Sleep and Breathing (2021) 25:737–748
https://doi.org/10.1007/s11325-020-02171-5

SLEEP BREATHING PHYSIOLOGY AND DISORDERS • ORIGINAL ARTICLE

Toward standardizing the clinical testing protocols of point-of-care devices for obstructive sleep apnea diagnosis

Vivek Tangudu[1] · Kahkashan Afrin[1] · Sandy Reddy[2] · Nicolaas E.P. Deutz[3] · Steven Woltering[4] · Satish T. S. Bukkapatnam[1,2]

Received: 19 April 2020 / Revised: 4 August 2020 / Accepted: 12 August 2020 / Published online: 31 August 2020

Abstract

Purpose In recent years, point-of-care (POC) devices, especially smart wearables, have been introduced to provide a cost-effective, comfortable, and accessible alternative to polysomnography (PSG)—the current gold standard—for the monitoring, screening, and diagnosis of obstructive sleep apnea (OSA). Thorough validation and human subject testing are essential steps in the translation of these device technologies to the market. However, every device development group tests their device in their own way. No standard guidelines exist for assessing the performance of these POC devices. The purpose of this paper is to critically distill the key aspects of the various protocols reported in the literature and present a protocol that unifies the best practices for testing wearable and other POC devices for OSA.

Methods A limited review and graphical descriptive analytics of literature—including journal articles, web sources, and clinical manuscripts by authoritative agencies in sleep medicine—are performed to glean the testing and validation methods employed for POC devices, specifically for OSA.

Results The analysis suggests that the extent of heterogeneity of the demographics, the performance metrics, subject survey, hypotheses, and statistical analyses need to be carefully considered in a systematic protocol for testing POC devices for OSA.

Conclusion We provide a systematic method and list specific recommendations to extensively assess various performance criteria for human subject testing of POC devices. A rating scale of 1–3 is provided to encourage studies to put a focus on addressing the key elements of a testing protocol.

Keywords Obstructive sleep apnea diagnosis · Portable monitors · Wearable sensor · Protocol · Human subject testing

Introduction

Obstructive sleep apnea (OSA) is a sleep disorder characterized by recurrent cessations of breathing. Patients regularly report functional impairments and daytime fatigue, which affect cognitive performance and overall quality of life [1, 2]. In fact, motor vehicle accidents are three to seven times more likely in drivers with untreated sleep apnea [3]. Untreated OSA also accounts for approximately $3.4 billion to health care costs in the USA [4, 5]. It has been epidemiologically at high risk of comorbidity with diseases such as stroke, obesity, and hypertension [6–9]. The development of an accurate, affordable, and convenient diagnostic device is necessary to improve worldwide accessibility to OSA assessment. While this disorder is prevalent among the population, studies show that 75 to 80% of OSA patients still remain undiagnosed [10–12].

Polysomnography (PSG), which must be performed in the sleep laboratory setting, is the only clinical gold standard for diagnosing OSA. However, with costs as high as $6000 per

Presentation at a conference: None

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11325-020-02171-5) contains supplementary material, which is available to authorized users.

✉ Satish T. S. Bukkapatnam
satish@tamu.edu

[1] Department of Industrial & Systems Engineering, Texas A&M University, 101 Bizzell St, College Station, TX 77843, USA

[2] Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA

[3] Center for the Translation Research in Aging and Longevity, Human Clinical Research Facility, Texas A&M University, College Station, TX 77843, USA

[4] Department of Educational Psychology, Texas A&M University, College Station, TX 77843, USA

(overnight) test, a frequent need to conduct more than two nights of study, and an average appointment wait time of about 1 year, PSG testing proves difficult to access [13–16]. This concern is highlighted by the fact that countries outside of the USA struggle to supply even conservative numbers of PSG devices and sleep laboratories necessary for fast, extensive diagnosis of all patients. This results in a lack of opportunity for diagnosis for potential OSA patients. PSG is also marked with its own demerits in regard to accuracy in diagnosis. Without adequate technician editing, the autoscoring methods can produce varying results compared with the alternative manual paper scoring [17, 18]. Additionally, a subject often finds it hard to fall asleep during the first night of a PSG test due to the difference in the sleep environment as well as continuous contact with multiple wires and sensors. This introduces a significant bias and possible variability among the outcomes from different nights of testing [19]. These differing results, attributed to the night-to-night effect, are particularly important in milder cases when OSA is detected by a small margin [20]. For other patients, the limited ergonomic features of PSG will make an impact on diagnoses. The poor level of sleep in the laboratory compared with home environments will impact the metrics used to diagnose OSA, such as the apnea-hypopnea index (AHI) and respiratory disturbance index (RDI), particularly for children and the elderly [21].

Home sleep testing (HST) is an alternative with greater convenience to patients but a different set of disadvantages [22–25]. Equipment is limited in resources and still results in expenses of over $700 in rental purchases. Furthermore, diagnostic quality may be impacted by a lack of patient supervision during self-hookup and potentially inadequate system sensitivity. While HST can provide an indication of a sleep disorder, negative results on a symptomatic patient do not necessarily rule out the possibility that one exists [26, 27].

With recent technological advancements, there has been a significant increase in the efforts to develop alternative point-of-care, wearable devices for OSA monitoring, diagnosis, and treatment. Portable monitors, activity monitors, and non-contact sensors are just some of the recent developments in the field of point-of-care (POC) device technologies for OSA. Upholding major initiatives to improve comfort, health outcomes, and affordable options, POC testing devices also offer the versatility to be utilized in several different sleeping environments [28, 29]. The ability to remotely monitor and collect medical data proves to be even more beneficial for chronically ill and elderly patients [30]. Moreover, professionals become capable of making quicker decisions and addressing concerns more proactively. In turn, this can improve general wellness of the patient and potentially reduce the duration of care. Such data integration can not only be incorporated through daily home monitoring but also alert the appropriate medical and personal caregivers with up-to-date information on a patient's condition. Also, POC diagnosis rapidly improves operating efficiency by reducing hospital visits and promoting accurate assessments in emergency care.

Several healthcare providers have implemented POC portable devices (e.g., EarlySense [31], HealthPatch [32], Watch PAT_100 [33]) with the intent to provide an alternative for OSA detection. However, each paper validating the quality of devices utilizes a different experimental setup and methods to compare to the gold standard. Such heterogeneity in the study design and result reporting puts the relative performance of these devices into question. Methods such as inter-scorer reliability, for example, should be universally implemented to ultimately improve the homogeneity of these validation studies. Without a proper baseline of assessment for each device, there is no certainty that a device can address practice guidelines, accreditation standards, or management principles.

Therefore, streamlining and standardization of the testing procedures and metrics for assessing the performance of POC devices for OSA is highly desirable. The purpose of this paper is to distill the best elements of prior testing methods and thereby establish necessary guidelines for the design, implementation, and reporting of studies on POC wearable devices for OSA. This distilled protocol serves to provide specificity on the level of evidence required from future studies. It is anticipated that streamlining and standardization of protocols can enhance the adoption of POC diagnostic devices for OSA. The remainder of the paper is organized as follows: the method employed to conduct a critical review and graphical analysis of the related literature is presented in "Materials and methods"; the resulting set of protocols for POC devices, distilled from the best elements of the methods reported in the literature, is presented in "Results"; and "Discussion" presents a brief discussion of the results and concluding remarks.

Materials and methods

The main objectives of this paper are to identify the critical elements of the current OSA device validation methods reported in the literature and to apply a graphical descriptive analytics method to establish the non-uniformity among the current methods for validation of OSA devices. The following sections describe the methodology in more detail.

Identification and selection of relevant literature

As summarized in Fig. 1, five databases were used to identify articles on POC devices for OSA: Wiley, Springer, Elsevier, IEEE Xplore, and Google Scholar. Articles resulting from the search came from a number of sources, including web pages, book chapters, and journals. The keywords include "Point-of-care," "Portable," "Sleep Apnea," "Validation," "Protocol," "Assessment," and "Literature Review." Papers were assessed by their title, abstract, and content to see if they were related to

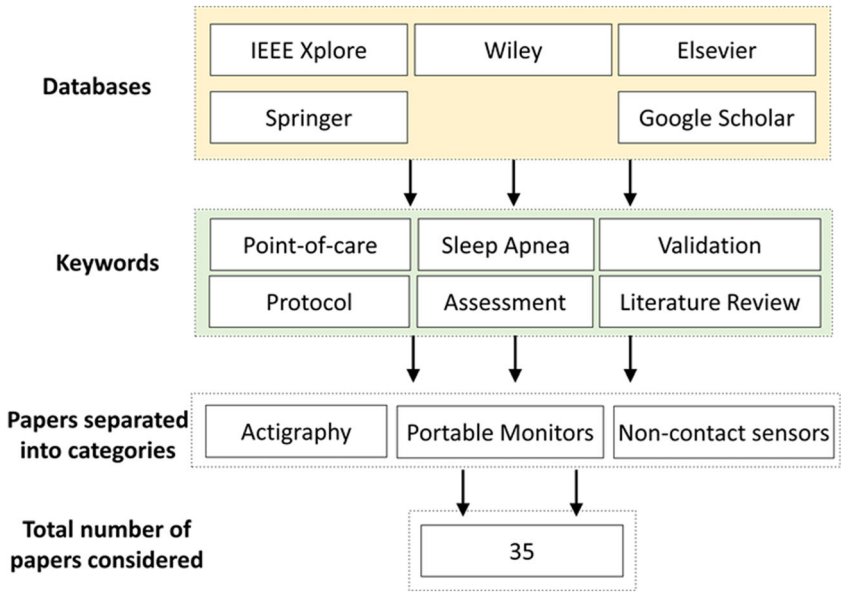

Fig. 1 Framework for article inclusion

the search terms. Articles sought for this review had to address the validation of a POC device with a primary target of OSA patients. The diagnostic results from the listed device also required comparison with a gold standard. There was also preference for selecting literature with a greater number of citations. The intent of this paper is to encourage future adoption of the protocol; by reviewing and analyzing the most popular papers, these standardized measures would be taken with a higher level of credibility. A total of 35 articles were finally reviewed after thorough collection of the literature.

Selection of articles for review

In an effort to fully display the dissimilarities among as well as some missing details in studies reported, we summarized literature into two tables. Table 1 presents demographic data of the five most commonly cited papers in our review of POC device validation for OSA. As evident from the table, these studies show a lack of consistency in all aspects, including the number of subjects and Epworth Sleepiness Scale (ESS). These are two critical factors for validation of an OSA device, since they affect variability in results and reflect the careful selection of subjects. This bird's-eye view of the table also serves to quickly provide insight into the irregularities present in the literature. The majority of the other papers reviewed also followed this manner of non-uniformity in validation of OSA devices, thus underlying the importance of a standardized protocol. Table A1 included as part of the Supplementary section summarizes all 35 reviewed papers to fully explain which key points of information were present and not present. In particular, the table displays the lack of consistency in every paper's "methods" and "results" sections. For example, only about 50% of papers in this review presented methods of feature extraction (i.e., manual scoring, epoch-by-epoch analysis, etc.).

Table 1 Subject demographic data in the five most-cited papers among those included in the present review

	Patients and age (years)	Number of patients (M/F)	Body mass index (BMI) (kg/m^2)	Epworth Sleepiness Scale (0–24)
MESEM 4 [34]	47.5	56 (46/10)	27	Not listed
Watch-PAT100 [25]	41.4	185 (162/2-3)	34.7	16.4
Actillume recorder [35]	86.4	10 (2/8)	Not listed	Not listed
Actiwatch-L [36]	Not listed	20	Not Listed	Not listed
Mini Motionlogger Actigraph [37]	25.5	21 (7/14)	Not listed	Not listed

Inclusion criteria

- Studies that seek to validate a POC device for obstructive sleep apnea patients
- Studies that specifically proposed a POC device in the form of portable monitors, actigraphy, or non-contact sensors
- Studies published in the English language

Exclusion criteria

- Studies that did not report a validation study or device comparison study

- Studies that did not compare the proposed POC device with a customary device

Key terms

In this paper, "obstructive sleep apnea" (OSA) is defined by recurrent cessations of breathing. Specifically, apneas are defined as a cessation of airflow ≥ 10 s and hypopneas as reduction of respiratory signals ≥ 10 s associated with oxygen desaturation > 4% and/or arousal. "Apnea-hypopnea index" (AHI) refers to the number of apneas and hypopneas per hour of sleep time and is considered the main metric for diagnosing OSA. "Respiratory disturbance index" (RDI) similarly refers to the number of apneas and hypopneas per hour of recording time, and "oxygen desaturation index (ODI)" measures the number of times that the oxygen level in blood drops by 4% or more during sleep. "Point-of-care" (POC) refers to any device that is considered accessible and relatively comfortable for a patient to wear (a.k.a. wearables). In this paper, POC devices include portable monitors (PM), actigraphy, and non-contact sensors. "Polysomnography" (PSG) is the diagnostic tool for sleep, widely considered the gold standard. "Validation paper" refers to a paper that provides evidence backing a proposed device when compared with a gold standard device. In this context, validation papers often serve to compare a POC device with PSG.

Figure 2 provides an encapsulated summary of the key elements of the OSA device validation methods reported in the literature. It serves to visually display Table A1 and provide takeaways regarding the frequency and correlation of different variables, or aspects, in each paper. Each color represents a unique category of a paper, and each node is sized according to the frequency in which it is utilized. For example, some of the larger nodes indicate that more than 60% of papers validate their device by using AHI and total sleep time (TST) as metrics for further analysis by Bland-Altman plots and statistical tests. Some categories of the graphic, such as metrics and quantitative validation, only display the most common aspects gleaned from the literature. The devices that are currently in the market are represented as bright (yellow) colored nodes under the *ground truth instrument* column. Nodes shaded with a gray color indicate that these devices are not available in the market.

In addition to visually representing each node, the graphic uses uniquely colored lines to correspond to each sample size range. This extra level of visibility brings out a number of meaningful insights. For one, it can be noted that studies containing more than 51 subjects were more likely to implement a wider range of metrics. Every range containing at least 51 subjects—represented by a red, blue, or green line—used each of the metrics. On the other hand, studies containing less than 25 subjects had only used TST or Sleep Efficiency (SE) as metrics, while implementing correlation as the only means of

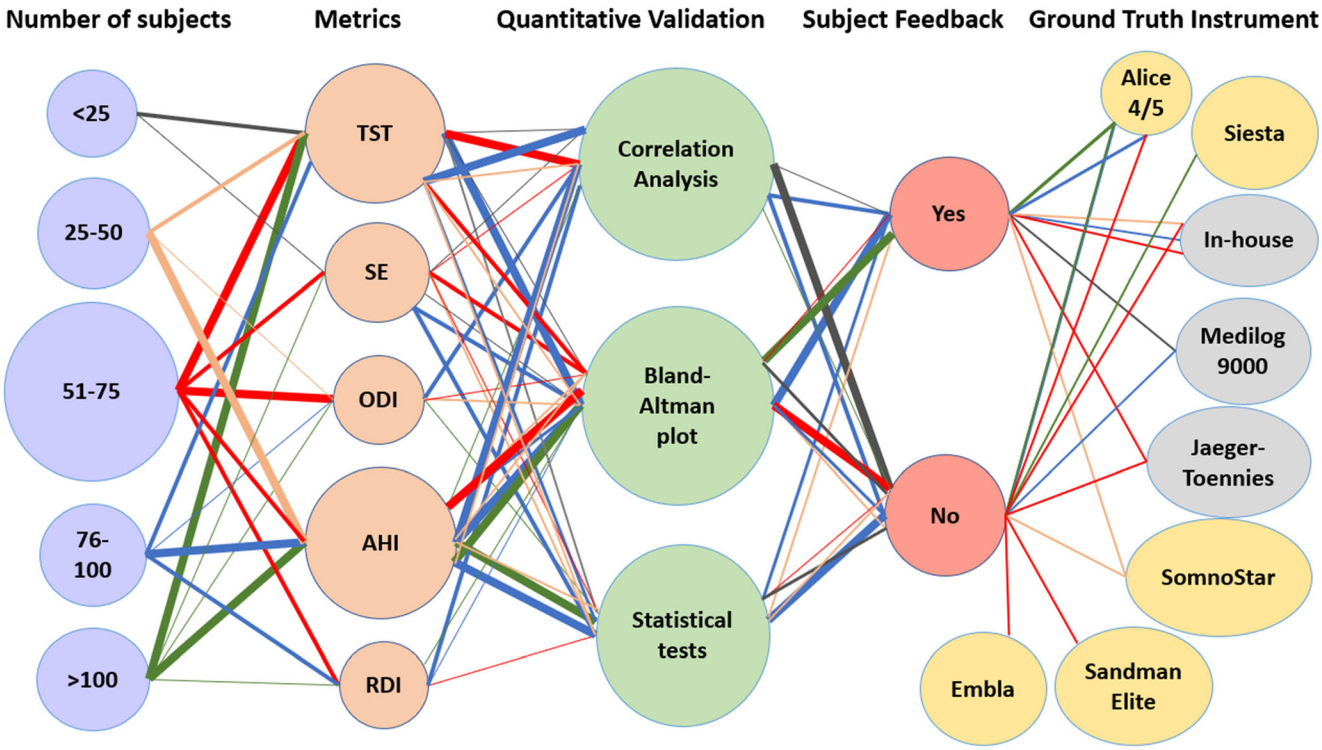

Fig. 2 Interrelations among key variables of OSA device testing papers reviewed [22, 25, 38–77]. Here, TST: total sleep time; SE: sleep efficiency; ODI: oxygen desaturation index; AHI: apnea–hypopnea index; RDI: respiratory disturbance index

quantitative validation. The graphic also uses the thickness of a line to show how frequently there is correlation between two attribute values (i.e., the number of tests in which these attribute values were chosen). Upon closer analysis of the path of these lines, studies were only shown to have consistently used all types of metrics and quantitative validation techniques if they had between 76 and 100 subjects. While every quantitative validation technique was used in conjunction with every metric at least once, statistical testing was more commonly used with AHI and correlation was more commonly used with TST.

This graphic thus brings out not just the extent of correlations between key variables but also visually displays the lack of consistency among papers examining subject feedback and/or using other metrics such as RDI. Following this extensive review of the literature, we present systematic recommendations on the test protocols that will combine the best practices and common elements of all papers and enhance diagnostic quality of devices.

Results

Based on our narrative review and employing procedures from the protocols of other devices [38], we identified a distilled protocol that incorporates seven key aspects of studies on sleep apnea diagnosis devices to ensure valid assessment and comparison. Table 2 provides a summary of these criteria. Following this list of recommendations and underscored observations is a proposed rating system. This rating system, seen in Table 3, assesses the performance of any given study and provides a summarized list of what will be mentioned in this section.

Device

In this section, we will outline each type of device that was reviewed in this paper and present expectations for the performance of each device. While certain reviews have been more inclusive of options for diagnosis (e.g., sleep switch device, Peripheral Arterial Tone (PAT)) [39], this paper, as noted earlier serves to highlight the divergence among the various device validation methods reported in the literature, and provide a protocol that aims to capture and combine their salient aspects.

Table 2 Key aspects of POC device studies for OSA

Key criteria	Summary
Device	Device type and no. of channels
Subjects	Background information
Study design	Detail on study protocol
Signals	Collected data during diagnosis
Metrics/metric extraction	Taken from features
Gold standard	Reference standard used
Validation and comparison	Quantitative and qualitative assessment

Type of devices

The main diagnostic solutions covered in this paper are portable monitors, activity monitors, and non-contact sensors. Actigraphy only requires one channel and sensors have none. As per American Academy of Sleep Medicine (AASM) criteria, sleep monitoring devices exist in four variants depending on the number of channels [40]. Only type III and type IV were considered a valid POC device for this paper since they can be performed at home and are not related to polysomnography (PSG). Type III utilizes at least four channels for diagnosis, while type IV utilizes one or two.

Device requirements and recommendations

The devices covered in this review are largely either of type III or type IV, each measuring a different kind and number of biometric variables. Interesting, one of the devices reviewed had a portable device with the option of an additional auxiliary channel [41]. While this addition was not implemented in the study, the use of additional, optional channels could be explored to better enhance diagnostic accuracy. Portable monitors have also been known to generate many false negatives, thereby placing potential risk for patients not getting diagnosed for OSA. The variability of such evidence tends to increase with the fewer number of channels the monitor has. In an effort to minimalize such a risk, the AASM recommends these systems record at least airflow, respiratory effort, and blood oxygenation [42]. Following these requirements of the protocol can ensure that all devices in the market have a baseline number of channels. An established baseline promotes higher device quality in the long run.

We also review studies not utilizing portable monitors, including activity monitors and non-contact sensors [43–53]. Studies validating activity monitors were more thoroughly reviewed in large part due to these devices' success in the market already. Studies are, however, still recommended to list the device components in use. Few clarify the way motion is being measured (e.g., linear accelerometer) [35, 36]. Even fewer discuss how various image and light signals are recorded. Studies regarding non-contact sensors are not expected to further detail the instruments and equipment employed for diagnosis. Most of these devices intuitively detect body and respiratory movements through the use of sensors not in contact with the subject. Since they are wireless, they may not directly comply with AASM recommendations of recording airflow, respiratory effort, and blood oxygenation. This does

Table 3 Listed categories and subcategories of proposed rating system

Study characteristic	Ratings		
	1	2	3
Blinded readers	One blinded reader	One blinded reader	Two or more blinded readers
Subject conditions	Either the subjects are healthy or diagnosed to some extent of SDB	One group of subjects is healthy and the other is diagnosed to some extent of SDB	One group of subjects is healthy and the other is diagnosed to some extent of SDB
Subject data	Includes age, sex, BMI, ESS	Includes age, sex, BMI, ESS and/or other subject history	Includes age, sex, BMI, ESS, and/or other subject history
Restrictions	No alcohol or caffeine from past 24 h	No alcohol or caffeine from past 24 h	No caffeine from past 48 h and no alcohol from past 72 h
Exclusionary criteria	Subjects follow the ESS criteria	Subjects follow the ESS criteria	Subjects follow the ESS criteria
Self-application of home-based POC device	Informational video	Informational video	Live instruction and informational video provided
Sleep metrics	Lists TST and AHI	Lists TST, AHI, and SE/RDI/ODI	Lists TST, SE, AHI, RDI, ODI
Metric extraction	Lists algorithms/analyses	Lists algorithms/analyses	Explains algorithms/analysis
Quantitative methods	Utilizes Bland-Altman, correlation, and one statistical test	Utilizes Bland-Altman, correlation, specificity, sensitivity, and one statistical test	Utilizes Bland-Altman, correlation, specificity, sensitivity, and more than one statistical test
Qualitative methods	Provided post-sleep questionnaire	Provided post-sleep questionnaire	Provided post lab questionnaire provided and noted number of adjustments
Data protection and security	Ensures compliance with institutional human subject testing standards	Ensures compliance with the institutional, as well as Health Insurance Portability and Accountability Act (HIPAA) and other international standards including the General Data Protection Regulation (GDPR)	Ensures compliance with international standards, and incorporates additional complex data protection, encryption and security to preserve the integrity of data and its management

AHI apnea-hypopnea index, *BMI* body mass index; *ESS* Epworth Sleepiness Scale, *GDPR*: General Data Protection Regulation, *HIPAA*: Health Insurance Portability and Accountability Act, *ODI* oxygen desaturation index, *RDI* respiratory disturbance index, *SDB* sleep-disordered breathing, *SE* sleep efficiency, *TST* total sleep time

not necessarily determine efficacy, but papers should make a note of such limitations (see [52] for additional discussion of this matter).

Subjects

Studies covered subject background in varying detail. Papers often showed an ad hoc approach to implementing exclusionary criteria and displaying demographic tables. The number of subjects in all studies ranged from 6 to 3924. One paper used the regression coefficient to estimate the number of subjects required for testing [44], but that can also be achieved using a desired significance level of 95% or power of 80% [54]. While this a minimum threshold, a report from the World Health Organization [55] indicates that a health study should include a high number of subjects. Figure 2 also suggests that papers containing more than 50 subjects tend to use more metrics and techniques for validation.

Depending on the device validation method, study subjects were either healthy or had some form of sleep disordered breathing. While most studies chose a population including only one or the other, combining both will allow for a broader, comprehensive range of results implying strength in diagnosis. One paper had subjects who were healthy volunteers or from a sleep disorders center; the healthy volunteers served as true negatives in the lower spectrum of results [56]. Implementing these different subject populations will explain the results of a paper far clearer, therefore cementing its credibility. In an effort to increase transparency in study design, all papers should provide subject data regarding age, sex, ESS, and body mass index (BMI) scores. Other recommendations include restrictions prior to testing. As people who drink alcohol [45] or smoke are more likely to have a sleep disorder [57], it is encouraged that subjects refrain from such behavior for at least 24 h before partaking in the study. Only two recorded studies in this review maintain such standards [37, 45]. Enforcing these restrictions can improve test integrity.

Finally, completion of the ESS will guarantee studies have listed exclusionary material for subjects who cannot be eligible for participation. While most studies do contain some exclusionary material, they are varying in detail. It is recommended that the specificity in subjects is kept in a reasonable balance. For instance, consider a POC device that fails to provide consistent, accurate diagnoses for patients with a mild to moderate likelihood for OSA. Studies validating that device might potentially attempt to exclude those specific subjects in order to make the results appear more conclusive. The ESS can serve as a tool to increase transparency and allow for more representative results in this case. While it is only used by less than 40% of literature (see Fig. 2, Tables 2, and AI), the questionnaire ensures that subject criteria are both explicit and wide-ranging. A standardization in exclusionary criteria would significantly reduce variability in results.

Study design

All studies must be approved by an accredited review board to ensure ethical procedures are being followed. In an effort to also remove any potential bias/ conflict of interest, blinded technician recordings are also recommended. Utilizing additional blinded readers can better account for inter-reader variability, as one study did [22]. It is also beneficial to require a sleep technologist to apply or educate the patient on how to apply the sensors. Compared with providing subjects a video for self-instruction, expert guidance in self-applying the device amounts to higher overall success [25, 58]. Factors such as inadequately applying the device also serve to explain why studies should include a home group of subjects. The home environment is far more representative of the potential challenges including the burden that can come from wearing the device, and reduced signal fidelity from a wireless device (e.g., attenuating signal intensity) [53, 56]. Following this protocol section, the rating system includes the aforementioned recommendations to have blinded technician recordings, informational videos, and/or live instruction.

Furthermore, any validation paper must include the ideal populations for said device. Many studies are already intended for patients with a particular ethnicity or disability [35, 50], but few will clarify on the device's optimal performance for a certain age demographic. Limited numbers of POC devices are currently clinically used. To at least provide diagnoses for certain populations, studies are recommended to specify prime subject demographics in the conclusion of the paper. These improvements in study design will further demonstrate the practicality of any given point-of-care device if it enters the market. The aforementioned improved practices would also increase the chances of Food and Drug Administration (FDA) approval since more effort was put into the study.

Signals

Sleep is recorded in a variety of methods for diagnosis, most commonly through body position, thoracic and abdominal respiration, oxygen saturation, and oronasal flow. Type III portable monitors measure all of the aforementioned channels, while type IV portable monitors record oxygen saturation. This data is then analyzed to gather sleep indices required for validation, such as AHI, ODI, and RDI. While type III portable monitors are recommended for enhanced diagnostic accuracy, type IV monitors have slowly proven credence and are recommended as a cheaper alternative [59]. One consideration for improving analysis would be to implement additional means of data collection (i.e., higher number of channels). For example, one paper determined snoring levels from acoustic microphones and head movement/ position from accelerometers [56].

Activity monitors primarily measure body movements, whereas non-contact sensors often measure body movement as well as heart rate and respiration. These sensors serve to detect sleep stages of the subject. Metrics such as TST and total wake time (TWT) are accordingly estimated from these sleep stage estimates. Similar to portable monitors, creating further methods for data collection will improve the power of diagnoses. Activity monitors often fail to adequately detect wake; determining additional device components or channels for data collection can possibly bridge the gap.

Metrics/metrics extraction

The primary outcome variables generated for algorithm and epoch-based analysis, the AHI and RDI, determine severity of sleep disordered breathing from the patient's experiences. All studies should implement one or both of these metrics to ensure reputability of the device. Although actigraphy and non-contact sensors are not equipped with a lot of channels, it is recommended that relevant sleep indices be generated by any means possible. The most widely recognized method for AHI determination is manual scoring, where an operator determines the number of apnea and hypopnea episodes in a designated time frame. Methods of automated scoring are also widely popular among literature. According to the AASM recommendations, sleep recordings should be scored manually, and automatic scoring algorithms are not yet considered as a reference. Therefore, POC devices should be validated against manual scoring and should provide signals that could be scored manually, regardless of the reliability of their automated scoring. Automated scoring systems can be used wherever possible, only to assist manual scoring.

Aside from manual scoring, devices often generate metrics with differing methods of automated scoring. One particular paper noted the lack of standardization of scoring from actigraphy [49]. Certain studies covering portable monitors

did not even mention any methods of data analytics for arriving at the sleep patterns and indices [41, 60]. Some papers elaborated on the specific algorithms derived for assessment or even provided a flow chart [37, 53, 61], yet the majority of studies validating those devices simply listed the methods or left the readers to assume [39]. Many other studies perform epoch-by-epoch analyses and explain the process in a similar fashion [43, 45, 46, 62, 63]. While it may be acceptable if these widely implemented analyses are not elaborated upon, less popular approaches (e.g., the statistical analysis of snoring) require a much deeper, and rigorous investigation [22]. This practice of explaining analyses increases transparency and provides more reasoning for the resulting diagnosis. Homogeneity in the reporting of these analyses also allows for more comparison between devices.

Although not required by the standard guidelines, all POC device testing protocols are recommended to be tested on metrics such as TST, TWT, sleep efficacy (SE), wake after sleep onset (WASO), and rapid eye movement (REM). Providing such sleep figures confirm if the POC device is precisely accurate in detecting sleep and wake stages. Considering actigraphy's main weakness of differentiating silent wake from sleep, these said factors are particularly crucial. Also, a common source of error for many of these devices is the inability to distinguish quiet wakefulness from sleep. Since these limitations strongly affect TST and SE, algorithms involving such variables are recommended to be appropriately adjusted. On the other hand, algorithms which estimate AHI have shown to incorrectly categorize false positives from false negatives. Additional methods of event-by-event detection and estimation of event duration have proven to curtail such an issue [53].

Pertinently, our review (see Fig. 2 and Table AI) suggests that almost 50% of reviewed papers do not report metrics such as SE, RDI, and ODI. However, these metrics can be reported and validated in much of the same fashion as AHI and TST. While our rating system expects that all papers include AHI and TST as metrics for validation, our literature review shows that many highly cited papers also report SE, RDI, and ODI [, , 49, 56, 59, 62].

Gold standard

Polysomnography (PSG) is the undisputed gold standard to compare POC devices for diagnosing OSA. Although still prone to a number of sleep study-related concerns such as discomfort, night-to-night variability, and unconventional sleeping environments, PSG is commonly referred to as the gold standard due to the sheer reality that it produces the most comprehensive diagnoses. However, studies should list the respective PSG system or software being utilized. It must also be noted that a PSG gold standard is not required to validate POC devices that are of type III and type IV. As a consequence, metrics such as TST that cannot be gathered from non-PSG gold standards may not be used for validation. In fact, based on Fig. 2, it is evident that prior works have employed metrics that can be gathered from non-PSG gold standard devices for performance assessment. At the outset, a POC device posting comparable results to a well-known PSG system receives considerably more attention. Papers that explicitly state the name of the PSG device will be providing more transparency and context to the statistical analyses as a result.

Validation and comparison

Nearing the end of a clinical POC device validation paper come the quantitative and qualitative methods of comparison. These include statistical methods to display strong concordance as well as patient ratings to imply higher satisfaction. The rating system, shown in Table 3, makes recommendations in regard to both methods for validation.

Quantitative methods

Quantitative validation methods are taken to compare the strength of diagnoses of samples undergoing PSG and POC devices. For example, over 80% of works reported in the literature (see Table A1) use correlation coefficients and the Bland-Altman technique to display the degree of association and differences with both aforementioned. Sensitivity and specificity, used to show accuracy of sleep states, should be further implemented to validate quality in terms of true wake and sleep. While correlation can only suggest the strength of relation, these approaches to detecting true positives and negatives properly measure concordance [37]. The receiver operating characteristic (ROC) curve is also recommended to demonstrate viable evidence for distinguishing between true and false positives and finding the optimal AHI cutoffs [64]. ROC analysis can also be used to develop metrics that effectively compare diagnostic results of a subject group after they have been collected from multiple devices [64].

Studies also utilized a variety of statistical tests to compare samples of PSG and POC devices for OSA. Figure 3 displays which metrics were analyzed through statistical methods. Analysis of variance (ANOVA), single/paired t test, and chi-squared test were all commonly implemented to analyze the differences between samples. Other studies utilized tests such as Mann-Whitney U test, Shapiro-Wilk test, Kruskal-Wallis test, and analysis of covariance (ANCOVA). Through the use of more than one statistical approach, a paper removes bias or possibilities for misleading agreement [49]. Our ratings system therefore encourages that studies utilize more than one statistical test.

As previously mentioned, metrics like AHI and/or RDI are strongly recommended for POC device validation. Sleep indices hold the greatest weight for inter-device comparison. Quantitative methods of validation for these metrics are

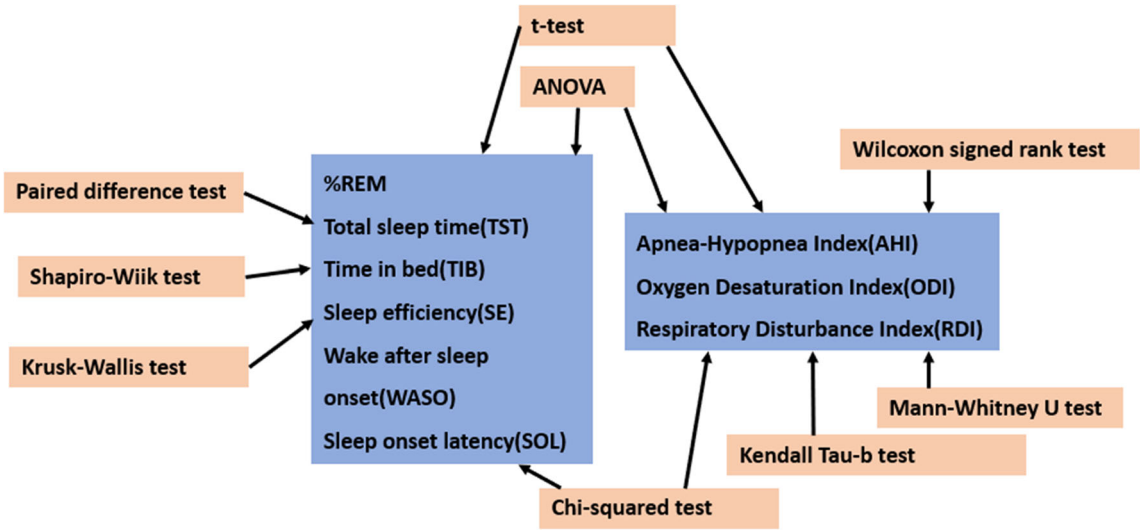

Fig. 3 Relationship between metrics and quantitative validation methods

therefore considered to be the utmost necessity. Ancillary metrics and features such as TST also require attention for means of validation, as shown in the figure. Additionally, reliability and agreement between a POC device and a gold standard (here, the OSA episode annotations from the POC device versus those from a gold standard) can be captured in terms of Cohen's kappa coefficients [65]. Kappa coefficients can provide an additional statistical quantification to validate the reliability of POC devices. The continued streamlined reporting of quantitative validation methods is paramount to the validity in results.

Qualitative methods

POC devices have been introduced as an alternative to PSG with the so-called merits of ease and affordability. While the benefits in cost are clear, measures of comfort require more validation. Hence, the use of post-sleep questionnaires comes into discussion. Since only one paper used such a feedback form to assess comfort of its device [66], pushing for such criteria will allow for more standardized, ergonomic comparison between different POC devices. These feedback forms could potentially try to survey the subject's overall experience with the device and accept possible complaints with the procedure. Our ratings system strongly encourages post-lab questionnaires. To improve transparency with the POC device's comfort, it is also recommended that sleep technicians record the number of adjustments taken for the subject's satisfaction. Although few studies focus on the ergonomics of the device, enhancing patient well-being during testing can further reduce the night-to-night variability [67]. It is therefore crucial that all studies identify and report the subject's experiences with the device.

Future implementation of the rating system

Table 3 outlines the main requirements when assessing the quality of a paper validating point-of-care (POC) devices for OSA. An appropriate rating system provides rationale for selecting one POC device over the other for reliable OSA diagnosis. Evaluation of each feature will be determined by a ratings system that is numbered 1 to 3. A rating of 1 is the minimum expectation for any paper, while a rating of 2 or 3 implies that a paper is taking necessary steps for homogeneity. These ratings, established after a comprehensive review of all literature, do not cater to the specific interests of certain studies. A paper relating to a POC device for OSA patients with dementia [], for instance, will only objectively be measured by the rating system in regard to the aforementioned seven groupings. From a high-level perspective, this rating system serves to reward papers that reduce variability of results and increase transparency and integrity of the study. It also encourages the standardization of future papers validating a POC device for OSA.

Discussion

Currently, there is a push toward using wearable technologies for sleep testing and diagnostics. Human subject testing of these devices is essential for translating these devices from laboratories into commercial practice. While there have been mature practices for testing sleep apnea devices, there have been significant non-uniformities as well as ad hoc-ism involving the current practice of testing wearable devices for sleep. The aforementioned narrative review outlines the lack of homogeneity, from study design to qualitative validation. This paper has distilled the various best practices from the

disparate publications featuring human subject testing for sleep. A protocol bridges the issue of non-uniformity by suggesting specific areas for standardization. By following the protocol effectively, devices would end up having stronger quality of diagnoses and studies would be designed more methodically.

Delineating the non-uniformities in literature while also providing recommendations is necessary to making improvements to the status quo. Through the course of review, statistical analyses implied to compare the performance of a wearable device to a gold standard have been shown to be widely heterogeneous. Even crucial aspects of metric extraction were performed in methods ranging from linear discriminant analysis to the statistical analysis of snoring, thereby making it harder to compare a certain data point between different papers.

This protocol brings together the key elements of previous, disparate works. It identifies different areas for improvement, thereby making a stronger case for implementation. It was designed into groups and subgroups to cohesively break down each aspect of a paper and suggest opportunities for growth. Altogether, this streamlined approach will improve the quality of future POC devices for OSA as well as provide efficient structure of future literature in the field.

Additionally, such a systematic protocol can benefit a device manufacturer with reporting consistent metrics toward complying with clearing FDA regulations. In order to be cleared for commercial distribution, the submitter must receive a letter finding the device to be "substantially equivalent" to another lawfully marketed device. Having a standardized protocol allows consistency in the validation of, and performance comparison among, POC devices. We also anticipate that this could lead to a higher worldwide acceptance of POC devices for OSA monitoring and diagnostics.

Author contributions Vivek Tangudu was the primary author of the paper. Kahkashan Afrin provided guidance to the primary author on the specific sources of literature and organization of the contents into tabular forms, and was responsible for detailing the contents of various sections in the manuscript. Sandy Reddy was responsible for conceptualizing the graphical representation as well as detailing the contents of the various sections in the manuscript. Nicolas Deutz and Steven Woltering contributed to the human subject testing and the review methodology. Satish Bukkapatnam is responsible for the organization, structure, and contents of the paper. All the authors were involved with the revision and editing process.

Funding This study was funded by NSF-PFI-AIR-TT 1543226.

Compliance with ethical standards

Disclosure All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Engleman H, Douglas N (2004) Sleep 4: sleepiness, cognitive function, and quality of life in obstructive sleep apnoea/hypopnoea syndrome. Thorax 59(7):618–622
2. Giles TL et al (2006) Continuous positive airways pressure for obstructive sleep apnoea in adults. Cochrane Database Syst Rev 1
3. Terán-Santos J, Jimenez-Gomez A, Cordero-Guevara J (1999) The association between sleep apnea and the risk of traffic accidents. N Engl J Med 340(11):847–851
4. Lee W, Nagubadi S, Kryger MH, Mokhlesi B (2008) Epidemiology of obstructive sleep apnea: a population-based perspective. Expert Rev Respir Med 2(3):349–364
5. Kapur V, Blough DK, Sandblom RE, Hert R, de Maine JB, Sullivan SD, Psaty BM (1999) The medical cost of undiagnosed sleep apnea. Sleep 22(6):749–755
6. Partinen M, Palomaki H (1985) Snoring and cerebral infarction. Lancet 326(8468):1325–1326
7. Koskenvuo M, Kaprio J, Telakivi T, Partinen M, Heikkila K, Sarna S (1987) Snoring as a risk factor for ischaemic heart disease and stroke in men. Br Med J (Clin Res Ed) 294(6563):16–19
8. Hung J, Whitford EG, Hillman DR, Parsons RW (1990) Association of sleep apnoea with myocardial infarction in men. Lancet 336(8710):261–264
9. Kripke DF, Ancoli-Israel S, Klauber MR, Wingard DL, Mason WJ, Mullaney DJ (1997) Prevalence of sleep-disordered breathing in ages 40-64 years: a population-based survey. Sleep 20(1):65–76
10. Duran J et al (2001) Obstructive sleep apnea–hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 Yr. Am J Respir Crit Care Med 163(3):685–689
11. Young T et al (2002) Predictors of sleep-disordered breathing in community-dwelling adults: the Sleep Heart Health Study. Arch Intern Med 162(8):893–900
12. Young T, Skatrud J, Peppard PE (2004) Risk factors for obstructive sleep apnea in adults. Jama 291(16):2013–2016
13. Mazzocchi S How much does a sleep study cost? Well, $600 or $5, 070. 2013; Available from: https://clearhealthcosts.com/blog/2013/04/how-much-does-a-sleep-study-cost-well-600-or-5070/. Accessed April 2020
14. Deutsch PA, Simmons MS, Wallace JM (2006) Cost-effectiveness of split-night polysomnography and home studies in the evaluation of obstructive sleep apnea syndrome. J Clin Sleep Med 2(2):145–153
15. Ahmadi N, Shapiro GK, Chung SA, Shapiro CM (2009) Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography. Sleep Breath 13(3):221–226
16. Flemons WW, Douglas NJ, Kuna ST, Rodenstein DO, Wheatley J (2004) Access to diagnosis and treatment of patients with suspected sleep apnea. Am J Respir Crit Care Med 169(6):668–672
17. Collop NA (2002) Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 3(1):43–47
18. White DP, Gibb TJ (1998) Evaluation of a computerized polysomnographic system. Sleep 21:188–196
19. Newell J, Mairesse O, Verbanck P, Neu D (2012) Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among

different clinical population samples. Psychiatry Res 200(2):795–801
20. Mosko SS, Dickel MJ, Ashurst J (1988) Night-to-night variability in sleep apnea and sleep-related periodic leg movements in the elderly. Sleep 11(4):340–348
21. Chesson AL Jr, Berry RB, Pack A (2003) Practice parameters for the use of portable monitoring devices in the investigation of suspected obstructive sleep apnea in adults. Sleep 26(7):907–913
22. Su S, Baroody FM, Kohrman M, Suskind D (2004) A comparison of polysomnography and a portable home sleep study in the diagnosis of obstructive sleep apnea syndrome. Otolaryngol Head Neck Surg 131(6):844–850
23. Masa JF, Corral J, Pereira R, Duran-Cantolla J, Cabello M, Hernandez-Blasco L, Monasterio C, Alonso A, Chiner E, Rubio M, Garcia-Ledesma E, Cacelo L, Carpizo R, Sacristan L, Salord N, Carrera M, Sancho-Chust JN, Embid C, Vazquez-Polo FJ, Negrin MA, Montserrat JM (2011) Effectiveness of home respiratory polygraphy for the diagnosis of sleep apnoea and hypopnoea syndrome. Thorax 66(7):567–573
24. Kim RD, Kapur VK, Redline-Bruch J, Rueschman M, Auckley DH, Benca RM, Foldvary-Schafer NR, Iber C, Zee PC, Rosen CL, Redline S, Ramsey SD (2015) An economic evaluation of home versus laboratory-based diagnosis of obstructive sleep apnea. Sleep 38(7):1027–1037
25. Berry RB et al (2008) Portable monitoring and autotitration versus polysomnography for the diagnosis and treatment of sleep apnea. Sleep 31(10):1423–1431
26. Victor LD (1999) Obstructive sleep apnea. Am Fam Physician 60(8):2279–2286
27. Whittle AT, Finch SP, Mortimore IL, MacKay TW, Douglas NJ (1997) Use of home sleep studies for diagnosis of the sleep apnoea/hypopnoea syndrome. Thorax 52(12):1068–1073
28. Price CP, Kricka LJ (2007) Improving healthcare accessibility through point-of-care technologies. Clin Chem 53(9):1665–1675
29. Anderson DA, Crowe SM, Garcia M (2011) Point-of-care testing. Current HIV/aids Reports 8(1):31–37
30. Ventola CL (2014) Mobile devices and apps for health care professionals: uses and benefits. Pharmacy and Therapeutics 39(5):356–364
31. EarlySense. Cited 2018 June 15; Available from: https://www.earlysense.com/. Accessed April 2020
32. VitalConnect. Cited 2018 June 15; Available from: https://vitalconnect.com/solutions/vitalpatch/. Accessed April 2020
33. Itamar-Medical . Cited 2018 June 15; Available from: http://www.itamar-medical.com/watchpat-main/. Accessed April 2020
34. Stoohs R, Guilleminault C (1992) MESAM 4: an ambulatory device for the detection of patients at risk for obstructive sleep apnea syndrome (OSAS). Chest 101(5):1221–1227
35. Ancoli-Israel S, Clopton P, Klauber MR, Fell R, Mason W (1997) Use of wrist activity for monitoring sleep/wake in demented nursing-home patients. Sleep 20(1):24–27
36. Paquet J, Kawinska A, Carrier J (2007) Wake detection capacity of actigraphy during sleep. Sleep 30(10):1362–1369
37. de Souza L, Benedito-Silva AA, Pires MLN, Poyares D, Tufik S, Calil HM (2003) Further validation of actigraphy for sleep studies. Sleep 26(1):81–85
38. Beniczky S, Ryvlin P (2018) Standards for testing and clinical validation of seizure detection devices. Epilepsia 59(Suppl 1):9–13
39. Van de Water AT, Holmes A, Hurley DA (2011) Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—a systematic review. J Sleep Res 20(1 Pt 2):183–200
40. Kushida CA, Littner MR, Morgenthaler T, Alessi CA, Bailey D, Coleman J Jr, Friedman L, Hirshkowitz M, Kapen S, Kramer M, Lee-Chiong T, Loube DL, Owens J, Pancer JP, Wise M (2005) Practice parameters for the indications for polysomnography and related procedures: an update for 2005. Sleep 28(4):499–523
41. Driver HS, Pereira EJ, Bjerring K, Toop F, Stewart SC, Munt PW, Fitzpatrick MF (2011) Validation of the MediByte(®) type 3 portable monitor compared with polysomnography for screening of obstructive sleep apnea. Can Respir J 18(3):137–143
42. Adult Obstructive Sleep Apnea Task Force of the American Academy of Sleep, M (2009) Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J Clin Sleep Med **5**(3):263–276
43. Tal A, Shinar Z, Shaki D, Codish S, Goldbart A (2017) Validation of contact-free sleep monitoring device with comparison to polysomnography. J Clin Sleep Med 13(3):517–522
44. O'Driscoll DM et al (2013) Energy expenditure in obstructive sleep apnea: validation of a multiple physiological sensor for determination of sleep and wake. Sleep Breath 17(1):139–146
45. Kanady JC, Drummond SPA, Mednick SC (2011) Actigraphic assessment of a polysomnographic-recorded nap: a validation study. Journal of Sleep Research **20**(1pt2):214–222
46. Shambroom JR, Fabregas SE, Johnstone J (2012) Validation of an automated wireless system to monitor sleep in healthy adults. J Sleep Res 21(2):221–230
47. de Zambotti M, Baker FC, Colrain IM (2015) Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep 38(9):1461–1468
48. Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD (2001) Sleep detection with an accelerometer actigraph: comparisons with polysomnography. Physiol Behav 72(1):21–28
49. Meltzer LJ et al (2012) Direct comparison of two new actigraphs and polysomnography in children and adolescents. Sleep 35(1):159–166
50. Meng L, Xu H, Guan J, Yi H, Wu H, Yin S (2016) Validation of a novel sleep-monitoring system for diagnosing obstructive sleep apnea: a comparison with polysomnography. Exp Ther Med 12(5):2937–2941
51. Zaffaroni A, Kent B, O'Hare E, Heneghan C, Boyle P, O'Connell G, Pallin M, de Chazal P, McNicholas WT (2013) Assessment of sleep-disordered breathing using a non-contact bio-motion sensor. J Sleep Res 22(2):231–236
52. Zhou Y, Shu D, Xu H, Qiu Y, Zhou P, Ruan W, Qin G, Jin J, Zhu H, Ying K, Zhang W, Chen E (2020) Validation of novel automatic ultra-wideband radar for sleep apnea detection. J Thorac Dis 12(4):1286–1295
53. Saha S, Kabir M, Montazeri Ghahjaverestan N, Hafezi M, Gavrilovic B, Zhu K, Alshaer H, Yadollahi A (2020) Portable diagnosis of sleep apnea with the validation of individual event detection. Sleep Med 69:51–57
54. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31(4):337–350
55. Lwanga SK, Lemeshow S, W.H. Organization (1991) Sample size determination in health studies: a practical manual
56. Ayappa I et al (2008) Validation of a self-applied unattended monitor for sleep disordered breathing. J Clin Sleep Med 4(1):26–37
57. Barger LK, Rajaratnam SMW, Wang W, O'Brien CS, Sullivan JP, Qadri S, Lockley SW, Czeisler CA, for the Harvard Work Hours, Health and Safety Group (2015) Common sleep disorders increase risk of motor vehicle crashes and adverse health outcomes in firefighters. J Clin Sleep Med 11(03):233–240
58. Xu L, Han F, Keenan BT, Kneeland-Szanto E, Yan H, Dong X, Chang Y, Zhao L, Zhang X, Li J, Pack AI, Kuna ST (2017) Validation of the Nox-T3 portable monitor for diagnosis of obstructive sleep apnea in Chinese adults. J Clin Sleep Med 13(5):675–683
59. Chen H, Lowe AA, Bai Y, Hamilton P, Fleetham JA, Almeida FR (2009) Evaluation of a portable recording device (ApneaLink) for

case selection of obstructive sleep apnea. Sleep Breath 13(3):213–219
60. Stehling F, Keull J, Olivier M, Große-Onnebrink J, Mellies U, Stuck BA (2017) Validation of the screening tool ApneaLink® in comparison to polysomnography for the diagnosis of sleep-disordered breathing in children and adolescents. Sleep Med 37: 13–18
61. Le TQ et al (2013) Wireless wearable multisensory suite and real-time prediction of obstructive sleep apnea episodes. IEEE J Transl Eng Health Med 1:2700109–2700109
62. Zou D, Grote L, Peker Y, Lindblad U, Hedner J (2006) Validation a portable monitoring device for sleep apnea diagnosis in a population based cohort using synchronized home polysomnography. Sleep 29(3):367–374
63. Heneghan C et al (2008) A portable automated assessment tool for sleep apnea using a combined Holter-Oximeter. Sleep 31(10): 1432–1439
64. Hajian-Tilaki K (2013) Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 4(2):627–635
65. Guggenmoos-Holzmann I (1996) The meaning of kappa: probabilistic concepts of reliability and validity revisited. J Clin Epidemiol 49(7):775–782
66. Cairns A, Wickwire E, Schaefer E, Nyanjom D (2014) A pilot validation study for the NOX T3TM portable monitor for the detection of OSA. Sleep Breath 18(3):609–614
67. Aloia MS, Arnedt JT, Stanchina M, Millman RP (2007) How early in treatment is PAP adherence established? Revisiting night-to-night variability. Behav Sleep Med 5(3):229–240
68. Zucconi M, Ferini-Strambi L, Castronovo V, Oldani A, Smirne S (1996) An unattended device for sleep-related breathing disorders: validation study in suspected obstructive sleep apnoea syndrome. Eur Respir J 9(6):1251–1256
69. To KW et al (2009) Validation study of a portable monitoring device for identifying OSA in a symptomatic patient population. Respirology 14(2):270–275
70. Garde A, Dehkordi P, Karlen W, Wensley D, Ansermino JM, Dumont GA (2014) Development of a screening tool for sleep disordered breathing in children using the phone Oximeter. PLoS One 9(11):e112959
71. O'Driscoll DM, Turton AR, Copland JM, Strauss BJ, Hamilton GS (2013) Energy expenditure in obstructive sleep apnea: validation of a multiple physiological sensor for determination of sleep and wake. Sleep Breath 17(1):139–146
72. Wongdhamma W, Le TQ, Bukkapatnam STS (2013) Wireless wearable multi-sensory system for monitoring of sleep apnea and other cardiorespiratory disorders. in 2013 IEEE International Conference on Automation Science and Engineering (CASE).
73. Le TQ, et al. (2013) Prediction of sleep apnea episodes from a wireless wearable multisensor suite. In 2013 IEEE Point-of-Care Healthcare Technologies (PHT). IEEE
74. Bukkapatnam ST, Le T, Wongdhamma W, Device and method for predicting and preventing obstructive sleep apnea (osa) episodes. 2014, Google Patents
75. Xu H, Zhao X, Shi Y, Li X, Qian Y, Zou J, Yi H, Huang H, Guan J, Yin S (2019) Development and validation of a simple-to-use clinical nomogram for predicting obstructive sleep apnea. BMC Pulm Med 19(1):18
76. Lu M, Fang F, Sanderson JE, Ma C, Wang Q, Zhan X, Xie F, Xiao L, Liu H, Liu H, Wei Y (2019) Validation of a portable monitoring device for the diagnosis of obstructive sleep apnea: electrocardiogram-based cardiopulmonary coupling. Sleep Breath 23(4):1371–1378
77. Fassbender P et al (2019) Validation of a photoplethysmography device for detection of obstructive sleep apnea in the perioperative setting. J Clin Monit Comput 33(2):341–345