Received: 6 October 2022 | Revised: 16 December 2022 | Accepted: 18 December 2022

DOI: 10.1111/jsr.13819

REVIEW ARTICLE

Consumer sleep technology for the screening of obstructive sleep apnea and snoring: current status and a protocol for a systematic review and meta-analysis of diagnostic test accuracy

Gabriel Natan Pires[1,2] | Erna Sif Arnardóttir[3,4] | Anna Sigridur Islind[3,5] | Timo Leppänen[6,7,8] | Walter T. McNicholas[9]

[1]Departamento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, Brazil

[2]European Sleep Research Society (ESRS), Regensburg, Germany

[3]Reykjavik University Sleep Institute, Reykjavik University, Reykjavik, Iceland

[4]Landspitali—The National University Hospital of Iceland, Reykjavik, Iceland

[5]Department of Computer Science, Reykjavik University, Reykjavik, Iceland

[6]Department of Technical Physics, University of Eastern Finland, Kuopio, Finland

[7]Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland

[8]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

[9]Department of Respiratory and Sleep Medicine, St Vincent's Hospital Group, School of Medicine, University College Dublin, Dublin, Ireland

Correspondence
Gabriel Natan Pires, Departamento de Psicobiologia—Universidade Federal de São Paulo, Rua Napoleão de Barros, 925—CEP: 04024-002—São Paulo, Brazil.
Email: gnspires@gmail.com, gabriel.pires@unifesp.br

Funding information
European Union's Horizon 2020 research and innovation program, Grant/Award Number: 965417; Academy of Finland, Grant/Award Number: 323536; Kuopio University Hospital Catchment Area for the State Research Funding, Grant/Award Number: 5041794; NordForsk, Grant/Award Number: 90458

Summary

There are concerns about the validation and accuracy of currently available consumer sleep technology for sleep-disordered breathing. The present report provides a background review of existing consumer sleep technologies and discloses the methods and procedures for a systematic review and meta-analysis of diagnostic test accuracy of these devices and apps for the detection of obstructive sleep apnea and snoring in comparison with polysomnography. The search will be performed in four databases (PubMed, Scopus, Web of Science, and the Cochrane Library). Studies will be selected in two steps, first by an analysis of abstracts followed by full-text analysis, and two independent reviewers will perform both phases. Primary outcomes include apnea–hypopnea index, respiratory disturbance index, respiratory event index, oxygen desaturation index, and snoring duration for both index and reference tests, as well as the number of true positives, false positives, true negatives, and false negatives for each threshold, as well as for epoch-by-epoch and event-by-event results, which will be considered for the calculation of surrogate measures (including sensitivity, specificity, and accuracy). Diagnostic test accuracy meta-analyses will be performed using the Chu and Cole bivariate binomial model. Mean difference meta-analysis will be performed for continuous outcomes using the DerSimonian and Laird random-effects model. Analyses will be performed independently for each outcome. Subgroup and sensitivity analyses will evaluate the effects of the types (wearables, nearables, bed sensors, smartphone applications), technologies (e.g., oximeter, microphone, arterial tonometry, accelerometer), the role of manufacturers, and the representativeness of the samples.

KEYWORDS
digital health, digital medicine, mobile applications, sleep trackers, smartphones, smartwatches, wearables

1 | INTRODUCTION

Sleep-disordered breathing (SDB), especially obstructive sleep apnea (OSA), is extremely prevalent and has been estimated to affect up to 1 billion adults worldwide (Benjafield et al., 2019). The traditional diagnosis approach has been overnight polysomnography (PSG) in a sleep laboratory, although there has been an increasing focus in recent years on ambulatory studies, often using simplified recording techniques (Arnardottir, Islind, & Óskarsdóttir, 2021). Furthermore, there has been increasing interest in consumer sleep technology (CST), which refers to any kind of technology or equipment (usually wearables, nearables, bed sensors, or mobile applications running on smartphones [apps]) that are marketed directly to consumers, without a need for prescription by a health professional, allowing individuals to self-monitor or track their sleep, or to manage or improve a certain sleep-related condition (Khosla et al., 2018; Schutte-Rodin et al., 2021). A large portion of those have not been designed with clinical use in mind (Schmitz et al., 2022). Several CSTs are related to the screening of SDB, and the number of devices and apps available for this purpose has been increasing consistently in recent years. These CSTs are usually directed at detecting either snoring or OSA, and their sensors and presentations vary considerably (O'Mahony, Garvey, & McNicholas, 2020). This includes the evaluation of pulse oximetry, heart rate and heart rate variability, respiratory rate, breathing-related sounds, and body movements, among others (O'Mahony et al., 2020; Perez-Pozuelo et al., 2020). The present paper provides a background review of existing CST devices related to OSA and snoring, and proposes a protocol for a systematic review and meta-analysis of existing CST devices regarding their diagnostic test accuracy (DTA).

2 | BACKGROUND

2.1 | The rise of CST for SDB screening

The increasing use of innovative technologies for SDB screening (mainly OSA and snoring) might be explained by several factors, from epidemiological, diagnostic, and commercial aspects. From an epidemiological perspective, OSA is an increasingly prevalent condition, ranging from 9% to 42% (Senaratna et al., 2017) and affecting ~936 million people worldwide (Benjafield et al., 2019). From a diagnostic perspective, the 'gold-standard' diagnosis of OSA is an in-laboratory PSG recording, which encompasses important limitations related to its costs, availability, data variability, and patient experience (Box 1). Also, the diagnosis is limited to a few diagnostic metrics, such as the apnea–hypopnea index (AHI) and the respiratory disturbance index (RDI), which often fail to capture all the aspects of OSA severity (Pevernagie et al., 2020). Finally, the market for consumer-oriented sleep products has been growing considerably, reaching $2 billion (American dollars) and growing 18.5% per year (Nester, 2019).

In short, the high prevalence of OSA and snoring creates the demand for screening and diagnosis, while the limitations of the traditional PSG diagnosis raise the importance of not only focusing on new screening tools, but also on patient- or user-experience (including comfort, usability availability, and affordability). These factors led the sleep-related market to grow, which stimulates companies to invest, innovate, design, develop, and improve sleep technologies directly to consumers, bypassing the role of health professionals, and in some cases even patients, in the screening process.

2.2 | Overview of technologies available

Type-I PSG is considered to be the gold-standard method for OSA diagnosis (Kapur et al., 2017). It consists of an in-laboratory overnight sleep study, which is prepared and monitored by a sleep technologist. A type-I PSG requires the acquisition and analysis of at least eight biological signals: an electroencephalogram (EEG), electro-oculogram (EOG), chin and legs electromyogram (EMG), airflow signals, respiratory effort, oxygen saturation, body position, and electrocardiogram, as described in the regularly updated American Academy of Sleep Medicine (AASM) manual (Berry et al., 2020).

The scoring of respiratory events in regular type-I PSG depends on two fundamental aspects: (i) sleep staging and (ii) detection and quantification of respiratory events. Sleep staging is important to prevent scoring respiratory events during wakefulness and to assess whether they are associated with a specific sleep stage. In addition, for some respiratory events, the EEG is required to be able to detect associated arousals or sleep fragmentation. Regarding the scoring of respiratory events, current guidelines (Berry et al., 2020) require four different types of sensors: an oronasal thermal sensor, a nasal pressure transducer, pulse oximetry, and some measure of respiratory effort (usually thoracoabdominal belts).

The need for all these sensors that require expert setup reinforces the limitations of type-I PSG (as disclosed in Box 1). Therefore, alternative ways to screen for SDB have been used and proposed, including the advent of home sleep apnea tests (HSATs), which encompasses sleep study types II–IV. A type-II PSG uses the same sensors and montage as a type-I but is thought to be performed unattended and without real-time supervision of a healthcare professional, therefore allowing the sleep study to be performed outside of a medical facility (Kapur et al., 2017). Although it might represent some improvement in the patient-experience in comparison with type-I PSG, the maintenance of all sensors still represents an important limitation.

A common strategy to overcome these issues is reducing the number of sensors, such as in portable cardiorespiratory sleep monitors (including type-III and type-IV sleep studies) (Kapur et al., 2017). These devices are usually restricted to the monitoring of cardiorespiratory variables, typically not including EEG, EOG, or EMG sensors, therefore not allowing the performance of sleep staging. Although reducing the number of sensors has practical benefits, it might lead to a reduction in diagnostic sensitivity, reducing the ability to rule out OSA (Caples, Anderson, Calero, Howell, & Hashmi, 2021). It also precludes the possibility of evaluating other sleep disorders (such as

BOX 1 Limitations on polysomnography-based diagnosis of sleep disorders.

Availability	**PSG beds might not be available in many medical centres, especially out of big cities and in rural areas, as it requires specialised healthcare professionals and an adequate laboratory setting. The higher prevalence of sleep disorders associated with the unavailability of sleep medicine centres increases the likelihood of underdiagnosis of OSA.**
Costs	Even when PSG is available it might not be affordable to many patients. It is usually an expensive medical examination, as its price must encompass costs related to devices, health professionals, and sleep laboratory maintenance. Costs-related concerns also justify the limited availability of PSG on public health systems and healthcare insurance plans.
Data variability	OSA is subjected to an important night-to-night variability. The variation in the AHI is >10 events/h in 65% of the individuals undergoing PSGs on sequential nights (Bittencourt et al., 2001). As the diagnosis of OSA is usually performed with a single-night PSG, there is a risk of misclassification due to data variability, which might affect diagnosis, treatment, and prevalence estimates.
Limitations related to manual analysis	Manual analysis of a PSG recording is still the 'gold standard' method to analyse and score it. It requires a sleep technologist to overview the whole recording to perform sleep staging and to score other sleep-associated events (e.g., respiratory events, arousals, leg movements). This process has three main limitations: 1. Time-consuming: the manual analysis of a PSG usually requires ~1.5 h of work from a sleep technologist (Fischer et al., 2012). 2. Prone to human errors: although good agreement rates among experienced sleep technologists have been reported (Kuna et al., 2013; Lee et al., 2022; Magalang et al., 2016), the manual analysis might be subjected to a significant amount of imprecision, especially among unexperienced scorers. 3. Costs: the need for sleep technologists scoring the PSG increases its costs, contributing to its limited affordability. These limitations could be overcome by improved semi-automatic analysis or by automatic algorithms (as used by many wearables/nearables devices).
Patient experience	Sleeping at a laboratory under constant monitoring might be an uncomfortable experience for many patients. Among the several aspects that might reduce the patient experience while undergoing a PSG are: 1. Sleeping out of their own rooms with bed and pillows they are not used to. 2. Subjected to environmental conditions different from what they are familiar with (including light, noise, and companion). 3. Unable to follow a usual pre-sleep routine. 4. Different timing for going to bed and waking up than normally. 5. Dealing with the discomfort that the PSG devices might cause; and 6. Being monitored by healthcare professionals at a medical facility. All these conditions might lead to altered sleep patterns, which are caused by environmental conditions rather than by a sleep disorder. These effects are especially observed in a first PSG ('first-night effect', Ding, Chen, Dai, & Li, 2022), contributing to the data variability often seen in PSGs.

Abbreviations: AHI, apnea–hypopnea index; OSA, Obstructive sleep apnea; PSG, polysomnography.

periodic limb movement disorder), thus impairing a proper differential diagnosis. Therefore, simplified sleep studies are useful in cases of well-grounded clinical suspicion of moderate-to-severe OSA, with no comorbid medical disorders or risk of other sleep disorders (Collop et al., 2007; Kapur et al., 2017). Also, just as for any sleep study, its results alone (i.e., without proper clinical evaluation by a health provider) are not sufficient for diagnosis, evaluation of clinical efficacy, and treatment decision (Rosen et al., 2018).

The CST measurement for OSA and snoring screening also embrace the idea of reducing the number of sensors to the minimum necessary for accurate results. Of note, as CSTs are marketed directly to the consumers bypassing the role of the medical professional and patients, it is more appropriate to consider them as screening devices, rather than as diagnostic tools at this particular time. In any case, the evaluation of the accuracy of CSTs in comparison to proper diagnostic tests is important, in order to assure their reliability.

The first widespread CST options for SDB screening were probably smartphone apps for snoring detection. Their technology is simple, especially for the apps that have no intention to diagnose or correlate it with OSA severity, as their functions are usually restricted to the use of a microphone. The apps using a microphone as the main sensor appear to perform well in the detection of snoring and provide stable data with overall good accuracy (Camacho et al., 2015; Chiang et al., 2022; Figueras-Alvarez et al., 2020; Klaus, Stummer, & Ruf, 2021). However, the specificity might be low in a real-world scenario, as the apps might confound snoring from the bed partner, other respiratory sounds from the user, and background noise with actual snoring sounds from the user (Camacho et al., 2015; Stippig, Hübers, & Emerich, 2015).

Although snoring detection has some clinical usefulness (Camacho et al., 2015), many companies have tried to improve the screening capabilities of their CST by estimating OSA based on the snore events. The use of respiratory sounds and movements has also been used for this purpose, employing more refined data analyses (such as spectral analysis of respiratory sounds or using the smartphone as a sonar for detecting respiratory movements). The respiratory flow or pattern is estimated from it, and changes to background patterns are interpreted as possible obstructive events. Although they perform well in some cases, sensitivity and specificity are usually <90%, being as low as 60% in some cases (Cho et al., 2022; Nakano et al., 2014; Narayan et al., 2019; Tiron et al., 2020).

Several other physiological measurements are currently being used for the portable assessment of OSA, some of them being included in CSTs. These include ultrasound and radiofrequency sensors, airflow analysis, pulse oximetry, arterial tonometry, photoplethysmography, and heart rate variability, among others (Behar, Roebuck, Domingos, Gederi, & Clifford, 2013; Penzel, Dietz-Terjung, Woehrle, & Schöbel, 2021; Uddin, Chow, & Su, 2018). Airflow and pulse oximetry seem to be the most logical variables to be analysed in CST OSA monitoring (Uddin et al., 2018), as they are more closely related to OSA pathophysiology. While devices based on direct airflow analyses are not often seen, oximetry-based analyses became more common with the advent of fitness trackers, smartwatches, and rings. The oximeters embedded into wearable devices seem to be accurate in detecting hypoxia in multiple conditions, including during daily life activities, during sleep, and in experimentally induced hypoxia (Jung et al., 2022; Marinari et al., 2022; Santos et al., 2022; Zhao et al., 2022).

The incorporation of additional data to pulse oximetry analyses appears to increase the accuracy, with movement, sound, and heart rate being the most commonly used parameters. Taking that all into account, the common sense is that single signal-based OSA detection is less accurate, being only able to differentiate between the presence or absence of OSA, while multi-signal detection is more accurate, being useful for detecting different levels of disease severity (Uddin et al., 2018). However, this might change as technology and data analysis evolve. As an example, a recent study using artificial neural network analysis of oxygen saturation (SpO_2) led to a median absolute error in the estimation of the AHI of ~1 event/h (Nikkonen, Afara, Leppänen, & Töyräs, 2020).

For consumer-based OSA screening, the most common formats are smartphone apps and wearables. The usefulness of smartphone apps depends on a combination of the smartphone apps' characteristics (and the algorithms embedded in them) and the sensors, which are embedded in the smartphone (or tablet), with variable quality depending on the model. There are 100s of smartphone apps available for OSA detection, but only ~3% of them provide proper validation studies (Baptista et al., 2022). A recent meta-analysis (Kim, Kim, & Hwang, 2022) concluded that the sensitivity of smartphone-based tools for the screening of OSA is >80% in all cases, regardless of the sensors being used. Regarding wearables, their sensitivity and specificity tend to be higher than what is observed in smartphones due to the higher number of sensors and variables being analysed. However, their actual diagnostic accuracy, sensitivity, and specificity are subject to great variability, ranging from ~40% to 90%, depending on the device, manufacturer, sensor type, and data analysis strategy used (Chen, Wang, Guo, Zhang, & Xie, 2021; John, Nundy, Cardiff, & John, 2021; Mokhtaran et al., 2022; Papini et al., 2020).

More recently, some innovative devices have been proposed to screen for OSA in the home setting, including new bed sensor devices, nearables, and wearables (Óskarsdóttir et al., 2022). Their accuracy might vary depending on the type of the device, the reference test, and the OSA classification threshold being considered, with sensitivity estimates ranging from 45.0% to 97.6% and specificity ranging from 51.3% to 97.8% (Rosa, Bellardi, Viana, Ma, & Capasso, 2018). Another characteristic of these devices is the improvement of the sensors used, both in their technology and the position where they are located. Such innovation seems to arise from a concomitant concern related to inventiveness, patentability, and diagnostic accuracy. Regarding the position of these devices, fingertip oximeters and rings are among the most common (Gu et al., 2020; Zhao et al., 2022), but they also include devices based on neck collars, mandibular movement monitors (Pepin et al., 2022; Pépin et al., 2020), and surface acoustic wave sensors (Jin et al., 2017).

2.3 | Problems and concerns regarding CST for SDB screening

Although a user-centred approach has benefits, there are several concerns regarding the validation and accuracy of CSTs. The first concern regards the reduction in the number of sensors. Type-I PSG has been developed in a way that respiratory events can be identified from different perspectives, approaching different pathophysiological manifestations of apneas and hypopneas (including airflow limitation, respiratory effort, desaturations, and arousals). Arguably, there is a trade-off between a reduced number of sensors and a decrease in accuracy, which might reflect in things such as variable sensitivity to detect hypopneas (especially when not associated with desaturation), and inability to differentiate obstructive and central events, among others.

The second problem relates to the indirectness in the assessment of sleep-related parameters. The more indirect a given measure, the higher the chance of this measure not being accurate in the detection of a given event. The most evident case of indirectness in CST regards sleep staging, which is primarily a neurobiological variable. CSTs that do not include an EEG might try to infer sleep stages based on other variables. Body movements and heart rate variability are the most frequently used variables to approximate sleep stages. Some indirectness is also observed in the detection of apneas or hypopneas, which is directly measured via airflow. Current CSTs use variables like respiratory movements, snore sounds, oxygen saturation, and mandibular movements to approximate respiratory events. However, although they are intimately related to respiratory events, they are not sufficient to diagnose them according to clinical standards (Berry et al., 2020).

A third problem relates to how data are gathered, stored, transmitted, and analysed using CSTs (Perez-Pozuelo et al., 2020). Patient-generated health data are not standardised among these technologies and the algorithms used to analyse data are not frequently available (Arnardottir et al., 2021; Khosla et al., 2018), so clinicians do not have a clear picture of how a certain result is reached.

Fourth, most available CSTs have not been tested in comparison with gold-standard methods nor have been approved by health regulatory agencies (Behar et al., 2013; Fino & Mazzetti, 2019; Khosla et al., 2018), and the sensitivity and specificity are uncertain in many

cases (Khosla et al., 2018; Schutte-Rodin et al., 2021). The lack of standards on the validation, proposal, and registration of CSTs causes a large accuracy variability, as well as uncertainties about their actual usefulness (Baptista et al., 2022; Fino & Mazzetti, 2019). It has been argued that some CSTs perform poorly in clinical samples in comparison to healthy populations (Baron et al., 2018), which is due to the lack of proper validation and confirmation studies in samples of individuals with OSA.

Fifth, many validation studies performed do not provide epoch-by-epoch or event-by-event analysis, disclosing only whole night overall statistics (de Zambotti et al., 2022; Menghini, Cellini, Goldstone, Baker, & de Zambotti, 2021). In these cases, similar metrics might eventually be reached between a CST and a PSG, although they might be labelling and scoring different events. This analysis would be important for a proper performance evaluation of CST, but they seem rather uncommon (Menghini et al., 2021). Protocols and recommendations for the evaluation of epoch-by-epoch and event-by-event analysis have already been published (Borsky, Serwatko, Arnardottir, & Mallett, 2022; Menghini et al., 2021).

Finally, the commercial potential of CSTs and the role of companies in their development might lead to publication bias and selective outcome reporting, therefore resulting in a partial and biased appraisal of data reliability (named as 'industry sponsorship bias'; Holman, Bero, & Mintzes, 2019). This concern is certainly not true for all new technologies. However, previous studies have already demonstrated a negative effect of the involvement of industry on the evaluation of the efficacy of drugs and medical devices (Lundh, Lexchin, Mintzes, Schroll, & Bero, 2017; Xie & Zhou, 2022), and the same might happen to CST.

Considering all these problems, limitations, and uncertainties, a comprehensive data reassessment of the accuracy of CSTs for the screening of OSA and snoring is needed, and it could be achieved by means of a systematic review and meta-analysis. This approach would help to understand the actual accuracy of new CSTs, being also able to detect which sensors and outcome variables are the most suitable for proper screening of OSA and snoring. Therefore, the present protocol discloses the methods and procedures for a systematic review and meta-analysis of DTA of CST for the screening of OSA and snoring.

3 | METHODS

3.1 | Reporting and registration standards

This protocol was prepared according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) – protocol extension (PRISMA-P) (Moher et al., 2015) and was registered at the International Prospective Register of Systematic Reviews (PROSPERO: CRD42022362186). The final report will be written according to the PRISMA-DTA (Salameh et al., 2020).

3.2 | Research question and basic definitions

The basic definitions of the included articles are specified according to the PI(R)T strategy (Participants, Index test, Reference test, Target condition), an adapted version of PICO (Population, Intervention, Comparison, Outcome) for DTA studies (Leeflang, Davenport, & Bossuyt, 2022). These four strategy items are defined below and more details on how each of these items will be addressed and analysed can be found in the 'Inclusion and exclusion criteria' section.

- *Participants*: individuals aged ≥18 years, regardless of diagnosis or suspicion of OSA, other sleep disorders, and co-morbidities.
- *Index test*: consumer-based technology (including devices or apps) for screening of OSA and/or snoring according to the classification proposed by the AASM (Schutte-Rodin et al., 2021).
- *Reference test*: full night type-I or type-II PSG, performed according to the AASM recommendations (Berry et al., 2020) or equivalent guidelines.
- *Target Condition*: OSA and snoring.

Based on these definitions, a list of PI(R)T questions was prepared, by a combination of the target conditions and index tests of interest, considering two different meta-analytical approaches: DTA and mean difference meta-analyses (as properly explained in the 'data synthesis and analyses section'). The list of PI(R)T research questions is presented in Table 1.

Of note, although we acknowledge that CSTs cannot be considered proper diagnostic tools, but rather OSA screening devices, we prefer to keep using the term 'diagnosis' on the research questions and on the statistical analyses. '*DTA meta-analysis*' (DTAMA) is an established meta-analytical approach that has been used whenever the performance of an index and a reference test are compared. The same is true for other statistical terms used throughout the manuscript, including *'diagnostic threshold'* and *'diagnostic odds ratio'* (DOR).

3.3 | Search strategy

A bibliographic search will be performed in four different databases: PubMed, Scopus, Web of Science, and the Cochrane Library. The primary search strategy was developed for PubMed and will be adapted to the syntax and search engines of the other databases.

This search strategy was composed of the combination of two domains: OSA and snoring (as the target condition) and consumer-based technology (as the index test). Both search domains were built by combining Medical Subject Headings (MeSH) terms (available at www.ncbi.nlm.nih.gov/mesh/) and relevant free terms, including spelling variations, alternative nomenclature, and plural forms. No search domain regarding PSG (as a reference test) was included, in order to increase search sensitivity. Any possible search strategy related to PSG would be redundant with the OSA domain, decreasing the search

TABLE 1 Population, Index test, Reference test, Target condition (PI(R)T) questions

#	Outcome	Index tests	Research questions
1.	Diagnostic test accuracy meta-analyses – Diagnosis of OSA and snoring		
1.1	Obstructive sleep apnea (OSA)[a]		
1.1.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the screening of OSA in comparison with PSG?
1.1.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the screening of OSA in comparison with PSG?
1.1.3		Nearable devices	What is the diagnostic accuracy of nearable devices in the screening of OSA in comparison with PSG?
1.1.4		Bed sensors	What is the diagnostic accuracy of bed sensors in the screening of OSA in comparison with PSG?
1.1.5		Apps	What is the diagnostic accuracy of apps in the screening of OSA in comparison with PSG?
1.2	Snoring		
1.2.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the detection of snoring in comparison with PSG?
1.2.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the detection of snoring in comparison with PSG?
1.2.3		Nearable devices	What is the diagnostic accuracy of nearable devices on the detection of snoring in comparison with PSG?
1.2.4		Bed sensors	What is the diagnostic accuracy of bed sensors on the detection of snoring in comparison with PSG?
1.2.5		Apps	What is the diagnostic accuracy of apps on the detection of snoring in comparison with PSG?
2.	Diagnostic test accuracy meta-analyses – epoch-by-epoch and event-by-event accuracy[b]		
2.1	Apneas and hypopneas combined		
2.1.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the detection of epochs and events of apneas and hypopneas in comparison with PSG?
2.1.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the detection of epochs and events of apneas and hypopneas in comparison with PSG?
2.1.3		Nearable devices	What is the diagnostic accuracy of nearable devices on the detection of epochs and events of apneas and hypopneas in comparison with PSG?
2.1.4		Bed sensors	What is the diagnostic accuracy of bed sensors on the detection of epochs and events of apneas and hypopneas in comparison with PSG?
2.1.5		Apps	What is the diagnostic accuracy of apps on the detection of epochs and events of apneas and hypopneas in comparison with PSG?
2.2	Apneas		
2.2.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the detection of epochs and events of apneas in comparison with PSG?
2.2.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the detection of epochs and events of apneas in comparison with PSG?
2.2.3		Nearable devices	What is the diagnostic accuracy of nearable devices on the detection of epochs and events of apneas in comparison with PSG?

TABLE 1 (Continued)

#	Outcome	Index tests	Research questions
2.2.4		Bed sensors	What is the diagnostic accuracy of bed sensors on the detection of epochs and events of apneas in comparison with PSG?
2.2.5		Apps	What is the diagnostic accuracy of apps on the detection of epochs and events of apneas in comparison with PSG?
2.3	Hypopneas		
2.3.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the detection of epochs and events of hypopneas in comparison with PSG?
2.3.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the detection of epochs and events of hypopneas in comparison with PSG?
2.3.3		Nearable devices	What is the diagnostic accuracy of nearable devices on the detection of epochs and events of hypopneas in comparison with PSG?
2.3.4		Bed sensors	What is the diagnostic accuracy of bed sensors on the detection of epochs and events of hypopneas in comparison with PSG?
2.3.5		Apps	What is the diagnostic accuracy of apps on the detection of epochs and events of hypopneas in comparison with PSG?
2.4	Snoring		
2.4.1		Consumer-based technology	What is the diagnostic accuracy of consumer-based technology (not further specified) in the detection of epochs and events of snoring in comparison with PSG?
2.4.2		Wearable devices	What is the diagnostic accuracy of wearable devices in the detection of epochs and events of snoring in comparison with PSG?
2.4.3		Nearable devices	What is the diagnostic accuracy of nearable devices on the detection of epochs and events of snoring in comparison with PSG?
2.4.4		Bed sensors	What is the diagnostic accuracy of bed sensors on the detection of epochs and events of snoring in comparison with PSG?
2.4.5		Apps	What is the diagnostic accuracy of apps on the detection of epochs and events of snoring in comparison with PSG?
3.	Mean difference		
3.1	Apnea–hypopnea index (AHI)		
3.1.1		Consumer-based technology	What is the estimated mean difference in the AHI values between consumer-based technology (not further specified) in comparison with PSG?
3.1.2		Wearable devices	What is the estimated mean difference in the AHI values between wearable devices in comparison with PSG?
3.1.3		Nearable devices	What is the estimated mean difference in the AHI values between nearable devices in comparison with PSG?
3.1.4		Bed sensors	What is the estimated mean difference in the AHI values between bed sensors in comparison with PSG?

(Continues)

TABLE 1 (Continued)

#	Outcome	Index tests	Research questions
3.1.5		Apps	What is the estimated mean difference in the AHI values between apps in comparison with PSG?
3.2	Respiratory disturbance index (RDI)		
3.2.1		Consumer-based technology	What is the estimated mean difference in the RDI values between consumer-based technology (not further specified) in comparison with PSG?
3.2.2		Wearable devices	What is the estimated mean difference in the RDI values between wearable devices in comparison with PSG?
3.2.3		Nearable devices	What is the estimated mean difference in the RDI values between nearable devices in comparison with PSG?
3.2.4		Bed sensors	What is the estimated mean difference in the RDI values between bed sensors in comparison with PSG?
3.2.5		Apps	What is the estimated mean difference in the RDI values between apps in comparison with PSG?
3.3	Respiratory event index (REI)		
3.3.1		Consumer-based technology	What is the estimated mean difference in the REI values between consumer-based technology (not further specified) in comparison with PSG?
3.3.2		Wearable devices	What is the estimated mean difference in the REI values between wearable devices in comparison with PSG?
3.3.3		Nearable devices	What is the estimated mean difference in the REI values between nearable devices in comparison with PSG?
3.3.4		Bed sensors	What is the estimated mean difference in the REI values between bed sensors in comparison with PSG?
3.3.5		Apps	What is the estimated mean difference in the REI values between apps in comparison with PSG?
3.4	Oxygen desaturation index (ODI)		
3.4.1		Consumer-based technology	What is the estimated mean difference in the ODI values between consumer-based technology (not further specified) in comparison with PSG?
3.4.2		Wearable devices	What is the estimated mean difference in the ODI values between wearable devices in comparison with PSG?
3.4.3		Nearable devices	What is the estimated mean difference in the ODI values between nearable devices in comparison with PSG?
3.4.4		Bed sensors	What is the estimated mean difference in the ODI values between bed sensors in comparison with PSG?
3.4.5		Apps	What is the estimated mean difference in the ODI values between apps in comparison with PSG?
3.5	Snoring duration		
3.5.1		Consumer-based technology	What is the estimated mean difference in the snoring duration between consumer-based technology (not further specified) in comparison with PSG?
3.5.2		Wearable devices	What is the estimated mean difference in the snoring duration between wearable devices in comparison with PSG?

TABLE 1 (Continued)

#	Outcome	Index tests	Research questions
3.5.3		Nearable devices	What is the estimated mean difference in the snoring duration between nearable devices in comparison with PSG?
3.5.4		Bed sensors	What is the estimated mean difference in the snoring duration between bed sensors in comparison with PSG?
3.5.5		Apps	What is the estimated mean difference in the snoring duration between apps in comparison with PSG?
3.6	Snoring frequency		
3.6.1		Consumer-based technology	What is the estimated mean difference in the snoring frequency between consumer-based technology (not further specified) in comparison with PSG?
3.6.2		Wearable devices	What is the estimated mean difference in the snoring frequency between wearable devices in comparison with PSG?
3.6.3		Nearable devices	What is the estimated mean difference in the snoring frequency between nearable devices in comparison with PSG?
3.6.4		Bed sensors	What is the estimated mean difference in the snoring frequency between bed sensors in comparison with PSG?
3.6.5		Apps	What is the estimated mean difference in the snoring frequency between apps in comparison with PSG?
3.x	Other outcomes		Other research questions might arise if other continuous outcomes related to OSA or snoring are identified (e.g., total absolute number and indices of respiratory effort-related arousals (RERAs), apneas, hypopneas, central events, obstructive events, or time spent with oxygen saturation <90%)

[a]All research questions in the 1.1 level will consider six independent diagnostic thresholds (AHI or RDI ≥5, ≥15, and ≥30 events/h).
[b]Although merged into the research questions, epoch-by-epoch and event-by-event analysis will be performed independently whenever possible.

sensitivity, as many studies employing PSG might not use this term in their titles, abstracts, or keywords. We used a search related to SDB in general rather than specifically to OSA to increase search sensitivity. The search strategy for PubMed is disclosed below.

- Search domain #1 (OSA):
  - "Sleep Apnea Syndromes"[mh] OR Snoring[mh] OR "sleep-disordered breathing" OR "sleep-related breathing disorders" OR (sleep AND (apnea* OR hypopnea* OR apnoea* OR hypopnoea*)) OR (OSA AND sleep) OR (OSAS AND sleep) OR (OSAHS AND sleep) OR "Cheyne-Stokes" OR "Upper airway resistance syndrome" OR snoring* OR snore* OR "apnea-hypopnea index" OR "apnoea-hypopnoea index" OR (IAH and (sleep OR apnea OR apnoea)) OR "respiratory disturbance index" OR (RDI AND (sleep OR apnea or apnoea)) OR "respiratory effort related arousal" OR (RERA AND (sleep OR apnea or apnoea))
- Search domains #2 (consumer-based technologies):
  - "wearable electronic devices"[mh] OR "mobile applications"[mh] OR "software"[mh] OR "smartphone"[mh] OR "computers, handheld"[mh] OR "remote sensing technology" OR "wireless technology"[mh] OR (consumer AND sleep[tiab]) OR portable* OR wearable* OR nearable* OR mobile OR smartphone* OR "smart phone*" OR smartwatch* OR tablet* OR app OR apps OR application* OR "bed sensor*" OR "consumer-based" OR "consumer grade"

Secondary data search includes: (i) checking reference lists of the included articles, (ii) contacting CST companies, and (iii) searching grey literature. Regarding reference lists analysis, the list of references of all included articles will be screened for additional studies not retrieved in the primary search. Regarding contacting CST companies, this is a strategy intended to retrieve undergoing, unpublished, or unretrieved studies that support or were used in the registration of devices and applications already commercially available. We will shortlist technology companies and start-ups related to OSA and/or snoring from two sources: the list of sponsors and exhibitors at the last three editions of the European Sleep Research Society (ESRS), World Sleep Society (WSS), and AASM congresses, and the devices listed in the AASM #SleepTechnology resource. The list of CST companies to be contacted will be properly disclosed and the results provided will be further explored by means of sensitivity analyses (more

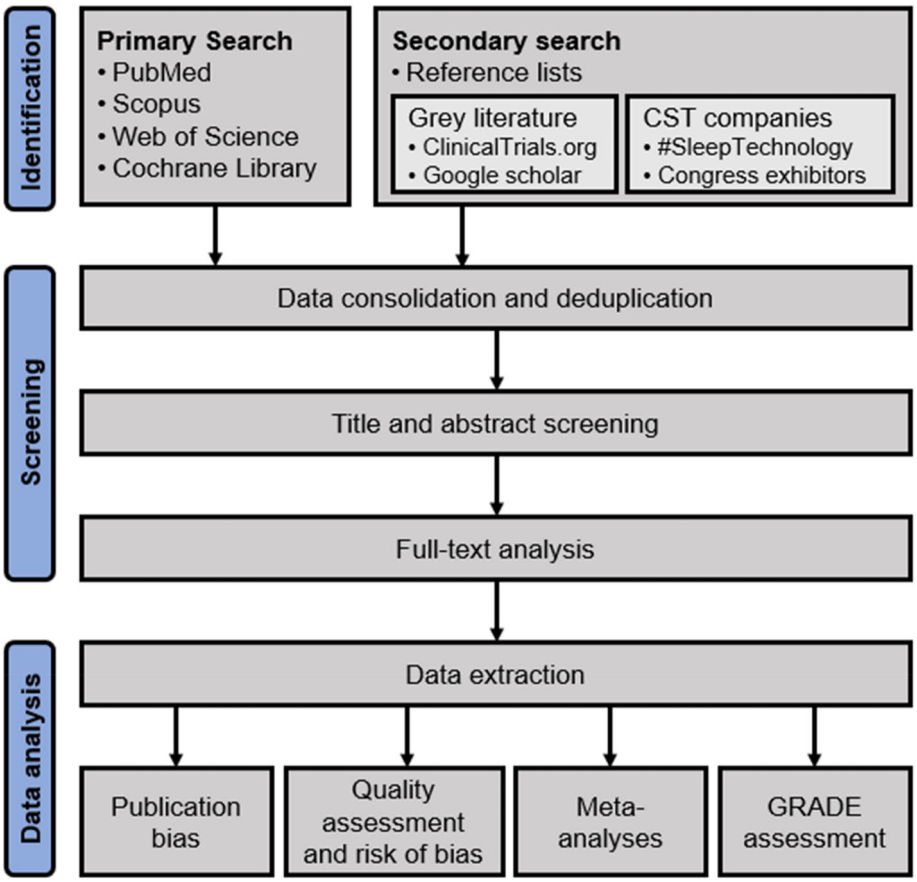

FIGURE 1 Flow diagram disclosing the sources of data, screening procedures, and data analysis. Both title and abstract screening and full-text analysis will be performed by two independent reviewers. The meta-analyses include both diagnostic test accuracy and mean-difference analyses (including the respective subgroup and sensitivity analyses). CST, consumer sleep technology. GRADE, Grading of Recommendations Assessment, Development and Evaluation.

information on the 'Analysis plan, subgroup analysis, and sensitivity analysis' section). Regarding searching grey literature, which encompasses literature available out of the main databases and often in non-final format, we will search ClinicalTrials.gov and Google Scholar (first 200 records).

All sources of data and the subsequent procedures, including data screening, extraction, and analysis disclosed in Figure 1.

3.4 | Evaluation and selection strategies

The records retrieved from the search in the four primary databases (PubMed, Scopus, Web of Science, and the Cochrane Library) will be exported to Covidence, where deduplication and eligibility analysis will be performed. Duplicated records will be automatically excluded. Two independent reviewers will evaluate all non-duplicated records, in a process based on two steps. The first step consists of reviewing titles and abstracts only, while the second encompasses full-text analysis. In each step, the decision considering each article as eligible or not relies on consensus between both reviewers' decisions. Disagreements between reviewers will be solved by a consensus, and if the discordance persists, a third reviewer will be consulted (GNP). Both reviewing steps will be conducted considering the inclusion and exclusion criteria disclosed below.

By the beginning of the evaluation process, a calibration round of the eligibility analysis will be performed, and the agreement rates between each reviewer in comparison with a senior reviewer (GNP) will be measured. Both the senior reviewer and each of the independent reviewers will analyse the titles and abstracts of 200 records. The Cohen's kappa index between each reviewer and the senior reviewer will be calculated, and the analysis will continue only if a kappa value of ≥0.8 is reached (considered as a strong agreement). If this threshold is not reached, meetings for discussing the inclusion and exclusion criteria and a new calibration round will be performed until a 0.8 index is achieved.

3.5 | Inclusion and exclusion criteria

Only studies evaluating the accuracy of consumer-based technologies related to OSA and snoring screening will be considered eligible, based on smartphone apps, wearables, nearables, or bed sensors. There will be no restrictions regarding publication date and language. The authors are able to handle records published in English, Portuguese, Spanish, Finnish, and Icelandic. If studies in other languages are retrieved, Google Translate will be used for abstract screening and native speakers will be contacted for assistance with full-text analysis.

The eligibility analysis will be based on the criteria below:

- *Abstracts*
  - Inclusion: articles that present an abstract, regardless of the language.
  - Exclusion: articles that do not present an abstract.
- *Source type*
  - Inclusion: full original/primary studies, reports, and datasets. Studies published in non-peer-reviewed sources will also be considered eligible if the data collection and analysis are complete and precisely described. Uncompleted data sources (e.g., congress abstracts, patents, and protocols) will be considered for the systematic review, but not for meta-analyses. In these cases, the authors will be contacted and inquired about the availability of the full study report or a complete dataset. In the case of redundant publications (i.e., secondary or derivative studies coming from the same dataset or source study), only the one with the biggest sample size will be considered eligible.
  - Exclusion: theoretical articles (including editorials, narrative reviews, and letters to the editors), systematic reviews, meta-analyses, and meta-epidemiological studies.
- *Participants*
  - Inclusion: individuals aged ≥18 years, with no restriction regarding sex and presence of concurrent sleep disorders or comorbidities. Studies encompassing paediatric samples will be considered eligible if it is possible to extract data from a subgroup of those aged ≥18 years in an independent and unbiased way.
  - Exclusion: studies with a sample of participants aged <18 years, or in cases in which it is not possible to dissociate a subgroup of those aged ≥18 years.
- *Index test*
  - Inclusion: CSTs intended for screening of OSA or snoring (Schutte-Rodin et al., 2021). In order to be categorised as a consumer-grade technology, technology should be accessible without a prescription, should be compatible with domestic use (i.e., does not require a clinical facility to function), should be used unattended (i.e., be able to function regardless of professional monitoring or scoring), and should be able to deliver a result directly to the user without the need for professional manual analysis and scoring. This might include, but is not limited to, wearable devices, nearable devices, bed sensors, and apps.
  - Exclusion: devices and apps that fail to be considered as a CST according to the definitions provided above, including HSATs.
- *Reference test*
  - Inclusion: full night type-I or type-II PSG, performed according to the AASM requirements (Berry et al., 2020).
  - Exclusion: studies with no PSG or performed with any type of sleep study other than full-night type-I of type-II (including sleep studies type-III, type-IV). Such a requirement is needed to assure that a CST will be compared to gold-standard diagnostic technologies only.
- *Study design*
  - Inclusion: cross-sectional, within-subject paired, and non-interventional studies in which both the reference test (type-I or type-II PSG) and the index test (CST) have been used in a group of participants in the same period and under the same setting conditions. Longitudinal studies (including cohorts, case–control studies, and clinical trials) can be considered eligible if there is a baseline measure complying with the previous requirements.
  - Exclusion: studies in which the reference and index tests have not been used in the same participants, period, or settings. Intervention studies in which the participants are subject to any type of intervention (including for OSA treatment, e.g., continuous positive airway pressure or intraoral devices).
- *Target condition and outcome measures*
  - Inclusion: OSA or snoring, measured and diagnosed in a full night type-I or type-II PSG and reporting at least one of the main outcomes (OSA or snoring diagnosis, AHI, RDI, REI, ODI, snoring duration, or snoring frequency). Any diagnostic threshold is considered eligible and the association with symptoms or consequences is not mandatory for the diagnosis of mild OSA.
  - Exclusion: no information regarding any of the main outcomes. Samples in which all individuals are diagnosed with OSA, or without OSA, cases in which a proper evaluation of specificity is not possible. No definition of a diagnostic threshold and no possibility to infer so based on the available data.

Exclusion criteria should be prioritised according to the order below (based on the criteria above): (i) no abstract, (ii) non-original article, (iii) wrong population, (iv) wrong index test (not a consumer-based OSA technology), (v) wrong reference test (not a full-night type-I or type-II PSG), (vi) no primary outcomes were reported, (vii) no full text was retrieved, and (viii) redundant study.

3.6 | Data extraction

Data extraction will be performed using Covidence by two independent reviewers and checked by a senior reviewer (GNP). Before actual data extraction, both reviewers will undergo a training session with the senior reviewer. Disagreements in the data extraction between both reviewers will be solved by consensus, and if discordance persists, the senior reviewer will be consulted (GNP). If the senior reviewer is not able to solve the discordance, the article's authors might be contacted.

Numeric outcome data will be extracted as mean ± standard deviation (SD). When a study discloses the standard error of the mean (SEM) rather than SD, SD will be calculated by multiplying the SEM by the square root of the sample size. When it is not possible to determine if data dispersion is displayed in SEM or SD, we will assume them as SD.

The only mandatory items for the analyses are metadata, sample size, type of device/app, and at least one main outcome. When needed, data will be extracted from graphs using a digital ruler (Plot Digitizer, plotdigitizer.sourceforge.net/). In case of missing data not extractable from charts or doubts regarding any specific result or methodological aspect, authors will be contacted and will be asked to

provide information about their protocol, results, or raw data. Two contact attempts will be made and if no successful response is obtained, the article might be excluded from the sample (when they fail to provide one of the mandatory items) or from a specific subgroup analysis.

In the case of longitudinal studies, data from all available nights will be extracted, provided that both the reference and the index tests were used concomitantly. The unit of analysis in this systematic review is not the articles, but studies within the articles. Therefore, when an article has two or more separate and fully independent studies, data may be extracted more than once from the same article. In studies evaluating two or more index tests, each of them will be considered as a separate study, even considering that they are compared with the same reference test group (data corrections might be needed for continuous outcomes).

The following variables will be extracted:

Descriptive information

- *Metadata*: full reference string, including first author, title, publication year, and publication source (journal).
- *Country*: defined as the country in which the sample was recruited. In international multicentric studies, the proportion of participants from each country will be extracted. In these cases, and for descriptive purposes, the study will be considered as belonging to the country that contributed the most to the sample.

Participants

- *Sample size*: considering only the final sample, composed of those participants subjected to both tests.
- *Sex*: it could be filled as 'men', 'women', 'both' or 'not disclosed'. In the case of 'both', the proportion of each sex in the final sample will be extracted.
- *Age*: both the recruitment age range and the mean age (±SD) will be extracted.
- *Body mass index*.
- *Self-reported ethnicity and Fitzpatrick skin colour scale*. The proportion of each ethnicity and phototype will be extracted, whenever available.
- *Exclusion criteria*: every criterion is taken for considering a potential participant as ineligible.
- *Concurrent health conditions*: any health condition or descriptive characteristic considered as part of the population description in a study (i.e., only individuals presenting a specific disease were included). For example, studies evaluating the accuracy of OSA apps among individuals with hypertension or with morbid obesity.
- *Pre-test assessment OSA risk*: average score and the number of individuals considered as having a high risk of OSA according to screening questionnaires for OSA risk (Małolepsza et al., 2021), namely the Berlin (Chung et al., 2008b), the STOP-BANG (Chung et al., 2008a) or the NoSAS (Marti-Soler et al., 2016) questionnaires.

Study design and description

- *Sample representativeness*: population-based study, probabilistic sampling, non-clinical convenience sample, or clinical convenience sample.

Index test:

- *Commercial name* (when available).
- *Manufacturer* (when available).
- *Version of the device/app* (when available).
- *Type of device/app*: it could be filled as 'wearable', 'nearable', 'bed sensors', 'app', or 'other'. Wearable devices are defined as any device that is worn or used in close and conditional contact with the participant's body. For technologies considered wearables, additional information about their mode of use and presentation will be extracted (e.g., ring, wristband, smartwatch, fitness tracker, headband, or chest belt). Nearables are considered as any device positioned close to the participant, but with no contact with its body, not being worn nor composing the bed and bed linen. For devices considered as nearables, information regarding their position will be extracted (e.g., on the nightstand, below the bed, or the headboard). Bed sensor devices include all devices integrated into the linen and other fabric materials that are not worn by the participants, but that compose the sleeping environment. For devices considered bed sensors, they will be categorised according to their use (e.g., mattresses, pillows and pillowcases, linen, or blanket). Apps are defined as software integrated and conditionally used in a smartphone or tablet. Although it might be seen as a nearable (as the smartphone or tablet should be positioned near the individual), its main feature is the software, while the nearables have the hardware as their main feature. Any other device will be categorised as 'other', and a further explanation will be provided. New categories not foreseen in this protocol might be considered depending on the characteristics of the retrieved studies.
- *Variables detected to measure OSA or snoring*: any physiological or environmental variable used to detect OSA and snoring (and to measure sleep time or sleep stage when those data are integrated into the detecting algorithm). For example, oxygen saturation, heart rate variability, respiratory rate, body temperature, movements, sound, and chest movements.
- *Equipment used to measure the intended variables*: the embedded technology or sensors that are primarily responsible for data acquisition. For example, an oximeter, microphone, arterial tonometry, and accelerometer.
- *Diagnostic threshold*: any diagnostic threshold using any variable to diagnose OSA or categorise its severity. More than one diagnostic threshold can be extracted per study.

Reference test:

- *Diagnostic criteria for apneas*: the exact definition of the event (e.g., ≥90% reduction in oronasal thermal sensor amplitude, lasting at least 10 s) and/or the scoring guideline considered (e.g., the AASM manual 2020).
- *Diagnostic criteria for hypopneas*: the exact definition of the event (e.g., ≥30% reduction in nasal pressure amplitude, lasting at least 10 s and associated with either a ≥3% desaturation or an arousal) and/or the scoring guideline considered (e.g., the AASM manual version 2.6 – recommended criteria).
- *Diagnostic threshold*: any diagnostic threshold using any variable used to diagnose OSA and snoring, or categorise its severity. More

FIGURE 2 Contingency table with hypothetical values for different obstructive sleep apnea (OSA) diagnostic thresholds between polysomnography (PSG) and an index test. (a) Contingency table with results between PSG and an index test considering all common classification thresholds. (b–d) Derived contingency tables, converting values into different classification thresholds. This figure allows comparing the actual sensitivity and specificity according to different classification thresholds. In this hypothetical example, the sensitivity is high and stable in all classification thresholds defined based on the apnea–hypopnea index (AHI) (i.e., there is a low risk of false negatives), while the specificity grows from mild to severe OSA. mod, moderate.

(a) Contingency table including all OSA severity levels

	Reference test			
	Severe OSA	Moderate OSA	Mild OSA	No OSA
Severe OSA	48	2	1	36
Moderate OSA	1	45	8	10
Mild OSA	1	2	35	4
No OSA	0	1	6	0

(b) Contingency table to diagnose OSA grouping all severity levels (AHI≥5)

	Reference test		
	OSA	No OSA	
OSA	143	14	Sensitivity: 0.95
No OSA	7	36	Specificity: 0.72

(c) Contingency table to diagnose moderate and severe OSA (AHI≥15)

	Reference test		
	Mod-severe OSA	No-mild OSA	
Mod-severe OSA	96	13	Sensitivity: 0.96
No-mild OSA	4	87	Specificity: 0.87

(d) Contingency table to diagnose severe OSA (AHI≥30)

	Reference test		
	Severe OSA	No-mod OSA	
Severe OSA	48	3	Sensitivity: 0.96
No-mod OSA	2	147	Specificity: 0.98

than one diagnostic threshold can be extracted per study. For the PSG-based diagnosis, the expected thresholds include AHI or RDI of ≥5, ≥15, or ≥30 events/h, but the least used thresholds (such as AHI ≥20 events/h) or diagnosis based on variables other than AHI or RDI will also be extracted.

- *PSG scoring approach*: although the AASM recommendations require manual scoring, the use of automatic approaches is becoming increasingly common. The PSG scoring approach will be extracted, being categorised as manual or automatic.
- *Number of technologists scoring the PSG recording*: the accuracy of a diagnostic test depends not only upon the sensitivity and specificity of the index test but also on the precision of the reference test. PSGs are always subjected to inter-rater variability (Kuna et al., 2013; Lee, Lee, Cho, & Choi, 2022; Magalang et al., 2016), and the bigger it is, the harder it will be for an index test to reach high accuracy. Having more than one technician scoring each PSG is a strategy to increase diagnostic accuracy within the reference test.

Main outcomes:

- Number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each diagnostic threshold. These variables are traditionally used to compose contingency tables from which specificity and sensitivity and other derivate measures are calculated. When these variables are not provided, they will be calculated or estimated, as suggested by Taylor et al. (Taylor, Mahtani, & Aronson, 2021) and Dinnes et al. (Dinnes, Deeks, Leeflang, & Li, 2021). When these data are provided in contingency tables larger than 2 × 2 (such as disclosing all OSA severity groups), data will be grouped in a way that 2 × 2 tables can be built for each diagnostic threshold (as demonstrated in Figure 2).
- Number of TPs, FPs, TNs, and FNs for epoch-by-epoch and event-by-event detection of obstructive events, apneas, hypopneas, and snoring.
- AHI for both index and reference tests (mean ± SD).
- RDI for both index and reference tests (mean ± SD).
- REI for both index and reference tests (mean ± SD).
- ODI for both index and reference tests (mean ± SD).
- Snoring duration for both index and reference tests (mean duration [s] ± SD).
- Snoring frequency for both index and reference tests (mean [s] ± SD).

Secondary outcomes

- Any other sleep-related respiratory variable reported in the article, including but not limited to absolute number and indices of respiratory effort-related arousals (RERA), apneas, hypopneas, central events, obstructive events; time spent with SpO_2 >90% (mean ± SD).

Role of sponsors:

- Study commissioned or sponsored by a device/app manufacturer (yes/no)
- One or more authors directly affiliated with the device/app manufacturer (yes/no)

3.7 | Publication bias

Publication bias will be assessed using Deeks' test, which was specifically designed for systematic reviews of DTA (Deeks, Macaskill, & Irwig, 2005) and performs better than the Begg and Egger test in these cases (van Enst, Ochodo, Scholten, Hooft, & Leeflang, 2014). It is based on plotting the DOR in natural logarithmic form (lnDOR) against the inverse of the effective sample size.

3.8 | Quality assessment and risk of bias

Quality assessment within the included studies will be evaluated using the revised version of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) (Whiting et al., 2011). This tool was specifically designed for quality assessment of diagnostic accuracy studies and is the most recommended option for risk assessment in systematic reviews. It consists of four domains, related to participant selection, index test, reference standard, and flow and timing. Each of these domains is evaluated in two independent ways: risk of bias and concerns regarding applicability (except for 'flow and timing', which is judged only about the risk of bias). Each item is judged as having low risk, high risk, or unclear risk. As common for quality and bias assessment, no summary statistics or final scores are provided. The results will be displayed in tables, disclosing the results of the assessment for each included article, and in charts, disclosing the percentage of low, high, and unclear risk for each item.

Additionally, the procedures used for the validation of point-of-care OSA screening devices in each included article will be evaluated by using the rating system proposed by Tangudu et al. (2021). This method evaluates 11 aspects related to validation studies: the number of PSG readers/scorers, subject conditions (SDB diagnosis), subject data (patient history), caffeine and alcohol restrictions, daytime sleepiness evaluation, instructions for self-application of home-based devices, sleep metrics under analysis, methods for data extraction, methods for quantitative analysis, methods for qualitative analysis, and data protection and security issues. Each item is rated from 1 to 3 based on the appropriateness of the methods employed in each study, leading to a final score ranging from 1 to 33.

3.9 | Data synthesis and analyses

Meta-analyses will be performed whenever three or more studies can be grouped using the same outcome measure. Two different types of meta-analysis will be performed in this study: DTA meta-analyses and mean difference meta-analyses.

3.9.1 | Diagnostic test accuracy meta-analyses

The DTA meta-analyses will be performed for outcomes for which 2×2 contingency tables were extracted. It includes both studies assessing diagnostic accuracy (PI(R)T questions #1) and studies employing epoch-by-epoch or event-by-event analyses (PI(R)T questions #2). Independent analyses will be performed for each diagnostic threshold, and no analyses will be performed by adding data from different diagnostic thresholds. For each study, the number of TP, FN, FP, and TN will be used, and summary statistics will be calculated using the random effects bivariate binomial model of Chu and Cole (Chu & Cole, 2006). This model allows the calculation of a summary estimate for both sensitivity and specificity among the whole sample. The estimated sensitivity and specificity (and their 95% confidence interval [95% CI]) for each included study will be displayed in Forest plots, and the summary estimate for the whole sample will be displayed in a summary receiver operating characteristics (ROC) curve (SROC curve), plotted using the sensitivity against the FP rate (1-specificity). No heterogeneity tests will be performed (such as the I^2 index), as they do not perform well in DTA meta-analyses (Leeflang, 2014). All DTA meta-analyses will be performed using the MetaDTA application (Freeman et al., 2019; Patel, Cooper, Freeman, & Sutton, 2021).

3.9.2 | Mean difference meta-analysis

Mean difference meta-analysis will be performed for continuous outcomes (mean ± SD), in cases in which both the index and the reference tests provide results in the outcome and use the same unit of measurement (PI(R)T questions #3). These meta-analyses are not commonly performed in systematic reviews of diagnostic accuracy, but in the present study, they will be used to explore the concordance of the index and the reference tests in detecting a continuous numeric variable used for screening purposes, regardless of the diagnostic threshold. For each study, the mean difference ($M_{\text{index test}} - M_{\text{reference test}}$) will be calculated for each included study. Meta-analyses will be performed using the DerSimonian and Laird random-effects model. Heterogeneity will be assessed using both the I^2 index and the Cochran's Q test. Data will be presented as effect size ± 95% CI in Forest plots. Statistically significant results ($p \leq 0.05$) with effect size ± 95% CI greater than zero will be interpreted as cases in which the index test overestimates the reference test measure, while effect size ± 95% CI less than zero and $p \leq 0.05$ will be interpreted as an underestimation of the index test in comparison with the reference test. Non-significant results ($p > 0.05$) with effect size ± 95% CI crossing the zero line will be interpreted as an equivalence between the index and the reference tests for a given outcome. All mean difference meta-analysis will be performed using the Comprehensive Meta-Analysis software.

3.9.3 | Analysis plan, subgroup analysis, and sensitivity analysis

The primary level analysis will include all possible studies for each given outcome, regardless of the device/app category and technology used. Although it is likely to result in highly heterogeneous analyses, it

will be useful to conclude the general accuracy of OSA and snoring screening devices and apps.

Second-to-fourth-level analyses correspond to subgroup analysis with an increasing level of methodological homogeneity. Second-level analyses will group studies by the type of devices/app (wearables, nearables, bed sensors, smartphone apps). Third-level analyses will group studies by the technology and equipment used to detect OSA or snoring (e.g., oximeter, microphone, arterial tonometry, and accelerometer). Fourth-level analyses will group studies by the exact commercial presentation (including commercial name and manufacturer).

To evaluate whether the accuracies of CSTs have increased over time, the stratified analysis will be performed according to the publication year, grouping studies in blocks of 5 years. To evaluate the accuracies of CSTs between individuals with different skin colours, subgroup analysis will be performed considering self-reported ethnicity and phototypes (based on the Fitzpatrick skin phototype scale whenever available). To evaluate the possible differential accuracy results of CSTs validated against manual scoring, stratified analyses will consider the PSG-scoring approach (manual, semi-automatic, or automatic). Other subgroup analyses might be performed depending on the number and characteristics of the included studies.

Sensitivity analyses will be performed considering the representativeness of the samples (excluding all convenience samples), the role of the manufacturer (excluding cases in which the manufacturer sponsored the study or when authors are directly affiliated with the manufacturer), source of data (excluding articles provided by manufacturers), concurrent health conditions (excluding samples with populations with comorbidities), pre-test assessment of OSA (excluding studies in which the sample was considered as high risk of OSA at the baseline). Other sensitivity analyses might be performed depending on the number and characteristics of the included studies.

3.10 | Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessment

The GRADE system is a methodology increasingly used to assess the certainty of evidence and to decide about the strength of recommendations in systematic reviews and guideline development (Guyatt et al., 2008), especially when related to therapeutic questions. Initial methodological suggestions have been made to adapt the GRADE system to questions related to DTA (Brozek et al., 2009; Schünemann et al., 2008). However, the use of the GRADE system has been proven to be challenging, mostly due to the lack of proper and explicit guidance on how to perform it (Gopalakrishna et al., 2014). More recent guidelines are being implemented, which will be used in this systematic review (Schünemann et al., 2019; Schünemann et al., 2020a, 2020b).

The GRADE assessment in this review will apply only to the categorical outcomes assessed in terms of their accuracy (including TP, FN, FP, and TN), as the continuous outcomes are analysed from a more exploratory perspective. For each question, sensitivity (TP and FN grouped) and specificity (TN and FP grouped) outcomes will be assessed independently. The certainty of the evidence for each outcome can be considered as high, moderate, low, or very low. Cross-sectional within-subject paired studies will start being considered as high-certainty evidence, as this design can be considered appropriate to assess test accuracy (Schünemann et al., 2020a). Based on this initial assessment, the level-of-evidence certainty can be decreased based on five criteria (risk of bias, indirectness, inconsistency, imprecision, and publication bias) (Schünemann et al., 2020b). Certainty of evidence can also increase based on three criteria (consistent sensitivity–specificity relationship, large estimates of test accuracy, and minimal plausible bias and confounding). However, rating up the certainty of the evidence is discouraged for test accuracy outcomes, as there is no consensus regarding this procedure for DTA systematic reviews and it still warrants further methodological development. All GRADE assessments will be performed with GRADEpro GDT (https://www.gradepro.org/).

4 | DISCUSSION AND EXPECTED RESULTS

Several CSTs related to OSA and snoring screening are commercially available and are becoming increasingly popular. As these technologies are designed for being used directly by the customers, usually without supervision or assistance from medical professionals, it is important to assure their results and reports are reliable and accurate. One of the main concerns regards sensitivity, as FN results would refrain a user from seeking professional assistance and treatment when it is needed.

The present protocol describes the methods and procedures for a systematic review and meta-analysis, which will evaluate the accuracy of CSTs for the screening of OSA and snoring. Other systematic reviews have already been performed to evaluate CSTs for other sleep-related conditions, such as for digital cognitive behavioural therapy for insomnia (Cheng & Dizon, 2012; Edinger et al., 2021; Seyffert et al., 2016; Ye et al., 2016; Zachariae, Lyby, Ritterband, & O'Toole, 2016) and sleep scoring (Haghayegh, Khoshnevis, Smolensky, Diller, & Castriotta, 2019). In all cases, the meta-analyses have been useful to assure a proper assessment of the validity, usefulness, and reliability of CSTs, although in conditions different from OSA.

The use of CST for OSA screening has already been reviewed and subjected to comprehensive theoretical analyses (Baptista et al., 2022; Baron et al., 2018; Fino & Mazzetti, 2019; Korkalainen et al., 2021; Uddin et al., 2018). At least two meta-analyses regarding CST for the screening of OSA have already been published (Kim et al., 2022; Rosa et al., 2018). Rosa et al. (2018) evaluated a sample of 18 studies including all CST types. The authors concluded that both contactless devices and bed-mattresses devices have the greatest sensitivity to detect OSA, especially in moderate and severe cases, while other devices showed low sensitivity and specificity. Kim et al. (2022) evaluated the accuracy of smartphones in the detection of OSA based on a sample of 11 studies and reached a sensitivity of >80% in all cases. Both studies were well performed and were

meritorious in evaluating the usefulness of these technologies. Although we acknowledge an overlap between these two previous meta-analyses and the present protocol, we understand they can be complementary. In addition, the present systematic review improves the knowledge by addressing points that were not addressed in the previous studies:

- **Updated analysis:** CSTs are improved at a fast pace and new technologies are constantly being developed, which includes new sensors, algorithms, and devices. Therefore, constant monitoring and evaluation of the reliability of these tools are needed, until safe and stable conclusions regarding their usefulness are reached. Rosa et al. (2018) was published 5 years ago and substantial development in CSTs has been made seen since then. Kim et al. (2022) is more recent, but their narrower focus led to not analysing the relevant improvements in technologies other than those linked to smartphones.
- **Broadness on the definition of CST:** sleep technologies are presented in many formats, and the reliability of one type of device might not be extrapolated to others. Rosa et al. (2018) have used a similar approach to ours, intending to include CSTs of multiple types, but it might have missed devices that were developed in the last couple of years. On the other hand, the recent study by Kim et al. (2022) focused on smartphone detection only, therefore not allowing us to conclude anything regarding wearables and nearables.
- **Detailed strategy to analyse data:** as the technology on these tools and devices varies considerably, a solid strategy to analyse data is needed to encompass the most important sources of variability and heterogeneity. In our analysis plan, we have included analyses related to different outcome measures (including but not limited to AHI, RDI, REI, ODI, snoring duration, and snoring frequency), categories of devices (wearables, nearables, bed sensors, and smartphone apps), technologies (e.g., oximeter, microphone, arterial tonometry, and accelerometer), sample representativeness, and the role of manufacturer, among others.
- **Concerns regarding sponsorship:** most research on new sleep technologies are directly sponsored or even primarily performed by the manufacturers. This is an important source of potential bias, increasing the likelihood of publication bias and selective outcome reporting. For this reason, we intend to broaden our search strategy by contacting manufacturers directly. Also, the role of the manufacturers will be included in subgroup analyses.
- **Specific methodology for DTA meta-analysis:** this is a very specific type of meta-analysis, for which the methodology is under constant improvement. The previous meta-analyses have encompassed some of the DTA-specific methodologies, but a few aspects might have been overlooked. The present protocol aims to encompass the most recent methodology for DTA meta-analysis.
- **Sample size:** the previous meta-analyses have been performed with a limited number of studies, especially when subgroup analyses are considered. We believe that with our enlarged search strategy and by contacting manufacturers and companies directly, we might have a larger sample of studies, therefore increasing our external validity on conclusive potential.

As currently understood, CSTs are not intended to be used for diagnosis, but rather for screening of sleep disorders, as most of them are marketed to be used autonomously by a consumer/user without medical prescription or supervision. However, two movements in the development of new CSTs have been observed. First, their overall diagnostic accuracy seems to be increasing, as new technologies are used, and more refined algorithms are implemented. Second, transitional technologies, which lie somehow in between CSTs and clinical-grade devices, are becoming increasingly common. Our results will help to assess whether the diagnostic accuracy of OSA-related CSTs is adequate, therefore bringing screening and diagnosis closer one to another. However, it should be kept in mind that diagnosis is not restricted to the measurement of certain diagnostic measures (such as AHI, RDI, or ODI), and it might involve a proper differential diagnosis or the assessment of comorbidities and other concomitant conditions. Both issues are highly dependent on a throughout clinical evaluation, therefore being overlooked when a device is used directly by the consumer regardless of a medical professional.

The meta-analyses resulting from this protocol will help to direct future technologies, assisting in the process of continuous technological development in the field of sleep medicine. This growth should be achieved both by focusing on consumer needs and data reliability. However, meta-analyses such as the one proposed here are limited by the fact that they analyse previously published data, therefore serving as a post hoc appraisal tool. It also focuses specifically on the accuracy of diagnostic accuracy of studies, not intending to evaluate other aspects related to the reliability of CSTs, such as personal data security, data protection, and data storage. The knowledge arriving from the present and other meta-analyses, as well as from the mutual collaboration between manufacturers, healthcare professionals, and sleep researchers, should be reverted into practical achievements and definitions that should be implemented before new devices become commercially available, impacting the way they are designed, developed, registered, and evaluated by health agencies, and promoted to the general public.

AUTHOR CONTRIBUTION

Gabriel Natan Pires: conceptualisation, data curation, investigation, methodology, project administration, writing – original draft, writing – review and editing. Erna Sif Arnardóttir: conceptualisation, funding acquisition, methodology, validation, writing – review and editing. Anna Sigridur Islind: conceptualisation, methodology, validation, writing – review and editing. Timo Leppänen: conceptualisation, methodology, validation, writing – review and editing. Walter T. McNicholas: conceptualisation, funding acquisition, methodology, supervision, validation, writing – review and editing.

ACKNOWLEDGEMENTS

This work has received research funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417. Timo Leppänen reports additional funding from NordForsk (NordSleep project 90458) via Business Finland (5133/31/2018), the Academy of Finland (323536), and the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (5041794). Erna Sif Arnardóttir and Anna Sigridur Islind report additional funding from NordForsk (NordSleep project 90458) via the Icelandic Research Fund.

CONFLICT OF INTEREST

Gabriel Natan Pires is a shareholder at SleepUp© and founder of P&P Metanálises. Erna Sif Arnardóttir discloses lecture fees from Nox Medical, Philips, ResMed, Jazz Pharmaceuticals, Linde Healthcare, Alcoa – Fjardaral, Wink Sleep and Novo Nordisk (via Vistor). Erna Sif Arnardóttir is also a member of the Philips Sleep Medicine and Innovation Medical Advisory Board. The other authors have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Gabriel Natan Pires https://orcid.org/0000-0003-0411-0111
Erna Sif Arnardóttir https://orcid.org/0000-0003-0877-3529
Timo Leppänen https://orcid.org/0000-0003-4017-821X
Walter T. McNicholas https://orcid.org/0000-0001-5927-2738

REFERENCES

Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The future of sleep measurements: A review and perspective. *Sleep Medicine Clinics*, *16*(3), 447–464. https://doi.org/10.1016/j.jsmc.2021.05.004

Baptista, P. M., Martin, F., Ross, H., O'Connor Reina, C., Plaza, G., & Casale, M. (2022). A systematic review of smartphone applications and devices for obstructive sleep apnea. *Brazilian Journal of Otorhinolaryngology*, *88*, S188–S197. https://doi.org/10.1016/j.bjorl.2022.01.004

Baron, K. G., Duffecy, J., Berendsen, M. A., Cheung Mason, I., Lattie, E. G., & Manalo, N. C. (2018). Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. *Sleep Medicine Reviews*, *40*, 151–159. https://doi.org/10.1016/j.smrv.2017.12.002

Behar, J., Roebuck, A., Domingos, J. S., Gederi, E., & Clifford, G. D. (2013). A review of current sleep screening applications for smartphones. *Physiological Measurement*, *34*(7), R29–R46. https://doi.org/10.1088/0967-3334/34/7/R29

Benjafield, A. V., Ayas, N. T., Eastwood, P. R., Heinzer, R., Ip, M. S. M., Morrell, M. J., … Malhotra, A. (2019). Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis. *The Lancet Respiratory Medicine*, *7*(8), 687–698. https://doi.org/10.1016/S2213-2600(19)30198-5

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., & Vaughn, B. V. (2020). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications, version 2.6*. American Academy of Sleep Medicine.

Bittencourt, L. R., Suchecki, D., Tufik, S., Peres, C., Togeiro, S. M., Bagnato, M. C., & Nery, L. E. (2001). The variability of the apnoea-hypopnoea index. *Journal of Sleep Research*, *10*(3), 245–251. https://doi.org/10.1046/j.1365-2869.2001.00255.x

Borsky, M., Serwatko, M., Arnardottir, E. S., & Mallett, J. (2022). Toward sleep study automation: Detection evaluation of respiratory-related events. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 3418–3426. https://doi.org/10.1109/JBHI.2022.3159727

Brozek, J. L., Akl, E. A., Jaeschke, R., Lang, D. M., Bossuyt, P., Glasziou, P., … GRADE Working Group. (2009). Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy*, *64*(8), 1109–1116. https://doi.org/10.1111/j.1398-9995.2009.02083.x

Camacho, M., Robertson, M., Abdullatif, J., Certal, V., Kram, Y. A., Ruoff, C. M., … Capasso, R. (2015). Smartphone apps for snoring. *The Journal of Laryngology and Otology*, *129*(10), 974–979. https://doi.org/10.1017/S0022215115001978

Caples, S. M., Anderson, W. M., Calero, K., Howell, M., & Hashmi, S. D. (2021). Use of polysomnography and home sleep apnea tests for the longitudinal management of obstructive sleep apnea in adults: An American Academy of sleep Medicine clinical guidance statement. *Journal of Clinical Sleep Medicine*, *17*(6), 1287–1293. https://doi.org/10.5664/jcsm.9240

Chen, Y., Wang, W., Guo, Y., Zhang, H., & Xie, L. (2021). A single-center validation of the accuracy of a Photoplethysmography-based smartwatch for screening obstructive sleep apnea. *Nature and Science of Sleep*, *13*, 1533–1544. https://doi.org/10.2147/NSS.S323286

Cheng, S. K., & Dizon, J. (2012). Computerised cognitive behavioural therapy for insomnia: A systematic review and meta-analysis. *Psychotherapy and Psychosomatics*, *81*(4), 206–216. https://doi.org/10.1159/000335379

Chiang, J. K., Lin, Y. C., Lin, C. W., Ting, C. S., Chiang, Y. Y., & Kao, Y. H. (2022). Validation of snoring detection using a smartphone app. *Sleep & Breathing*, *26*(1), 81–87. https://doi.org/10.1007/s11325-021-02359-3

Cho, S. W., Jung, S. J., Shin, J. H., Won, T. B., Rhee, C. S., & Kim, J. W. (2022). Evaluating prediction models of sleep apnea from smartphone-recorded sleep breathing sounds. *JAMA Otolaryngology. Head & Neck Surgery*, *148*(6), 515–521. https://doi.org/10.1001/jamaoto.2022.0244

Chu, H., & Cole, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: A generalized linear mixed model approach. *Journal of Clinical Epidemiology*, *59*(12), 1331–1332; author reply 1332-1333. https://doi.org/10.1016/j.jclinepi.2006.06.011

Chung, F., Yegneswaran, B., Liao, P., Chung, S. A., Vairavanathan, S., Islam, S., … Shapiro, C. M. (2008a). STOP questionnaire: A tool to screen patients for obstructive sleep apnea. *Anesthesiology*, *108*(5), 812–821. https://doi.org/10.1097/ALN.0b013e31816d83e4

Chung, F., Yegneswaran, B., Liao, P., Chung, S. A., Vairavanathan, S., Islam, S., … Shapiro, C. M. (2008b). Validation of the Berlin questionnaire and American Society of Anesthesiologists checklist as screening tools for obstructive sleep apnea in surgical patients. *Anesthesiology*, *108*(5), 822–830. https://doi.org/10.1097/ALN.0b013e31816d91b5

Collop, N. A., Anderson, W. M., Boehlecke, B., Claman, D., Goldberg, R., Gottlieb, D. J., … Portable Monitoring Task Force of the American Academy of Sleep Medicine. (2007). Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. Portable monitoring task force of the American Academy of sleep Medicine. *Journal of Clinical Sleep Medicine*, *3*(7), 737–747.

de Zambotti, M., Menghini, L., Grandner, M. A., Redline, S., Zhang, Y., Wallace, M. L., & Buxton, O. M. (2022). Rigorous performance evaluation (previously, "validation") for informed use of new technologies for

sleep health measurement. *Sleep Health*, *8*(3), 263–269. https://doi.org/10.1016/j.sleh.2022.02.006

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, *58*(9), 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016

Ding, L., Chen, B., Dai, Y., & Li, Y. (2022). A meta-analysis of the first-night effect in healthy individuals for the full age spectrum. *Sleep Medicine*, *89*, 159–165. https://doi.org/10.1016/j.sleep.2021.12.007

Dinnes, J., Deeks, J., Leeflang, M., & Li, T. (2021). Chapter 9: Collecting data. Draft version (20 May 2021). In J. Deeks, P. Bossuyt, M. Leeflang, & Y. Takwoingi (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy* (2nd ed.). Cochrane.

Edinger, J. D., Arnedt, J. T., Bertisch, S. M., Carney, C. E., Harrington, J. J., Lichstein, K. L., … Martin, J. L. (2021). Behavioral and psychological treatments for chronic insomnia disorder in adults: An American Academy of sleep Medicine systematic review, meta-analysis, and GRADE assessment. *Journal of Clinical Sleep Medicine*, *17*(2), 263–298. https://doi.org/10.5664/jcsm.8988

Figueras-Alvarez, O., Cantó-Navés, O., Cabratosa-Termes, J., Roig-Cayón, M., Felipe-Spada, N., & Tomàs-Aliberas, J. (2020). Snoring intensity assessment with three different smartphones using the SnoreLab application in one participant. *Journal of Clinical Sleep Medicine*, *16*(11), 1971–1974. https://doi.org/10.5664/jcsm.8676

Fino, E., & Mazzetti, M. (2019). Monitoring healthy and disturbed sleep through smartphone applications: A review of experimental evidence. *Sleep & Breathing*, *23*(1), 13–24. https://doi.org/10.1007/s11325-018-1661-3

Fischer, J., Dogas, Z., Bassetti, C. L., Berg, S., Grote, L., Jennum, P., … Board of the European Sleep Research Society (ESRS). (2012). Standard procedures for adults in accredited sleep medicine centres in Europe. *Journal of Sleep Research*, *21*(4), 357–368. https://doi.org/10.1111/j.1365-2869.2011.00987.x

Freeman, S. C., Kerby, C. R., Patel, A., Cooper, N. J., Quinn, T., & Sutton, A. J. (2019). Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Medical Research Methodology*, *19*(1), 81. https://doi.org/10.1186/s12874-019-0724-x

Gopalakrishna, G., Mustafa, R. A., Davenport, C., Scholten, R. J., Hyde, C., Brozek, J., … Langendam, M. W. (2014). Applying grading of recommendations assessment, development and evaluation (GRADE) to diagnostic tests was challenging but doable. *Journal of Clinical Epidemiology*, *67*(7), 760–768. https://doi.org/10.1016/j.jclinepi.2014.01.006

Gu, W., Leung, L., Kwok, K. C., Wu, I. C., Folz, R. J., & Chiang, A. A. (2020). Belun ring platform: A novel home sleep apnea testing system for assessment of obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, *16*(9), 1611–1617. https://doi.org/10.5664/jcsm.8592

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., … GRADE Working Group. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., & Castriotta, R. J. (2019). Accuracy of wristband Fitbit models in assessing sleep: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *21*(11), e16273. https://doi.org/10.2196/16273

Holman, B., Bero, L., & Mintzes, B. (2019). Industry sponshorship bias. In *Catalogue of Bias—Center for Evidence-Based Medicine*. University of Oxford.

Jin, H., Tao, X., Dong, S., Qin, Y., Yu, L., Luo, J., & Deen, M. J. (2017). Flexible surface acoustic wave respiration sensor for monitoring obstructive sleep apnea syndrome. *Journal of Micromechanics and Microengineering*, *27*, 115006.

John, A., Nundy, K. K., Cardiff, B., & John, D. (2021). SomnNET: An SpO2 based deep learning network for sleep apnea detection in smartwatches. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *2021*, 1961–1964. https://doi.org/10.1109/EMBC46164.2021.9631037

Jung, H., Kim, D., Lee, W., Seo, H., Seo, J., Choi, J., & Joo, E. Y. (2022). Performance evaluation of a wrist-worn reflectance pulse oximeter during sleep. *Sleep Health*, *706*, 420–428. https://doi.org/10.1016/j.sleh.2022.04.003

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., & Harrod, C. G. (2017). Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American Academy of sleep Medicine clinical practice guideline. *Journal of Clinical Sleep Medicine*, *13*(3), 479–504. https://doi.org/10.5664/jcsm.6506

Khosla, S., Deak, M. C., Gault, D., Goldstein, C. A., Hwang, D., Kwon, Y., … American Academy of Sleep Medicine Board of Directors. (2018). Consumer sleep technology: An American Academy of sleep Medicine position statement. *Journal of Clinical Sleep Medicine*, *14*(5), 877–880. https://doi.org/10.5664/jcsm.7128

Kim, D. H., Kim, S. W., & Hwang, S. H. (2022). Diagnostic value of smartphone in obstructive sleep apnea syndrome: A systematic review and meta-analysis. *PLoS One*, *17*(5), e0268585. https://doi.org/10.1371/journal.pone.0268585

Klaus, K., Stummer, A. L., & Ruf, S. (2021). Accuracy of a smartphone application measuring snoring in adults-how smart is it actually? *International Journal of Environmental Research and Public Health*, *18*(14), 7326. https://doi.org/10.3390/ijerph18147326

Korkalainen, H., Nikkonen, S., Kainulainen, S., Dwivedi, A. K., Myllymaa, S., Leppänen, T., & Töyräs, J. (2021). Self-applied home sleep recordings: The future of sleep Medicine. *Sleep Medicine Clinics*, *16*(4), 545–556. https://doi.org/10.1016/j.jsmc.2021.07.003

Kuna, S. T., Benca, R., Kushida, C. A., Walsh, J., Younes, M., Staley, B., … Malhotra, A. (2013). Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*, *36*(4), 583–589. https://doi.org/10.5665/sleep.2550

Lee, Y. J., Lee, J. Y., Cho, J. H., & Choi, J. H. (2022). Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine*, *18*(1), 193–202. https://doi.org/10.5664/jcsm.9538

Leeflang, M. M. (2014). Systematic reviews and meta-analyses of diagnostic test accuracy. *Clinical Microbiology and Infection*, *20*(2), 105–113. https://doi.org/10.1111/1469-0691.12474

Leeflang, M. M., Davenport, C., & Bossuyt, P. (2022). Chapter 6: Defining the review question. Draft version (14 January 2022). In J. J. Deeks, P. M. Bossuyt, M. M. Leeflang, & Y. Takwoingi (Eds.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (2nd ed.). Cochrane.

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., & Bero, L. (2017). Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews*, *2*, MR000033. https://doi.org/10.1002/14651858.MR000033.pub3

Magalang, U. J., Arnardottir, E. S., Chen, N. H., Cistulli, P. A., Gíslason, T., Lim, D., … Investigators, S. (2016). Agreement in the scoring of respiratory events among international sleep centers for home sleep testing. *Journal of Clinical Sleep Medicine*, *12*(1), 71–77. https://doi.org/10.5664/jcsm.5398

Małolepsza, A., Kudrycka, A., Karwowska, U., Hoshino, T., Wibowo, E., Pál Böjti, P., … Kuczyński, W. (2021). The role of screening questionnaires in the assessment of risk and severity of obstructive sleep apnea—polysomnography versus polygraphy. *Advances in Respiratory Medicine*, *89*(2), 188–196. https://doi.org/10.5603/ARM.a2021.0038

Marinari, S., Volpe, P., Simoni, M., Aventaggiato, M., De Benedetto, F., Nardini, S., … Palange, P. (2022). Accuracy of a new pulse oximetry in detection of arterial oxygen saturation and heart rate measurements: The SOMBRERO study. *Sensors (Basel, Switzerland)*, *22*(13), 5031. https://doi.org/10.3390/s22135031

Marti-Soler, H., Hirotsu, C., Marques-Vidal, P., Vollenweider, P., Waeber, G., Preisig, M., … Heinzer, R. (2016). The NoSAS score for screening of sleep-disordered breathing: A derivation and validation study. *The Lancet Respiratory Medicine*, *4*(9), 742–748. https://doi.org/10.1016/S2213-2600(16)30075-3

Menghini, L., Cellini, N., Goldstone, A., Baker, F. C., & de Zambotti, M. (2021). A standardized framework for testing the performance of sleep-tracking technology: Step-by-step guidelines and open-source code. *Sleep*, *44*(2), zsaa170. https://doi.org/10.1093/sleep/zsaa170

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1. https://doi.org/10.1186/2046-4053-4-1

Mokhtaran, M., Sacchi, L., Tibollo, V., Risi, I., Ramella, V., Quaglini, S., & Fanfulla, F. (2022). Obstructive sleep apnea home-monitoring using a commercial wearable device. *Studies in Health Technology and Informatics*, *290*, 522–525. https://doi.org/10.3233/SHTI220131

Nakano, H., Hirayama, K., Sadamitsu, Y., Toshimitsu, A., Fujita, H., Shin, S., & Tanigawa, T. (2014). Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: Proof of concept. *Journal of Clinical Sleep Medicine*, *10*(1), 73–78. https://doi.org/10.5664/jcsm.3364

Narayan, S., Shivdare, P., Niranjan, T., Williams, K., Freudman, J., & Sehra, R. (2019). Noncontact identification of sleep-disturbed breathing from smartphone-recorded sounds validated by polysomnography. *Sleep & Breathing*, *23*(1), 269–279. https://doi.org/10.1007/s11325-018-1695-6

Nester, R. (2019). Smart Sleep Tracking Device Market: Global Demand Analysis & Opportunity Outlook 2024.

Nikkonen, S., Afara, I. O., Leppänen, T., & Töyräs, J. (2020). Author correction: Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. *Scientific Reports*, *10*(1), 4977. https://doi.org/10.1038/s41598-020-62003-0

O'Mahony, A. M., Garvey, J. F., & McNicholas, W. T. (2020). Technologic advances in the assessment and management of obstructive sleep apnoea beyond the apnoea-hypopnoea index: A narrative review. *Journal of Thoracic Disease*, *12*(9), 5020–5038. https://doi.org/10.21037/jtd-sleep-2020-003

Óskarsdóttir, M., Islind, A. S., August, E., Arnardóttir, E. S., Patou, F., & Maier, A. M. (2022). Importance of getting enough sleep and daily activity data to assess variability: Longitudinal observational study. *JMIR Formative Research*, *6*(2), e31807. https://doi.org/10.2196/31807

Papini, G. B., Fonseca, P., van Gilst, M. M., Bergmans, J. W. M., Vullings, R., & Overeem, S. (2020). Wearable monitoring of sleep-disordered breathing: Estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography. *Scientific Reports*, *10*(1), 13512. https://doi.org/10.1038/s41598-020-69935-7

Patel, A., Cooper, N., Freeman, S., & Sutton, A. (2021). Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. *Research Synthesis Methods*, *12*(1), 34–44. https://doi.org/10.1002/jrsm.1439

Penzel, T., Dietz-Terjung, S., Woehrle, H., & Schöbel, C. (2021). New paths in respiratory sleep Medicine: Consumer devices, e-health, and digital health measurements. *Sleep Medicine Clinics*, *16*(4), 619–634. https://doi.org/10.1016/j.jsmc.2021.08.006

Pepin, J. L., Le-Dong, N. N., Cuthbert, V., Coumans, N., Tamisier, R., Malhotra, A., & Martinot, J. B. (2022). Mandibular movements are a reliable noninvasive alternative to esophageal pressure for measuring respiratory effort in patients with sleep apnea syndrome. *Nature and science of sleep*, *14*, 635–644. https://doi.org/10.2147/NSS.S346229

Pépin, J. L., Letesson, C., Le-Dong, N. N., Dedave, A., Denison, S., Cuthbert, V., … Gozal, D. (2020). Assessment of mandibular movement monitoring with machine learning analysis for the diagnosis of obstructive sleep apnea. *JAMA Network Open*, *3*(1), e1919657. https://doi.org/10.1001/jamanetworkopen.2019.19657

Perez-Pozuelo, I., Zhai, B., Palotti, J., Mall, R., Aupetit, M., Garcia-Gomez, J. M., … Fernandez-Luque, L. (2020). The future of sleep health: A data-driven revolution in sleep science and medicine. *NPJ Digital Medicine*, *3*, 42. https://doi.org/10.1038/s41746-020-0244-4

Pevernagie, D. A., Gnidovec-Strazisar, B., Grote, L., Heinzer, R., McNicholas, W. T., Penzel, T., … Arnardottir, E. S. (2020). On the rise and fall of the apnea-hypopnea index: A historical review and critical appraisal. *Journal of Sleep Research*, *29*(4), e13066. https://doi.org/10.1111/jsr.13066

Rosa, T., Bellardi, K., Viana, A., Ma, Y., & Capasso, R. (2018). Digital health and sleep-disordered breathing: A systematic review and meta-analysis. *Journal of Clinical Sleep Medicine*, *14*(9), 1605–1620. https://doi.org/10.5664/jcsm.7346

Rosen, I. M., Kirsch, D. B., Carden, K. A., Malhotra, R. K., Ramar, K., Aurora, R. N., … American Academy of Sleep Medicine Board of Directors. (2018). Clinical use of a home sleep apnea test: An updated American Academy of sleep Medicine position statement. *Journal of Clinical Sleep Medicine*, *14*(12), 2075–2077. https://doi.org/10.5664/jcsm.7540

Salameh, J. P., Bossuyt, P. M., McGrath, T. A., Thombs, B. D., Hyde, C. J., Macaskill, P., … McInnes, M. D. F. (2020). Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): Explanation, elaboration, and checklist. *BMJ*, *370*, m2632. https://doi.org/10.1136/bmj.m2632

Santos, M., Vollam, S., Pimentel, M. A., Areia, C., Young, L., Roman, C., … Watkinson, P. (2022). The use of wearable pulse oximeters in the prompt detection of hypoxemia and during movement: Diagnostic accuracy study. *Journal of Medical Internet Research*, *24*(2), e28890. https://doi.org/10.2196/28890

Schmitz, L., Sveinbjarnarson, B., Gunnarsson, G., Davíðsson, Ó., Davíðsson, Þ., Arnardóttir, E., … Islind, A. (2022). Towards a Digital Sleep Diary Standard. In *Proceedings of 28th Americas conference on information systems (AMCIS)*. Association for Information Systems.

Schünemann, H. J., Mustafa, R. A., Brozek, J., Santesso, N., Bossuyt, P. M., Steingart, K. R., … GRADE Working Group. (2019). GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. *Journal of Clinical Epidemiology*, *111*, 69–82. https://doi.org/10.1016/j.jclinepi.2019.02.003

Schünemann, H. J., Mustafa, R. A., Brozek, J., Steingart, K. R., Leeflang, M., Murad, M. H., … GRADE Working Group. (2020a). GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. *Journal of Clinical Epidemiology*, *122*, 129–141. https://doi.org/10.1016/j.jclinepi.2019.12.020

Schünemann, H. J., Mustafa, R. A., Brozek, J., Steingart, K. R., Leeflang, M., Murad, M. H., … GRADE Working Group. (2020b). GRADE guidelines: 21 part 2. Test accuracy: Inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *122*, 142–152. https://doi.org/10.1016/j.jclinepi.2019.12.021

Schünemann, H. J., Schünemann, A. H., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., … GRADE Working Group. (2008). Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*, *336*(7653), 1106–1110. https://doi.org/10.1136/bmj.39500.677199.AE

Schutte-Rodin, S., Deak, M. C., Khosla, S., Goldstein, C. A., Yurcheshen, M., Chiang, A., … Ramar, K. (2021). Evaluating consumer and clinical sleep technologies: An American Academy of sleep Medicine update. *Journal of Clinical Sleep Medicine*, *17*(11), 2275–2282. https://doi.org/10.5664/jcsm.9580

Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., … Dharmage, S. C. (2017). Prevalence of obstructive sleep apnea in the general population: A systematic review. *Sleep Medicine Reviews*, *34*, 70–81. https://doi.org/10.1016/j.smrv.2016.07.002

Seyffert, M., Lagisetty, P., Landgraf, J., Chopra, V., Pfeiffer, P. N., Conte, M. L., & Rogers, M. A. (2016). Internet-delivered cognitive behavioral therapy to treat insomnia: A systematic review and meta-analysis. *PLoS One*, *11*(2), e0149139. https://doi.org/10.1371/journal.pone.0149139

Stippig, A., Hübers, U., & Emerich, M. (2015). Apps in sleep medicine. *Sleep & Breathing*, *19*(1), 411–417. https://doi.org/10.1007/s11325-014-1009-6

Tangudu, V., Afrin, K., Reddy, S., Deutz, N. E. P., Woltering, S., & Bukkapatnam, S. T. S. (2021). Toward standardizing the clinical testing protocols of point-of-care devices for obstructive sleep apnea diagnosis. *Sleep & Breathing*, *25*(2), 737–748. https://doi.org/10.1007/s11325-020-02171-5

Taylor, K. S., Mahtani, K. R., & Aronson, J. K. (2021). Extracting data from diagnostic test accuracy studies for meta-analysis. *BMJ Evidence-Based Medicine*, *26*(1), 19–21. https://doi.org/10.1136/bmjebm-2020-111650

Tiron, R., Lyon, G., Kilroy, H., Osman, A., Kelly, N., O'Mahony, N., … Penzel, T. (2020). Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. *Journal of Thoracic Disease*, *12*(8), 4476–4495. https://doi.org/10.21037/jtd-20-804

Uddin, M. B., Chow, C. M., & Su, S. W. (2018). Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: A systematic review. *Physiological Measurement*, *39*(3), 3TR01. https://doi.org/10.1088/1361-6579/aaafb8

van Enst, W. A., Ochodo, E., Scholten, R. J., Hooft, L., & Leeflang, M. M. (2014). Investigation of publication bias in meta-analyses of diagnostic test accuracy: A meta-epidemiological study. *BMC Medical Research Methodology*, *14*, 70. https://doi.org/10.1186/1471-2288-14-70

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., … Group, Q. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Xie, F., & Zhou, T. (2022). Industry sponsorship bias in cost effectiveness analysis: Registry based analysis. *BMJ*, *377*, e069573. https://doi.org/10.1136/bmj-2021-069573

Ye, Y. Y., Chen, N. K., Chen, J., Liu, J., Lin, L., Liu, Y. Z., … Jiang, X. J. (2016). Internet-based cognitive-behavioural therapy for insomnia (ICBT-i): A meta-analysis of randomised controlled trials. *BMJ Open*, *6*(11), e010707. https://doi.org/10.1136/bmjopen-2015-010707

Zachariae, R., Lyby, M. S., Ritterband, L. M., & O'Toole, M. S. (2016). Efficacy of internet-delivered cognitive-behavioral therapy for insomnia—A systematic review and meta-analysis of randomized controlled trials. *Sleep Medicine Reviews*, *30*, 1–10. https://doi.org/10.1016/j.smrv.2015.10.004

Zhao, R., Xue, J., Zhang, X., Peng, M., Li, J., Zhou, B., … Han, F. (2022). Comparison of ring pulse oximetry using reflective Photoplethysmography and PSG in the detection of OSA in Chinese adults: A pilot study. *Nature and Science of Sleep*, *14*, 1427–1436. https://doi.org/10.2147/NSS.S367400

How to cite this article: Pires, G. N., Arnardóttir, E. S., Islind, A. S., Leppänen, T., & McNicholas, W. T. (2023). Consumer sleep technology for the screening of obstructive sleep apnea and snoring: current status and a protocol for a systematic review and meta-analysis of diagnostic test accuracy. *Journal of Sleep Research*, e13819. https://doi.org/10.1111/jsr.13819