – 100.0)**
	Local – Expert	95.8 (93.1 – 98.5)	96.1 (93.6 – 98.6)	98.7 (97.2 – 100.0)
Accuracy (%)	**PPG – Expert**	**93.1 (89.9 – 96.3)**	**97.6 (95.7 – 99.5)**	**96.7 (94.5 – 98.9)**
	Local – Expert	92.2 (88.8 – 95.6)	95.5 (92.9 – 98.1)	97.1 (95.0 – 99.2)
Cohen's Kappa	**PPG – Expert**	**0.71 (0.57 – 0.84)**	**0.84 (0.71 – 0.97)**	**0.65 (0.41 – 0.89)**
	Local – Expert	0.66 (0.52 – 0.81)	0.66 (0.47 – 0.86)	0.68 (0.45 – 0.91)
ROC AUC	**PPG – Expert**	**0.94**	**0.98**	**0.98**
	Local – Expert	NA	NA	NA
Precision Recall AUC	**PPG – Expert**	**0.80**	**0.79**	**0.62**
	Local – Expert	NA	NA	NA

Table 9: Performance Endpoints. 95% confidence interval estimates for the performance endpoints are shown in brackets. PPG – Expert refers to the predictions of the PPG-based method compared to Expert Analysis. Local – Expert refers to the Local Analysis compared to the Expert Analysis. AUC = area under the curve; NPV = negative predictive value; PPV = positive predictive value; ROC = receiver operating characteristic.

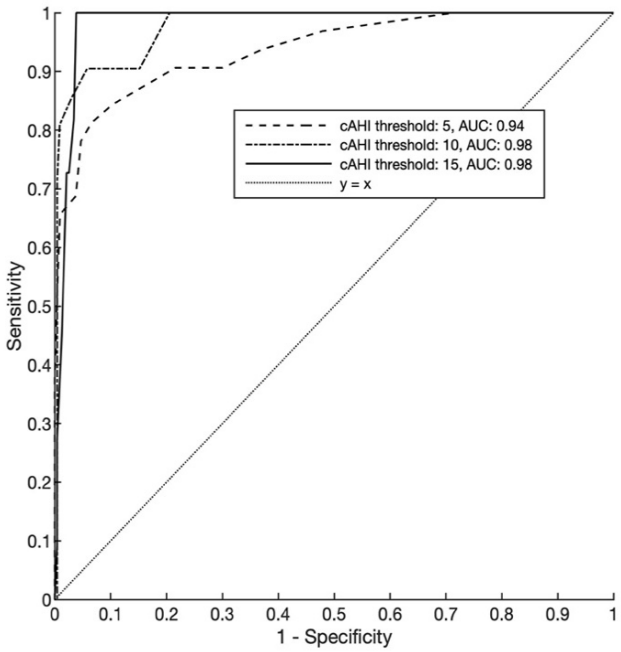

Figure 21: ROC curves for cAHI cutoffs of 5, 10, and 15 of the PPG-based method. AUC = area under the curve; ROC =receiver operating characteristic.

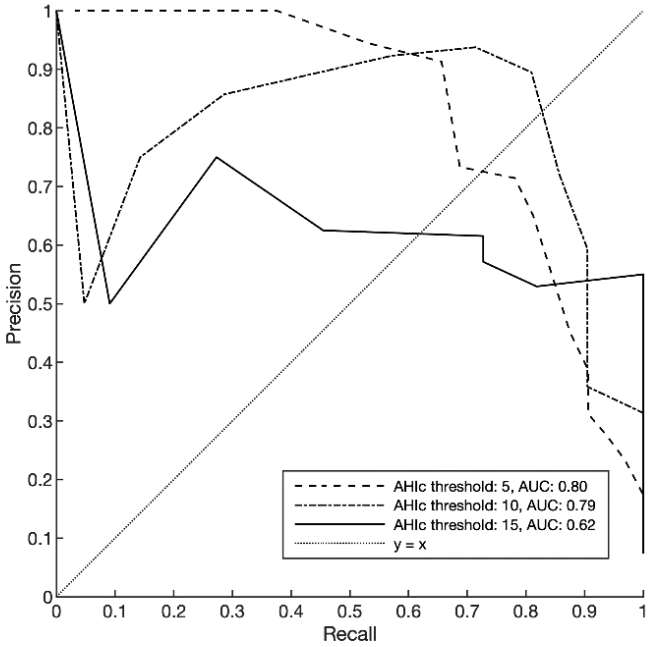

Figure 22: Precision Recall curves for cAHI cutoffs of 5, 10, and 15 of the PPG-based method. AUC = area under the curve.

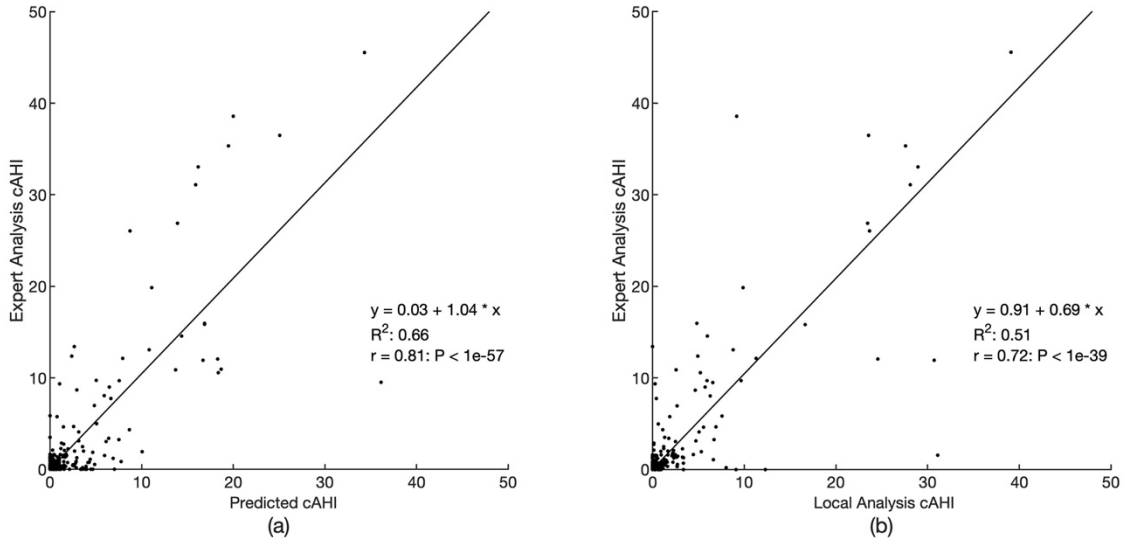

Figure 23: (Left) Scatterplot of Expert Analysis cAHI vs PPG-based method's predicted cAHI. (Right) Scatterplot of Expert Analysis cAHI vs Local Analysis cAHI. P = p-value; r = Pearson correlation coefficient; R^2 = R-squared value; y = a + b ∗ x = the equation of the least-squares fitted line.

		Predicted Class							
		cAHI < 5		*5 ≤ cAHI < 10*		*10 ≤ cAHI < 15*		*cAHI ≥ 15*	
		PPG – Expert	Local – Expert	**PPG – Expert**	Local – Expert	**PPG – Expert**	Local – Expert	**PPG – Expert**	Local – Expert
True Class	*cAHI < 5*	**203**	203	**9**	8	**1**	1	**0**	1
	5 ≤ cAHI < 10	**5**	5	**5**	6	**0**	0	**1**	0
	10 ≤ cAHI < 15	**2**	3	**1**	3	**3**	1	**4**	3
	cAHI ≥ 15	**0**	1	**1**	2	**2**	0	**8**	8

Table 10: Confusion matrix. PPG – Expert refers to the predictions of the PPG-based method compared to Expert Analysis. Local – Expert refers to the Local Analysis compared to the Expert Analysis.

Figure 23 shows scatterplots of the cAHI predicted by the PPG-based method and Local Analysis' cAHI against the Expert Analysis' cAHI. The correlation was 0.81 for PPG-based predictions (p-value < 0.001) and 0.72 for the Local Analysis (p-value < 0.001). Table 10 shows the confusion

matrix of PPG-based predictions for four cAHI intervals, and the same analysis is used for the Local Analysis. The 4-class accuracy for cAHI prediction derived from this matrix was 89.4% for PPG-based predictions and 89.0% for the Local Analysis. Cohen's Kappa was 0.573 for PPG-based predictions and 0.544 for the Local Analysis.

Feature importance

Table 11 lists the 5 most important features, which capture 53.1% of the total feature set importance. The best-performing feature is the minimum range of the breathing signal calculated on a 7-second window sliding over the duration of the respiratory event with a shift of 1 s. The other features in the top 5 capture similar information. The local maxima features are the lowest-ranked features.

	Feature importance (%)	Cumulative sum (%)
Minimum of range values (window = 7s)	19.1	19.1
Minimum of variance values (window = 11s)	12.4	31.5
Average of three lowest IQR values (window = 11s)	7.7	39.2
Minimum of range values (window = 5s)	7.0	46.2
10th percentile of range values (window = 5s)	6.9	53.1

Table 11: Ranked feature importance of the top 5 features. IQR = inter-quartile range.

Discussion

General

The method described in this paper allows for accurate cAHI prediction based on PPG data obtained by a finger probe and does so without the need for any additional sensor modalities such as a chest probe.

PAT HSAT performance

The performance of our method was found to be comparable to that of the Local Analysis of the concurrently acquired PSG.

Another comparison can be made between the proposed method and the PAT HSAT cAHI detection performance described by Pillar et al.[64] (note that, as stated earlier, this method combines PPG and chest accelerometery modalities). The study population described by Pillar et al.[64] was preselected for CSA. As such, PPV and NPV endpoints could not be compared due to the large prevalence discrepancies between their study and ours.

The method described by Pillar et al.[64] achieved a sensitivity of 71.4% and specificity of 98.6% for a cAHI cutoff of 10. For cAHI cutoff of 15, it achieved a sensitivity of 66.7% and a specificity of 100%. In comparison, our method's performance endpoints listed in Table 9 show an outperformance in the sum of sensitivity and specificity for both cAHI cutoffs.

Selecting the cAHI cutoff for CSA screening; clinical implications and trade-offs

The choice of the most desirable cAHI cutoff value for CSA screening should take into consideration the trade-off between NPV and PPV. Since patients flagged by the method will be referred to a full in-lab PSG, having a high PPV is important to avoid unnecessary follow-up examinations, whereas a high NPV is required to avoid false negative tests.

Figure 24 shows the PPV, NPV, and their sum for varying cAHI cutoffs. For cAHI cutoffs between 2.5 and 15, the NPV remains relatively stable at around 95%, reaching a maximum of 99.1% at cAHI cutoff 13.5. The PPV increases from 57.8% at cAHI cutoff 2.5 to 94.7% at cAHI cutoff 9.5. At cutoff

10.5, it starts to decrease approximately monotonically. The cAHI cutoff corresponding to the highest sum of NPV and PPV of 1.93 was located at cAHI cutoff 10.5.

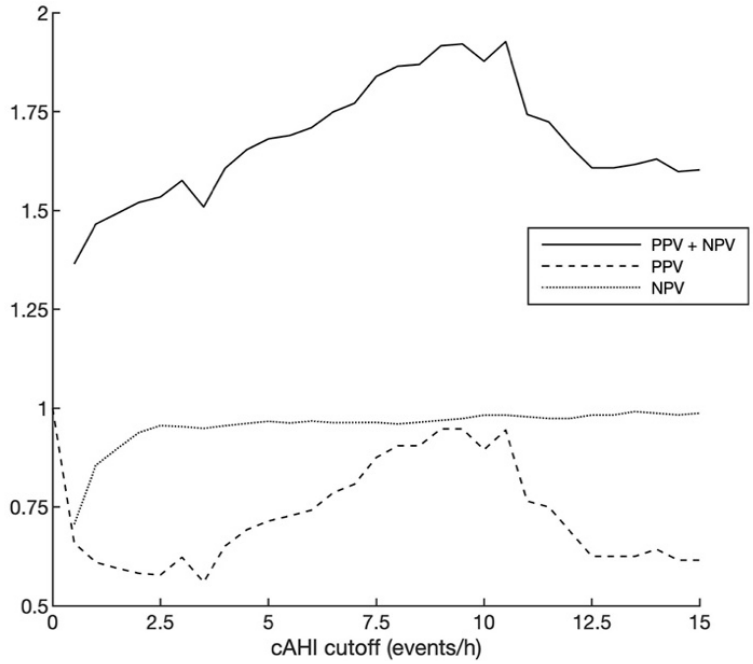

Figure 24: PPV, NPV, and the sum of their values for the PPG-based method for cAHI cutoffs in the range of 0 to 15. PPV = Positive Predictive Value; NPV = Negative Predictive Value.

Figure 23 shows that the PAT HSAT systematically underestimates cAHI at higher cAHI values (≥ 20 events/h). However, the main goal of flagging potential central apnea patients is still achieved since each of the patients with cAHI ≥ 30 is categorised as cAHI ≥ 15 by the PAT HSAT.

Another consideration when selecting the cAHI, besides maximising NPV and PPV, is the selection of the desired sensitivity for detecting plausible CSA predominance. Predominant CSA occurs when 50% or more of the events are central in nature. As such, if we wish to detect patients with predominant CSA with an AHI as low as 5, we need to place the cAHI cutoff at 2.5. If we only wish to detect patients with predominant CSA with an AHI of 15 or higher, we need to place the cAHI cutoff at 7.5. The most performant cAHI cutoff of approximately 10 is optimised to flag predominant CSA in patients with an AHI of 20 or higher. Considering the above, the graph in Figure 24 reveals

the flexibility available to clinicians to select the cAHI cutoff based on this desired minimum AHI for which predominant CSA can be flagged. In the range of cAHI cutoffs between 2.5 and 10, a lower cAHI cutoff is generally associated with a lower PPV, and as such, a higher rate of unnecessary follow-up examinations.

Limitations and future work

One limitation of this study was the low number of predominant CSA patients (only 4 patients in the study dataset) and CSA patients with cAHI ≥ 30 (only 6 patients in the study dataset). The limited sample made the study unreliable in analysing the method's performance of CSA predominance or cAHI ≥ 30 prediction.

Another limitation is that, although the study population matches the intended target population of the study device, some anthropometric and medical background data are missing. For instance, no fine-grained Fitzpatrick scales or information on present comorbidities or the medication use of participants is available. This limits the generalisability of the research findings.

Future work should address these limitations by externally validating the method on a study population with a higher incidence of severe CSA and a more comprehensive medical background record. This will allow us to analyse whether performance degradation may occur in certain subpopulations.

Aside from additional external validation, another possible improvement of the method would be utilising multi-night PAT HSAT data for a more accurate CSA prediction. Taking advantage of multi-night data might also decrease the number of patient recordings that were rejected due to large event rejection proportions.

Conclusion

This work presents a novel cAHI prediction method that uses only finger photoplethysmography data. The results suggest that this method could be incorporated into PAT HSAT for minimally obtrusive CSA screening. In this study population of patients with suspected sleep apnea, the method performed similarly to the inter-rater variability of PSG-derived cAHI estimates.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	Central Sleep Apnea Detection by Means of Finger Photoplethysmography
Publication status:	*Published*
Publication details:	Massie F, Vits S, Khachatryan A, Van Pee B, Verbraecken J, Bergmann J. Central Sleep Apnea Detection by Means of Finger Photoplethysmography. IEEE J Transl Eng Health Med. 2023 Jan 12;11:126-136. doi: 10.1109/JTEHM.2023.3236393.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed CSA algorithm development and evaluation, clinical study design and writeup, and publication design and writeup. S vits performed CSA algorithm evaluation, clinical study data acquisition, and publication editing. A Khachatryan, B Van Pee, and J Verbraecken provided publication editing. J Bergmann provided supervisory guidance and publication editing.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Chapter 4: Enhancing the ability of PAT HSAT to measure sleep accurately

Research Aim 3.1: Evaluating and improving REM OSA phenotyping

Published as: Massie F, Van Pee B, Vits S, Verbraecken J, Bergmann J. Phenotyping REM OSA using peripheral arterial tone-based home sleep apnea testing and polysomnography: A critical assessment of the sensitivity and specificity of both methods. *Journal of Sleep Research*. 2022 Apr;31(2):e13481. doi: 10.1111/jsr.13481.

Reiteration of research aim and introductory context

The gold standard method for staging sleep, as performed by PSG, is based on biopotential (EEG, EOG, and EMG) analysis. Repeated efforts have been made to use PAT HSAT's PPG signal for sleep staging. However, the accuracy achieved has consistently been inferior to that of PSG[36–39]. Accurately identifying sleep stages and, in addition, identifying the tendency for respiratory events to cluster within certain sleep stages (especially REM sleep) has relevant clinical implications[40,41]. The REM OSA phenotype is an important subcategorisation of the sleep apnea condition, characterised by the tendency for respiratory events to increase in frequency or severity during REM sleep[42]. In the paper introduced in the following sections, we first briefly discuss the performance of our algorithmic approach to PAT HSAT-based REM sleep detection. Secondly, we compare the accuracy of PAT HSAT-based to PSG-based REM OSA phenotyping. Finally, we evaluate to what degree a multi-night (repeat) testing protocol can enhance the accuracy of PAT HSAT-based REM OSA phenotyping.

Abstract

Study objectives

The clinical relevance of REM OSA is supported by its associated adverse health outcomes and impact on optimal treatment strategies. To date, no assessment of REM OSA phenotyping performance has been conducted for any type of sleep testing technology. The first objective of this study was to assess this for PSG and PAT HSAT. Subsequently, we assessed the potential of a multi-night PAT HSAT testing protocol to increase the performance of REM OSA phenotype detection.

Methods

In a dataset comprising 261 participants, the sensitivity and specificity of the agreement on REM OSA phenotyping were assessed for two independent scorings of polysomnography and a synchronously administered PAT HSAT.

Results

The sensitivity and specificity of REM OSA phenotyping were 0.87 and 0.89, respectively, for the polysomnography inter-scorer comparison, and 0.68 and 0.97 for the PAT HSAT on a single-night basis, using the conventional minimum required rapid eye movement sleep time of 30 min.

Conclusions

PSG-based REM OSA phenotyping was found to be sensitive and specific even for a single-night testing protocol. PAT-based REM OSA phenotyping showed a lower sensitivity, but a slightly higher specificity compared to polysomnography. To increase performance and conclusiveness of PAT-based REM OSA phenotyping, a multi-night protocol of 2 to 5 nights could be considered. Finally,

the minimum required rapid eye movement sleep time could be lowered from the conventional 30 min to 15 min without significantly lowering REM OSA phenotyping sensitivity and specificity, while increasing the level of phenotyping conclusiveness.

Introduction

It has been well studied that for a significant subgroup of people with OSA, the condition is more severe during rapid eye movement (REM) sleep[42]. Indeed, during REM sleep, muscle tone decreases, which increases the propensity for the tissues surrounding the upper airway to collapse[95]. Hypoxic and hypercapnic ventilatory drive also tends to reduce during REM sleep[96]. These phenomena may result in more frequent or more extended respiratory events, and deeper oxygen desaturations[42]. The sympathetic discharge near the end-stage of respiratory events is exacerbated during REM sleep[97], which causes repetitive surges in blood pressure and heart rate[98]. The heightened severity of OSA during REM sleep likely contributes to an increased risk and severity of its established consequences, including endothelial dysfunction[99], systemic hypertension[100], insulin resistance[101], atherosclerosis[102], adverse neurocognitive outcomes[103], and cardiovascular mortality[101]. The clinical relevance of identifying REM OSA as a distinct OSA phenotype is supported not only by its adverse health outcomes but also by its impact on optimal treatment strategies. It is well known that the majority of REM sleep epochs tend to occur during the last part of a patient's sleep[42]. Once treated with CPAP, patients with increased severity of OSA during REM sleep could benefit from ensuring that the CPAP is also active throughout the final hours before awakening[104]. Since many patients cannot achieve such high levels of CPAP adherence, it is imperative to explore alternative treatment strategies or even combine less effective therapeutic approaches (e.g., oral appliance plus nasal Expiratory-PAP) in order to achieve clinical efficacy in lowering the REM apnea-hypopnea index (REM AHI)[42]. Knowledge of a patient's REM OSA phenotype and its adverse impact on comorbidities might equally influence treatment strategies for these comorbidities.

To date, no consensus has been established on the clinical criteria used to define REM OSA[41]. Nevertheless, most papers use the criterion set out below, which comprises three independent requirements. The first requirement is that the REM AHI must be at least two times higher than the AHI during non-REM sleep stages (NREM AHI)[105]. Second, to ensure REM OSA is only labelled when there is sufficiently clinically relevant OSA, and to ensure that the results are numerically stable, the overall AHI must be larger than 5[104,106–108]. Third, the phenotyping is rendered inconclusive for participants for whom the total REM sleep time is less than 30 minutes[109,110], since short REM sleep time might lead to imprecise results[41] or an overestimation of the clinical effects of REM OSA[40]. The three criteria combined can be formulated as:

REM OSA is present if REM AHI / NREM AHI > 2 and AHI > 5, and REM OSA is inconclusive if REM sleep time < 30 minutes

Duce et al.[110] criticised the minimum REM sleep time requirement, finding that employing a minimum required REM sleep time risks underdiagnosing REM OSA. The first-night effect might cause an underestimation of REM sleep time during a single-night sleep test, further adding to this risk[110]. Regardless, fixing this cutoff at exactly 30 minutes appears to be a rule of thumb without the support of a statistical analysis. In this study, we therefore assessed what minimum REM sleep time is required to obtain an adequately sensitive and specific REM OSA phenotyping.

The value of OSA phenotyping for the development of personalised OSA management has been well recognised[111], and the body of research on PSG-based OSA phenotyping is rapidly growing[45]. Coinciding with the prominent role of PSG for such phenotyping, there is an increased utilisation of reduced-channel HSAT[112], which is a trend fuelled by the impact of COVID-19 on the admittance of patients in the sleep lab[113]. These reduced-channel HSATs have more restrictive capabilities in determining OSA phenotypes, such as REM OSA, primarily due to their lack of an EEG for sleep recording. Despite the absence of EEG, the AASM Scoring Manual reports one particular category

of HSAT devices that is required to demonstrate REM sleep, namely the PAT-based devices[1]. These devices monitor changes in peripheral arterial tone related to sympathetic surges, alongside surges in pulse rate and oxygen desaturations[1], and do so through the registration of a conditioned PPG signal. As sympathetic tone increases during REM sleep, its manifestations on the peripheral arterial tone and the pulse rate can be used to identify episodes of REM sleep. At the time of this publication, two PAT-based HSAT devices have been cleared by the FDA: WatchPAT (Itamar Medical, Israel)[50] and NightOwl (ResMed, USA)[21], the latter of which was used in this study. For both devices, there currently exist no studies reporting on their ability to accurately determine the REM OSA according to any conventional definition; hence, this was evaluated in this study.

To fairly assess the ability of an HSAT device to estimate a specific clinical parameter, it is important to compare its performance against the inter-scorer agreement of PSG for the same parameter. While substantial inter-scorer variability has been demonstrated for REM sleep time[32], no similar exercise has been performed to date for REM OSA phenotyping.

Therefore, the primary aim of our study was to compare the REM OSA phenotyping performance of PAT HSAT to PSG.

The results of the study may help to provide insight into whether the trend towards lower complexity sleep testing is reconcilable with an increased understanding of the importance of REM OSA phenotyping.

Methods

Dataset

The dataset analysed in this manuscript comprises 261 participants with suspected OSA, who were prospectively recruited in a consecutive cohort across four different centres, namely three in

the USA (where all centres were part of the United Health Systems Group in Miami, Florida), and one located in Belgium (Ziekenhuis Oost Limburg, ZOL, Genk). All participants were scheduled for one overnight in-lab PSG. The USA branch of the study was approved by Aspire IRB, part of the WIRB-Copernicus Group. The Belgian branch of the study was approved by the Ethics Committee of ZOL. Underaged or cognitively impaired participants were excluded from the study. For all participants, gender, age, and BMI were recorded. Persons of diverse racial and ethnic backgrounds were included.

Protocol and devices

A graphical representation of the study setup is provided in Figure 25. Routine PSG was performed for all study participants. Qualified lab technicians at each participating study centre were responsible for setting up the equipment and capturing PSG data. During the setup of PSG, the PAT HSAT was attached to the middle finger of the hand to which the pulse oximeter of PSG was applied. All PSG data were double-scored by two independent centres which were blinded from one-another's analysis.

Polysomnography

For the European and USA centres, respectively, the Alice 6 PSG (Philips Respironics, USA) and Cadwell Easy PSG (Cadwell, USA) were used.

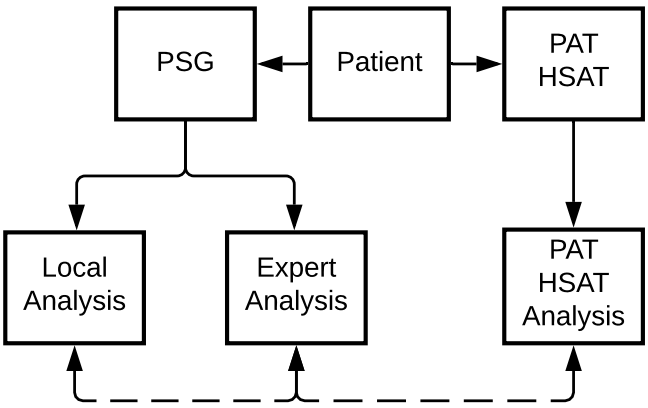

Figure 25 Diagram of the data acquisition set-up. This diagram depicts that for each study participant, PSG and PAT HSAT were concurrently administered. PSG was analysed by two independent scoring centres (expert and local), and PAT HSAT analysis were compared with Expert Analysis (illustrated by the dashed line).

Each PSG was scored by two independent scoring centres. The first scoring was performed by the team of sleep technicians of the centre where the patient was admitted (further referred to as Local Analysis).

Another independent scoring was performed by the scorers of Cerebra Medical (Canada). This company provides computer-aided sleep scoring services to support PSG scoring for clinical centres and clinical trials. The studies were first analysed by their proprietary Michele Sleep Scoring System (MSSS) and were subsequently complemented with complete manual rescoring by an expert technologist.

Malhotra et al.[32] confirmed in a multicentric trial that the MSSS, complemented with manual editing by an expert scorer, is more robust than the results of a single expert scorer. Because of these conclusions, Cerebra Medical's analysis served as the expert benchmark (Expert Analysis**)** to which

the Local Analysis and PAT HSAT's analysis were compared. All PSG data were scored according to the 2012 AASM scoring rules using the Recommended 1A (3%) rule for hypopnea scoring [1].

PAT HSAT and its principle of operation

PAT HSAT is built around a fingertip-mounted PPG probe. PPG operates on simple optical technology to detect blood volume changes in the tissue microvascular bed. From the PPG measurement, the SpO_2, PR, and PAT are derived. The latter signal reflects changes in the blood volume due to peripheral artery constriction following an increase in sympathetic tone at the end-stage of a respiratory event[21]. In line with the AASM's definition of PAT technology, the combination of a decrease in PAT, an SpO_2 desaturation, and an increase in PR is inspected to detect respiratory events[1]. The probe also contains an accelerometer for the detection of limb movement to determine sleep/wake based on actigraphy[21].

REM sleep time estimation based on PAT

Bursts of rapid eye movements have been related to phasic vasoconstrictions throughout REM episodes[114]. The temporal pattern of these vasoconstrictions, as detected by PAT, was found to be different from that observed in NREM sleep[115]. Furthermore, PR low-frequency power has been shown to increase[116] during REM, and oxygen desaturation dynamics are significantly different during REM sleep compared to NREM sleep[117]. The study device automatically analyses the abovementioned phenomena to identify whether a sleep epoch comprises REM sleep. For the dataset of this study, the ICC(A,1) of the PAT HSAT REM sleep time with the Expert PSG Analysis was 0.50, and the LoA spanned 118.9 minutes. These results were comparable to the results reported by Zhang et al.[24,36] who found a Spearman correlation coefficient of 0.48 and a LoA span of 133.7 minutes in a dataset of 170 participants comparing WatchPAT with PSG.

Statistical analysis

General

Statistical analysis was performed using MATLAB (version 2019a, MathWorks, USA). For all sensitivity and specificity endpoint parameters, 95% confidence intervals were computed. Statistical significance was defined as $p \leq 0.05$.

Data synchronisation

PSG and PAT HSAT data were algorithmically synchronised by matching the instantaneous heart rate traces derived from the ECG trace of PSG and the PR trace of PAT HSAT. This synchronisation method resulted in a maximum desynchronisation tolerance of 3 seconds. Data epochs that were of insufficient quality to be interpreted by the sleep technician or PAT HSAT were rejected from PSG and the PAT HSAT traces. This resulted in a median rejection rate of 10% of data epochs per recording.

Data adequacy

All 261 participants included in the dataset met the following data adequacy criteria: all PSG channels could be interpreted by the technicians (e.g., there were no detachments of the nasal cannula or pulse oximeter), and at least 4 hours of analysable signal could be obtained for the PAT HSAT, as recommended by the AASM[33].

Participants with missing patient characteristics, such as age and gender data, were omitted from the analysis of population demographic statistics. These participants were still included for the calculation of endpoint parameters.

Endpoint analysis

Primary study aim: assessing REM OSA phenotyping agreement

The criterion used to determine REM OSA was:

> **REM OSA is present if REM AHI / NREM AHI > 2 and AHI > 5, and REM OSA is inconclusive if REM sleep time < 30 minutes**

For both the PAT HSAT analysis and the Local Analysis, the sensitivity and specificity of the agreement with the Expert Analysis were calculated. To describe the agreement of REM OSA phenotyping by just a single parameter, the sum of the sensitivity and specificity was also calculated[118].

Auxiliary study aim: assessing the minimum required REM sleep time

To better understand the impact of the minimum required REM sleep time (further referred to as the REM exclusion criterion, or REC), it was investigated how the REM OSA phenotyping performance changed as the REC varied from 1 minute to the conventional 30 minutes. Participants for whom the REM sleep time was less than the REC according to either the Expert Analysis or the comparator analysis (i.e., PAT HSAT or Local Analysis) were labelled as inconclusive. As such, conclusiveness was defined as the percentage of participants for whom the REM sleep time was larger than the REC. Statistical significance of changes in sensitivity by varying the REC was determined by means of a *two-proportion z-test*.

Results

Demographic information

Participants were predominantly male (60%), of middle age (mean 54 yrs., SD 14), and overweight (mean BMI 30.0 kg/m^2 and SD 5.9). The mean AHI was 31.9 (SD 25.6). 29 participants had no clinical sleep apnea, 59 participants had mild sleep apnea, 60 participants had moderate sleep apnea, and 113 participants had severe sleep apnea. More details on patient demographics are provided in Table 12.

	Mean	SD	Min	Max	p25	p50	p75
Age	54	14	21	84	45	55	66
AHI	31.9	25.6	0.0	118.7	10.0	25.5	48.5
BMI	30.0	5.9	18.2	53.8	25.8	29.3	33.2

Table 12: Demographic and clinical characteristics of participants in the dataset. p25 = 25th percentile value; p50 = median; p75 = 75th percentile value; SD = standard deviation.

REM OSA phenotyping performance and influence of minimum required REM sleep time

The prevalence of REM OSA in this dataset, calculated at a REC of 30 minutes, was 26.1%. The sensitivity and the specificity of the REM OSA phenotyping by the Local Analysis and PAT HSAT are visualised in Figure 26 for REC values ranging from 1 to 30 minutes. The red segment of the sensitivity line of the PAT HSAT highlights those REC values for which the sensitivity was statistically significantly lower than at the conventional REC of 30 minutes. An increase in sensitivity can be observed as REC increases from 1 minute to 30 minutes. For PAT HSAT, as REC values exceeded 10 minutes, sensitivities were no longer significantly lower than the sensitivity at REC 30. For the Local Analysis, no statistically significant reduction in sensitivity was observed when reducing the REC from 30 minutes to 1 minute. There was no REC value in the investigated

range for which the specificity of either comparator was statistically significantly lower than that at a REC of 30 minutes.

As stated in the methods section, only assessments for which both the Expert Analysis and the comparator determined a REM sleep above the REC were labelled as conclusive. The percentage of participants for whom there was conclusive phenotyping is visualised in Figure 26 for each REC value. As the REC increases, the percentage of conclusiveness drops. A decline from more than 85% conclusiveness at a REC of 1 minute to approximately 45% conclusiveness at a REC of 30 minutes was observed. The numeric results presented in Figure 26 are listed in Table 13 for RECs of 15 and 30 minutes. For both REC values, the PAT HSAT sensitivity is substantially lower than the Local Analysis, while its specificity is slightly higher. For both comparators, a marked difference in conclusiveness from approximately 45% at a REC of 30 minutes to more than 65% at a REC of 15 minutes was found.

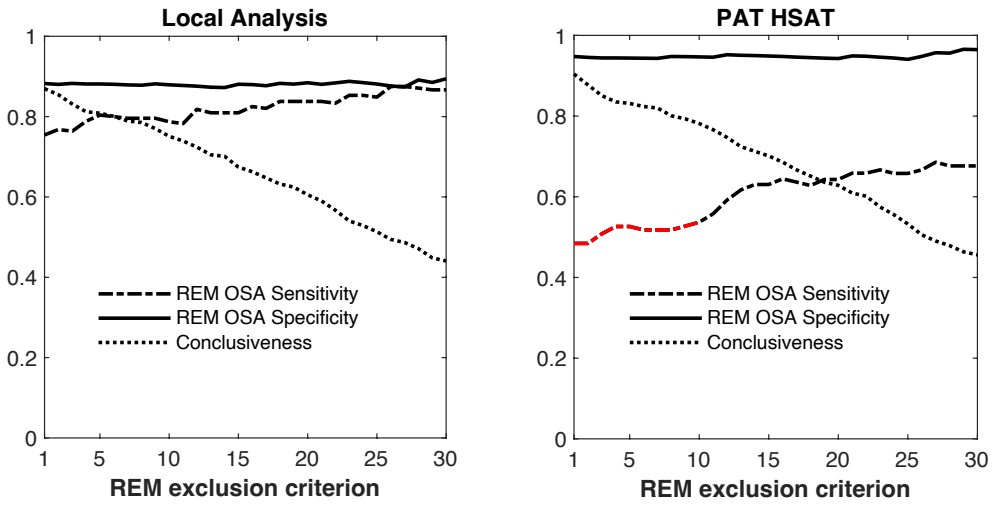

Figure 26: Depiction of how the REM OSA phenotyping sensitivity and specificity increase, and phenotyping conclusiveness decrease with increasing minimum required REM time (REM exclusion criterion) in minutes. The red line segments highlight zones where the phenotyping sensitivity was statistically significantly lower than that at a REM exclusion criterion of 30 minutes.

REC	Comparator	Sensitivity	Specificity	Sum	Conclusiveness
15 min	*PAT HSAT*	0.63 (±0.14)	0.95 (±0.04)	1.58 (±0.10)	0.70 (±0.06)
	Local PSG	0.81 (±0.12)	0.88 (±0.06)	1.69 (±0.12)	0.67 (±0.06)
30 min	*PAT HSAT*	0.68 (±0.16)	0.97 (±0.04)	1.64 (±0.12)	0.46 (±0.06)
	Local PSG	0.87 (±0.12)	0.89 (±0.07)	1.76 (±0.13)	0.44 (±0.06)

Table 13: Tabulation of REM OSA sensitivity, specificity, their sum, and conclusiveness for REM exclusion criterion (REC) of 15 and 30 minutes. 95% confidence intervals around the parameters are provided within the brackets. REC = REM exclusion criterion.

Discussion

Minimum required REM sleep time when phenotyping REM OSA: the current consensus might be too conservative

The findings presented in this manuscript support that a reduction of the minimum required REM sleep time for REM OSA phenotyping from 30 minutes to 10 minutes can be permitted before the reduction in phenotyping sensitivity becomes statistically significant.

By lowering the REC from 30 to 15 minutes, the conclusiveness percentage increases from 44% to 67% for the Local PSG Analysis, and from 46% to 70% for the PAT HSAT.

While Mokhlesi et al.[41] argue that a minimum REM sleep time of 30 minutes would be required to reduce the possibility of overestimating the effects of REM OSA in individuals with short REM duration, we hold the opinion that considering prevalent inconclusiveness, it would rather be more prudent to avoid inconclusive REM OSA phenotyping. In view of steep gains in conclusiveness while experiencing only a non-significant reduction in sensitivity and specificity resulting from a reduction of the minimum required REM sleep time from 30 minutes to 15 minutes, we propose the use of a minimum required REM sleep time of 15 minutes instead of 30 when electing to phenotype REM OSA. Although a reduction of the minimum required REM sleep time to 10 minutes would be

supported by the data, we chose to add a margin of safety of 5 minutes to this minimum. In contrast to Mokhlesi et al.[41], we propose that when REM OSA is assessed on the basis of a short REM sleep time, it is preferable to pursue multi-night phenotyping rather than avoid phenotyping altogether.

N	AHI Cutoff	Sensitivity	Specificity	Sum
261	*5*	0.94 (±0.03)	0.76 (±0.16)	1.70 (±0.10)
	15	0.93 (±0.04)	0.77 (±0.09)	1.70 (±0.08)

Table 14: Tabulation of the sensitivity and specificity of the Local Analysis' OSA diagnosis at AHI cutoffs 5 and 15 when taking the Expert Analysis as reference for the dataset of 261 participants. 95% confidence intervals around the parameters are provided.

Phenotyping REM OSA: PAT HSAT may fall short

Since this work presents the first reporting on REM OSA phenotyping performance of PSG and PAT HSAT according to a conventional REM OSA criterion, no consensus exists on an adequate phenotyping sensitivity and specificity.

In clinical practice, a certain margin of error for the diagnosis of OSA has been accepted, as implied from the inherent residual inter-scorer variability of PSG-derived AHI[32]. It can be argued that the accepted clinical performance of REM OSA phenotyping ought not to be expected to exceed the accepted clinical performance of the OSA diagnosis itself, as the latter has generally more impactful treatment implications than the former. Therefore, a conservative performance target for REM OSA sensitivity and specificity would be to take the inter-rater agreement level of the diagnosis of OSA.

As highlighted by Bianchi[118], the agreement between a comparator and a reference can be characterised by the sum of its sensitivity and specificity at a binary cut-off. To assess the clinical performance of OSA diagnosis on the same patient demographic for which the REM OSA performance was determined in this study, the sensitivity and specificity of diagnosing OSA by the

Local PSG Analysis were calculated from the same dataset of 261 double scored PSGs, while the Expert PSG Analysis was again designated as the reference.

The AHI-cutoffs most important to the diagnosis of OSA are 5 and 15, depending on the presence of OSA-related symptoms[43]. As displayed in Table 14, the sum of the sensitivity and specificity was 1.70 for both AHI cut-offs. To support the generalisability of these results, it was investigated whether the inter-rater agreement of this dataset was in line with those reported in the literature. Since inter-scorer variability literature reports mostly on a correlation coefficient, despite its obvious shortcomings as a parameter for disease severity accuracy, the ICC(A,1) between the two AHI estimates was calculated. The correlation coefficient between the two AHIs was 0.91, which is close to the across-site ICC of 0.95 reported by Malhotra et al.[32] As displayed in Table 13 of the results section, the sum of the sensitivity and specificity for determining REM OSA when respectively employing REC values of 15 and 30 minutes were 1.69 and 1.76, which was close to, or larger than, the target of 1.70. Therefore, it can be concluded that the inter-rater agreement of phenotyping REM OSA using PSG is similar to the inter-rater agreement of diagnosing OSA. For PAT HSAT, the same sums calculated for REC values of 15 and 30 minutes were 1.58 and 1.64, respectively, which were lower than the target of 1.70. Hence, though it has not yet been formally explored, this type of test might benefit from multi-night testing to increase its sensitivity and specificity.

A theoretical model of multi-night REM OSA phenotyping: it may move the needle

It could be concluded that, contrary to PSG, REM OSA phenotyping performance of PAT HSAT did not meet the proposed target sum of sensitivity and specificity on a single-night basis. We subsequently numerically investigated whether multi-night testing could improve the performance above the target level. To this end, multi-night testing was modelled as a binomial process in which each single-night test outputs a REM OSA phenotype assessment with a sensitivity and specificity

equal to those introduced in Table 13. This model can be interpreted as a voting model in which a single vote is cast after every single-night test, judging whether REM OSA was present during that particular night. The global vote on the presence of REM OSA is cast when at least k out of n votes were positive. The model parameter n is to be interpreted as the number of single-night tests performed for which the REM OSA vote was conclusive. In the event of one or more inconclusive single-night tests, repeat testing continues until n conclusive single-night phenotype determinations are obtained. Therefore, the actual number of single-night tests would be substantially dependent on the choice of the REC value, since that parameter was demonstrated to meaningfully influence the conclusiveness of phenotyping. The model parameter k is to be interpreted as the minimum number of positive single-night REM OSA votes required to render the global REM OSA phenotype determination positive.

From the voting model formulation follows that the sensitivity and specificity of the global vote under the assumptions laid out above can be estimated as set out in equations 1 (EQ1) and 2 (EQ2).

$$se(k,n) = 1 - \sum_{i=0}^{k-1} \frac{n!}{i!(n-i)!} se(1,1)^i (1 - se(1,1))^{n-i} \quad \text{EQ1}$$

$$sp(k,n) = 1 - \sum_{i=0}^{n-k} \frac{n!}{i!(n-i)!} sp(1,1)^i (1 - sp(1,1))^{n-i} \quad \text{EQ2}$$

The resulting $se(k,n)$ and $sp(k,n)$ parameters are respectively the sensitivity and specificity of the global REM OSA phenotype determination (i.e., the global vote) for any given choice of model parameters k and n. The resulting sum of $se(k,n)$ and $sp(k,n)$ for different values of k and n is displayed in Figure 27. The figure illustrates how extending the number of nights markedly improves the performance, with a second night of testing crossing the 1.70 performance target for k equal to 1. As n exceeds 3 nights of testing for k equal to 1, the performance starts to decrease, and it drops below the 1.70 performance target as n exceeds 4 nights of testing. This is to be expected, since the requirement of only one positive REM OSA vote out (with k = 1) of three or more single-night tests

would likely result in a high false positive rate of the global vote. Extending the number of test nights to 4 results in marked performance surges for $k = 2$. By increasing the number of test nights to 5, a performance increase from 1.58 to 1.91 was simulated. Increasing the number of test nights beyond 5 shows diminishing incremental performance gains. For a test duration exceeding 5 nights, one might argue that the incremental burden for the patient to test for an additional night starts to outweigh the limited incremental performance gain.

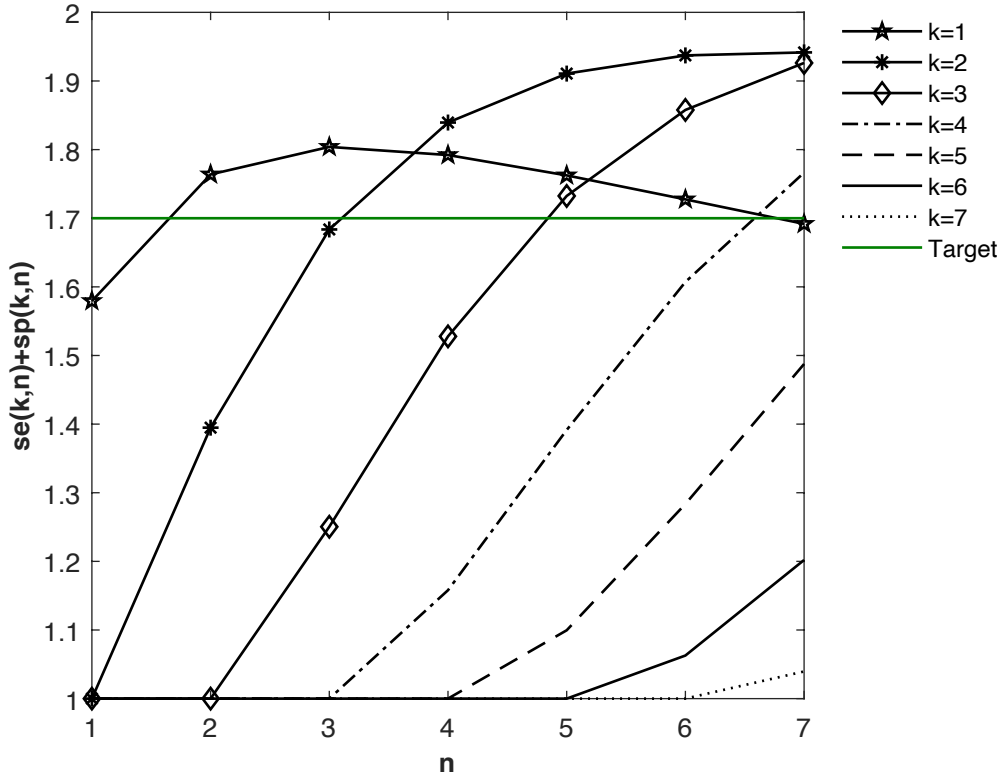

Figure 27: Depiction of the theoretical sum of REM OSA phenotyping sensitivity and specificity of PAT HSAT for multi-night testing of n nights, with k the minimum number of single-night positive REM OSA phenotype determinations to render the global phenotype determination positive. The green line highlights the target REM OSA phenotyping performance expressed as the sum of phenotyping sensitivity and specificity. k = minimum required single-night positive REM OSA phenotype determinations to render global phenotype determination positive; n = number of HSAT nights; se = sensitivity; sp = specificity.

Aside from the above discussion, one should note that multi-night testing has the additional benefit of further mitigating inconclusiveness from which both PAT HSAT and PSG still suffer, even after reducing the REC from 30 to 15 minutes.

This model has several limitations. First, the model assumes that the REM OSA phenotype is stable throughout the multi-night testing protocol. The validity of this assumption requires further research. Though it has been well established that the AHI is subject to significant night-to-night variability[18], it stands to reason that the REM OSA phenotype itself might remain mostly stable over time. Indeed, whereas the AHI might be influenced by the sleeping position and the REM sleep time, among others, which may vary significantly from night to night, the underlying pathophysiology driving the REM predominance of the OSA is likely to remain stable. Second, the sensitivity and specificity values cannot be determined with perfect accuracy since we have shown that PSG itself suffers from significant inter-rater variability when phenotyping REM OSA. A more stable sensitivity and specificity estimate could be obtained by averaging PSG data scored by a large number of expert sleep technicians. Third, a binomial model assumes statistical independence between the repeated measurements. However, a degree of dependence between individual single-night REM OSA phenotyping errors within a multi-night protocol is expected. For example, when a patient's REM episodes are difficult to estimate from the PAT measurement because of sustained and abnormally depressed sympathetic activity, a systematic underestimation of REM sleep time may occur, which may result in repeated false negative REM OSA phenotyping.

Nevertheless, statistical dependence between repeated phenotyping measurements would reduce the apparent gain from additional nights, and thus underestimate the number of nights required to surpass the target REM OSA phenotyping performance. Hence, even if the assumption of independence does not hold, the overall conclusion remains that a multi-night protocol is desirable to increase the REM OSA phenotyping performance of PAT HSAT.

Strengths and weaknesses of the study

The strength of this study lies in its analysis of a large and independently double-scored dataset, which includes both PSG data and synchronously acquired PAT HSAT data.

A number of weaknesses of the study should be noted. Firstly, only one embodiment of PAT HSAT was included. Investigating the generalisability of the results to other PAT HSAT devices could be the subject of future research. Finally, the simulation used to estimate the potential of multi-night testing by PAT HSAT relies on several assumptions. Multi-night clinical trial data would be required to test the validity of these assumptions conclusively.

Conclusion

This work presents the first assessment of the REM OSA phenotyping performance of either PSG or PAT HSAT. Firstly, it was found that PSG-based REM OSA phenotyping was sensitive and specific even for a single-night test. The PAT-based REM OSA phenotyping showed a lower sensitivity but a slightly higher specificity than PSG. To increase the performance of PAT-based REM OSA phenotyping to a level comparable to the single-night PSG, a multi-night testing protocol should be considered. A multi-night test duration of 2 to 5 nights was simulated to yield the most substantial incremental performance gains. Secondly, it was suggested that to increase the level of conclusiveness of REM OSA phenotyping, the minimum required REM sleep time can be lowered from the conventional 30 minutes to 15 minutes without significantly impacting the sensitivity and specificity of the phenotyping. A multi-night testing protocol would also further increase the conclusiveness for both PSG and PAT HSAT.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	Phenotyping REM OSA by means of peripheral arterial tone-based home sleep apnea testing and polysomnography: A critical assessment of the sensitivity and specificity of both methods.
Publication status:	*Published*
Publication details:	Massie F, Van Pee B, Vits S, Verbraecken J, Bergmann J. Phenotyping REM OSA by means of peripheral arterial tone-based home sleep apnea testing and polysomnography: A critical assessment of the sensitivity and specificity of both methods. J Sleep Res. 2022 Apr;31(2):e13481. doi: 10.1111/jsr.13481.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed REM algorithm development and evaluation, clinical study design and writeup, and publication design and writeup. S vits performed REM algorithm evaluation, clinical study data acquisition, and publication editing. B Van Pee, and J Verbraecken provided publication editing. J Bergmann provided supervisory guidance and publication editing.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Research Aim 3.2: Enhancing PAT HSAT sleep staging by incorporating a single-lead biopotential measurement

Published as: Massie F, Vits S, Verbraecken J, Bergmann J. The evaluation of a novel single-lead biopotential device for home sleep testing. *SLEEP*. 2025 Apr 11;48(4):zsae248. doi: 10.1093/sleep/zsae248.

Reiteration of research aim and introductory context

In the previous subchapter, our own findings corroborated the consistently reported underperformance of PAT HSAT-based REM sleep detection compared to PSG[36,59,114,119]. With the aim of improving the sleep staging of PAT HSAT, we introduced a single-lead biopotential unit into the test setup that captures a superposition of EEG, EOG, and EMG signals. Standard PSG determines sleep stages using multiple biopotential leads, most of which are positioned within the patient's hair, and this configuration can substantially inconvenience patients[1]. A major factor driving the widespread clinical adoption of PAT HSAT is its minimal obtrusiveness compared to alternative technologies[28]. Therefore, we only considered approaches to enhance its sleep staging capability that preserve the minimally obtrusive character of the original method.

Single-lead biopotential devices can be designed in a highly miniaturised form factor and positioned without placing electrodes in the hair. We hypothesised that, with careful consideration of electrode placement, such a device could meaningfully improve sleep stage classification accuracy in PAT HSAT systems. Our preliminary investigation in healthy individuals indicated that placing the reference electrode at the centre of the forehead and the measurement electrode below the eye yields promising single-lead sleep stage classification performance. To test this hypothesis, we developed sleep scoring algorithms tailored to this single-lead biopotential signal and evaluated the

performance of the extended PAT HSAT system in a clinical study. The findings of this research are described in the publication included in the following paragraphs.

Abstract

Study objectives

This paper reports on the clinical evaluation of the sleep staging performance of a novel single-lead biopotential extension to PAT HSAT.

Methods

133 patients suspected of obstructive sleep apnea were included in a multi-site cohort. All patients underwent polysomnography and received the study device, a single-lead biopotential measurement device attached to the forehead. Clinical endpoint parameters were selected to evaluate the device's ability to determine sleep stages. Finally, the device's performance was compared to the clinical study results of comparable devices.

Results

Concurrent PSG and study device data were successfully acquired for 106 of the 133 included patients. The results of this study demonstrated significant similarity in overall sleep staging performance (5-stage Cohen’s Kappa of 0.70) to the best-performing reduced-lead biopotential device to which it was compared (5-stage Cohen’s Kappa of 0.73). Contrary to the comparator devices, the study device reported a higher Cohen’s Kappa for REM (0.78) compared to N3 (0.61), which can be explained by its measuring electrode placement (diagonally across the lateral cross-section of the eye). This placement was optimised to ensure the polarity of rapid eye movements

could be adequately captured, enhancing the capacity to differentiate between N3 and REM sleep when using only a single-lead setup.

Conclusions

The results of this study demonstrate the feasibility of incorporating a single-lead biopotential extension in a reduced-channel home sleep apnea testing setup. Such incorporation could narrow the gap in the functionality of reduced-channel home sleep testing and in-lab polysomnography without compromising the patient's ease of use and comfort.

Introduction

In-hospital PSG has for decades been the gold standard for the evaluation of sleep architecture and common sleep disorders[14]. One of PSG's main strengths is its inclusion of multiple sensing modalities, which allows for a comprehensive assessment of the sleep condition. The PSG setup incorporates multi-channel EEG, EOG, and EMG, from which it can accurately determine sleep stages, thereby classifying sleep into one of five stages: Wakefulness, REM sleep, and three stages of non-REM sleep (N1, N2, N3). N1 and N2 are referred to as light sleep, whereas N3 is denominated as deep sleep[1]. The progression of sleep stages throughout the night, scored in thirty-second epochs, is called the hypnogram. Assessing these sleep characteristics and other derivations, such as sleep onset and latencies, is crucial in sleep disorder diagnostics. For example, the diagnosis of OSA relies on the estimation of the TST to calculate the AHI. Furthermore, interest in comprehensive OSA phenotyping for developing personalised OSA management is growing[120]. REM-predominant OSA, for instance, is one phenotype that can inform healthcare providers about choosing an optimal therapy strategy[121].

However, EEG analysis from PSG is not without limitations. It is a costly and time-consuming procedure, requiring highly trained personnel's assistance for sensor application, patient monitoring, and data scoring. The high density of attached sensors and electrodes, combined with an unfamiliar sleep environment, may cause discomfort and alter the patient's natural sleep patterns. Factors such as night-to-night variability and the first-night effect can further decrease the reliability of the PSG analysis[122]. This is why, in recent years, ambulatory versions of PSG that support in-bedroom, multi-night assessment of sleep have gained popularity. Notably, reduced-channel HSAT devices, such as PAT HSAT, are well-positioned to address this need[66]. PAT HSAT devices typically consist of a finger-based PPG and accelerometery sensing module from which signal modalities, such as SpO_2, PR, activity, and PAT, can be derived and analysed to estimate the AHI[1]. These devices' lack of EEG, EOG, and EMG sensing capabilities is an important limitation, prohibiting the conventional and accurate assessment of sleep stages.

Therefore, we have developed an optional extension to a PAT HSAT system, which is worn on the forehead and contains a biopotential measurement sensor (further referred to as the 'study device'). While the study device is designed to operate in conjunction with a PAT HSAT system, it can also be used in a standalone fashion. The base of the device, comprising the reference electrode, is attached to the centre of the forehead using an adhesive electrode patch to which it is attached via a snap connector. The measurement electrode is connected via an electric wire to the device's base and is attached below the eye via another adhesive electrode patch (see Figure 28). From the biopotential measurement sensor, single-lead EEG, EOG, and EMG data is derived by filtering the raw biopotential data. The device's base also contains an accelerometer to determine the head orientation (supine versus non-supine) and a microphone to record nocturnal sound levels.

A multi-channel sleep lab EEG setup is usually derived from the standard 10-20 protocol proposed by Herbert Jasper[123]. This system identifies a set of anatomical landmarks to guide electrode placement. Electrodes are placed at 10% or 20% intervals along the anatomical directions with the

ground electrode. Different electrodes present distinct sleep signatures depending on their position. The AASM recommends using frontal electrodes to detect K-complexes and delta wave activity, central electrodes to detect sleep spindles, and occipital electrodes to detect alpha waves[1]. These sleep signatures are used to identify sleep stages and are crucial for visual sleep scoring. Since the study device consists of a single-lead biopotential module, which records a superposition of frontal EEG, EOG, and EMG, it does not unambiguously present each of the typical sleep signatures. Advanced machine learning techniques, like deep learning methods, can overcome this challenge by learning and interpreting sleep-related patterns that are visually less straightforward to discern. Deep learning networks have some advantages over traditional machine learning algorithms in that they do not require domain knowledge for manual feature extraction and can, therefore, easily self-adapt to different biopotential setups[124]. The study device uses a deep learning network framework consisting of epoch and sequence-based Recurrent Neural Net (RNN) layers. RNNs are particularly effective in capturing temporal dependencies in time series, such as biopotential data[124]. In this paper, we evaluate its automated sleep staging performance.

Study objective

Incorporating a single-lead biopotential extension in an HSAT setup will widen its applicability, improve its TST estimation accuracy, and allow for extended phenotyping of sleep apnea. It is necessary to attain functionality similar to in-lab PSG while optimising ease of use, comfort, and cost. This study aimed to compare the automated sleep staging detection of the study device to the standard visual sleep scoring of PSG. To this end, a population with suspicion of OSA was recruited in a multi-centre clinical study to undergo routine PSG with simultaneous attachment of the study device.

Methods

Patients

One hundred thirty-three patients with suspicion of OSA were consecutively included in a cohort across three different centres located in the United States. All centres were part of the United Sleep Diagnostics group in Florida. All patients were scheduled for one overnight in-lab PSG. Patients were asked for informed consent. Aspire IRB, part of the WIRB-Copernicus Group, approved the study Protocol, Patient Information, and Consent Form. Cognitively impaired patients or patients below the age of thirteen were excluded from participation in the study. No patients were excluded based on medical conditions or drug treatments. Recruitment occurred between October 5^{th,} 2021, and June 2^{nd,} 2022. For all patients, gender, age, and BMI were recorded. Patients also completed the FDA's *self-completion questionnaire for ethnicity and race.*

Protocol and devices

Routine PSG was performed for all patients. Qualified lab technicians at each participating study centre were responsible for setting up the equipment and capturing PSG data. The study device was applied using two adhesive electrode patches: one positioned centrally on the forehead and one below the eye. A snap connector attached the reference and measurement electrodes to these patches (Figure 28). To allow for data synchronisation, a PAT HSAT (NightOwl, ResMed, USA) was placed on the finger next to the pulse oximeter of PSG and was time-synchronised via Bluetooth to the study device.

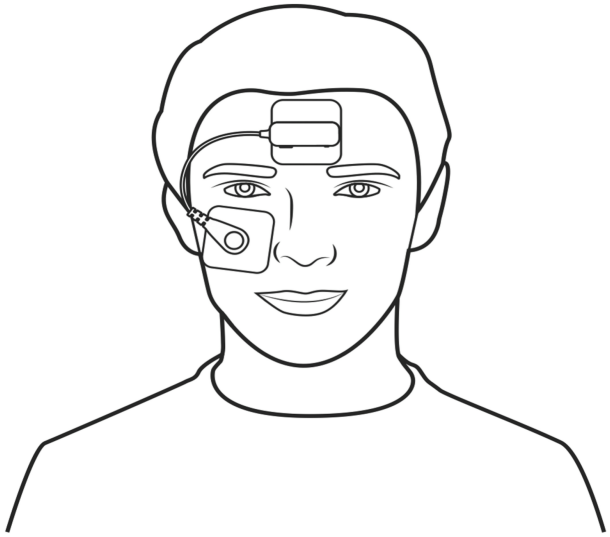

Figure 28: Pictogram of the study device. The reference electrode is attached via an adhesive patch to the centre of the forehead. Similarly, the measuring electrode is attached below the eye.

Polysomnography

The Compumedics E-Series (Compumedics, Australia) PSG was used in all study centres. PSG data were scored centrally by qualified sleep scorers employed by United Sleep Diagnostics, further referred to as the 'PSG Analysis.' All PSG data were scored according to the 2020 AASM sleep staging rules[1].

Statistical analysis

General

Statistical formulas and calculations were implemented using the MATLAB programming language (version 2019a, MathWorks, USA). The study device was compared to the PSG analysis outcome to compute the endpoint parameters. For all endpoint parameters, 95% confidence intervals were calculated. Confidence intervals were computed for proportion-based endpoints by approximating

the error distribution around a binomially distributed observation with a normal distribution. Significance levels were set at a p-value of 0.05.

Data synchronisation

Synchronisation between the study device and PSG was achieved by shifting and linearly stretching the PR trace of the PAT HSAT until the closest matching overlap with the PR trace of PSG was found. This allowed the synchronisation of all other traces since the hypnogram of PSG is precisely synchronised to the PR trace of PSG, as they rely on the same hardware clock, and since the PAT HSAT and the study device's Bluetooth data streams are synchronised by design. This method achieved a synchronisation error of less than 100 milliseconds. Data epochs of the study device with high lead impedance and Bluetooth disconnections were rejected from the analysis. Epochs from PSG were rejected whenever the sleep technician deemed the data unreliable for scoring.

Data adequacy

A test was considered technically adequate if at least 30 minutes of overlapping total sleep time between the study device and PSG could be obtained. This cutoff criterion of technical adequacy was employed to include patients' data in the dataset for which the endpoint parameters were determined. Other causes for a technical failure might be an early interruption of data acquisition due to acquisition software crashes or other connectivity issues with the data acquisition app. When the PSG recording was technically inadequate, for instance, when one of the channels could not be interpreted by the technicians, the patient was excluded from the analysis. Similarly, when PSG data or any annotations were missing due to administrative errors, the patient was removed from further analysis. Patients with missing patient characteristics, such as age and gender data, were omitted from the study of population demographic statistics but were still included in the endpoint statistics.

Performance endpoint selection

Since there are no standardised performance targets established for sleep stage determination, endpoints that are consistently reported on in other validation studies of wearable biopotential-based sleep staging devices were selected because they facilitate a transparent comparison between those studies. Starting from Imtiaz's 'A Systematic Review of Sensing Technologies for Wearable Sleep Staging'[125], we identified validation studies from peer-reviewed articles on wearable biopotential devices using at least EEG suited for at-home sleep staging analysis. As such, validation studies based solely on biopotential sensing modalities were included. Studies reporting on automated sleep staging that relied on single-channel data extracted from PSG itself, such as those proposed by Berthomier et al.[126] and Popovic et al.[127] were excluded. In-ear EEG studies were also excluded due to their fundamental differences from conventional EEG technologies, such as their proximity to reference and measurement electrode placement. Other device studies referenced in the study papers included in Imtiaz's review that fulfilled these requirements were also included. Finally, the dataset was extended by searching in Google Scholar using the search term 'EEG sleep staging validation.' We identified validation studies on six biopotential devices:

1. Zeo (Axon Labs, USA), three frontal electrodes, two studies available[128,129]

2. Dreem (Beacon Biosignals, USA), five electrodes (O1, O2, FpZ, F7, F8), one study available[130]

3. SOMNOwatch (SOMNOmedics, Germany), ten electrodes including EEG, EOG, and EMG, two studies available[131,132]

4. SleepProfiler (Advanced Brain Monitoring, USA), three frontal electrodes, two studies available[133,134]

5. Zmachine (General Sleep, USA), single-lead (A1-A2), one study available[135]

6. Cognionics study device (Cognionics, USA), two-lead (location unspecified)[136]

Robust endpoint parameters most consistently reported in these studies were stage-wise and overall accuracy, stage-wise and overall Cohen's Kappa (κ), and stage-wise sensitivity and specificity. Another commonly reported outcome is the sleep staging confusion matrices. When these parameters are not readily available, they can still be calculated if the study reports on confusion matrices. All included studies report on endpoint parameter values derived from the standard 30-second epoch hypnogram. Equally, the study device's sleep staging endpoint parameter values were derived from its 30-second epoch hypnogram. As sensitivity and specificity are 2-class parameters, they can only be calculated stage-wise. Accuracy and Cohen's kappa, by contrast, can be calculated either for individual sleep stages or across all stages combined. Cohen's Kappa is a statistic used to measure inter-rater reliability for categorical items. It is generally considered a more robust measure than the simple percent agreement calculation, as κ considers the possibility of the agreement occurring by chance[72]. 95% confidence intervals were calculated for each of the agreement measures.

Additionally, we compared parameters that are commonly derived from the sleep stages to the PSG analysis, specifically the TST, Sleep Efficiency (SE), Sleep Onset Latency (SOL), Wake After Sleep Onset (WASO), and time spent in each of the sleep stages.

The Total Recording Time (TRT) is the total sum of non-rejected epochs, and TST is the sum of those epochs spent in sleep. SE is then calculated as TST/TRT x 100 %. The SOL is the time from the start of the recording until the first three consecutive epochs of sleep. WASO is the time spent in wake after the SOL point. For each of these parameters and the time spent in each sleep stage (Wake, N1, N2, N3, REM), we calculated the mean and bias, together with their 95% confidence intervals. Bland-Altman analysis plots were used to set out the differences between PSG versus study device estimates for each parameter, while also indicating the mean difference and 95% limits of agreement.

It facilitates discussion of underestimation or overestimation biases of the study device and the variability of the accuracy of the study device estimates.

Sample size determination

Statistical power was determined by postulating that the maximum width of the 95% confidence interval of Cohen's Kappa of the least prevalent sleep stage (N1) should be at most 0.1. Assuming a baseline Cohen's Kappa parameter value of 0.3 and an equal rate of false positives and negatives, a minimum sample size of 20 patients was found under the conservative assumption that the average time spent in N1 per patient would be 15 minutes for an average total sleep time of 5 hours.

Algorithm

Data pre-processing

The raw biopotential data were bandpass filtered between 0.5 and 35 Hz and resampled from 256 Hz to 100 Hz to accelerate the overall processing time. Next, each patient's biopotential trace was normalised by subtracting the median and dividing by the interquartile range. Since the sleep staging model (SSM) makes use of a recurrent neural network (RNN) that expects a two-dimensional matrix as input, the biopotential signal was transformed into a time-frequency image X using a Short-Time Fourier Transform (STFT). As input parameters for the STFT, we have selected a Hamming window with 50% overlap. The window size was 2 seconds. Logarithmic scaling was used to obtain the log-power spectrum applied: $X_z = 20 \times log_{10}|X|$. These power spectrograms could then be divided in epochs to serve as input data to the deep learning models.

Neural network architecture

The central model architecture is an end-to-end hierarchical RNN adapted from Phan et al.[137] Its different layers can be summarised as follows:

1. A parallel filter bank designed to extract the most informative frequency sub bands.

2. An epoch-level bidirectional attention-based Recurrent Neural Network (RNN) for sequential feature extraction.

3. A sequence-level bidirectional RNN for longer-term modelling of sequences of epoch-based features.

4. A softmax layer that predicts probabilities for each of the input classes.

This approach enables an end-to-end learning strategy with a global optimisation solution while leveraging the sequential characteristics of the sleep data. A high-level overview of the neural network architecture and pre-processing steps is provided in Figure 29.

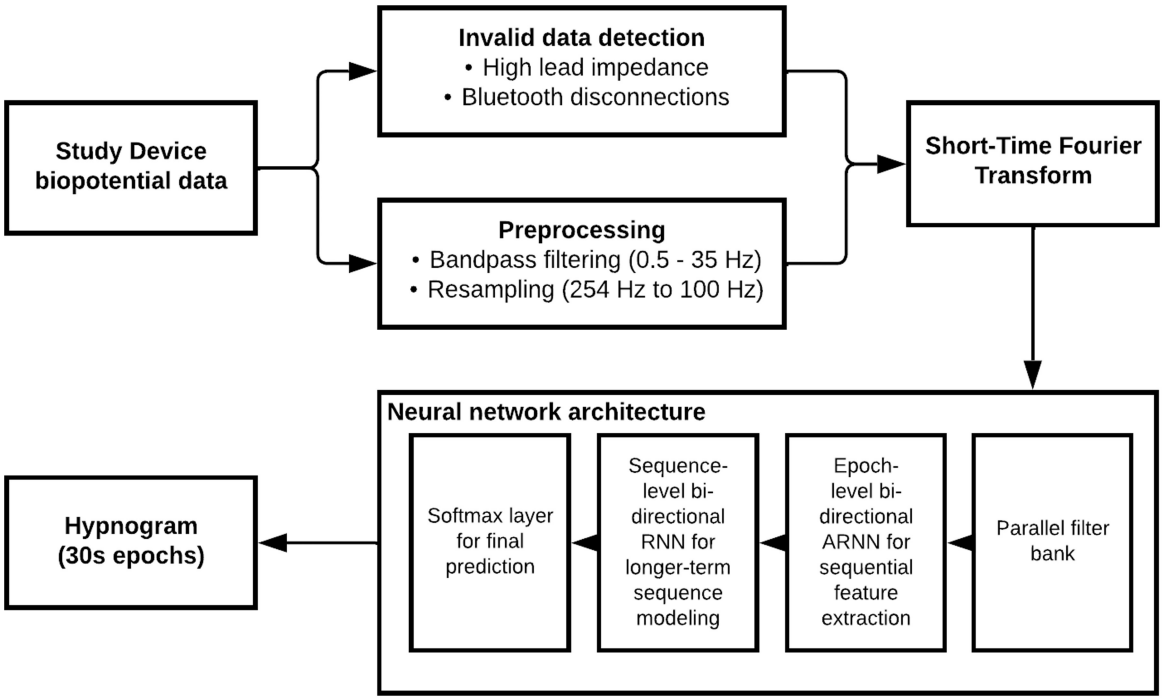

Figure 29: Overview of sleep staging model, including the preprocessing steps and neural network architecture.

Sleep staging model

Transfer learning

Due to the high number of parameters to be trained and the variability in training data, building performant deep neural networks usually requires large amounts of data. Transfer learning solutions can help improve the generalisability and training efficiency of these networks. With transfer learning, the goal is to transfer model knowledge from an already learned source task to the learning of a related target task[138].

To this end, a base model was first trained on a large source dataset. The trained model parameters then served as a starting point for further training and fine-tuning using the target dataset.

The Sleep Heart Health Study (SHHS), a multicentric cohort study to investigate cardiovascular and other effects of sleep-disordered breathing, was adopted as the source dataset[139,140]. The dataset

consists of labelled PSG data of 6441 individuals, including C3/M2 and C4/M1 EEGs at 125 Hz. A transfer learning base model using the described neural network architecture is trained on the C3/M2 channel of 200 patients. The base model training input data were extracted by transforming the EEG channel of each patient into its time-frequency spectrum using the SSM pre-processing settings and dividing the obtained power spectrum into 30-second epochs.

Training and cross-validation

The base model was further fine-tuned using the target study device dataset. Again, each patient's pre-processed biopotential channel was transformed into its power spectrum and divided into 30-second epochs. The power spectrum entries of rejected epochs were set to zero, and the corresponding labels were set to 'Invalid.' By introducing this additional label, the model automatically learned to deal with rejected epochs while still preserving the data's sequential nature and time dependence.

To maximise the utilisation of available data, training and evaluation were performed via a patient-wise 10-fold cross-validation. For each fold, the base model was fine-tuned using 90% of the available patients and was subsequently evaluated on the remaining 10%. This process was repeated ten times to obtain the total output for each patient independently. This final set of validation outcomes was then used to calculate the endpoint parameters.

Results

Patient inclusion and population statistics

Figure 30 shows the total number of recruited patients and a breakdown of included and excluded patients, together with the reason for failure. Out of the 133 patients who gave informed consent, concurrent PSG and study device data were successfully acquired for 106 patients. For 21 patients,

there was an issue with study device data acquisition due to Bluetooth connectivity instability. For two patients, a technical failure of PSG occurred, resulting in one or more uninterpretable channels. Due to administrative errors, no corresponding match between PSG and study device recording could be found for three patients. For one patient, the finger sensor device was applied with the optical sensor facing away from the finger, making it impossible to synchronise the data to the PSG recording. The demographic characteristics of the 106 included patients can be found in Table 15. The patients were predominantly male (57%), of middle age (mean 58 years, SD 15), and obese (mean BMI 34, SD 9). The mean AHI was 24 (SD 23). OSA classifications were evenly distributed, with 24% of patients having no OSA, mild OSA, or moderate OSA, and 27% having severe OSA. 28% of patients self-identified as black, 59% as white, and 43% as Hispanic, Latino, or Spanish.

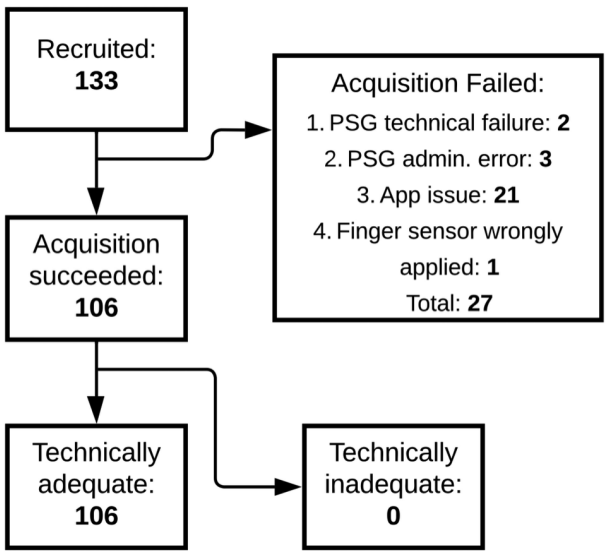

Figure 30: Flowchart of patient recruitment. This flowchart depicts how many patients gave informed consent (133), for how many patients PSG and study device data acquisition was successful (106), and for how many patients the study device data were technically adequate. Additionally, the flowchart summarises the reasons for data acquisition failure.

	Mean	SD	Min	Max
Age	58	15	22	82
BMI	34	9	18	68
AHI	24	23	0	114
	N (%)			
No OSA	26 (24%)			
Mild OSA	26 (24%)			
Moderate OSA	25 (24%)			
Severe OSA	29 (28%)			
Black	30 (28%)			
White	63 (59%)			
Other/No data	13 (13%)			
Hispanic/ Latino/Spanish	46 (43%)			
No data	1 (1%)			

Table 15: Demographic and clinical characteristics of patients in the dataset. N = number of participants; SD = standard deviation.

Endpoint parameters

Table 16 shows the sensitivity, specificity, accuracy, and Cohen's Kappa for each sleep stage, as well as the overall accuracy and Cohen's Kappa, including 95% confidence intervals. These endpoints were listed for both the study device and for each of the validation studies found in the performance endpoint selection section. Figure 31 shows the same Cohen's Kappa values from Table 2 to simplify comparing the devices' relative performance. For the Dreem, Cognionics, Zeo 2, and Zmachine studies we calculated all endpoint parameter values from the confusion matrix. Endpoint parameters for SleepProfiler 1 and Zeo 1 were taken directly from the paper when available. Several studies did not report on N1 and N2 separately but combined these stages into light sleep. For easier comparison, the other studies' overall and light sleep agreement parameters were also calculated by

combining the N1 and N2 stages. The other stage-wise agreement parameters (Wake, N3, and REM) were unaffected. The SleepProfiler 2 study was excluded because the authors combined N2 and N3, making a granular comparison to other studies impossible. Both SOMNOwatch studies were excluded because they did not report on stage-wise agreements and did not provide any data that allowed calculating the overall accuracy parameters. Parameters reporting on all five and four stages with N1 and N2 combined were respectively called *5-stage* and *4-stage* classification parameters.

Device	Stage	Sensitivity	Specificity	Accuracy	Kappa	# Rec.	Ref.
Study Device	Overall 5-stage	NA	NA	0.797 ± 0.003	0.696 ± 0.003	106	NA
	Overall 4-stage	NA	NA	0.820 ± 0.003	0.708 ± 0.003		
	Wake	0.842 ± 0.005	0.938 ± 0.002	0.915 ± 0.002	0.765 ± 0.003		
	N1	0.472 ± 0.015	0.971 ± 0.001	0.945 ± 0.002	0.444 ± 0.003		
	N2	0.815 ± 0.004	0.875 ± 0.003	0.844 ± 0.003	0.688 ± 0.003		
	N3	0.758 ± 0.011	0.951 ± 0.002	0.936 ± 0.002	0.615 ± 0.003		
	REM	0.795 ± 0.008	0.976 ± 0.001	0.954 ± 0.001	0.780 ± 0.003		
	Light Sleep	0.824 ± 0.004	0.849 ± 0.004	0.835 ± 0.003	0.666 ± 0.003		
Dreem	Overall 5-stage	NA	NA	0.810	0.727	25	[130]
	Overall 4-stage	NA	NA	0.840	0.747		
	Wake	0.776	0.972	0.947	0.757		
	N1	0.476	0.961	0.930	0.431		
	N2	0.827	0.899	0.863	0.726		
	N3	0.878	0.960	0.950	0.783		
	REM	0.860	0.947	0.931	0.781		
	Light Sleep	0.840	0.870	0.853	0.705		
SleepProfiler 1	Overall 5-stage	NA	NA	NA	NA	14	[133]
	Wake	0.442	0.962	0.920	0.420		
	N1	0.270	0.906	0.890	0.070		
	N2	0.718	0.833	0.780	0.550		
	N3	0.603	0.960	0.880	0.620		
	REM	0.724	0.885	0.850	0.580		
Cognionics	Overall 5-stage	NA	NA	0.738	0.636	12	[136]
	Overall 4-stage	NA	NA	0.792	0.681		
	Wake	0.816	0.944	0.915	0.757		
	N1	0.280	0.933	0.887	0.199		
	N2	0.738	0.905	0.826	0.648		

	N3	0.868	0.958	0.948	0.761		
	REM	0.734	0.921	0.899	0.574		
	Light Sleep	0.778	0.874	0.822	0.644		
Zeo 1	Overall 4-stage	NA	NA	0.726	0.560	21	[128]
	Wake	0.408	NA	NA	NA		
	Light Sleep	0.800	NA	NA	NA		
	N3	0.621	NA	NA	NA		
	REM	0.737	NA	NA	NA		
Zeo 2	Overall 4-stage	NA	NA	0.811	0.699	26	[129]
	Wake	0.639	0.980	0.929	0.689		
	Light Sleep	0.865	0.932	0.919	0.750		
	N3	0.862	0.824	0.845	0.687		
	REM	0.710	0.959	0.930	0.661		
Zmachine	Overall 4-stage	NA	NA	0.822	0.716	99	[135]
	Wake	0.843	0.977	0.948	0.842		
	Light Sleep	0.848	0.806	0.829	0.655		
	N3	0.782	0.978	0.963	0.740		
	REM	0.732	0.940	0.904	0.668		

Table 16: Tabulation of endpoint parameter values and their confidence intervals (after the '±' symbol) for the study device and the included similar devices. Kappa = Cohen's Kappa; NA = not available; Ref. = reference.

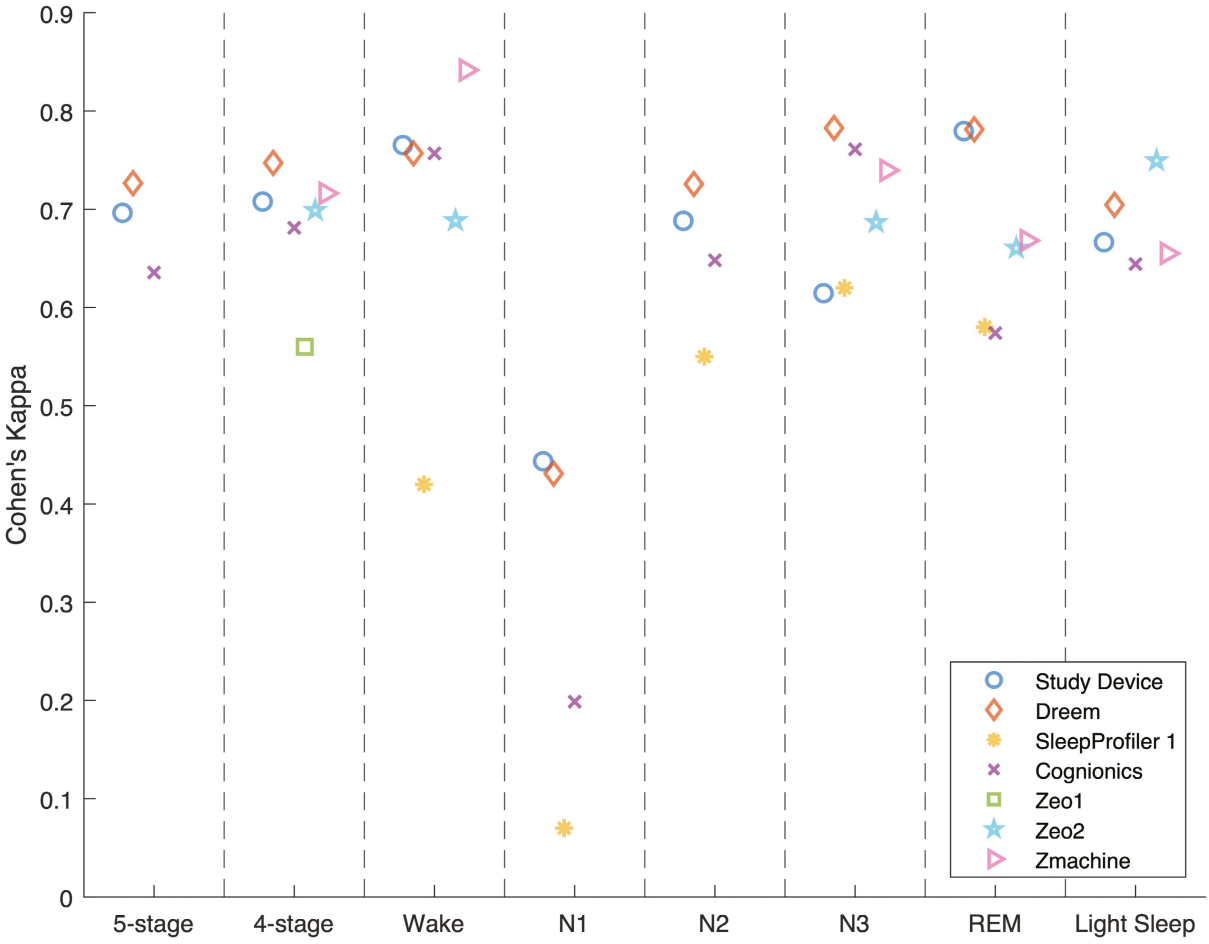

Figure 31: Classification performance plot showing the Cohen's Kappa for five-stage and four-stage classification and each sleep stage separately. Each marker type represents the same study for the categorisations labelled on the *x*-axis.

The study device reported overall κ of 0.696 and 0.708 for 5-stage and 4-stage classification, respectively. The Dreem study was the only one reporting statistically significantly higher κ than the study device for both 5-stage and 4-stage classification, with κ of 0.727 and 0.747, respectively.

With κ ranging from 0.070 (SleepProfiler 1) to 0.444 (study device), it can be observed that N1 was the most difficult sleep stage to classify. This is consistent with general findings from the scientific literature[141]. As such, the N1 κ reported for the study was considerably lower than the κ of other stages (of which N3 was the second lowest with a κ of 0.615).

Unlike all other studies, the study device's κ for REM (0.780) was higher than its κ for N3 (0.615).

Across all studies, it could be observed that, except for N3 for the Zeo 2 study, the specificity was higher than the sensitivity for all five individual sleep stages. For all studies, the specificity for N1 exceeded 0.9, while the sensitivity ranged from 0.270 (SleepProfiler 1) to 0.4765 (Dreem). This can be explained by the lowest prevalence and, hence, the lowest prior probability of encountering an N1 epoch of all five stages, the typically short duration of N1 stages, and the documented difficulty in scoring N1[141]. As a result, autoscoring systems and human scorers will frequently incorrectly score N1 as either Wake or N2.

Out of all studies, our study device had the highest statistical power, with 106 successfully included patients, followed by Zmachine, with 99 included patients.

The mean, bias, and Pearson correlation of the aggregated sleep continuity and sleep architecture parameters are set out in Table 17. The results indicate a low bias for each of the estimated parameters, apart from N3, which the study device overestimates by about 30% compared to PSG analyses. The correlation coefficient of N3 was the lowest of the listed parameters, which is consistent with the findings from Table 16. The Bland-Altman plots in Figure 32 and Figure 33 provide more detail to these findings. The confusion matrix, from which the endpoints in Table 16 were derived, is provided in Table 18.

	PSG mean (min)	Study Device mean (min)	Bias (min)	ρ
TST	296.09 ± 17.64	291.49 ± 17.07	4.60 ± 6.15	0.938 (0.910 – 0.957)
SE	0.762 ± 0.039	0.750 ± 0.037	0.012 ± 0.017	0.905 (0.864 – 0.935)
SOL	33.33 ± 5.96	26.55 ± 5.48	6.78 ± 3.20	0.847 (0.783 – 0.893)
WASO	62.82 ± 12.32	73.94 ± 12.81	-11.12 ± 5.60	0.901 (0.858 – 0.932)
Wake	89.69 ± 14.75	94.29 ± 14.17	-4.60 ± 6.15	0.910 (0.871 – 0.938)
N1	20.10 ± 3.21	20.03 ± 2.99	0.07 ± 2.19	0.753 (0.657 – 0.825)
N2	199.14 ± 13.86	185.76 ± 14.33	13.38 ± 7.99	0.840 (0.772 – 0.888)
N3	30.09 ± 5.51	40.21 ± 7.40	-10.11 ± 5.55	0.667 (0.546 – 0.761)
REM	46.75 ± 6.38	45.49 ± 7.10	4.76 ± 0.76	0.755 (0.660 – 0.827)

Table 17: Displays all endpoint parameter values and their confidence intervals (after the '±' symbol or between brackets) for the study device and PSG analysis. ρ= Pearson correlation coefficient; SE = sleep efficiency; SOL = sleep onset latency; WASO = wake after sleep onset.

		Study Device				
		Wake	N1	N2	N3	REM
PSG	Wake	**16017**	1240	1331	57	370
	N1	964	**2010**	1084	18	186
	N2	2329	785	**34417**	3529	1157
	N3	93	5	1395	**4836**	51
	REM	587	207	1154	84	**7880**

Table 18: Displays the confusion matrix comparing the sleep staging category determined by the PSG analysis to the study device.

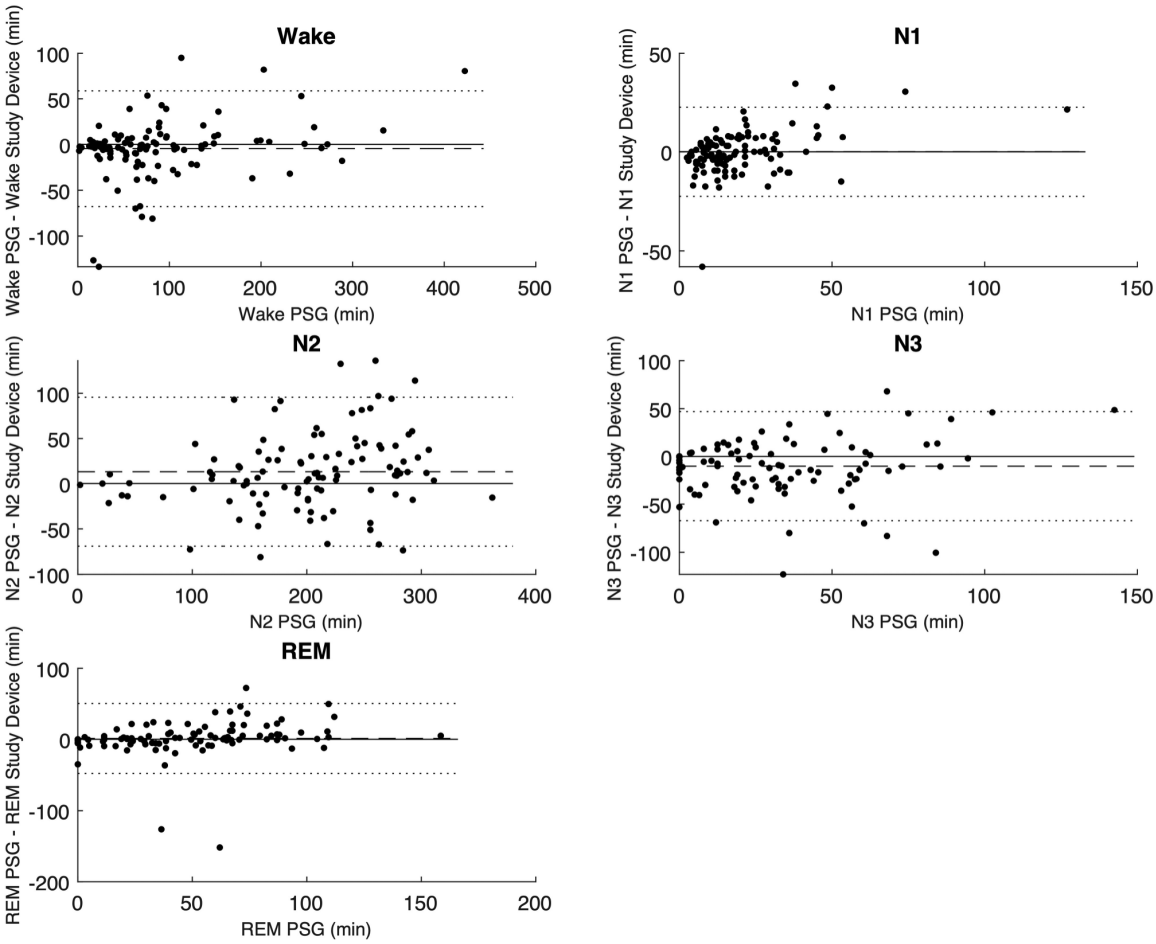

Figure 32: Bland-Altman plots comparing time spent in the five sleep stages between the PSG analysis and the study device. The dotted lines represent the Limits of Agreement. The dashed line represents the mean difference (bias).

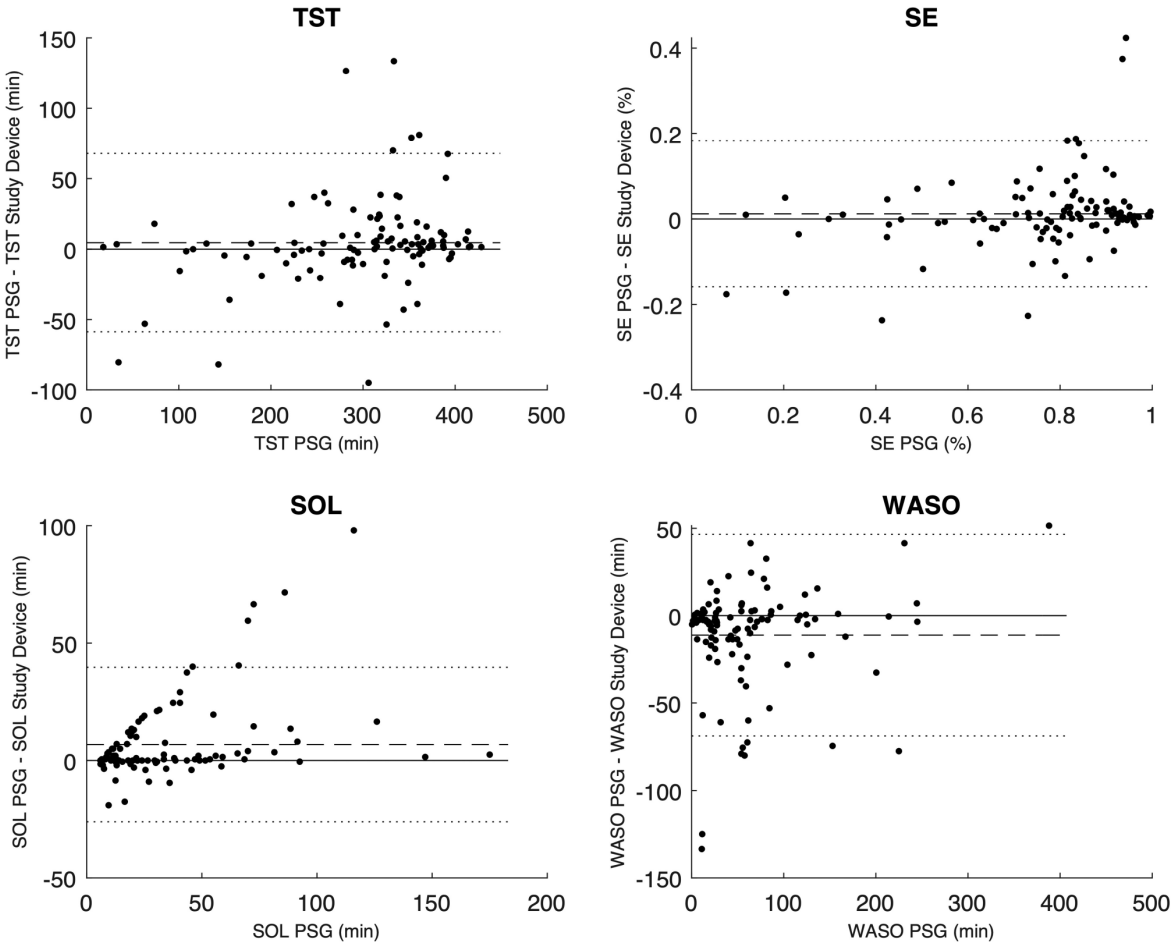

Figure 33: Bland-Altman plots comparing the TST, SE, SOL, and WASO between the PSG analysis and the study device. The dotted lines represent the Limits of Agreement. The dashed line represents the mean difference or bias. SE = sleep efficiency; SOL = sleep onset latency; WASO = wake after sleep onset.

Discussion

Endpoint analysis highlights and discussion

The most notable discrepancy from the other studies was the study device's higher REM detection accuracy compared to N3. The other studies reported either higher accuracy for N3 than for REM, or similar accuracies for both stages. This can be explained by the specific choice of electrode placement of the study device. We chose to place the electrodes diagonally across the lateral cross-

section of the eye so that the direction of rapid eye movements would be clearly reflected in the signal as deflections of opposite polarity, thereby greatly improving REM sleep detection. Because this placement precluded the use of the conventional M1/M2 (mastoid) electrode positions, the resulting delta-wave amplitude was reduced, which in turn diminishes N3 detectability. However, this setup enhances the capability to differentiate between N3 and REM sleep when using only a single-lead setup, improving the overall 5-stage performance compared to alternative electrode placements.

The results of our study demonstrate similarity in overall sleep staging performance and the trade-off of specificity and sensitivity for each of the individual sleep stages when compared to the Dreem device, which achieved the highest overall 5-stage and 4-stage κ. In terms of 5-stage and 4-stage classification performance, the study device ranked third among the included studies (with Sleep Profiler 1 reporting on neither the 5-stage nor 4-stage classification). Only the Dreem device's 5-stage and 4-stage κ values were significantly higher. The outperforming Dreem device uses five electrodes positioned above the hairline. By contrast, the study device was deliberately designed as a compact, low-burden system that avoids above-hairline placement to minimise patient discomfort. The electrode configuration was specifically optimised to maximise overall 5-stage classification accuracy under the strict form-factor requirements.

Strengths and limitations of the study

Our study has some strengths and limitations that warrant discussion. The first strength of this study is its adequately powered multicentric design, resulting in a population size of 106 patients. The Zmachine study discussed in this paper reported on 99 recruited patients, whereas the other studies recruited between 12 and 44 patients. The second strength of this study is its inclusion of a comparison of the study device's sleep staging performance to that obtained from clinical studies of comparable devices.

The first limitation is the inclusion of patients with suspected OSA. While this population matches the primary intended-use population for home sleep testing, assessing the device's performance in individuals with a low pre-test probability of OSA and in populations with a higher prevalence of non-respiratory sleep disorders would also be valuable. Such cohorts often present with sleep fragmentation, atypical arousal patterns, periodic limb movements, insomnia-related cortical hyperarousal, parasomnias, or circadian rhythm disturbances which can alter the distribution or stability of sleep stages. This may challenge automated sleep staging or REM-detection algorithms that rely on a limited single-channel signal. Evaluating performance in these groups would therefore clarify the robustness of the algorithm to biopotential patterns unrelated to respiratory pathology. The second limitation of the study is the relatively large device failure rate due to resolvable instabilities in the Bluetooth connectivity of the prototype study device. The third limitation is the single-scored nature of the reference PSG data. Scoring the PSG data by multiple qualified sleep scorers, ideally from independent centres, could reveal the inter-scorer agreement on the PSG data to which the study device's hypnogram can subsequently be compared. Additionally, bundling multiple scorers' hypnograms, for example, by using a majority voting scheme for each epoch, may serve as a more optimal training dataset, which may further enhance the device's performance. Finally, like PSG, the study device relies on EEG, EMG, and EOG data for sleep staging. As the device typically operates as an optional extension to a PAT HSAT, it would be valuable to explore whether including PAT, SpO_2, and actigraphy of the PAT HSAT, which are not traditionally used for sleep staging, could further enhance the performance.

Conclusion

This multi-site clinical evaluation was conducted to assess the study device's capability to automatically determine sleep stages. Comparison with similar devices evaluated in prior clinical studies indicates favourable performance, and only the five-channel Dreem device demonstrated

significantly higher accuracy. These findings support the feasibility of incorporating a single-lead biopotential extension, which uses a performance- and comfort-optimised electrode placement, into an HSAT setup. Such an extension could broaden clinical applicability, improve the accuracy of sleep stage-derived parameters such as total sleep time (thereby enhancing AHI estimation), and enable more detailed phenotyping of sleep apnea. As such, this enhanced HSAT configuration could narrow the gap between reduced-channel HSAT and in-lab PSG while maintaining patient comfort and ease of use.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	The evaluation of a novel single-lead biopotential device for home sleep testing.
Publication status:	*Published*
Publication details:	Massie F, Vits S, Verbraecken J, Bergmann J. The evaluation of a novel single-lead biopotential device for home sleep testing. Sleep. 2025 Apr 11;48(4):zsae248. doi: 10.1093/sleep/zsae248. PMID: 39441980; PMCID: PMC11985384.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed sleep staging algorithm development and evaluation, clinical study design, writeup, and data acquisition, and publication design and writeup. S vits, B Van Pee, and J Verbraecken provided publication editing. J Bergmann provided supervisory guidance and publication editing.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Chapter 5: Rendering PAT HSAT autoscoring context-aware

Research Aim 4.1: Investigating context-aware PAT HSAT autoscoring

Published as: Massie F, Vits S, Verbraecken J, Bergmann J. Context-aware analysis enhances autoscoring accuracy of home sleep apnea testing. *Journal of Clinical Sleep Medicine*. 2025 May 1;21(5):789-804. doi: 10.5664/jcsm.11534.

Reiteration of research aim and introductory context

Current respiratory event scoring guidelines for PSG as well as HSAT autoscoring algorithms do not explicitly consider the wealth of information that can be derived from the multitude of physiological phenomena contained within signal episodes occurring before and after the event that is being scored (the event context). This is especially relevant since respiratory events tend to occur in highly repetitive patterns with similar morphological characteristics[45,46]. In the paper introduced in the following sections, we report on the performance of our context-aware PAT HSAT autoscoring algorithm, which addresses these scoring limitations and further reduces the performance gap to PSG.

Abstract

Study objectives

PAT HSAT is increasingly being deployed because of its ability to test for multiple nights. However, these tests do not have access to modalities such as airflow and cortical arousals and instead rely on alternative sources of information to detect respiratory events. This results in an a-priori performance

disadvantage. In this study, we describe the Panorama algorithm, which aims to reduce this disadvantage by incorporating information from characteristically repetitive sequences in physiological changes associated with respiratory events. These include changes in PAT, pulse rate, and oxygen saturation. The method was designed to facilitate manual review by providing the scoring rationale for each respiratory event.

Methods

The method was developed and evaluated using a dataset of 266 participants from a multicentric cohort suspected of having OSA. All participants underwent simultaneous PSG and PAT HSAT, and all PSG data were double-scored. Scoring was performed according to the 3% and 4% rules for hypopnea scoring. Clinical endpoint parameters, including the OSA severity categorisation accuracy and Cohen's Kappa, were selected to compare the algorithm to a conventional context-unaware algorithm. Data analysis and reporting followed the TRIPOD+AI[142] reporting guidance for prediction models that use machine learning.

Results

Regarding OSA severity categorisation accuracy, the Panorama algorithm outperformed context-unaware autoscoring by 9% using 3% rule scoring and 7% using 4% rule scoring.

Conclusions

The context-aware method improves the performance of PAT HSAT while still facilitating scoring review by providing event-specific scoring rationale.

Introduction

PAT HSAT is increasingly being deployed because of its ability to test for multiple nights due to its low cost, low-touch setup, and availability in an entirely disposable format[24,26,27,66]. A primary disadvantage of PAT HSAT compared to polysomnography, the gold standard method for diagnosing sleep apnea, is its lack of an airflow channel and assessment of sleep by EEG[25]. While it can measure oxygen desaturations from its oximetry channel, its reduced-channel design does imply that PAT HSAT must rely on alternative sources of information to indirectly infer whether the patient experiences an airflow reduction or cortical arousal to detect both apnea and hypopnea accurately. This burdens PAT HSAT with an a-priori performance disadvantage. Traditionally, only information contained within one isolated event episode of the recording is considered when scoring the event[36]. As such, PAT HSAT devices do not explicitly consider the information contained within neighbouring signal episodes (the event context)[44]. This is especially relevant since respiratory events frequently occur in highly repetitive patterns with similar morphological characteristics (see Figure 34)[46,143]. To understand how increasing the scoring field of view may improve accuracy, consider a hypopnea with associated cortical arousal but an oxygen desaturation of less than 3%. A PAT HSAT cannot directly rely on airflow reduction and cortical arousal information, making these events challenging to score. Investigating whether this event occurs in a repetitive sequence of clear respiratory events or isolation might help determine whether this event is truly a hypopnea or, for instance, a motion artefact. Although a human scorer will inevitably consider, to some degree, the event's context without this being stated in the scoring rules, the same does not hold for a context-unaware computer algorithm.

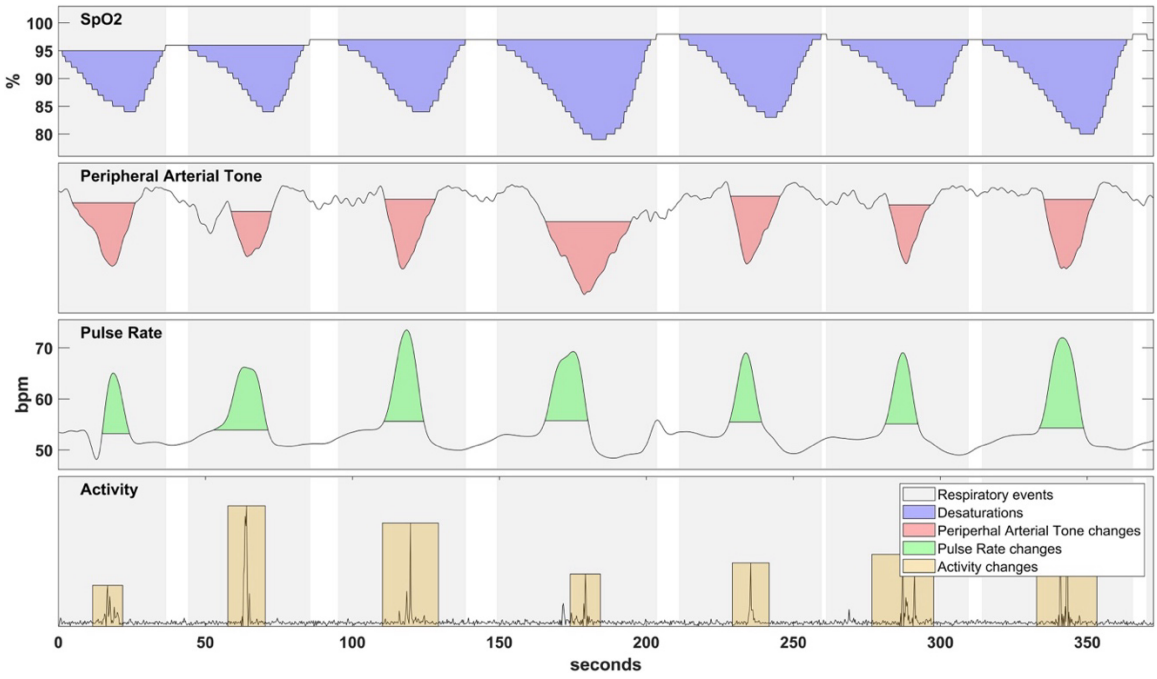

Figure 34: A sequence of respiratory events. Repetitive blood oxygen desaturations, PAT amplitude decreases (reflecting vasoconstrictions), pulse rate surges, and activity surges (due to sudden movement after autonomic arousal) that characterise a sequence of respiratory events. Each signal-type event is indicated in a different colour. The grey bars mark the area around which a respiratory event is occurring (as scored by PSG).

While incorporating event context can potentially increase (auto)scoring accuracy, the resulting algorithm might become so complex as to become a 'black box,' meaning its inference method becomes unintelligible to humans. However, as is the subject of this study, such issues can be resolved by restricting the choice of inference algorithms to human-interpretable models such as decision trees. For each event, this decision-making process can be rendered easily understandable and, more importantly, can be displayed to the human reviewer; this type of autoscoring may facilitate manual review rather than complicate it. To address the limitations mentioned above of the state-of-the-art of respiratory event scoring, we propose the Panorama algorithm for autoscoring PAT HSAT. This decision-tree-based algorithm incorporates extensive event context information to reduce the gap between PAT HSAT and the gold-standard PSG while simultaneously facilitating manual review by providing a concise human-interpretable rationale for why each event was scored. To illustrate the performance gains that can be achieved by rendering PAT HSAT autoscoring

'context-aware,' we will compare the performance of the Panorama algorithm to an alternative algorithm that scores the same HSAT data in a conventional context-unaware fashion, using a dataset of 266 patients acquired from clinical studies.

Methods

Working principles of Panorama

PAT-based algorithms are, congruent with the definition provided by the AASM, based on the concurrent analysis of changes in PAT, pulse rate[144], and SpO_2 (sometimes complemented with limb movement analysis based on actigraphy[145])[1]. In its most straightforward implementation (see Figure 35 and Figure 36 A), an event is scored by looking only at physiological changes directly associated with that event[44]. Panorama expands on this concept and incorporates temporal and morphological information about these physiological changes, as well as those occurring in the immediate and more remote vicinity of the event. As introduced earlier, it has been well described how respiratory events tend to occur in repetitive patterns with relatively stable periodicity[45,143]. In addition, there tends to be morphological similarity between physiological changes associated with each event within such repetition (for example, consecutive oxygen desaturations tend to show a similar rate of desaturation and resaturation)[146]. Incorporating a description of the temporal and morphological characteristics of such repetitive behaviour could increase the performance of PAT HSAT autoscoring. In what follows, we will introduce the working principles of the algorithm, which incorporates such information via six main processing steps. This overview is not meant to be a technically exhaustive algorithm description. Instead, it aims to explain its core working principles.

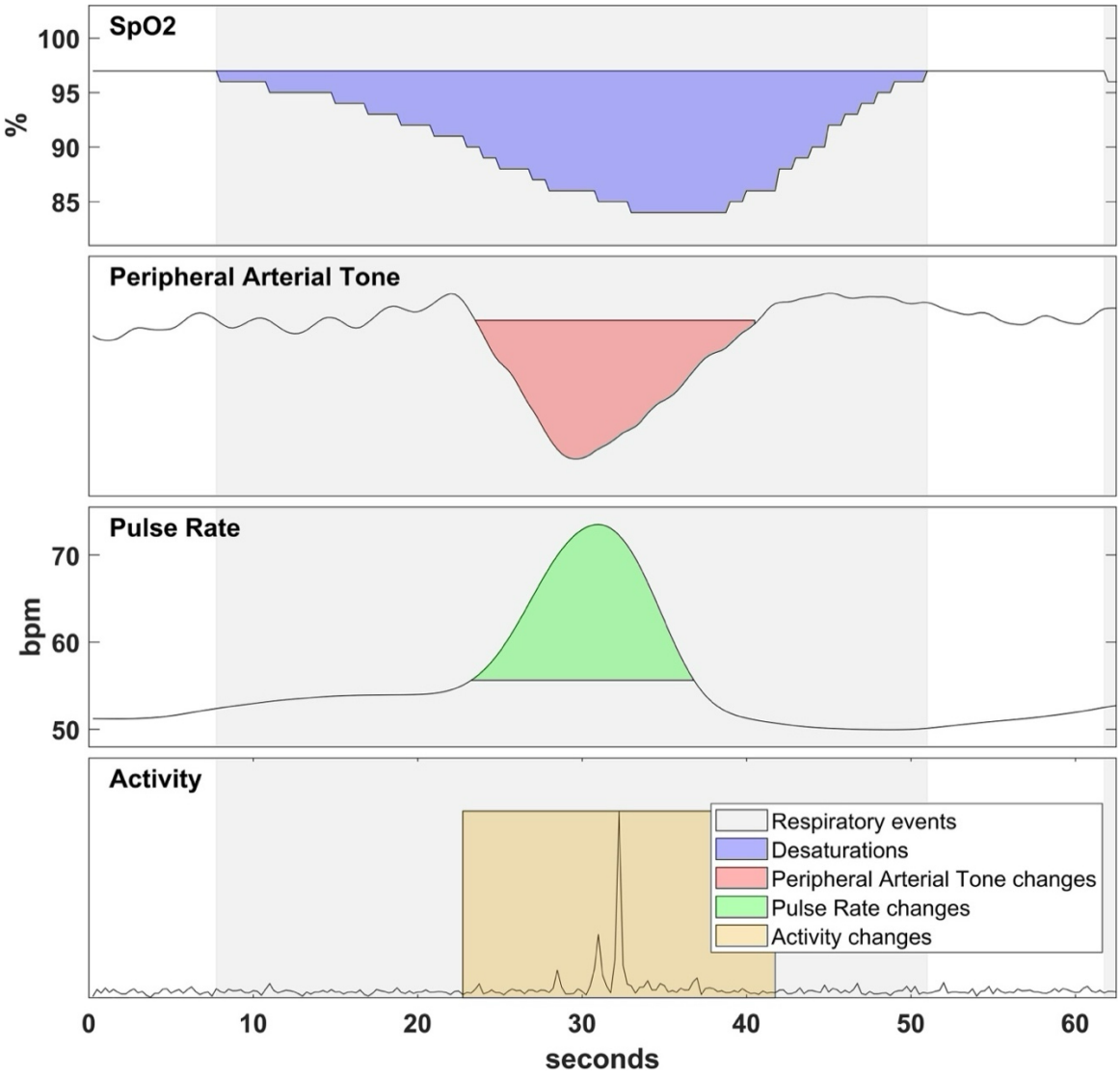

Figure 35: A close-up of a single respiratory event. A respiratory event (occurring around the location of the grey bar) is traditionally scored by detecting an oxygen desaturation alongside the concurrent increase in pulse rate and reduction in the PAT signal, signalling vasoconstriction, possibly also accompanied by an activity surge due to sudden movement after autonomic arousal. Each signal event type is indicated in a different colour.

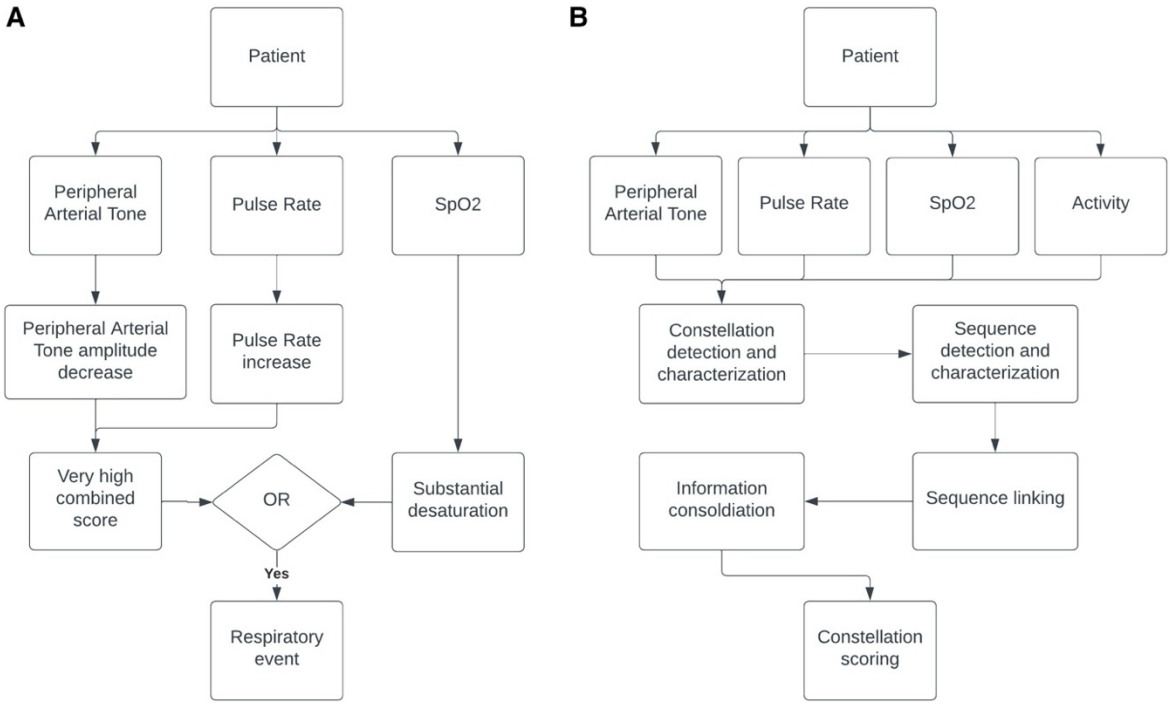

Figure 36: Event scoring flowcharts of different methods. (A) Regular context-unaware flowchart of respiratory event scoring by means of PAT HSAT. (B) The main steps of the context-aware scoring by the Panorama algorithm.

Step 1: Marker event detection

As introduced earlier (Figure 35), PAT HSAT analyses characteristic respiratory-event-related changes in pulse rate, PAT, SpO_2, and optionally limb movement (actigraphy). Such short-lived characteristic changes in a signal measured by PAT HSAT are what we refer to as marker events. Before any respiratory event scoring can be performed, we determine the location of all marker events as they comprise the building blocks for further inference. The black rectangular boxes $C_{2,SpO2}$, $C_{2,P}$, $C_{2,PR}$, and $C_{2,act}$ highlighted in Figure 37 provide an example of the four marker event types analysed by Panorama: an oxygen desaturation, a decrease in PAT signal amplitude, an increase in pulse rate and an increase in limb activity.

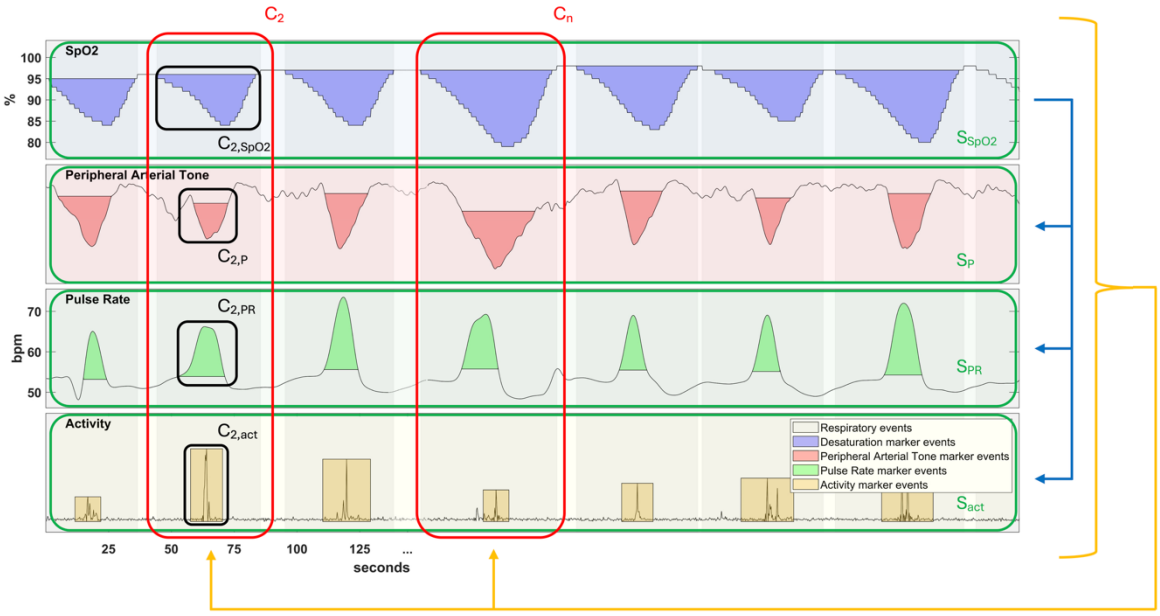

Figure 37: An illustration of the working principles of Panorama based on a sequence of respiratory events. *Step 1*, marker event detection, is indicated by the black boxes $C_{2,SpO2}$, $C_{2,P}$, $C_{2,PR}$, and $C_{2,act}$. Each box represents the respective SpO_2, PAT, pulse rate, and activity marker events for constellation number 2. *Step 2*, constellation detection and characterisation, is indicated by the red boxes C_2 and C_n, representing the constellations 2 and n. Each constellation undergoes morphological and temporal characterisation. This is done on individual marker level by using features such as desaturation depth, but also on the constellation level, for example by characterising the relative position of each marker event within the constellation. *Step 3*, marker event sequence determination, is indicated by the green boxes S_{SpO2}, S_P, S_{PR}, and S_{act}. Here, all sequences of marker events are again characterised in a similar fashion to single-event constellations, for example, by calculating the average desaturation depth of S_{SpO2}. *Step 4*, sequence linking and characterisation, is indicated by the blue arrows. Here, synchronisation characteristics are determined, including the degree of alignment of characteristic points in the sequences. Finally, in *Step 5*, information consolidation is indicated by the yellow arrows. This step links all the information gathered in the previous steps to the individual constellations to determine the final feature set.

Step 2: Constellation determination and characterisation

Determination

After detecting all marker events, they are grouped into what is further referred to as constellations. Constellations are groups of marker events that are likely associated with a single respiratory event. For example, the red box C_2 represents a constellation comprising an oxygen desaturation (marker

event $C_{2,SpO2}$), which coincides with a PAT amplitude reduction (marker event $C_{2,P}$), a pulse rate increase (marker event $C_{2,PR}$), and an increase in activity (marker event $C_{2,act}$). These four marker events can all be associated with one respiratory event, forming a single constellation. Each constellation is considered a 'respiratory event candidate' given that its marker events are a-priori plausibly manifestations of a single respiratory event. The information gathered in the following steps will determine whether each constellation should eventually be scored as a respiratory event.

Characterisation

After all constellations have been determined, they each undergo morphological and temporal feature characterisation. Table 19 lists the categories of features calculated during the constellation characterisation step as well as during the other algorithm steps, which are introduced in the subsequent paragraphs. Morphological characterisation means that several morphological features are determined for each marker event within the constellation. For oxygen desaturations, this includes the depth of the desaturation and the slope asymmetry between the desaturation and the resaturation phase. For the other marker events, this comprises similar characterisations, such as the relative increase or decrease of the event magnitude compared to the pre-event baseline value.

In addition to morphological characterisation, the constellation undergoes temporal characterisation. This means the relative position is determined for all marker events within the constellation and their distance to the neighbouring constellation's marker events of the same type. For example, it is determined by how many seconds the oxygen desaturation trails or precedes the peak of the pulse rate increase and PAT amplitude decrease. During temporal characterisation, the degree of alignment between the pulse rate local maximum and PAT local minimum is also considered. A near perfect alignment is an indicator of an autonomic arousal, a sympathetic activation that often occurs near the end stage of a respiratory event[21].

Algorithm step	Marker event/sequence type	Feature categories
Constellation characterisation	All marker event types (including desaturations)	*Temporal features* • Distances between peak/valley of marker events within constellation • Distance to neighbouring constellations' marker events of the same type *Morphological features* • Marker event magnitude/depth • Duration of marker events *Combination features* • First and second order statistics of all features of marker event and those of neighbouring constellations (e.g. the interquartile range of the depth of the desaturation and the depth of the desaturations of the four nearest neighbouring constellations on either side of the constellation)
	Desaturations	*Morphological features* • Asymmetry of desaturation shape
Sequence characterisation	All sequence types (including desaturation and activity sequences)	*Temporal features* • Sequence duration • Number of marker events in sequence *Combination features* • First and second order statistics of morphological and temporal features of each marker event within the sequence (e.g. the average duration of the PR marker events within the sequence)
	Desaturation sequences	*Morphological features* • Morphological similarity of desaturations within sequence • Percentage counts of desaturations within sequence of increasing minimum depth (e.g. the percentage of desaturations with a magnitude of at least 3% within the sequence)
	Activity sequences	*Morphological features* • Sequence sparsity (quantifies presence of concentrated areas of activity intensity within sequence) • Sequence energy (total activity intensity normalised by sequence duration)
Linked sequence characterisation	All linked sequence types	*Temporal features* • Temporal alignment statistics of the two types of marker events within the linked sequences (e.g. the median difference in desaturation minimum and PR peak locations of the linked desaturation and PR sequence) *Combination features*

		• First and second order statistics of morphological and temporal features of each marker event within the linked sequence (e.g. the median magnitude of the PR marker events within the linked PR and PAT sequence)
Information consolidation	Calculated for each constellation	• Aggregation of all features from marker events, sequences and linked sequences associated with each constellation

Table 19: Overview of feature categories derived in the different steps of the Panorama algorithm. Temporal features are related to the alignment of marker events, sequences, and linked sequences. Morphological features are related to the duration, shape, and intensity of marker events, either considered in isolation or within a sequence. Combination features contain both temporal and morphological information. IQR = interquartile range; PR = pulse rate.

Step 3: Marker event sequence determination and characterisation

Determination

As introduced in the beginning of this chapter, respiratory events, and by implication, their associated marker events, tend to occur in repetitive sequences. Detecting such marker event sequences is the central part of *Step 3* of the algorithm. Marker event sequences are determined by detecting chains of marker events that are sufficiently proximate to one another. The sequence ends as soon as a discontinuity of the marker events is found. A discontinuity is determined either by a sudden change in marker event intensity or when the next marker event is more than 90 seconds removed from the last event in the sequence. Figure 37 highlights such a sequence of respiratory events. The grey bars highlight areas around which a respiratory event occurred. The green rectangles delimit the location of sequences for each marker event type. It is important to note that the detection of sequences is not restricted by any time window. Sometimes, a sequence of respiratory events and their associated physiological markers can persist uninterruptedly for multiple hours.

Characterisation

After determining all sequences, they are morphologically and temporally characterised similarly to single-event-associated constellations. For example, for each sequence, the typical period between two neighbouring marker events is characterised, as well as the stability of this period. Similarly, the average event morphology of the marker events within a sequence is characterised, as is the stability of such morphology. A highly repetitive (i.e., having a highly stable period) sequence of marker events of a very similar morphology has a high likelihood of being associated with a sequence of respiratory events[46,143]. This concept can be appreciated from Figure 37, which highlights the morphological similarity and stable cadence of neighbouring marker events. Table 19 expands on the different categories of sequence characterisation.

Step 4: Sequence linking and characterisation

Linking

Once all sequences for each marker event type have been determined, overlapping sequences of different marker event types are defined. The blue arrows in Figure 37 highlight the linking of overlapping sequences of all four marker event types.

Characterisation

Relevant information is contained in the synchronisation between sequences of different marker event types. For example, a pulse rate sequence for which all pulse rate peaks near perfectly align with the valleys of the PAT amplitude reductions renders it more likely that all marker events in those sequences are manifestations of (a sequence of) respiratory events. In Figure 37, one can appreciate how the local minima or maxima of the marker events within the pulse rate and PAT sequences almost perfectly align, while the oxygen desaturation minimum location trails with a relatively constant delay. Table 19 expands on the different categories of linked sequence characterisation.

Step 5: Information consolidation

The stages up to this step comprise the detection of all marker events (*Step 1*), clustering of marker events into constellations and characterisation thereof (*Step 2*), detection of sequences of marker events and characterisation thereof (*Step 3*), and, finally, detection of overlapping sequences of marker events and characterisation thereof (*Step 4*). Before the final scoring of each constellation, we compile all characterisation information for the constellation, its associated sequences, and any linked sequences to which it belongs. This phase of linking all related information to the constellation is called information consolidation. This set of information, which is structured in a total set of 350 features, can then be used to inform the scoring of each individual constellation.

Step 6: Constellation scoring

For each constellation, the complete set of information consolidated in *Step 5* is used to infer whether it corresponds to a respiratory event. To this end, a decision tree was trained by means of supervised machine learning on a large dictionary of constellations for which the ground truth event scoring by means of independently double-scored PSG was available. Details about the dataset are provided in the following paragraphs. A (nested) leave-one-patient-out cross-validation (LOPOCV) method was used to safeguard against cross-contamination of training and test data and, as such, against the introduction of algorithm bias[147–149]. This means that the model was trained and tested as many times as there were participants included in the dataset (266 times). More specifically, to test the model for one participant, it was first trained on the 265 other participants. This process was repeated until the model was evaluated on all 266 participants. This round-robin training and testing method ensures that the model was never trained and tested on data from the same participants, while maximally utilising available data. A summary of all six steps presented above is provided in Figure 36 B of the introduction section.

The model is designed for explainability, i.e., to permit a straightforward review of the scoring rationale. For each constellation, the Panorama algorithm derives an event-specific scoring rationale directly from the same tree-based ensemble that performs the classification. First, the ensemble predicts the probability that the constellation represents a respiratory event using all 350 features. The algorithm then evaluates many reduced feature subsets by treating the excluded features as missing and recomputing the prediction. It ranks these subsets according to a combined objective that favours both a small number of features and a predicted probability that remains close to the original full-feature prediction. The subset with the best trade-off is interpreted as the minimal set of features that is sufficient to support the decision. Within this subset, the influence of each feature is quantified by temporarily removing it, recomputing the prediction, and measuring the resulting change in event-level probability. Features whose removal causes the largest reduction in probability are ranked highest. The ordered list of these dominant features (the scoring rationale) can finally be translated into human-readable statements (for example, 'constellation has deep desaturation' or 'constellation is part of a highly regular desaturation sequence') which are presented in Table 24 of the discussion section as the auto-generated scoring rationale for each constellation.

Comparing Panorama to context-unaware scoring

Dataset

The dataset analysed in this paper comprises 266 participants with suspicion of OSA, which were recruited in a consecutive cohort across seven different centres, of which one was located in Belgium (Ziekenhuis Oost Limburg, ZOL, Genk, Belgium) and six in the USA (three centres were part of the United Health Systems Group in Miami, Florida, USA and three other centres were part of the United Sleep Diagnostics group in Florida, USA). All participants were scheduled for one overnight in-lab PSG. Participants were asked for informed consent. The USA branch of the study was approved by

the Aspire IRB, part of the WIRB-Copernicus Group. The European branch of the study was approved by the Ethics Committee of ZOL. Underaged or cognitively impaired participants were excluded from participation in the study. For the European centre, recruitment took place between July 2018 and September 2018 (recruitment period 1). For the USA-based centres, recruitment took place between December 2019 and January 2020 (recruitment period 2) and between October 2021 and June 2022 (recruitment period 3). For all participants, gender, age, BMI, and racial/ethnic self-identification data were recorded.

Protocol and devices

During the setup of PSG, the NightOwl PAT HSAT (ResMed, San Diego, USA) was attached to the middle finger of the hand to which the pulse oximeter of PSG was applied. Each PSG was scored by two independent scoring centres. The first scoring was performed by the team of sleep technicians at the centre where the participant was admitted (further referred to as the 'Local PSG Analysis'). A second independent scoring was performed by the expert scorers of Cerebra Medical (Winnipeg, Canada)[32]. The analysis by CM is referred to as the 'Expert PSG Analysis' and is used as ground truth for endpoint parameter calculation. The data were scored according to the AASM scoring manual's rule 1A for the scoring of hypopnea (3% rule), requiring a 3% desaturation or cortical arousal for the scoring of hypopnea, as well as the alternative rule 1B (4% rule) for the scoring of hypopnea[1]. Details on the study protocol for recruitment periods 1 and 2 were reported before[92]. For recruitment period 3, the same protocol was repeated.

Scoring comparison

The results described in Chapter 2 relating to the multicentric study comprising the NightOwl PAT HSAT were derived from the outputs of a context-unaware version of the autoscoring algorithm. As such, we compared the performance of this context-unaware algorithm version to the Panorama

algorithm and additionally compared both to the Local PSG Analysis. In each instance the Expert PSG Analysis served as the ground truth.

Comparison endpoints

For both algorithm versions and the Local PSG Analysis, the OSA severity categorisation accuracy (or 4-way categorisation accuracy), Cohen's Kappa (κ)[72], and the confusion matrix were reported. Additionally, for each of the OSA severity cutoffs (5, 15, and 30), the sensitivity (Se), specificity (Sp), NPV, PPV, cutoff agreement (Acc), and cutoff Cohen's Kappa were calculated. The results are provided for both the 3% rule and the 4% rule scoring. The results are provided with and without near-boundary double-labelling (NBL). Additionally, a Bland-Altman and correlation analysis was performed, and the bias, ICC(A,1)[74], upper limits of agreement (ULA) and lower limits of agreement (LLA)[73,150] were also reported. These metrics were also calculated by considering only AHIs below 30, as extreme AHIs are known to inflate the performances of correlation coefficients[30]. Chapter 2 elaborates in detail on the rationale behind the selection of these endpoint parameters.

Statistical analysis

General

Data analysis and reporting followed the TRIPOD+AI guidance for reporting clinical prediction models that use regression or machine learning methods[142]. Statistical analysis was performed using MATLAB (version 2019a, MathWorks, USA). For all endpoint parameters, 95% confidence intervals were computed. Significance levels were determined using an alpha level of 0.05. Differences between two proportions were assessed by a *two-proportion z-test*.

Sample size determination

Statistical power was determined by postulating that a 10% decrease in OSA severity categorisation accuracy should be detected as statistically significant with an alpha level of 0.05 and a power of 0.8 using patient-wise cross-validation for model evaluation. Assuming a minimum 4-way categorisation accuracy parameter value of 0.75 for the Local Analysis, a minimum sample size of 165 participants was found.

Data adequacy

All participants retained in the dataset met the following data adequacy criteria: all PSG channels could be interpreted by the scorers (e.g., there were no detachments of the nasal cannula or pulse oximeter), and at least 4 hours of analysable signal and 1 hour of TST could be obtained for the PAT HSAT. Only participants whose data were technically adequate in both the context-unaware and Panorama versions of the algorithm were included.

Participants with missing participant characteristics, such as age and gender data, were omitted from the analysis of population demographic statistics. These participants were still included for the calculation of endpoint parameters. Figure 38 provides a flowchart highlighting the number of recruited participants as well as administrative and technical failures, including the reason for failure.

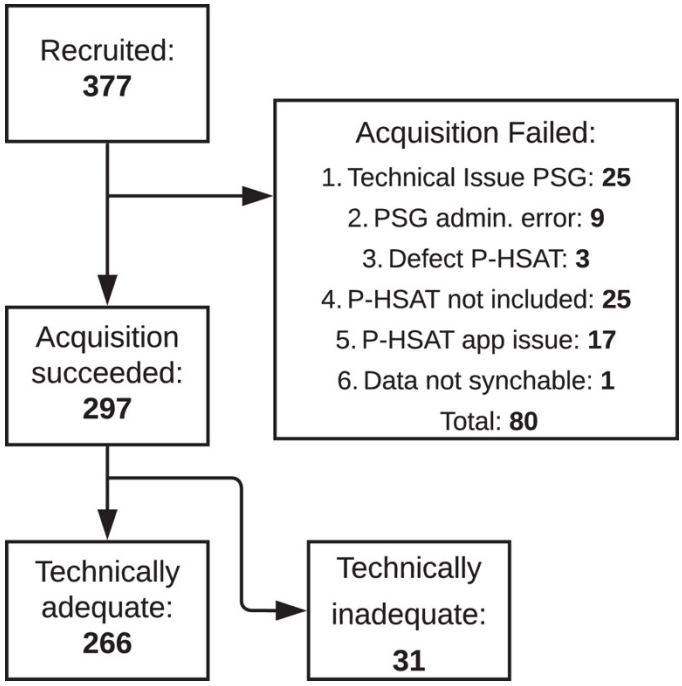

Figure 38: Flowchart of participant recruitment. The flowchart depicts how many participants gave informed consent (377), how many participants PSG and PAT HSAT data acquisition was successful (297), and how many participants the PAT HSAT was technically adequate (266). Additionally, the flowchart summarises the reasons for data acquisition failure.

Results

Demographic information

As listed in Table 20, participants were predominantly male, middle-aged, and overweight or obese. The mean AHI was 31.6. 39 (15%) participants had no sleep apnea, 59 (22%) participants had mild sleep apnea, 61 (23%) participants had moderate sleep apnea, and 107 (40%) participants had severe sleep apnea. 19% of participants identified as black, 80% as white, and 32% as Hispanic, Latino, or Spanish.

	Mean	SD	Min	Max	No data (%)
Age	56	15	20	87	6 (2%)
BMI	31.5	7.3	14.7	59.2	7 (3%)
AHI	31.6	27.9	0	117.8	NA
	N (%)				
Male	153 (58%)				
Female	108 (40%)				
No data	5 (2%)				
No OSA	39 (15%)				
Mild OSA	59 (22%)				
Moderate OSA	61 (23%)				
Severe OSA	107 (40%)				
Black	50 (19%)				
White	214 (80%)				
No data	2 (1%)				
Hispanic, Latino, or Spanish	84 (32%)				

Table 20: Demographic and clinical characteristics of participants in the dataset. NA = not applicable; SD = standard deviation.

Endpoint comparison

Table 21 reports on all the endpoint parameters, including the p-values for the hypothesis test that the context-unaware and Panorama endpoints are statistically significantly different. For the 3% rule and 4% rule, the Panorama algorithm increases the overall 4-way categorisation accuracy respectively from 67% to 76% and 78% to 85%. We observe larger performance gains for the 3% rule compared to the 4% rule. The 3% rule's non-desaturating hypopnea events (hypopnea with a cortical arousal but without significant desaturation) are comparatively harder to score than hypopnea defined by the 4% rule[151]. As such, it is expected that the benefit of incorporating event context would be more significant for 3% rule scoring. For both the 3% and 4% rule, we found a significant increase in specificity for AHI cutoff 5 (respectively from 0.59 to 0.80 and 0.84 to 0.91).

This gain is explained by the Panorama algorithm's comparatively greater ability to discard false positives, since incoherent and isolated constellations can more easily be discarded as artefacts when incorporating event context.

Endpoint parameter		NBL	3% Rule			4% Rule			p-values	
			PAT HSAT context-unaware	*PAT HSAT Panorama*	*Local PSG*	*PAT HSAT context-unaware*	*PAT HSAT Panorama*	*Local PSG*	*Context-unaware vs Panorama (3% Rule)*	*Context-unaware vs Panorama (4% Rule)*
4-way categorisation accuracy		*off*	0.67 (0.61 – 0.73)	0.76 (0.71 – 0.81)	0.74 (0.68 – 0.79)	0.78 (0.74 – 0.83)	0.85 (0.81 – 0.90)	0.81 (0.76 – 0.86)	0.020*	0.040*
		on	0.80 (0.75 – 0.85)	0.87 (0.83 – 0.91)	0.86 (0.82 – 0.90)	0.88 (0.84 – 0.92)	0.92 (0.89 – 0.95)	0.89 (0.85 – 0.93)	0.027*	0.108
4-way Cohen's Kappa		*off*	0.53 (0.45 – 0.61)	0.66 (0.59 – 0.74)	0.63 (0.56 – 0.70)	0.71 (0.64 – 0.77)	0.80 (0.74 – 0.86)	0.74 (0.68 – 0.81)	0.011*	0.058
		on	0.72 (0.65 – 0.78)	0.82 (0.76 – 0.87)	0.80 (0.74 – 0.86)	0.83 (0.78 – 0.89)	0.89 (0.85 – 0.94)	0.85 (0.80 – 0.90)	0.055	0.231
Cutoff AHI 5	**Se**	*off*	0.93 (0.90 – 0.96)	0.93 (0.90 – 0.96)	0.96 (0.94 – 0.99)	0.96 (0.93 – 0.98)	0.98 (0.96 – 1.00)	0.98 (0.96 – 1.00)	0.827	0.159
		on	0.97 (0.95 – 0.99)	0.96 (0.94 – 0.99)	0.99 (0.97 – 1.00)	0.98 (0.96 – 1.00)	0.98 (0.97 – 1.00)	0.99 (0.98 – 1.00)	0.751	0.659
	Sp	*off*	0.59 (0.53 – 0.65)	0.80 (0.75 – 0.85)	0.71 (0.66 – 0.77)	0.84 (0.80 – 0.88)	0.91 (0.88 – 0.95)	0.83 (0.78 – 0.87)	0.000*	0.009*
		on	0.84 (0.79 – 0.88)	0.93 (0.91 – 0.96)	0.88 (0.84 – 0.92)	0.97 (0.95 – 0.99)	1.00 (1.00 – 1.00)	0.96 (0.93 – 0.98)	0.000*	0.007*
	NPV	*off*	0.63 (0.57 – 0.69)	0.71 (0.65 – 0.76)	0.80 (0.75 – 0.85)	0.89 (0.86 – 0.93)	0.95 (0.92 – 0.98)	0.94 (0.92 – 0.97)	0.056	0.020*
		on	0.84 (0.79 – 0.88)	0.84 (0.80 – 0.89)	0.93 (0.89 – 0.96)	0.95 (0.92 – 0.97)	0.96 (0.94 – 0.98)	0.97 (0.95 – 0.99)	0.852	0.433
	PPV	*off*	0.92 (0.88 – 0.95)	0.96 (0.93 – 0.98)	0.94 (0.91 – 0.97)	0.93 (0.90 – 0.96)	0.96 (0.94 – 0.99)	0.93 (0.90 – 0.96)	0.040*	0.108
		on	0.97 (0.95 – 0.99)	0.99 (0.97 – 1.00)	0.98 (0.96 – 1.00)	0.99 (0.98 – 1.00)	1.00 (1.00 – 1.00)	0.98 (0.97 – 1.00)	0.177	0.093
	Acc	*off*	0.87 (0.83 – 0.91)	0.91 (0.88 – 0.94)	0.92 (0.89 – 0.95)	0.92 (0.89 – 0.95)	0.96 (0.93 – 0.98)	0.93 (0.90 – 0.96)	0.129	0.068
		on	0.95 (0.92 – 0.97)	0.96 (0.93 – 0.98)	0.97 (0.95 – 0.99)	0.98 (0.96 – 1.00)	0.99 (0.98 – 1.00)	0.98 (0.96 – 1.00)	0.539	0.313
	Kappa	*off*	0.53 (0.38 – 0.67)	0.70 (0.58 – 0.81)	0.71 (0.59 – 0.83)	0.81 (0.73 – 0.89)	0.90 (0.84 – 0.96)	0.83 (0.76 – 0.91)	0.055	0.295
		on	0.81 (0.71 – 0.90)	0.86 (0.78 – 0.94)	0.88 (0.81 – 0.96)	0.94 (0.90 – 0.99)	0.97 (0.94 – 1.00)	0.95 (0.91 – 0.99)	0.523	0.745
Cutoff AHI 15	**Se**	*off*	0.91 (0.87 – 0.94)	0.92 (0.89 – 0.95)	0.94 (0.91 – 0.97)	0.94 (0.91 – 0.97)	0.95 (0.93 – 0.98)	0.92 (0.88 – 0.95)	0.483	0.362
		on	0.94 (0.92 – 0.97)	0.96 (0.94 – 0.99)	0.97 (0.95 – 0.99)	0.95 (0.92 – 0.97)	0.97 (0.95 – 0.99)	0.94 (0.91 – 0.97)	0.272	0.169
	Sp	*off*	0.81 (0.76 – 0.85)	0.88 (0.84 – 0.92)	0.80 (0.75 – 0.85)	0.92 (0.89 – 0.95)	0.94 (0.91 – 0.97)	0.89 (0.86 – 0.93)	0.021*	0.390
		on	0.86 (0.82 – 0.90)	0.91 (0.88 – 0.95)	0.87 (0.83 – 0.91)	0.96 (0.93 – 0.98)	0.97 (0.95 – 0.99)	0.92 (0.89 – 0.96)	0.061	0.398
	NPV	*off*	0.85 (0.81 – 0.90)	0.87 (0.83 – 0.91)	0.89 (0.85 – 0.93)	0.94 (0.91 – 0.97)	0.95 (0.93 – 0.98)	0.92 (0.88 – 0.95)	0.569	0.500
		on	0.91 (0.88 – 0.95)	0.94 (0.91 – 0.97)	0.94 (0.92 – 0.97)	0.95 (0.92 – 0.98)	0.97 (0.95 – 0.99)	0.94 (0.91 – 0.97)	0.214	0.223
	PPV	*off*	0.87 (0.83 – 0.91)	0.93 (0.90 – 0.96)	0.89 (0.85 – 0.93)	0.92 (0.88 – 0.95)	0.94 (0.91 – 0.97)	0.90 (0.86 – 0.93)	0.032*	0.258
		on	0.91 (0.87 – 0.94)	0.95 (0.92 – 0.97)	0.93 (0.90 – 0.96)	0.95 (0.93 – 0.98)	0.97 (0.95 – 0.99)	0.93 (0.90 – 0.96)	0.098	0.318
	Acc	*off*	0.87 (0.82 – 0.91)	0.91 (0.87 – 0.94)	0.89 (0.85 – 0.93)	0.93 (0.90 – 0.96)	0.95 (0.92 – 0.97)	0.91 (0.87 – 0.94)	0.134	0.369
		on	0.91 (0.88 – 0.94)	0.94 (0.92 – 0.97)	0.93 (0.90 – 0.96)	0.95 (0.93 – 0.98)	0.97 (0.95 – 0.99)	0.93 (0.90 – 0.96)	0.134	0.266
	Kappa	*off*	0.72 (0.63 – 0.80)	0.80 (0.72 – 0.87)	0.75 (0.67 – 0.84)	0.86 (0.80 – 0.92)	0.89 (0.84 – 0.95)	0.81 (0.74 – 0.88)	0.343	0.664
		on	0.81 (0.74 – 0.88)	0.88 (0.82 – 0.94)	0.85 (0.79 – 0.92)	0.90 (0.85 – 0.95)	0.94 (0.90 – 0.98)	0.86 (0.80 – 0.93)	0.425	0.664
Cutoff AHI 30	**Se**	*off*	0.90 (0.86 – 0.93)	0.93 (0.89 – 0.96)	0.91 (0.87 – 0.94)	0.86 (0.82 – 0.90)	0.89 (0.85 – 0.93)	0.95 (0.92 – 0.98)	0.241	0.353
		on	0.94 (0.91 – 0.97)	0.97 (0.95 – 0.99)	0.95 (0.92 – 0.97)	0.91 (0.87 – 0.94)	0.94 (0.91 – 0.96)	0.99 (0.97 – 1.00)	0.093	0.244
	Sp	*off*	0.91 (0.87 – 0.94)	0.93 (0.89 – 0.96)	0.89 (0.86 – 0.93)	0.95 (0.93 – 0.98)	0.96 (0.94 – 0.99)	0.95 (0.93 – 0.98)	0.449	0.550
		on	0.93 (0.90 – 0.96)	0.96 (0.94 – 0.98)	0.94 (0.91 – 0.97)	0.96 (0.93 – 0.98)	0.97 (0.95 – 0.99)	0.96 (0.93 – 0.98)	0.119	0.526
	NPV	*off*	0.93 (0.90 – 0.96)	0.95 (0.92 – 0.98)	0.93 (0.90 – 0.96)	0.94 (0.91 – 0.97)	0.95 (0.93 – 0.98)	0.98 (0.96 – 1.00)	0.363	0.595
		on	0.96 (0.94 – 0.98)	0.98 (0.96 – 1.00)	0.96 (0.94 – 0.98)	0.96 (0.94 – 0.99)	0.97 (0.95 – 0.99)	0.99 (0.99 – 1.00)	0.188	0.493
	PPV	*off*	0.86 (0.82 – 0.91)	0.89 (0.86 – 0.93)	0.85 (0.81 – 0.89)	0.88 (0.84 – 0.92)	0.91 (0.88 – 0.94)	0.89 (0.86 – 0.93)	0.320	0.304
		on	0.90 (0.86 – 0.94)	0.95 (0.92 – 0.97)	0.92 (0.89 – 0.95)	0.90 (0.86 – 0.93)	0.92 (0.89 – 0.96)	0.91 (0.87 – 0.94)	0.047*	0.278
	Acc	*off*	0.90 (0.87 – 0.94)	0.93 (0.89 – 0.96)	0.90 (0.86 – 0.94)	0.92 (0.89 – 0.96)	0.94 (0.91 – 0.97)	0.95 (0.93 – 0.98)	0.355	0.490
		on	0.94 (0.91 – 0.97)	0.97 (0.94 – 0.99)	0.94 (0.92 – 0.97)	0.94 (0.92 – 0.97)	0.96 (0.93 – 0.98)	0.97 (0.94 – 0.99)	0.108	0.421
	Kappa	*off*	0.80 (0.72 – 0.87)	0.85 (0.78 – 0.91)	0.79 (0.72 – 0.87)	0.82 (0.74 – 0.90)	0.86 (0.79 – 0.92)	0.89 (0.83 – 0.95)	0.587	0.667
		on	0.87 (0.81 – 0.93)	0.93 (0.89 – 0.98)	0.88 (0.83 – 0.94)	0.86 (0.79 – 0.93)	0.90 (0.84 – 0.96)	0.92 (0.87 – 0.97)	0.470	0.666

Table 21: Comparison of all endpoint parameters. The comparison is made between the context-unaware algorithm, the Panorama algorithm, and the Local PSG Analysis, with the Expert PSG Analysis serving as ground truth. Confidence intervals are provided in brackets. The final two columns also indicate the p-value for the hypothesis test that the context-unaware and Panorama endpoints are statistically significantly different (*p < .05). The results are provided for the 3% rule and 4% rule, with and without NBL. Acc = accuracy; NBL = near-boundary double-labelling; NPV = negative predictive value; PPV = positive predictive value; Se = sensitivity; Sp = specificity.

Table 22 provides the OSA severity category confusion matrices for each of the estimates (PAT HSAT context-unaware, PAT HSAT Panorama, and Local PSG Analysis), compared to the Expert PSG Analysis as ground truth. The endpoint parameters in Table 21 can be directly calculated from these confusion matrices. Figure 39 and Figure 40, as well as Table 23 show the results of a Bland-Altman and correlation analysis. The range between the ULA and LLA is consistently narrower for the Panorama algorithm compared to the context-unaware version. The ICC(A,1) is also consistently higher for the former.

3% Rule		**PAT HSAT** context-unaware NBL off I NBL on				**PAT HSAT** Panorama NBL off I NBL on				**Local PSG** NBL off I NBL on			
		No	Mild	Mod.	Severe	No	Mild	Mod.	Severe	No	Mild	Mod.	Severe
Expert PSG	No	27 I 36	14 I 6	5 I 1	0 I 0	36 I 43	8 I 3	1 I 0	0 I 0	32 I 37	11 I 5	2 I 0	0 I 0
	Mild	15 I 7	31 I 44	15 I 14	1 I 0	13 I 8	30 I 40	11 I 9	0 I 0	6 I 1	30 I 41	16 I 13	2 I 0
	Mod.	1 I 0	13 I 9	24 I 33	14 I 11	2 I 0	9 I 6	37 I 43	12 I 6	0 I 1	8 I 3	37 I 46	15 I 9
	Severe	0 I 0	1 I 0	10 I 6	95 I 99	0 I 0	2 I 0	6 I 3	99 I 105	2 I 2	0 I 0	8 I 4	97 I 105

4% Rule		**PAT HSAT** context-unaware NBL off I NBL on				**PAT HSAT** Panorama NBL off I NBL on				**Local PSG** NBL off I NBL on			
		No	Mild	Mod.	Severe	No	Mild	Mod.	Severe	No	Mild	Mod.	Severe
Expert PSG	No	68 I 72	13 I 2	0 I 0	0 I 0	74 I 75	6 I 0	1 I 0	0 I 0	67 I 69	12 I 3	2 I 0	0 I 0
	Mild	7 I 3	40 I 52	11 I 6	0 I 0	3 I 2	42 I 50	7 I 4	0 I 0	2 I 0	38 I 50	12 I 10	0 I 0
	Mod.	1 I 1	6 I 6	32 I 40	9 I 8	1 I 1	5 I 3	40 I 48	7 I 6	1 I 1	9 I 6	34 I 41	9 I 8
	Severe	0 I 0	1 I 0	10 I 7	68 I 69	0 I 0	0 I 0	9 I 5	71 I 72	1 I 1	0 I 0	3 I 0	76 I 77

Table 22: OSA category confusion matrices. The confusion matrices are provided for the PAT HSAT context-unaware, the PAT HSAT Panorama, and the Local PSG Analysis, respectively, all vs the Expert PSG Analysis. Results are provided for the 3% rule and 4% rule, with and without NBL. Mod. = moderate; NBL = near-boundary double-labelling.

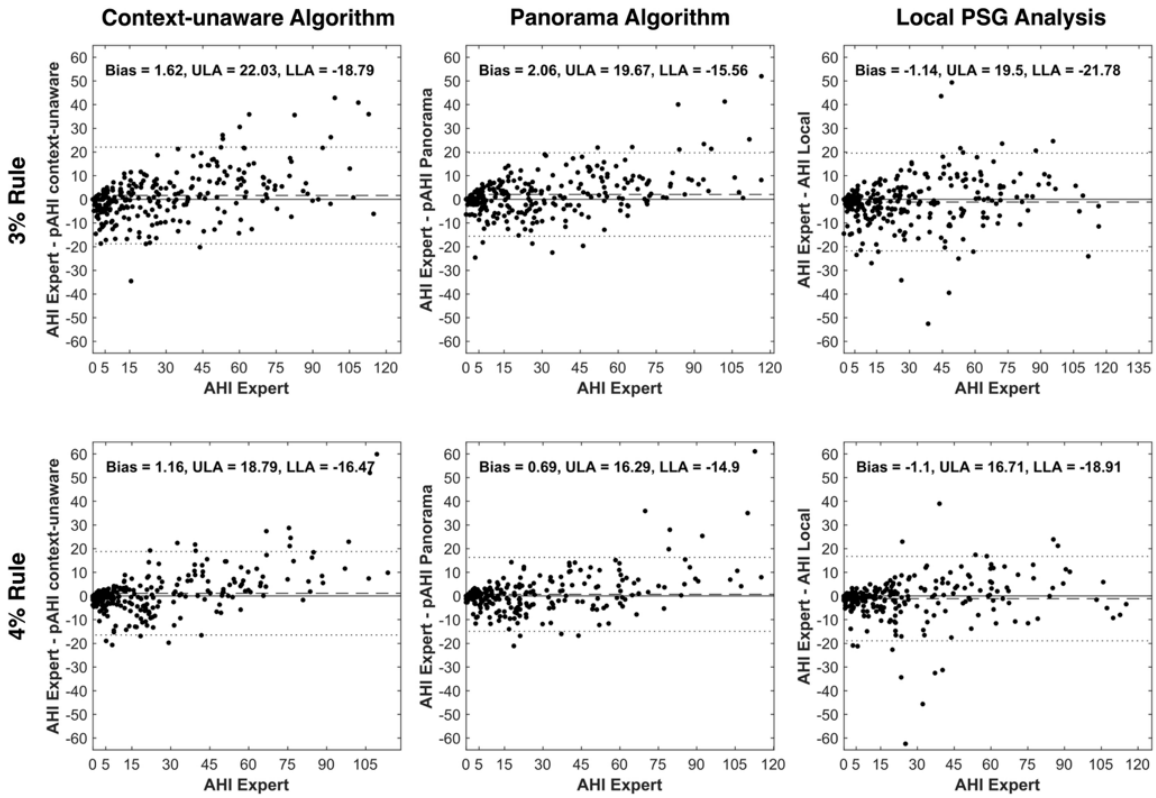

Figure 39: Bland-Altman (BA) plots. The top row contains BA plots using 3% rule scoring, the bottom row using 4% rule scoring. The left, middle, and right columns, respectively, contain BA plots where the Expert PSG Analysis is compared to context-unaware (i.e., traditional) PAT HSAT scoring (left), the context-aware Panorama Algorithm's PAT HSAT Scoring (middle), and the Local PSG Analysis (right). The dotted lines represent the mean limits of agreement. The dashed line represents the mean difference or bias. LLA = lower limit of agreement; pAHI = AHI estimated by PAT HSAT; ULA = upper limit of agreement.

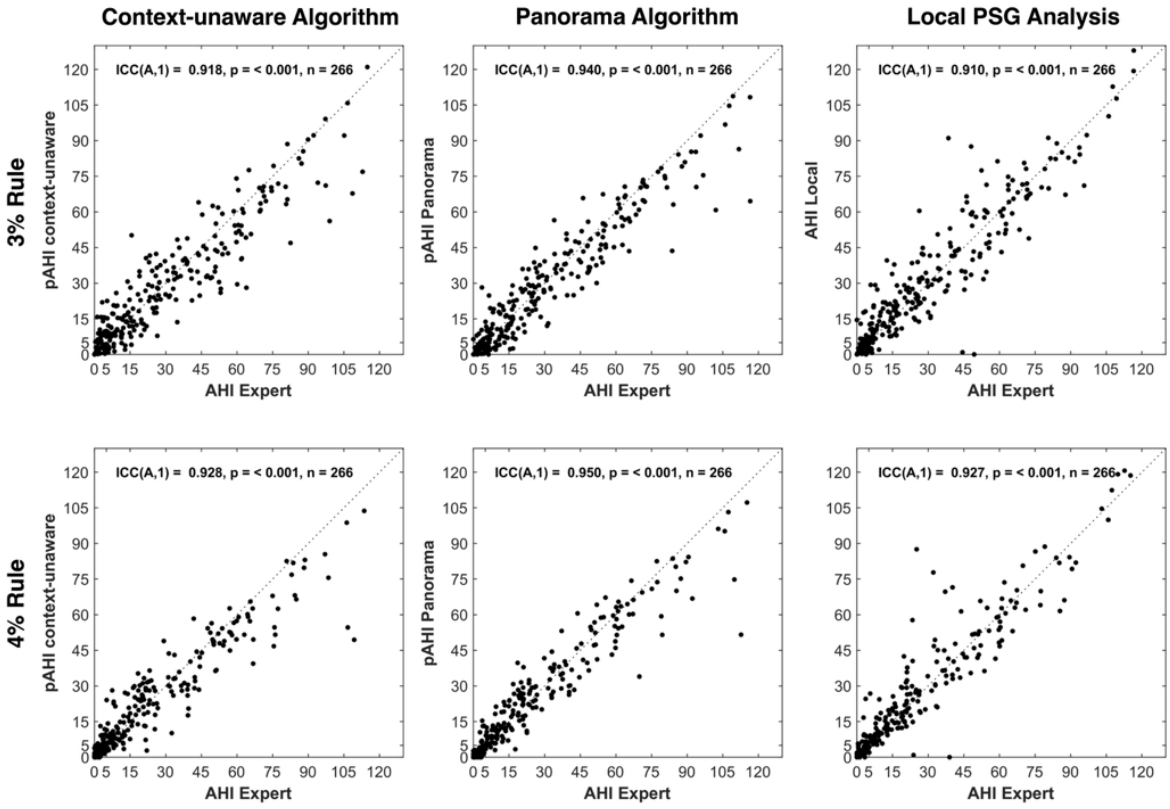

Figure 40: Correlation analysis plots. The top row contains correlation plots using 3% rule scoring, the bottom row using 4% rule scoring. The left, middle, and right columns, respectively, contain correlation plots where the Expert PSG Analysis is compared to context-unaware (i.e., traditional) PAT HSAT scoring (left), the context-aware Panorama Algorithm's PAT HSAT scoring (middle), and the Local PSG Analysis (right). The dotted line represents the points for which the y-axis values equal the x-axis values of the graph (identity line). ICC(A,1) = intraclass correlation coefficient of type (A,1); pAHI = apnea-hypopnea index estimated by PAT HSAT

		3% Rule			4% Rule		
		PAT HSAT context-unaware	*PAT HSAT Panorama*	*Local PSG*	*PAT HSAT context-unaware*	*PAT HSAT Panorama*	*Local PSG*
All AHIs	Bias	1.62	2.06	-1.14	1.16	0.69	-1.10
	ULA	22.03	19.67	19.50	18.79	16.29	16.71
	LLA	-18.79	-15.56	-21.78	-16.47	-14.9	-18.91
	ICC(A,1)	0.918 (0.916 - 0.920)	0.940 (0.939 - 0.942)	0.910 (0.907 - 0.912)	0.928 (0.926 - 0.930)	0.950 (0.949 - 0.951)	0.927 (0.925 - 0.928)
AHIs < 30	Bias	-1.87	-0.65	-2.72	-1.64	-1.12	-1.95
	ULA	12.45	11.52	11.43	9.34	8.18	12.33
	LLA	-16.20	-12.81	-16.87	-12.62	-10.42	-16.24
	ICC(A,1)	0.71 (0.68 - 0.74)	0.81 (0.79 - 0.82)	0.73 (0.70 - 0.76)	0.80 (0.79 - 0.82)	0.86 (0.85 - 0.87)	0.73 (0.70 - 0.75)

Table 23: Summary of Bland-Altman and correlation analysis of AHI. This table displays the bias, ULA, LLA, and ICC(A,1) from the Bland-Altman and correlation analysis comparing the pAHI of the context-unaware algorithm, the Panorama algorithm, and the Local PSG Analysis, with the Expert PSG Analysis serving as ground truth. When confidence intervals were calculated, they are displayed between brackets. These parameters are calculated for the 3% rule and the 4% rule and for all AHIs as well as only AHIs < 30 events/h (as determined by Expert PSG Analysis). ICC(A,1) = intraclass correlation coefficient of type (A,1); LLA = lower limits of agreement; pAHI = AHI estimated by PAT HSAT; ULA = upper limits of agreement.

Discussion

Illustrative examples

Comparison between the context-unaware algorithm and the Panorama algorithm demonstrates that a statistically significant 4-class OSA severity category accuracy increase of 9% (3% rule) and 7% (4% rule) can be realised by rendering the algorithm context-aware. The performance gain of the context-aware algorithm can further be illustrated by examples in which the context-unaware alternative under- or over-detected respiratory events because of its inherent limitations.

Figure 41 displays a sequence of respiratory events (highlighted by the grey bars) as scored by Panorama and the Expert PSG Analysis. The second and third respiratory events are associated with an oxygen desaturation of only 2% and were therefore missed by the context-unaware algorithm, as such a model typically requires a more pronounced (≥ 3%) oxygen desaturation to accumulate enough confidence that a standalone event should be scored. However, the Panorama algorithm was able to score this event with sufficient confidence, as the oxygen-desaturation, PAT, and pulse-rate

marker event sequences exhibited clear repetitiveness and synchronisation. As introduced earlier, the model is designed to permit a straightforward review of the scoring rationale. Table 24 illustrates this concept, by displaying the *scoring rationale* (of which the generation method is explained in the constellation scoring section of the methods section) for each of the events scored in Figure 41.

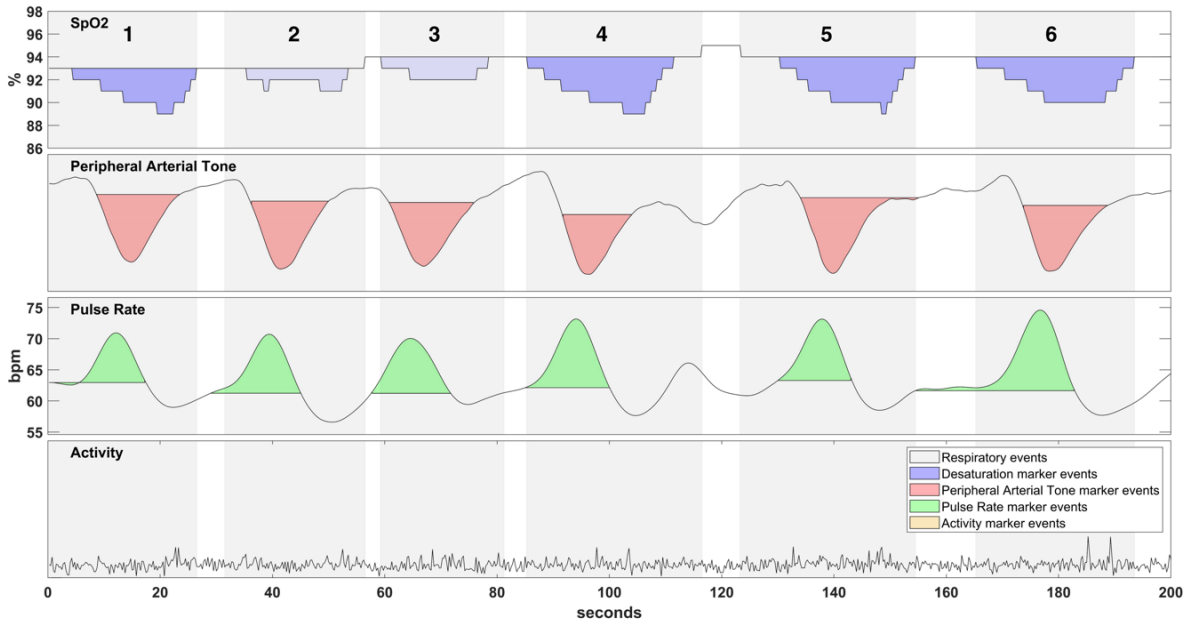

Figure 41: Difficult-to-score respiratory event sequence. The six respiratory event episodes, as scored by PSG, are highlighted in grey. The numbers in the figure correspond to the 6 constellations scored by the Panorama algorithm. The figure shows a sequence of respiratory events with two oxygen desaturations < 3%, indicated in a lighter blue than the other desaturation marker events (associated with constellations 2 and 3). These events correspond to hypopnea without substantial desaturation, but with cortical arousal. The context-unaware algorithm incorrectly did not score these constellations as respiratory events. The Panorama algorithm correctly marks the respiratory events (grey bars) by considering the clear repetitiveness of, and synchronisation between, oxygen desaturation, PAT, and pulse rate marker event sequences.

Scoring rationale Figure 41

1	1. constellation has deep desaturation 2. constellation has strong overlap of desaturation, PR, and PAT marker events 3. constellation is proximate to desaturations of sufficient depth 4. constellation is part of desaturation sequence with high regularity and depth 5. constellation is part of linked sequence of PR and PAT marker events with strong periodicity and overlap
2	1. constellation has shallow desaturation 2. constellation has strong overlap of desaturation and PR marker events 3. constellation is proximate to desaturations of sufficient depth and strong periodicity 4. constellation is part of desaturation sequence with high regularity and depth 5. constellation is part of linked sequence of desaturations and PAT marker events with deep desaturations
3	1. constellation has shallow desaturation 2. constellation has strong overlap of desaturation, PR, and PAT marker events 3. constellation is part of desaturation sequence with high regularity and depth 4. constellation is part of linked sequence of PR and PAT marker events with strong periodicity and overlap 5. constellation is part of linked sequence of desaturations and PAT marker events with deep desaturations
4	1. constellation has deep desaturation 2. constellation has strong overlap of desaturation and PAT marker events 3. constellation is proximate to desaturations of sufficient depth 4. constellation is part of desaturation sequence with high regularity, shape similarity, and depth 5. constellation is part of linked sequence of PR and PAT marker events with strong overlap 6. constellation is part of linked sequence of desaturations and PAT marker events with deep desaturations
5	1. constellation has deep desaturation 2. constellation has strong overlap of desaturation and PAT marker events 3. constellation is proximate to desaturations of sufficient depth 4. constellation is part of desaturation sequence with high regularity and depth 5. constellation is part of linked sequence of PR and PAT marker events with strong periodicity and overlap

6	1. constellation has deep desaturation 2. constellation has strong overlap of desaturation, PR, and PAT marker events 3. constellation is proximate to desaturations of sufficient depth 4. constellation is part of desaturation sequence with high regularity and depth 5. constellation is part of linked sequence of PR and PAT marker events with strong periodicity and overlap 6. constellation is part of linked sequence of desaturations and PAT marker events with deep desaturations

Scoring rationale Figure 42

1	1. constellation has insufficient overlap of desaturation and PAT marker events 2. constellation is part of desaturation sequence with strong shape irregularity and short duration 3. constellation is part of PAT marker event sequence with irregular periodicity 4. constellation is part of linked sequence of desaturations and PAT marker events with insufficient overlap
2	1. constellation has insufficient overlap of desaturation and PAT marker events 2. constellation is part of desaturation sequence with moderate and irregular depth 3. constellation is part of PAT marker event sequence with irregular periodicity 4. constellation is part of linked sequence of desaturations and PAT marker events with insufficient overlap
3	1. constellation has insufficient overlap of desaturation and PAT marker events 2. constellation is part of desaturation sequence with moderate and irregular depth
4	1. constellation has insufficient overlap of desaturation and PAT marker events 2. constellation is part of desaturation sequence with moderate and irregular depth 3. constellation is part of PAT marker event sequence with irregular periodicity 4. constellation is part of linked sequence of desaturations and PAT marker events with insufficient overlap

Table 24: Listing of the auto-generated scoring rationale for each of the constellations scored as respiratory events in Figure 41 and the scoring rationale for each of the constellations rejected as respiratory events in Figure 42. The numbers in the leftmost column correspond to the constellation numbers highlighted in Figure 41 and Figure 42.

Figure 42 contains an example where the context-unaware algorithm incorrectly scored four respiratory events (the incorrect scoring was dominated by the presence of substantial oxygen desaturations). Reviewing the context of these false-positive events shows that the oxygen desaturations share little morphological similarity and that no repetitive pattern is present in the pulse

rate or PAT traces. This lack of consistent structure enabled the Panorama algorithm to correctly refrain from scoring the associated constellations as respiratory events. Table 24 illustrates the auto-generated scoring rationale for each constellation. As shown, this rationale can be examined both for constellations that were scored as respiratory events as well as for those that were rejected.

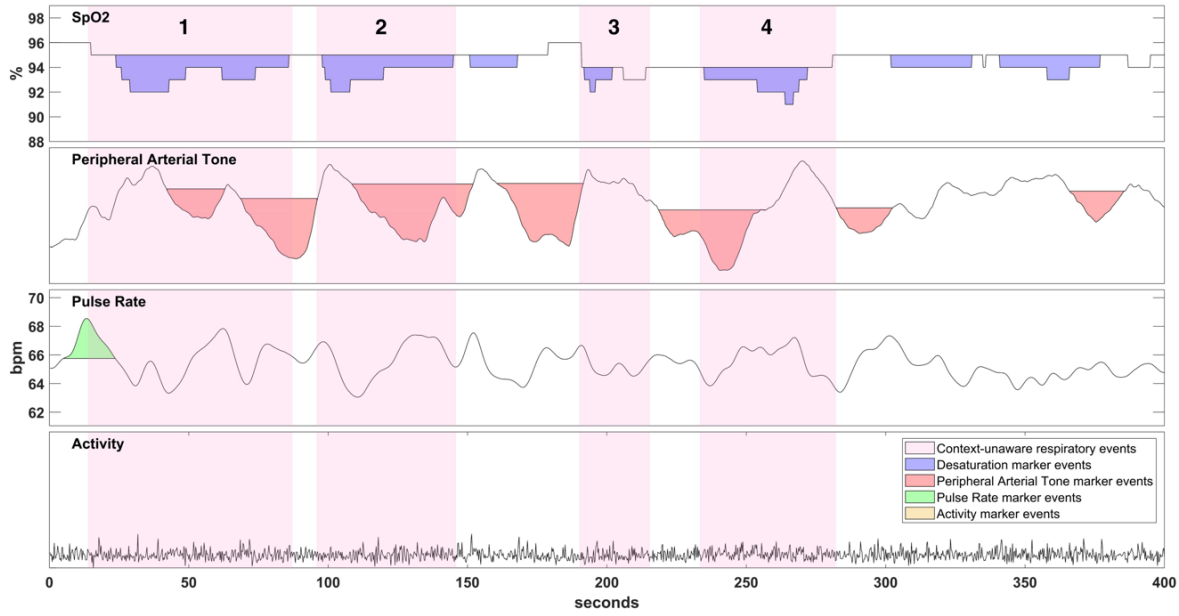

Figure 42: Irregular sequence of marker events. The numbers in the figure correspond to 4 false-positive respiratory events scored by the context-unaware method. The figure shows sequences of irregular oxygen desaturation marker events. Some desaturations are substantial (3% or 4%), and a context-unaware algorithm detected them as respiratory events (pink areas). However, these desaturations lack consistent morphology, and there is neither temporal nor morphological synchronisation with other signal marker events. Therefore, the Panorama algorithm did not score these constellations of marker events as being associated with respiratory events.

Strengths and limitations

The primary strength of this study is its double labelling of PSG data. This allowed the comparison of both algorithms to the PSG inter-scorer agreement of respiratory events found in this dataset. The first limitation of this study was the restriction of the analysis of PAT HSAT data acquired by just one type of device. Further research is required to confirm whether a similar benefit of context-aware

scoring can be generalised to other PAT HSATs. Secondly, further research is needed to investigate the repeatability of our findings and to investigate any consistent scoring bias, using additional independent multi-scored datasets. Finally, the Panorama algorithm does not differentiate between central and obstructive respiratory events. Further research is required to evaluate its performance to that end.

Conclusion

In this study, we proposed a reviewable and context-aware autoscoring algorithm which shows promise in reducing the performance disadvantage of PAT HSAT to PSG. We found that the algorithm significantly outperformed context-unaware PAT HSAT autoscoring by 9% (3% rule) and 7% (4% rule) in terms of OSA severity categorisation accuracy. These improvements directly address Limitation 4 by demonstrating that extending the analysis scope beyond isolated event windows can substantially enhance PAT HSAT autoscoring performance while retaining reviewability of the scoring rationale.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of paper:	Context-aware analysis enhances autoscoring accuracy of home sleep apnea testing.
Publication status:	*Published*
Publication details:	Massie F, Vits S, Verbraecken J, Bergmann J. Context-aware analysis enhances autoscoring accuracy of home sleep apnea testing. J Clin Sleep Med. 2025 May 1;21(5):789-804. doi: 10.5664/jcsm.11534. PMID: 39748512; PMCID: PMC12048324.

Student confirmation

Student name:	Frederik Massie
Contribution to the paper:	F Massie performed Panorama algorithm development and evaluation, clinical study design and writeup, and publication design and writeup. S vits performed clinical study data acquisition, Panorama algorithm evaluation, and publication editing. J Verbraecken provided publication editing. J Bergmann provided supervisory guidance and publication editing.
Signature and date:	 30-11-2025

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Jeroen Bergmann
Supervisor comments:	/
Signature and date:	 30-11-2025

Chapter 6: Discussion and conclusion

In this final chapter, we revisit the four substantial limitations of PAT HSAT outlined in Chapter 1 and show how the work in Chapters 2-5 addresses them. For each limitation, we relate the corresponding research aim(s) to the key findings and discuss their implications for the broader goal of improving the clinical utility of minimally obtrusive technologies for diagnosing sleep disorders.

Limitation 1: The absence of clinical validation reporting standards for HSAT leads to selective reporting and difficulties in reproducing or comparing results

Performance evaluations of HSAT devices have lacked a consistent and standardised framework. Many validation studies rely on statistical measures that are not well suited to assessing diagnostic agreement and that may overstate performance. A prominent example is the widespread use of simple AHI correlation coefficients between HSAT and PSG. As demonstrated in Chapter 2, correlation captures only the strength of a linear relationship and does not evaluate whether the two methods agree on clinically meaningful endpoints such as sleep apnea severity classification. On its own, correlation is therefore a poor indicator of diagnostic performance.

Research Aim 1.1 investigated this issue by systematically comparing reported AHI correlations with the corresponding diagnostic accuracy in published HSAT validation studies. Across the 20 included papers, we found a weighted average AHI correlation of 0.82 and a markedly lower weighted average diagnostic accuracy of 0.61, which shows that correlation can give a false impression of agreement where substantial misclassification is present. The analysis also revealed that only six studies explicitly reported accuracy values that could be directly retrieved. This pattern reflects a wider tendency to underreport clinically meaningful endpoints. Our analysis shows that correlation, when used on its own, can give a misleading impression of diagnostic agreement. These results therefore

argue for supplementing correlation coefficients with metrics that more directly assess diagnostic agreement.

Research Aim 1.2 extends these efforts to improve validation practice by addressing an additional and often overlooked issue. Even when more informative metrics are used, the PSG reference itself is subject to inter-scorer variability. Small differences in manually scored AHI values can push a recording across OSA severity boundaries, which may unfairly penalise or obscure HSAT performance. Chapter 2 introduced two approaches to address this limitation. First, HSAT performance was benchmarked against the inter-rater agreement between two independent PSG scorers, which provides a realistic performance target rather than assuming a perfect reference. Second, the near-boundary double labelling method was proposed to correct for AHI values that fall within ranges where scorer disagreement on the OSA severity is likely. This method prevents disproportionate penalisation of disagreements that lie inside typical PSG scoring variability and yields a more balanced assessment of HSAT performance. The clinical relevance of this approach has already been recognised by other authors, who have endorsed or adopted the NBL method in their own analyses, as noted in Chapter 2.

Research Aims 1.1 and *1.2* both address methodological weaknesses in HSAT validation. By outlining the limits of correlation-based reporting, showing how often accuracy metrics are missing, and introducing methods that account for imperfections in the PSG reference, this work provides a more transparent way to evaluate HSAT devices.

Limitation 2: PAT HSAT does not differentiate between CSA and OSA

In Chapter 3, we pursued *Research Aim 2.1*, which was to develop a method for CSA detection using only the PPG signals available from PAT HSAT. We have previously explained how PPG measures

fluctuations in local tissue perfusion. Based on the observation that perfusion in the extremities is affected by alterations in thoracic pressure throughout the respiratory cycle[93], we hypothesised that we could extract PPG-derived correlates of respiratory effort. We developed a signal-processing algorithm to extract these correlates. Subsequently, we created a machine learning pipeline that infers the absence or presence of respiratory effort during a respiratory event to classify each respiratory event detected by PAT HSAT as either central or obstructive. This, in turn, allowed us to estimate the central apnea-hypopnea index (cAHI).

The resulting algorithmic pipeline was evaluated on a clinical dataset with both OSA and CSA cases, and it achieved a promising level of accuracy in identifying patients with substantial central apnea. The method achieved a sensitivity of 81%, a specificity of 99%, a PPV of 89%, and an NPV of 98% at the cAHI cutoff of 10 events per hour of sleep, which was comparable to the inter-rater agreement of PSG.

These results are especially relevant considering that no traditional respiratory effort sensors (e.g., thoracic belts) or airflow measurements were used, as the model relies purely on the finger-worn sensor data. Our approach therefore represents a minimally obtrusive solution to a problem previously thought to require more complex instrumentation. It is worth noting that others have attempted to address CSA detection in HSAT by adding extra sensors. For example, an enhanced PAT HSAT system was validated with an added snoring and body position sensor[64].

By enabling PAT-based HSAT to flag the presence of central apnea, Limitation 2 has been directly addressed. Clinically, this means home testing could be used in a broader range of patients, including those with heart failure, opioid use, or other risk factors for central apnea, with reduced risk of missing a CSA diagnosis. This step moves PAT HSAT's comprehensiveness closer to parity with in-lab PSG.

Limitation 3: PAT HSAT does not measure sleep accurately

A key limitation of PAT HSAT is its inferior ability to determine sleep stages and total sleep time compared to PSG. Because PAT HSAT lacks EEG, EOG, and EMG channels, it must infer sleep stages from alternative peripheral physiological signals such as PAT and actigraphy. This constraint affects the accuracy of sleep staging as well as diagnostic outputs that depend on reliable sleep stage identification, such as REM-related OSA phenotyping and the accurate estimation of the apnea-hypopnea index.

Research Aim 3.1 addressed this limitation by evaluating the ability of PAT HSAT to identify REM sleep and to determine REM OSA phenotypes. By comparing synchronised PAT HSAT and PSG recordings that were independently double scored, we found that PAT HSAT shows lower single-night sensitivity for REM OSA phenotyping than PSG. The analysis also showed that PSG frequently yields inconclusive REM OSA classifications in nights with short REM duration. We therefore examined how changing the minimum required REM sleep time affected the ability to classify REM OSA. The results showed that the commonly used 30-minute threshold is unnecessarily restrictive, as lowering it to 15 minutes produced similar sensitivity and specificity while markedly increasing the proportion of conclusive assessments. We subsequently evaluated a theoretical multi-night testing model to improve PAT-based phenotyping. The simulations showed that a protocol involving two to five nights could bring PAT-based REM OSA phenotyping performance closer to the single-night PSG benchmark and could further reduce inconclusiveness. The results also showed that PAT HSAT cannot fully match PSG on a single night, but its low burden enables multi-night protocols that help compensate for the reduced accuracy of single-night measurements.

Research Aim 3.2 pursued a complementary strategy by evaluating whether the addition of a minimally obtrusive single-lead biopotential sensor could improve PAT HSAT sleep stage estimation. The clinical study showed that incorporating a compact forehead-to-cheek electrode configuration enables accurate automated sleep staging that approaches the performance reported for more complex wearable biopotential devices. The single-lead device achieved an overall five-stage Cohen's Kappa of 0.70 and showed particularly strong REM detection performance. This configuration avoids the discomfort of conventional multi-channel EEG setups and maintains the practical advantages of reduced-channel home testing.

Together, the findings from *Research Aims 3.1* and *3.2* show two complementary pathways for addressing the sleep-measurement limitations of PAT HSAT. Multi-night testing can strengthen REM OSA phenotyping without adding new sensors, while the integration of a single-lead biopotential device can substantially improve overall sleep stage detection. Both approaches narrow the performance gap between reduced-channel HSAT and in-lab PSG while preserving the accessibility, comfort, and scalability that have driven the widespread adoption of PAT HSAT.

Limitation 4: The limited analysis scope of PAT HSAT autoscoring algorithms restricts performance

Conventional PAT HSAT autoscoring algorithms evaluate respiratory events in isolation. They focus on short segments of data in which a desaturation, a PAT amplitude reduction, and a pulse rate increase occur concurrently. This narrow analysis window limits diagnostic accuracy, because respiratory events in sleep apnea typically occur in repetitive and morphologically consistent sequences. These patterns help distinguish true respiratory events from artefacts, but conventional scoring methods ignore this context.

Research Aim 4.1 therefore tested whether analysing event context could improve autoscoring performance. We developed the Panorama algorithm, which assesses each candidate event in the light of the surrounding physiological patterns. The algorithm identifies marker events in the PAT, oximetry, pulse rate, and activity signals, groups them into constellations representing candidate respiratory events, and characterises the repetitive structure of neighbouring events through the detection and analysis of marker event sequences. It also evaluates the temporal alignment and similarity between different marker event (sequence) types. All this contextual information is then consolidated and used to classify each constellation through a supervised decision tree model.

The results demonstrated that incorporating event context yields meaningful improvements in diagnostic performance. When validated against double-scored PSG data, the Panorama algorithm outperformed the conventional context-unaware algorithm across nearly all endpoint parameters. The OSA severity categorisation accuracy increased from 67% to 76% under the 3% rule and from 78% to 85% under the 4% rule. These improvements reflect the algorithm's greater ability to distinguish physiologically plausible respiratory events from isolated artefacts.

A further strength of the Panorama approach is its interpretability. The algorithm was designed to enable the generation of the scoring rationale for each event in natural language, which simplifies manual review and supports clinical transparency. This contrasts with more complex machine learning models that may improve accuracy but are difficult to interpret in a clinical setting.

Together, these findings show that context-aware autoscoring can substantially reduce the performance gap between PAT HSAT and PSG without compromising interpretability or practical usability.

Future research

Refinement and standardisation of validation protocols

Encouraged by endorsement and replication by other researchers[28,152–157], the proposed enhancements to HSAT validation reporting should be further refined and widely reviewed. Future studies can work on establishing consensus guidelines for HSAT validation reporting. For instance, an expert task force could define a core outcome set that every new validation study should include. One of the most widely adopted guiding checklist frameworks for evaluating the quality of studies reporting on diagnostic accuracy is the Standards for Reporting Diagnostic Accuracy (STARD) 2015 framework[158]. It consists of a 30-item checklist developed by an international group of methodologists, researchers, and editors. Recently, this checklist was updated to the STARD-AI checklist to incorporate the unique considerations associated with artificial intelligence-centred diagnostic test studies[159]. We could tailor and expand the STARD-AI checklist for the particularities of HSAT validation studies. Additionally, we may expand the validation framework to other sleep disorders.

Extended CSA model validation

The evaluation of the CSA detection algorithm could be strengthened by validating it in larger and more diverse patient cohorts. Such work is important because CSA encompasses several distinct physiological patterns. For instance, heart failure patients often experience CSA as Cheyne-Stokes respiration, which is characterised by a crescendo-decrescendo pattern of breathing, where tidal volume gradually increases and then decreases, culminating in central apnea or hypopnea. In contrast, chronic opioid therapy can induce central apnea through effects on ventilatory control, and additional CSA subtypes have their own characteristic manifestations[160]. These variations may require dedicated validation studies in specific clinical subpopulations to ensure that the method

generalises reliably across the full spectrum of CSA phenotypes. A larger patient cohort would also allow validation of the method on a population with cAHI > 30, of which the prevalence was low in our dataset.

Informed multi-night testing

Our findings on the benefits of multi-night testing indicate that future research should focus on designing optimal multi-night home testing protocols. Key questions include how many nights are needed before additional recordings provide diminishing returns and whether the required number of nights can be personalised based on the variability observed in the initial recordings. Such personalisation could support algorithms that decide, after each night of testing, whether further data are required to achieve adequate diagnostic confidence.

Recent work by Hynes et al.[161] also highlights the need for patient-specific metrics of night-to-night variability to guide optimal testing schedules. Prospective multi-night PAT HSAT studies with concurrent PSG in a subset of patients could provide the data needed to develop such metrics. These studies could also validate the theoretical voting model for REM OSA phenotyping and quantify the added value of multi-night data for CSA detection.

Deeper integration of biopotential monitoring in PAT HSAT

The encouraging results from the single-lead biopotential device suggest a future step in combining the biopotential-based sleep stage detection and the Panorama algorithm for respiratory event detection into a unified algorithm. Doing so could, for instance, enable stage-dependent scoring (e.g., using different event criteria depending on whether the patient is in REM or NREM sleep). It would also allow detection of cortical arousals linked to respiratory events, providing a measure of sleep fragmentation due to sleep apnea that PAT alone cannot fully capture[162]. Future work should also

assess the single-lead biopotential extension in populations with non-respiratory sleep disorders and in cohorts with a lower pre-test probability of OSA, to evaluate robustness to atypical EEG patterns and sleep fragmentation. In addition, using multi-scorer PSG reference hypnograms and integrating PAT-derived signals, SpO_2, and actigraphy into the staging model may further improve accuracy and generalisability.

Next-generation autoscoring

A natural next step would be to integrate CSA detection more tightly into the Panorama algorithm, allowing unified multi-class autoscoring that distinguishes obstructive and central events while drawing on temporal and morphological context. The underlying idea of context-aware analysis may also extend to applications beyond respiratory event scoring. One potential extension is the analysis of periodic limb movements in sleep, which often occur in coherent and repetitive bursts[163].

Chiang et al.[28] recently highlighted the need for novel (PAT) HSAT technologies that support intuitive and efficient editing of the scoring and provide the necessary user training. To investigate how well the Panorama algorithm's inherent explainability addresses this need, a future implementation study could assess how much it reduces manual editing time compared to traditional scoring algorithms.

Toward a fully comprehensive home sleep diagnostic

Each improvement introduced in this thesis brings HSAT closer to the diagnostic capabilities of PSG, yet a longer-term research objective is the development of a fully comprehensive home sleep monitoring system. Future work could examine the incorporation of additional minimally obtrusive sensors, such as nocturnal sound recording for the assessment of snoring or speech and bruxism-related sounds. The integration of such data with the multi-modal algorithms presented in this thesis

may enable home-based testing for a wider range of sleep disorders, including periodic limb movement, bruxism, REM behaviour disorder, and comorbid insomnia. Moreover, Chiang et al.[28] have noted that PAT-based devices may capture cardiovascular biomarkers, including heart rate variability and hypoxic burden, offering another route to broaden their clinical relevance.

More broadly, the future research directions proposed here focus on improving the clinical utility of home-based sleep diagnostics through continued progress in sensors, algorithms, and clinical evaluation.

Conclusion

This thesis contributes to the field of sleep medicine by enhancing the clinical utility of PAT-based home sleep apnea testing. By systematically addressing four key limitations throughout Chapters 2 to 5, it brings PAT HSAT closer to the diagnostic capabilities of in-lab polysomnography while retaining the convenience and patient-friendliness of a minimally obtrusive wearable home testing device.

These combined contributions advance the design and evaluation of HSAT systems, particularly PAT-based devices, making them more accurate and informative for diagnosing sleep apnea. These enhancements can help bridge the longstanding gap between home and laboratory sleep studies. Ultimately, by lowering barriers to accurate sleep apnea detection, this work may help reduce the number of patients whose sleep disorders remain undiagnosed or untreated due to practical testing limitations.

Note on affiliations

Besides my enrolment in the self-funded part-time DPhil programme at the Natural Interaction Lab, supervised by Prof. Jeroen Bergmann, to access data to support my research, I was retained as a scientific consultant for ResMed. ResMed has acquired various clinical datasets with its PAT HSAT platform, of which it is the sole owner. Adequate informed consent and IRB approval were obtained to permit its use for scientific research in a de-identified fashion. I have been provided access to these raw de-identified datasets to conduct my DPhil research.

List of abbreviations

AASM	American Academy of Sleep Medicine
Acc	Accuracy
AHI	Apnea-hypopnea index
AUC	Area under the curve
BA	Bland-Altman
BMI	Body mass index
cAHI	Central apnea-hypopnea index
CM	Cerebra Medical
CPAP	Continuous positive airway pressure
CSA	Central sleep apnea
ECG	Electrocardiography
EEG	Electroencephalography
EOG	Electrooculography
EMG	Electromyography
ERP	Event rejection proportion
FDA	Food and Drug Administration

HSAT	Home sleep apnea test/testing
ICC	Intraclass correlation coefficient
ICC(A,1)	Intraclass correlation coefficient of the type of *two-way fixed model with single measures of absolute agreement*
IQR	Interquartile range
IRB	Institutional review board
LLA	Lower limits of agreement
LoA	Limits of agreement
LOPOCV	Leave-one-patient-out cross-validation
LR	Likelihood ratio
LR+	Positive likelihood ratio
LR-	Negative likelihood ratio
MSSS	Michele Sleep Scoring System
N1	Stage 1 sleep
N2	Stage 2 sleep
N3	Stage 3 sleep
NBL	Near-boundary double-labelling
NBZ	Near-boundary zones

NREM	Non-REM sleep
NPV	Negative predictive value
OSA	Obstructive sleep apnea
PAT HSAT	Home sleep apnea test/testing based on peripheral arterial tonometry
P-HSAT	Home sleep apnea test/testing based on peripheral arterial tonometry
pAHI	AHI estimated by peripheral arterial tonometry
PAP	Positive airway pressure
PAT	Peripheral arterial tonometry or peripheral arterial tone
PD	Photodiode
PG	Polygraphy
PPG	Photoplethysmography
PPGDR	PPG-derived respiratory effort
PPV	Positive predictive value
PR	Pulse rate
PSG	Polysomnography
RDI	Respiratory disturbance index
REC	REM exclusion criterion
REM	Rapid eye movement sleep

RERA	Respiratory effort-related arousal
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SD	Standard deviation
SE	Sleep efficiency
SHHS	Sleep Heart Health Study
SOL	Sleep onset latency
SpO_2	Blood oxygen saturation
SSM	Sleep staging model
STFT	Short-time Fourier transform
TRT	Total recording time
TST	Total sleep time
ULA	Upper limits of agreement
WASO	Wake after sleep onset
ZOL	Ziekenhuis Oost Limburg

References

1. Berry RB, Quan SF, Abreu AR, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* Vol 13(5). Version 2.6. American Academy of Sleep Medicine; 2020. doi:10.5664/jcsm.6576

2. Remmers JE, deGroot WJ, Sauerland EK, Anch AM. Pathogenesis of upper airway occlusion during sleep. *J Appl Physiol*. 1978;44(6):931-938. doi:10.1152/jappl.1978.44.6.931

3. Morgenthaler TI, Kagramanov V, Hanak V, Decker PA. Complex sleep apnea syndrome: Is it a unique clinical syndrome? *Sleep*. 2006;29(9):1203-1209. doi:10.1093/sleep/29.9.1203

4. Malhotra A, Ayappa I, Ayas N, et al. Metrics of sleep apnea severity: beyond the apnea-hypopnea index. *Sleep*. 2021;44(7). doi:10.1093/sleep/zsab030

5. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;7(8):687-698. doi:10.1016/S2213-2600(19)30198-5

6. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. *Am J Epidemiol*. 2013;177(9):1006-1014. doi:10.1093/aje/kws342

7. Bonsignore MR, Baiamonte P, Mazzuca E, Castrogiovanni A, Marrone O. Obstructive sleep apnea and comorbidities: a dangerous liaison. *Multidiscip Respir Med*. 2019;14(1):8. doi:10.1186/s40248-019-0172-9

8. Iannella G, Pace A, Bellizzi MG, et al. The Global Burden of Obstructive Sleep Apnea. *Diagnostics*. 2025;15(9):1088. doi:10.3390/diagnostics15091088

9. Weatherly HLA, Griffin SC, Mc Daid C, et al. An economic analysis of continuous positive airway pressure for the treatment of obstructive sleep apnea-hypopnea syndrome. *Int J Technol Assess Health Care*. 2009;25(1):26-34. doi:10.1017/S0266462309090047

10. Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The Occurrence of Sleep-Disordered Breathing among Middle-Aged Adults. *N Engl J Med*. 1993;328(17):1230-1235. doi:10.1056/NEJM199304293281704

11. Patil SP, Ayappa IA, Caples SM, Kimoff RJ, Patel SR, Harrod CG. Treatment of Adult Obstructive Sleep Apnea with Positive Airway Pressure: An American Academy of Sleep Medicine Clinical Practice Guideline. *J Clin Sleep Med*. 2019;15(2):335-343. doi:10.5664/jcsm.7640

12. Weaver TE, Grunstein RR. Adherence to Continuous Positive Airway Pressure Therapy: The Challenge to Effective Treatment. *Proc Am Thorac Soc*. 2008;5(2):173-178. doi:10.1513/pats.200708-119MG

13. Sutherland K, Vanderveken OM, Tsuda H, et al. Oral Appliance Treatment for Obstructive Sleep Apnea: An Update. *J Clin Sleep Med*. 2014;10(2):215-227. doi:10.5664/jcsm.3460

14. Dolek T, Akpinar D, Bek S, Cantas Turkis F, Kutlu G. A comparative analysis of automatic and manual scoring methods in polysomnography. *Sleep*. 2025;48(10). doi:10.1093/sleep/zsaf137

15. Masoud AI, Patwari PP, Adavadkar PA, Arantes H, Park C, Carley DavidW. Validation of the MediByte Portable Monitor for the Diagnosis of Sleep Apnea in Pediatric Patients. *J Clin Sleep Med*. 2019;15(05):733-742. doi:10.5664/jcsm.7764

16. Campbell AJ, Neill AM. Home set-up polysomnography in the assessment of suspected obstructive sleep apnea. *J Sleep Res*. 2011;20(1pt2):207-213. doi:10.1111/j.1365-2869.2010.00854.x

17. Ramar K. The COVID-19 pandemic: reflections for the field of sleep medicine. *J Clin Sleep Med*. 2020;16(7):993-996. doi:10.5664/jcsm.8580

18. Punjabi NM, Patil S, Crainiceanu C, Aurora RN. Variability and Misclassification of Sleep Apnea Severity Based on Multi-Night Testing. *Chest*. 2020;158(1):365-373. doi:10.1016/j.chest.2020.01.039

19. Setty AR. Underestimation of Sleep Apnea With Home Sleep Apnea Testing Compared to In-Laboratory Sleep Testing. *J Clin Sleep Med*. 2017;13(4):531-532. doi:10.5664/jcsm.6534

20. Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas*. 2007;28(3):R1-R39. doi:10.1088/0967-3334/28/3/R01

21. Massie F, Mendes de Almeida D, Dreesen P, Thijs I, Vranken J, Klerkx S. An Evaluation of the NightOwl Home Sleep Apnea Testing System. *J Clin Sleep Med*. 2018;14(10):1791-1796. doi:10.5664/jcsm.7398

22. Dzedzickis A, Kaklauskas A, Bucinskas V. Human Emotion Recognition: Review of Sensors and Methods. *Sensors*. 2020;20(3):592. doi:10.3390/s20030592

23. Davis MJ, Hill MA. Signaling mechanisms underlying the vascular myogenic response. *Physiol Rev*. 1999;79(2):387-423. doi:10.1152/physrev.1999.79.2.387

24. Ramezani A, Sharafkhaneh A, BaHammam A, Kuna S, Razjouyan J. Effects of COVID-19 on Sleep Services Use and Its Recovery. *Nat Sci Sleep*. 2024;16:761-768. doi:10.2147/NSS.S456214

25. Zou D, Grote L, Peker Y, Lindblad U, Hedner J. Validation a Portable Monitoring Device for Sleep Apnea Diagnosis in a Population Based Cohort Using Synchronized Home Polysomnography. *Sleep*. 2006;29(3):367-374. doi:10.1093/sleep/29.3.367

26. Allam JS, Ioachimescu OC. Pulse Arterial Tonometry Evaluation of Reliability study. *J Clin Sleep Med*. 2021;17(6):1331-1332. doi:10.5664/jcsm.9272

27. Röcken J, Schumann DM, Herrmann MJ, et al. Peripheral arterial tonometry versus polysomnography in suspected obstructive sleep apnoea. *Eur J Med Res*. 2023;28(1):251. doi:10.1186/s40001-023-01164-w

28. Chiang AA, Jerkins E, Holfinger S, et al. OSA diagnosis goes wearable: are the latest devices ready to shine? *J Clin Sleep Med*. 2024;20(11):1823-1838. doi:10.5664/jcsm.11290

29. Collop NA, Tracy SL, Kapur V, et al. Obstructive Sleep Apnea Devices for Out-Of-Center (OOC) Testing: Technology Evaluation. *J Clin Sleep Med*. 2011;7(5):531-548. doi:10.5664/JCSM.1328

30. Massie F, Van Pee B, Bergmann J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. *J Clin Sleep Med*. 2022;18(3):871-876. doi:10.5664/jcsm.9744

31. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth Analg*. 2018;126(5):1763-1768. doi:10.1213/ANE.0000000000002864

32. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*. 2013;36(4):573-582. doi:10.5665/sleep.2548

33. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. *J Clin Sleep Med*. 2017;13(3):479-504. doi:10.5664/jcsm.6506

34. Caples SM, Gami AS, Somers VK. Obstructive Sleep Apnea. *Ann Intern Med*. 2005;142(3):187-197. doi:10.7326/0003-4819-142-3-200502010-00010

35. Aurora RN, Chowdhuri S, Ramar K, et al. The Treatment of Central Sleep Apnea Syndromes in Adults: Practice Parameters with an Evidence-Based Literature Review and Meta-Analyses. *Sleep*. 2012;35(1). doi:10.5665/sleep.1580

36. Zhang Z, Sowho M, Otvos T, et al. A comparison of automated and manual sleep staging and respiratory event recognition in a portable sleep diagnostic device with in-lab sleep study. *J Clin Sleep Med*. 2020;16(4):563-573. doi:10.5664/jcsm.8278

37. Hedner J, White DP, Malhotra A, et al. Sleep Staging Based on Autonomic Signals: A Multi-Center Validation Study. *J Clin Sleep Med*. 2011;07(03):301-306. doi:10.5664/JCSM.1078

38. O'Brien LM, Bullough AS, Shelgikar AV, Chames MC, Armitage R, Chervin RD. Validation of Watch-PAT-200 Against Polysomnography During Pregnancy. *J Clin Sleep Med*. 2012;08(03):287-294. doi:10.5664/jcsm.1916

39. Yuceege M, Firat H, Demir A, Ardic S. Reliability of the Watch-PAT 200 in Detecting Sleep Apnea in Highway Bus Drivers. *J Clin Sleep Med*. 2013;09(04):339-344. doi:10.5664/jcsm.2584

40. Mokhlesi B, Finn LA, Hagen EW, et al. Obstructive sleep apnea during REM sleep and hypertension: Results of the Wisconsin sleep cohort. *Am J Respir Crit Care Med*. 2014;190(10):1158-1167. doi:10.1164/rccm.201406-1136OC

41. Mokhlesi B, Punjabi NM. “REM-related” Obstructive Sleep Apnea: An Epiphenomenon or a Clinically Important Entity? *Sleep*. 2012;35(1):5-7. doi:10.5665/sleep.1570

42. Varga AW, Mokhlesi B. REM obstructive sleep apnea: risk for adverse health outcomes and novel treatments. *Sleep Breath*. 2019;23(2):413-423. doi:10.1007/s11325-018-1727-2

43. Epstein LJ, Kristo D, Strollo PJ, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J Clin Sleep Med*. 2009;5(3):263-276.

44. Zhang Z, Sowho M, Otvos T, et al. A comparison of automated and manual sleep staging and respiratory event recognition in a portable sleep diagnostic device with in-lab sleep study. *J Clin Sleep Med*. 2020;16(4):563-573. doi:10.5664/jcsm.8278

45. Finnsson E, Ólafsdóttir GH, Loftsdóttir DL, et al. A scalable method of determining physiological endotypes of sleep apnea from a polysomnographic sleep study. *Sleep*. 2021;44(1). doi:10.1093/sleep/zsaa168

46. Edwards BA, Wellman A, Sands SA, et al. Obstructive Sleep Apnea in Older Adults is a Distinctly Different Physiological Phenotype. *Sleep*. 2014;37(7):1227-1236. doi:10.5665/sleep.3844

47. Yalamanchali S, Farajian V, Hamilton C, Pott TR, Samuelson CG, Friedman M. Diagnosis of Obstructive Sleep Apnea by Peripheral Arterial Tonometry. *JAMA Otolaryngol Head Neck Surg*. 2013;139(12):1343. doi:10.1001/jamaoto.2013.5338

48. Flemons WW, Littner MR. Measuring Agreement Between Diagnostic Devices. *Chest*. 2003;124(4):1535-1542. doi:10.1378/chest.124.4.1535

49. Flemons WW, Littner MR, Rowley JA, et al. Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature. *Chest*. 2003;124(4):1543-1579. doi:10.1378/chest.124.4.1543

50. Ioachimescu OC, Allam JS, Samarghandi A, et al. Performance of peripheral arterial tonometry–based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort. *J Clin Sleep Med*. 2020;16(10):1663-1674. doi:10.5664/jcsm.8620

51. Kirsch D. Autopilot and algorithms: accidents, errors, and the current need for human oversight. *J Clin Sleep Med*. 2020;16(10):1651-1652. doi:10.5664/jcsm.8762

52. Holmedahl NH, Fjeldstad OM, Engan H, Saxvig IW, Grønli J. Validation of peripheral arterial tonometry as tool for sleep assessment in chronic obstructive pulmonary disease. *Sci Rep*. 2019;9(1):1-8. doi:10.1038/s41598-019-55958-2

53. Penzel T, Kesper K, Pinnow I, Becker HF, Vogelmeier C. Peripheral arterial tonometry, oximetry and actigraphy for ambulatory recording of sleep apnea. *Physiol Meas*. 2004;25(4):1025-1036. doi:10.1088/0967-3334/25/4/019

54. Gan YJ, Lim L, Chong YK. Validation study of WatchPat 200 for diagnosis of OSA in an Asian cohort. *European Archives of Oto-Rhino-Laryngology*. 2016;274(3):1741-1745. doi:10.1007/s00405-016-4351-4

55. Choi JH, Kim EJ, Kim YS, et al. Validation study of portable device for the diagnosis of obstructive sleep apnea according to the new AASM scoring criteria: Watch-PAT 100. *Acta Otolaryngol*. 2010;130(7):838-843. doi:10.3109/00016480903431139

56. Boyd SB, Upender R, Walters AS, et al. Effective apnea-hypopnea index ("effective ahi"): A new measure of effectiveness for positive airway pressure therapy. *Sleep*. 2016;39(11):1961-1972. doi:10.5665/sleep.6224

57. Weimin L, Rongguang W, Dongyan H, Xiaoli L, Wei J, Shiming Y. Assessment of a portable monitoring device WatchPAT 200 in the diagnosis of obstructive sleep apnea. *Eur Arch Otorhinolaryngol*. 2013;270(12):3099-3105. doi:10.1007/s00405-013-2555-4

58. Pinto JA, de Godoy LBM, Ribeiro RC, Mizoguchi EI, Hirsch LAM, Gomes LM. Accuracy of peripheral arterial tonometry in the diagnosis of obstructive sleep apnea. *Braz J Otorhinolaryngol*. 2015;81(5):473-478. doi:10.1016/j.bjorl.2015.07.005

59. Körkuyu E, Düzlü M, Karamert R, et al. The efficacy of Watch PAT in obstructive sleep apnea syndrome diagnosis. *Eur Arch Otorhinolaryngol*. 2015;272(1):111-116. doi:10.1007/s00405-014-3097-0

60. Ayas NT, Pittman S, MacDonald M, White DP. Assessment of a wrist-worn device in the detection of obstructive sleep apnea. *Sleep Med*. 2003;4(5):435-442. doi:10.1016/S1389-9457(03)00111-4

61. Jen R, Orr JE, Li Y, et al. Accuracy of WatchPAT for the Diagnosis of Obstructive Sleep Apnea in Patients with Chronic Obstructive Pulmonary Disease. *COPD*. 2020;17(1):34-39. doi:10.1080/15412555.2019.1707789

62. Choi JH, Lee B, Lee JY, Kim HJ. Validating the Watch-PAT for Diagnosing Obstructive Sleep Apnea in Adolescents. *J Clin Sleep Med*. 2018;14(10):1741-1747. doi:10.5664/jcsm.7386

63. Pittman SD, Pillar G, Berry RB, Malhotra A, MacDonald MM, White DP. Follow-up assessment of CPAP efficacy in patients with obstructive sleep apnea using an ambulatory device based on peripheral arterial tonometry. *Sleep Breath*. 2006;10(3):123-131. doi:10.1007/s11325-006-0058-x

64. Pillar G, Berall M, Berry R, et al. Detecting central sleep apnea in adult patients using WatchPAT—a multicenter validation study. *Sleep Breath*. 2020;24(1):387-398. doi:10.1007/s11325-019-01904-5

65. Tauman R, Berall M, Berry R, et al. Watch-pat is useful in the diagnosis of sleep apnea in patients with atrial fibrillation. *Nat Sci Sleep*. 2020;12:1115-1121. doi:10.2147/NSS.S278752

66. Zou D, Vits S, Egea C, et al. A new approach to streamline obstructive sleep apnea therapy access using peripheral arterial tone-based home sleep test devices. *Front Sleep*. 2023;2:1256078. doi:10.3389/frsle.2023.1256078

67. Hertzman AB. Photoelectric Plethysmography of the Fingers and Toes in Man. *Exp Biol Med*. 1937;37(3):529-534. doi:10.3181/00379727-37-9630

68. Kamal AA, Harness JB, Irving G, Mearns AJ. Skin photoplethysmography–a review. *Comput Methods Programs Biomed*. 1989;28(4):257-269. doi:10.1016/0169-2607(89)90159-4

69. Hertzman AB, Roth LW. The absence of vasoconstrictor reflexes in the forehead circulation. Effects of cold. *Am J Physiol Leg-Cont*. 1942;136(4):692-697. doi:10.1152/ajplegacy.1942.136.4.692

70. Lugaresi E, Coccagna G, Mantovani M, Lebrun R. Some periodic phenomena arising during drowsiness and sleep in man. *Electroencephalogr Clin Neurophysiol*. 1972;32(6):701-705. doi:10.1016/0013-4694(72)90106-X

71. Pillar G, Bar A, Shlitner A, Schnall R, Shefy J, Lavie P. Autonomic Arousal Index: an Automated Detection Based on Peripheral Arterial Tonometry. *Sleep*. 2002;25(5):541-547. doi:10.1093/sleep/25.5.541

72. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104

73. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310.

74. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30-46. doi:10.1037/1082-989X.1.1.30

75. Nelson LS. The Anderson-Darling Test for Normality. *J Qual Technol*. 1998;30(3):298-299. doi:10.1080/00224065.1998.11979858

76. Bonetti PO, Pumper GM, Higano ST, Holmes DR, Kuvin JT, Lerman A. Noninvasive identification of patients with early coronary atherosclerosis by assessment of digital reactive hyperemia. *J Am Coll Cardiol*. 2004;44(11):2137-2141. doi:10.1016/j.jacc.2004.08.062

77. Punjabi NM. The epidemiology of adult obstructive sleep apnea. *Proc Am Thorac Soc*. 2008;5(2):136-143. doi:10.1513/pats.200709-155MG

78. Dempsey JA, Veasey SC, Morgan BJ, O'Donnell CP. Pathophysiology of Sleep Apnea. *Physiol Rev*. 2010;90(1):47-112. doi:10.1152/physrev.00043.2008

79. Yen FC, Behbehani K, Lucas EA, Burk JR, Axe JR. A noninvasive technique for detecting obstructive and central sleep apnea. *IEEE Trans Biomed Eng*. 1997;44(12):1262-1268. doi:10.1109/10.649998

80. Smallwood RG, Vitiello M V, Giblin EC, Prinz PN. Sleep Apnea: Relationship to Age, Sex, and Alzheimer's Dementia. *Sleep*. 1983;6(1):16-22. doi:10.1093/sleep/6.1.16

81. Pevernagie DA, Gnidovec-Strazisar B, Grote L, et al. On the rise and fall of the apnea–hypopnea index: A historical review and critical appraisal. *J Sleep Res*. 2020;29(4):e13066. doi:10.1111/jsr.13066

82. Eckert DJ, Jordan AS, Merchia P, Malhotra A. Central Sleep Apnea. *Chest*. 2007;131(2):595-607. doi:10.1378/chest.06.2287

83. Javaheri S, Dempsey JA. Central Sleep Apnea. In: Prakash YS, ed. *Compr Physiol*. 1st ed. Wiley; 2013:141-163. doi:10.1002/cphy.c110057

84. Sullivan CE, Berthon-Jones M, Issa FG, Eves L. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet*. 1981;317(8225):862-865. doi:10.1016/S0140-6736(81)92140-1

85. Randerath WJ, Verbraecken J, Andreas S, et al. Non-CPAP therapies in obstructive sleep apnoea. *Eur Respir J*. 2011;37(5):1000-1028. doi:10.1183/09031936.00099710

86. Bloch KE. Polysomnography: a systematic review. *Technol Health Care*. 1997;5(4):285-305. doi:10.3233/THC-1997-5403

87. Kushida CA, Littner MR, Morgenthaler T, et al. Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep*. 2005;28(4):499-523. doi:10.1093/sleep/28.4.499

88. Rundo JV, Downey R. Polysomnography. In: *Handbook of Clinical Neurology*. Vol 160. Elsevier; 2019:381-392. doi:10.1016/B978-0-444-64032-1.00025-4

89. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. *Sleep*. 1999;22(5):667-689.

90. Lazazzera R, Deviaene M, Varon C, et al. Detection and Classification of Sleep Apnea and Hypopnea Using PPG and SpO2 Signals. *IEEE Trans Biomed Eng*. 2021;68(5):1496-1506. doi:10.1109/TBME.2020.3028041

91. Alian AA, Shelley KH. Photoplethysmography. *Best Pract Res Clin Anaesthesiol*. 2014;28(4):395-406. doi:10.1016/j.bpa.2014.08.006

92. Van Pee B, Massie F, Vits S, et al. A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry. *Sleep*. 2022;45(5). doi:10.1093/sleep/zsac028

93. Boccignone G, D'Amelio A, Ghezzi O, Grossi G, Lanzarotti R. An Evaluation of Non-Contact Photoplethysmography-Based Methods for Remote Respiratory Rate Estimation. *Sensors*. 2023;23(7):3387. doi:10.3390/s23073387

94. Feiner JR, Severinghaus JW, Bickler PE. Dark Skin Decreases the Accuracy of Pulse Oximeters at Low Oxygen Saturation: The Effects of Oximeter Probe Type and Gender. *Anesth Analg*. 2007;105(6):S18-S23. doi:10.1213/01.ane.0000285988.35174.d9

95. Zhou Y, Zhao D, Yin G, et al. Analysis of the myoelectric characteristics of genioglossus in REM sleep and its improvement by CPAP treatment in OSA patients. *Sleep Breath*. 2020;24(2):471-482. doi:10.1007/s11325-019-01875-7

96. Douglas NJ, White DP, Weil J V, Pickett CK, Zwillich CW. Hypercapnic ventilatory response in sleeping adults. *Am Rev Respir Dis*. 1982;126(5):758-762. doi:10.1164/arrd.1982.126.5.758

97. Somers VK, Dyken ME, Clary MP, Abboud FM. Sympathetic neural mechanisms in obstructive sleep apnea. *J Clin Invest*. 1995;96(4):1897-1904. doi:10.1172/JCI118235

98. Schächinger H, Weinbacher M, Kiss A, Ritz R, Langewitz W. Cardiovascular Indices of Peripheral and Central Sympathetic Activation. *Psychosom Med*. 2001;63(5):788-796. doi:10.1097/00006842-200109000-00012

99. Budhiraja R, Parthasarathy S, Quan SF. Endothelial dysfunction in obstructive sleep apnea. *J Clin Sleep Med*. 2007;3(4):409-415.

100. Dean DA, Wang R, Jacobs DR, et al. A systematic assessment of the association of polysomnographic indices with blood pressure: The multi-ethnic study of atherosclerosis (MESA). *Sleep*. 2015;38(4):587-596. doi:10.5665/sleep.4576

101. Budhiraja R, Quan SF. Sleep-disordered breathing and cardiovascular health. *Curr Opin Pulm Med*. 2005;11(6):501-506. doi:10.1097/01.mcp.0000183058.52924.70

102. Lévy P, Pépin JL, Arnaud C, Baguet JP, Dematteis M, Mach F. Obstructive Sleep Apnea and Atherosclerosis. *Prog Cardiovasc Dis*. 2009;51(5):400-410. doi:10.1016/j.pcad.2008.03.001

103. Stansbury RC, Strollo PJ. Clinical manifestations of sleep apnea. *J Thorac Dis*. 2015;7(9):E298-E310. doi:10.3978/j.issn.2072-1439.2015.09.13

104. Alzoubaidi M, Mokhlesi B. Obstructive sleep apnea during rapid eye movement sleep. *Curr Opin Pulm Med*. 2016;22(6):545-554. doi:10.1097/MCP.0000000000000319

105. Oksenberg A, Arons E, Nasser K, Vander T, Radwan H. REM-related Obstructive Sleep Apnea: The Effect of Body Position. *J Clin Sleep Med*. 2010;06(04):343-348. doi:10.5664/jcsm.27875

106. Haba-Rubio J, Janssens JP, Rochat T, Sforza E. Rapid eye movement-related disordered breathing: Clinical and polysomnographic features. *Chest*. 2005;128(5):3350-3357. doi:10.1378/chest.128.5.3350

107. Gabryelska A, Białasiewicz P. Association between excessive daytime sleepiness, REM phenotype and severity of obstructive sleep apnea. *Sci Rep*. 2020;10(1):1-6. doi:10.1038/s41598-019-56478-9

108. Su CS, Liu KT, Panjapornpon K, Andrews N, Foldvary-Schaefer N. Functional Outcomes in Patients with REM-Related Obstructive Sleep Apnea Treated with Positive Airway Pressure Therapy. *J Clin Sleep Med*. 2012;08(03):243-247. doi:10.5664/jcsm.1902

109. Nisha Aurora R, Crainiceanu C, Gottlieb DJ, Kim JS, Punjabi NM. Obstructive sleep apnea during REM sleep and cardiovascular disease. *Am J Respir Crit Care Med*. 2018;197(5):653-660. doi:10.1164/rccm.201706-1112OC

110. Duce B, Kulkas A, Langton C, Töyräs J, Hukins C. The prevalence of REM-related obstructive sleep apnoea is reduced by the AASM 2012 hypopnoea criteria. *Sleep Breath*. 2018;22(1):57-64. doi:10.1007/s11325-017-1526-1

111. Zinchuk A, Yaggi HK. Phenotypic Subtypes of OSA: A Challenge and Opportunity for Precision Medicine. *Chest*. 2020;157(2):403-420. doi:10.1016/j.chest.2019.09.002

112. Chiao W, Durr ML. Trends in sleep studies performed for Medicare beneficiaries. *Laryngoscope*. 2017;127(12):2891-2896. doi:10.1002/lary.26736

113. Johnson KG, Sullivan SS, Nti A, Rastegar V, Gurubhagavatula I. The impact of the COVID-19 pandemic on sleep medicine practices. *J Clin Sleep Med*. 2021;17(1):79-87. doi:10.5664/jcsm.8830

114. Herscovici S, Pe'er A, Papyan S, Lavie P. Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy. *Physiol Meas*. 2007;28(2):129-140. doi:10.1088/0967-3334/28/2/002

115. Dvir I, Adler Y, Freimark D, Lavie P. Evidence for fractal correlation properties in variations of peripheral arterial tone during REM sleep. *Am J Physiol Heart Circ Physiol*. 2002;283(1):H434–H439. doi:10.1152/ajpheart.00336.2001

116. Chouchou F, Desseilles M. Heart rate variability: A tool to explore the sleeping brain? *Front Neurosci*. 2014;8(DEC):1-9. doi:10.3389/fnins.2014.00402

117. Choi E, Park DH, Yu JH, Ryu SH, Ha JH. The severity of sleep disordered breathing induces different decrease in the oxygen saturation during rapid eye movement and non-rapid eye movement sleep. *Psychiatry Investig*. 2016;13(6):652-658. doi:10.4306/pi.2016.13.6.652

118. Bianchi MT. Bayes' theorem and the rule of 100: a commentary on 'Validity of administrative data for identification of obstructive sleep apnea.' *J Sleep Res*. 2017;26(3):401. doi:10.1111/jsr.12534

119. Onder NS, Akpinar ME, Yigit O, Gor AP. Watch peripheral arterial tonometry in the diagnosis of obstructive sleep apnea: Influence of aging. *Laryngoscope*. 2012;122(6):1409-1414. doi:10.1002/lary.23233

120. Zinchuk AV, Gentry MJ, Concato J, Yaggi HK. Phenotypes in obstructive sleep apnea: A definition, examples and evolution of approaches. *Sleep Med Rev*. 2017;35:113-123. doi:10.1016/j.smrv.2016.10.002

121. Massie F, Van Pee B, Vits S, Verbraecken J, Bergmann J. Phenotyping REM OSA by means of peripheral arterial tone-based home sleep apnea testing and polysomnography: A critical assessment of the sensitivity and specificity of both methods. *J Sleep Res*. 2022;31(2). doi:10.1111/jsr.13481

122. Newell J, Mairesse O, Verbanck P, Neu D. Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples. *Psychiatry Res*. 2012;200(2-3):795-801. doi:10.1016/j.psychres.2012.07.045

123. Jasper H. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*. 1958;10(2):370-375. doi:10.1016/0013-4694(58)90053-1

124. Eldele E, Chen Z, Liu C, et al. An Attention-Based Deep Learning Approach for Sleep Stage Classification with Single-Channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2021;29:809-818. doi:10.1109/TNSRE.2021.3076234

125. Imtiaz SA. A Systematic Review of Sensing Technologies for Wearable Sleep Staging. *Sensors*. 2021;21(5):1562. doi:10.3390/s21051562

126. Berthomier C, Drouot X, Herman-Stoïca M, et al. Automatic Analysis of Single-Channel Sleep EEG: Validation in Healthy Individuals. *Sleep*. 2007;30(11):1587-1595. doi:10.1093/sleep/30.11.1587

127. Popovic D, Khoo M, Westbrook P. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. *J Sleep Res*. 2014;23(2):211-221. doi:10.1111/jsr.12105

128. Griessenberger H, Heib DPJ, Kunz AB, Hoedlmoser K, Schabus M. Assessment of a wireless headband for automatic sleep scoring. *Sleep Breath*. 2013;17(2):747-752. doi:10.1007/s11325-012-0757-4

129. Shambroom JR, Fábregas SE, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. *J Sleep Res*. 2012;21(2):221-230. doi:10.1111/j.1365-2869.2011.00944.x

130. Arnal PJ, Thorey V, Debellemaniere E, et al. The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*. 2020;43(11). doi:10.1093/sleep/zsaa097

131. Voinescu BI, Wislowska M, Schabus M. Assessment of SOMNOwatch plus EEG for sleep monitoring in healthy individuals. *Physiol Behav*. 2014;132:73-78. doi:10.1016/j.physbeh.2014.04.052

132. Hof Zum Berge A, Kellmann M, Kallweit U, et al. Portable PSG for sleep stage monitoring in sports: Assessment of SOMNOwatch plus EEG. *Eur J Sport Sci*. 2020;20(6):713-721. doi:10.1080/17461391.2019.1659421

133. Finan PH, Richards JM, Gamaldo CE, et al. Validation of a Wireless, Self-Application, Ambulatory Electroencephalographic Sleep Monitoring Device in Healthy Volunteers. *J Clin Sleep Med*. 2016;12(11):1443-1451. doi:10.5664/jcsm.6262

134. Lucey BP, Mcleland JS, Toedebusch CD, et al. Comparison of a single-channel EEG sleep study to polysomnography. *J Sleep Res*. 2016;25(6):625-635. doi:10.1111/jsr.12417

135. Wang Y, Loparo KA, Kelly MR, Kaplan RF. Evaluation of an automated single-channel sleep staging algorithm. *Nat Sci Sleep*. 2015;7:101-111. doi:10.2147/NSS.S77888

136. Casciola AA, Carlucci SK, Kent BA, et al. A Deep Learning Strategy for Automatic Sleep Staging Based on Two-Channel EEG Headband Data. *Sensors*. 2021;21(10):3316. doi:10.3390/s21103316

137. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Trans Neural Syst Rehab Eng*. 2019;27(3):400-410. doi:10.1109/TNSRE.2019.2896659

138. Olivas ES, Guerrero JDM, Martinez-Sober M, Magdalena-Benedito JR, Serrano López AJ. *Handbook of Research on Machine Learning Applications and Trends*. (Olivas ES, Guerrero JDM, Martinez-Sober M, Magdalena-Benedito JR, Serrano López AJ, eds.). IGI Global; 2010. doi:10.4018/978-1-60566-766-9

139. Quan SF, Howard B V, Iber C, et al. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*. 1997;20(12):1077-1085.

140. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;25(10):1351-1358. doi:10.1093/jamia/ocy064

141. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *J Clin Sleep Med*. 2013;9(1):81-87. doi:10.5664/jcsm.2350

142. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385. doi:10.1136/bmj-2023-078378

143. Eckert DJ, Malhotra A. Pathophysiology of Adult Obstructive Sleep Apnea. *Proc Am Thorac Soc*. 2008;5(2):144-153. doi:10.1513/pats.200707-114MG

144. Stoohs R, Guilleminault C. Cardiovascular changes associated with obstructive sleep apnea syndrome. *J Appl Physiol*. 1992;72(2):583-589. doi:10.1152/jappl.1992.72.2.583

145. Kim MJ, Lee GH, Kim CS, et al. Comparison of three actigraphic algorithms used to evaluate sleep in patients with obstructive sleep apnea. *Sleep Breath*. 2013;17(1):297-304. doi:10.1007/s11325-012-0689-z

146. Schlotthauer G, Di Persia LE, Larrateguy LD, Milone DH. Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry. *Med Eng Phys*. 2014;36(8):1074-1080. doi:10.1016/j.medengphy.2014.05.008

147. de Rooij M, Weeda W. Cross-Validation: A Method Every Psychologist Should Know. *Adv Methods Pract Psychol Sci*. 2020;3(2):248-263. doi:10.1177/2515245919898466

148. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91. doi:10.1186/1471-2105-7-91

149. Kunjan S, Grummett TS, Pope KJ, et al. The Necessity of Leave One Subject Out (LOSO) Cross Validation for EEG Disease Diagnosis. In: Mahmud M, Kaiser MS, Vassanelli S, Dai Q, Zhong N, eds. *Brain Informatics*. Vol 12960. Springer International Publishing; 2021:558-567. doi:10.1007/978-3-030-86993-9_50

150. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160. doi:10.1177/096228029900800204

151. Mansukhani MP, Kolla BP, Wang Z, Morgenthaler TI. Effect of Varying Definitions of Hypopnea on the Diagnosis and Clinical Outcomes of Sleep-Disordered Breathing: A Systematic Review and Meta-Analysis. *J Clin Sleep Med*. 2019;15(5):687-696. doi:10.5664/jcsm.7750

152. Ioachimescu OC. On PAT, Patterns and Paths. *Sleep*. 2022;45(5):zsac057. doi:10.1093/sleep/zsac057

153. Martinot JB, Pépin JL, Malhotra A, Le-Dong NN. Near-boundary double-labeling-based classification: the new standard when evaluating performances of new sleep apnea diagnostic solutions against polysomnography? *Sleep*. 2022;45(10). doi:10.1093/sleep/zsac188

154. Xie J, Fonseca P, van Dijk J, Overeem S, Long X. A multi-task learning model using RR intervals and respiratory effort to assess sleep disordered breathing. *Biomed Eng Online*. 2024;23(1):45. doi:10.1186/s12938-024-01240-0

155. Xie J, Fonseca P, van Dijk JP, Long X, Overeem S. The Use of Respiratory Effort Improves an ECG-Based Deep Learning Algorithm to Assess Sleep-Disordered Breathing. *Diagnostics*. 2023;13(13):2146. doi:10.3390/diagnostics13132146

156. Xie J, Fonseca P, van Dijk JP, Overeem S, Long X. Multi-modal multi-task deep neural networks for sleep disordered breathing assessment using cardiac and audio signals. *Int J Med Inform*. 2025;201:105932. doi:10.1016/j.ijmedinf.2025.105932

157. Fonseca P, Ross M, Cerny A, et al. Estimating the Severity of Obstructive Sleep Apnea Using ECG, Respiratory Effort and Neural Networks. *IEEE J Biomed Health Inform*. 2024;28(7):3895-3906. doi:10.1109/JBHI.2024.3383240

158. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799

159. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283-3289. doi:10.1038/s41591-025-03953-8

160. Javaheri S, Randerath WJ, Safwan Badr M, Javaheri S. Medication-induced central sleep apnea: a unifying concept. *Sleep*. 2024;47(8):zsae038. doi:10.1093/sleep/zsae038

161. Hynes DJ, Mann DL, Landry SA, Joosten SA, Edwards BA, Hamilton GS. Night-to-night variability in obstructive sleep apnea severity, the physiological endotypes, and the frequency of flow limitation. *Sleep*. 2025;48(4). doi:10.1093/sleep/zsae295

162. Shahrbabaki SS, Linz D, Hartmann S, Redline S, Baumert M. Sleep arousal burden is associated with long-term all-cause and cardiovascular mortality in 8001 community-dwelling older men and women. *Eur Heart J*. 2021;42(21):2088-2099. doi:10.1093/eurheartj/ehab151

163. Provini F, Vetrugno R, Meletti S, et al. Motor pattern of periodic limb movements during sleep. *Neurology*. 2001;57(2):300-304. doi:10.1212/WNL.57.2.300

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Jeroen Bergmann, for his exemplary mentorship and advice, and kind-heartedness. My appreciation extends to all members of the Natural Interaction Lab for fostering a collaborative and stimulating research environment.

I am equally grateful to the research collaborators of my published work for their contributions, constructive feedback, and collegial spirit, which enriched both the quality and clarity of this thesis.

Special thanks go to the clinical teams and centres involved in the data collection for the studies presented in my work. Their efforts made the practical validation of the work possible.

I would also like to thank my Transfer and Confirmation assessors for their critical evaluation and valuable feedback at key stages of this doctoral journey.

Finally, I am deeply thankful to my friends and family for their unwavering support throughout this process.