Green, S. F., Frame, T., Banerjee , L. V., Gibson, A., Blackman, J., Morrison, H. D., Lloyd, K. E., Rudd, S., Fotherby, W. G. F., Bartsch, U., Purcell, S., Jones, M. W., & Coulthard, E. J. (2022). A systematic review of the validity of non-invasive sleep-measuring devices in mid-to-late life adults: Future utility for Alzheimer's disease research. *Sleep Medicine Reviews*, *65*, Article 101665. Advance online publication. https://doi.org/10.1016/j.smrv.2022.101665

A systematic review of the validity of non-invasive sleep-measuring devices in mid-to-late life adults: Future utility for Alzheimer's disease research

Sebastian Francis Green, Tory Frame, Luke Vikram Banerjee, Amy Gimson, Jonathan Blackman, Hamish Morrison, Katie Lloyd, Sarah Rudd, William George Frederick Fotherby, Ullrich Bartsch, Shaun Purcell, Matt Jones, Liz Coulthard

PII: S1087-0792(22)00078-8

DOI: https://doi.org/10.1016/j.smrv.2022.101665

Reference: YSMRV 101665

To appear in: *Sleep Medicine Reviews*

Received Date: 14 January 2022

Revised Date: 21 June 2022

Accepted Date: 23 June 2022

Please cite this article as: Green SF, Frame T, Banerjee LV, Gimson A, Blackman J, Morrison H, Lloyd K, Rudd S, Frederick Fotherby WG, Bartsch U, Purcell S, Jones M, Coulthard L, A systematic review of the validity of non-invasive sleep-measuring devices in mid-to-late life adults: Future utility for Alzheimer's disease research, *Sleep Medicine Reviews* (2022), doi: https://doi.org/10.1016/j.smrv.2022.101665.

Full title:

A systematic review of the validity of non-invasive sleep-measuring devices in mid-to-late life adults: Future utility for Alzheimer's disease research

Running head title:

Sleep-measuring devices and Alzheimer's disease

Authors:

Sebastian Francis Green[1,2], Tory Frame[3], Luke Vikram Banerjee[1], Amy Gimson[2], Jonathan Blackman[1,2], Hamish Morrison[1,2], Katie Lloyd[1,2], Sarah Rudd[4], William George Frederick Fotherby[5], Ullrich Bartsch[6], Shaun Purcell[7], Matt Jones[8], Liz Coulthard[1,2].

[1]Bristol Medical School, University of Bristol, Bristol, UK

[2]Neurology Department, Southmead Hospital, North Bristol NHS Trust, Bristol, UK

[3]Department of Computer Science, University of Bath, Bath, UK

[4]Library and Knowledge Service, Southmead Hospital, North Bristol NHS Trust, Bristol, UK

[5]Manchester Metropolitan University, Manchester, UK

[6]UK DRI Care Research & Technology at the University of Surrey & Surrey Sleep Research Centre, University of Surrey, UK

[7]Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Harvard University, USA

[8]School of Physiology, Pharmacology & Neuroscience, University of Bristol, Bristol, UK

Corresponding authors:

Dr Liz Coulthard

Associate Professor in Dementia Neurology &

Dr Sebastian Green

Academic Clinical Fellow in Neurology

Bristol Medical School

University of Bristol

BS2 8DZ

United Kingdom

Email: Elizabeth.coulthard@bristol.ac.uk & sg1743@bristol.ac.uk

Telephone: +44 (0) 117 414 7801

Conflicts of interest: None declared

Acknowledgements: None

Summary

Changes in sleep during mid-to-late life are associated with risk for Alzheimer's disease (AD). Mechanistic understanding of this association necessitates measurement tools able to quantify these sleep changes longitudinally and accurately. We conducted a systematic review with meta-analysis of validity studies of non-invasive sleep-measuring devices published since 2015 that record sleep metrics associated with AD in adults over 40 (mean 52.9, SD 6.1 years). We reviewed 52 studies, including 32 wearable and ten non-wearable single or multi-sensor devices validated against polysomnography (minimum one night). The apnoea hypopnoea index and oxygen desaturation index were accurately measured across devices. Total sleep time and sleep efficiency were significantly overestimated (p<0.001) by mean 33.2 minutes and 7.6%, respectively. Slow wave sleep duration was inaccurately measured except by a headband device with electroencephalography. There was no significance difference in accuracy between participants with and without sleep disorders. Studies were undermined by high risk of bias from closed-access algorithms and classification thresholds, and incomplete reporting of accuracy data. Only one study investigated slow wave activity, and none investigated sleep spindles. Nonetheless, we have identified devices that could be used in future studies of sleep and AD risk and discuss some of the limitations of available research.

Keywords

Sleep; Alzheimer's; NREM; REM; slow wave; apnoea; hypopnoea

Abbreviations

AD = Alzheimer's disease

AHI = Apnoea hypopnoea index

AUC = Area under curve

CC = Correlation coefficient

CHF = Congestive heart failure

CI = Confidence interval

COPD = Chronic obstructive pulmonary disease

ECG = Electrocardiography

EEG = Electroencephalography

ICC = Intraclass correlation

NISMD = Non-invasive sleep measuring device

NPV = Negative predictive value

NREM = Non rapid eye movement

OSA = Obstructive sleep apnoea

ODI = Oxygen desaturation index

PPV = Positive predictive value

PSG = Polysomnography

R = Range

RR = Reference range

REM = Rapid eye movement

SD = Sleep disorders

SDB = Sleep disordered breathing

SE = Sleep efficiency

SEr = Standard error

SWA = Slow wave activity

SWS = Slow wave sleep

QADAS-2 = Quality assessment of diagnostic accuracy studies-2

TST = Total sleep time

Introduction

The global prevalence of Alzheimer's disease (AD) is increasing rapidly [1], precipitating an enormous health and socioeconomic burden and making AD prevention critical. Sleep disturbances are emerging as important risk factors for AD. Cognitively healthy people with subjective or objectively measured sleep disturbance in mid [2, 3] and later life [3-9] have increased risk of cognitive decline and AD. Sleep disturbance is also associated with neuropathological changes of AD, including beta-amyloid and tau deposition, that gradually accrue over many years prior to noticeable cognitive symptoms [10, 11]. Given the contributions of sleep to both memory consolidation and AD pathogenesis [12, 13], changes in sleep may be important identifiers of early functional consequences of AD pathology and/or future AD risk, stratifying individuals for targeted disease prevention.

Changes in sleep macroarchitecture, microarchitecture and sleep disordered breathing (SDB) have been associated with AD [10, 14]. Commonly associated macroarchitectural changes include reduced total sleep time (TST) [2, 7], reduced sleep quality or sleep efficiency (SE) [7, 8, 15-17], disruption or reduction of slow wave sleep (SWS) [18, 19], and reduced rapid eye movement (REM) duration [5]. Associated microarchitectural changes include reduced sleep spindle duration, density and count [20], slow oscillation/sleep spindle coupling [21], and proportion of 0.6-1Hz [17, 21, 22] or 1-4Hz slow wave activity (SWA) [23, 24]. There is also a strong association with obstructive sleep apnoea (OSA) as measured by the apnoea hypopnoea index (AHI) or oxygen desaturation index (ODI) [10, 25].

To detect longitudinal changes in these sleep metrics on population levels for AD prevention research, objective, accurate, and non-invasive measurement tools are needed. Such measurement tools could elaborate on the nature of interrelationships between sleep disturbance and AD risk, such as when in the lifespan these sleep changes occur, whether they are causal, how they respond to intervention, how they associate with cognitive impairments, and how they compare with normal ageing. The gold-standard technique for measuring sleep is type one (fully attended in a laboratory setting) polysomnography (PSG). However, PSG requires multiple complex sensors, specialist setup and interpretation, and is recorded with participants sleeping in an unfamiliar setting. Considering the additional associated cost and time burdens, PSG is not an appropriate tool to capture longitudinal, non-invasive, and natural sleep data for the purpose of early detection and AD prevention studies.

The last few decades have seen an explosion in the interest and development of both commercial and research-grade non-invasive sleep-measuring devices (NISMDs) [26, 27], which could potentially be used as measurement tools for epidemiological research in sleep disturbance and AD risk. However, the usefulness of these devices depends on their capacity to accurately measure the features of sleep that change in people with AD neuropathology or who are at increased risk of incident clinical AD. We therefore conducted a systematic review of validity studies of NISMDs, which specifically measure sleep metrics most strongly associated with neuropathological AD or the incidence of clinical AD. We reviewed accuracy studies of NISMDs on healthy or comorbid (e.g., raised BMI, hypertensive, smokers) adults over 40; these adults represent cohorts at-risk for AD and constitute the mainstay of AD prevention research [28]. Through this objective, we aimed to determine a) which of the index sleep metrics have

been measured with available NISMDs in this population; b) how accurately NISMDs can record these metrics compared to type one PSG; c) highlight some of the current best NISMDs for recording each metric; d) discuss some of the limitations of research methodology in this field and of available device hardware and software. Considering the rapid pace of NISMD development, including the use of newer algorithms and multi-sensor devices, we limited our search to studies published since 2015. We hope our work will inform and improve future studies investigating the complex relationship between sleep disturbance and AD risk.

Methods

Choice of sleep metrics

We conducted an initial scoping search using Web of Science on 17th July 2021 using terms for "sleep" AND ("amyloid" OR "tau"; Table S1). We then extracted a list of studies describing changes in sleep associated with AD-like changes in beta-amyloid or tau on neuroimaging or in serum or cerebrospinal fluid. To keep this systematic review focussed and specific to AD, we used this list in collaboration with available epidemiological data of incident clinical AD risk [2, 5-9] to prioritise the most reported and/or most strongly associated objectively measured sleep metrics. We selected total sleep time (TST), sleep efficiency (SE), slow wave sleep (SWS) duration (NREM stage three), rapid eye movement (REM) duration, apnoea hypopnoea index (AHI), oxygen desaturation index (ODI), SWA and sleep spindles. These metrics are not exhaustive of all sleep metrics associated with AD.

Study design

This was a systematic review with meta-analysis of validity (diagnostic accuracy) studies. Our study conforms to the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy studies checklist [29] and is registered on PROSPERO (ID CRD42021249738).

Search

The search strategy, devised in collaboration with a medical librarian, used keyword and thesaurus terms relevant to the reference test (PSG), the index test (NISMD), and the target condition ("sleep"; Figure S1). The search was limited to humans and English language documents. Searches were conducted on Medline (Ovid), Embase (Ovid), Emcare (Ovid), PsycINFO (ProQuest) and CINAHL (Ebsco) (all from 1st January 2015 to 14th July 2021). Results were deduplicated automatically using Endnote X9 (Clarivate Analytics) and then manually using Rayyan [30].

Screening and selection criteria

Screening was conducted against our selection criteria (full details in Figure S2). Minimum sample size was 10 and minimum mean sample age was 40 years. We excluded studies that did not compare the device against attended level 1 PSG in a sleep laboratory or hospital. Devices were deemed 'non-invasive' where usable or used outside of a laboratory setting and not

requiring third-party specialist set-up. Considering that a range of physical and mental health comorbidities increase risk of AD [31], we included people with comorbidities. We excluded studies in acute hospital settings.

Records were equally distributed and independently screened by eight authors (SG, AG, JB, HM, KL, WF, LB, TF). Agreement on the first 10% of titles and abstracts was compared against decisions separately made by an additional second author (SG or HM). All disagreements were discussed in person and selection criteria were revised before reviewers completed the remaining 90%. A third author (JB) was involved where disagreements were not resolved. For the full text screening stage, each record was screened separately by two authors (from SG, JB, HM, KL, AF, TF), and a third author (SG or HM) was involved where there was an unresolved disagreement. Reviewers were blinded at each stage of screening.

Data collection and analysis

Manuscripts were equally distributed across three authors (SG, TF, LB) who independently tabulated accuracy data. The data extracted by each author was independently cross checked by one of the two other authors. The data recorded varied by device and sleep metric, and included accuracy, sensitivity, specificity, area under curve (AUC) from the receiver operating characteristic curve, positive predictive value (PPV), negative predictive value (NPV), mean difference including Bland-Altman plot bias and limits of agreements, intra-class correlations (ICC), correlation coefficients (CC), and more. Where possible, we derived additional accuracy metrics from available data (e.g., raw standard deviation using Bland-Altman mean bias and

limits of agreements). Data were extracted at the device unit level. For each metric, we have highlighted the most accurate devices, determined according to the mean difference relative to standard error, the number of participants that fell within one standard deviation of the mean and/or ICC, and/or sensitivity and specificity where these data were available. We conducted meta-analysis to provide summary measures for each sleep metric, before doing subgroup analysis. Subgroup analyses were grouped by device sensor and device type. It is important to include people with the range of sleep disorders that are encountered in dementia clinical practice, and we wanted to demonstrate whether results were applicable to patients with and without diagnosed sleep disorders. Therefore, we also carried out subgroup analysis in those with versus those without diagnosed sleep disorders. Group differences were analysed using unpaired *t* test or one-way ANOVA. These analyses grouped mean differences (device minus PSG) for a given sleep metric. We chose random effects models for meta-analysis due to the heterogeneity in cohort characteristics between these studies. Where several scorers or approaches were used to analyse the same dataset to produce accuracy estimates, we used the most accurate estimates for meta-analysis. We also ran one or two-way ANOVAs to investigate the differences in device performance by sensor type and the relationship between AHI severity and sensitivity or specificity. We used Microsoft Excel (Microsoft Corporation), SPSS (IBM Corporation) and GraphPad Prism (Graphpad Software) for data analysis and figure production.

In addition to accuracy data, we recorded data on study information (author, year, setting), sample characteristics (sample size, age, sex, disease group or healthy, BMI), study design (number of nights' sleep, PSG specification and scoring criteria), NISMD characteristics (device type, name, and location worn if wearable), algorithm used (proprietary, open access,

classification thresholds), and study sponsor. Authors were contacted by email where further data were sought.

Bias assessment was done (by SG, LB, AG) using the QUADAS-2 tool [32]. This tool provides a framework to judge the risk of bias in diagnostic accuracy studies according to a series of generic and study-specific signalling questions. This framework assesses four domains: patient selection, index test (NISMD), reference standard (PSG), and flow and timing (Figure S3). Three authors (SG, LB, AG) separately used the tool to assess the same first 10 studies with overall agreement (bias assessment across domains) of >90%.

Results

Study and device characteristics

We identified 52 studies (Figure 1, Table 1) of 42 NISMDs (Figure 2). We included ten non-wearable and 32 wearable devices. Sample size ranged from 16 to 500 (mean = 82.2), participant mean age from 42 to 69 years (summary mean = 52.9, Figure 3), mean BMI from 25 to 41 kg/m^2 (summary mean 29.7) and proportion of female participants from 15 to 100% (mean = 38.8%). Heterogeneity of these characteristics is displayed in Figure S4. The numbers of studies recording each sleep metric were as follows: AHI score or equivalent (n = 36), TST (n = 23), SE (n = 12), SWS (n = 8), REM (n = 7) and ODI (n = 5, Tables S2 and S3). Only one study reported the accuracy of a NISMD for measuring SWA [33]. No devices reported accuracy for measuring sleep spindles. Two studies presented device accuracy for sleep staging using a device with EEG

sensors which measured SWA and sleep spindle activity [33, 34]. Only two studies measured device accuracy across more than one night [35, 36]. Only six studies used algorithms that were open access [37-42]. We did not identify any studies including participants with mild cognitive impairment or AD.

We assessed whether there was a correlation between age or BMI and accuracy using correlation coefficient values derived from weighted mean differences. There was no significant correlation between the accuracy for recording each sleep metric and age or BMI. Next, we compared the grouped mean difference in sleep metrics (vs PSG) between those with and without diagnosed sleep disorders (Figure S5). There was no significant difference between these groups for AHI (t_{df} = -0.39$_{26}$, p = 0.703), ODI (t_{df} = 0.61$_{2}$, p = 0.605), TST (t_{df} = -0.11$_{10}$, p = 0.914), SE (t_{df} = 0.39$_{11}$, p = 0.420), REM (t_{df} = 0.33$_{3}$, p = 0.763), or SWS duration (t_{df} = -0.22$_{4}$, p = 0.836).

AHI scores and OSA detection

36 studies investigated the accuracy of NISMDs in measuring the AHI score or detecting mild (AHI 5-15), moderate (AHI 15-30), or severe (AHI >30) OSA (Table S2). These included eight non-wearable and 18 wearable devices, most commonly WatchPAT (n = 8), and the most common device sensors per study were actigraphy (n = 13), acoustic sensor (n = 14), oximetry (n = 19), plethysmography (n = 14), airflow (n = 8) and pressure sensors (n = 12). The mean age range of participants was 43 to 68 years and BMI range was 25.0 to 40.8 kg/m^2. 32 studies used cohorts where the majority had suspected OSA or SDB. Some studies compared equivalent or alternative measures to AHI scores from PSG including the respiratory event index [43-46], the

effective AHI [47], and the ODI [48]. Two studies did not publish mean differences [49, 50]. Five studies published Bland Altman plots but did not display mean difference values or a measure of variance [51-55].

We ran meta-analysis across all studies that presented mean difference data to produce a summary estimate (Figure 4) and ranked all devices by mean difference (Figure S6). Across devices, measurement of AHI did not significantly differ from zero ($Z = 1.872$; $p = 0.061$). When grouped by sensor type, we found that devices that used an airflow sensor significantly underestimated AHI ($p = 0.006$, Figure 5). Devices that used other sensors did not significantly over or underestimate AHI. There was no significant difference between single and multi-sensor devices for AHI mean difference versus PSG ($t_{df} = 0.468_{26}$, $p = 0.322$).

Next, we used meta-analysis to produce a summary estimate for AHI mean difference using the WatchPAT device (Figure 6). WatchPAT also did not significantly differ from zero (mean difference [95% CI] 1.56 [-0.44 to 3.56], $p = 0.13$). We note one study [56] substantially overestimated AHI compared to the other WatchPAT studies [36, 47, 57-60]. The reason for this difference is unclear, since this study used the same algorithm and had comparable cohort age, sex, and sample size, but may relate to a difference in BMI, which authors of this study did not specify.

No algorithm description (including classification thresholds) was specified for the device with the smallest mean difference in AHI score (unnamed multisystem setup [46]), which combined a finger oxygen saturation sensor with a bedsheet sensor incorporating multiple pressure sensors.

The next most accurate (by mean difference) devices were WatchPAT and BresoDx [42]. BresoDx uses breath sounds from a microphone. After normalizing and removing outliers from raw breath sounds, the BresoDx algorithm differentiates apnoea and hypopnoea from non-pathological variations in ventilation (e.g., snoring; Table 1). This pre-validated approach underestimated AHI by only -0.6 events per hour. The most accurate data for the WatchPAT device came from a study [58] which used the automatic, proprietary zzzPAT 4.3.62 algorithm. WatchPAT uses finger plethysmography to measure peripheral arterial volume changes associated with the increased sympathetic drive by apnoeas or hypopnoeas. This signal is combined with heart rate and oximetry to detect respiratory events according to thresholds of vasoconstriction or oxygen desaturation. These events are divided by sleep time which is differentiated from wake according to a background movement threshold and signal periodicity measured using actigraphy.

Next, we investigated the accuracy of NISMDs in diagnosing OSA. 23 studies reported sensitivity and specificity for diagnosing mild OSA, 27 for moderate OSA, and 21 for severe OSA. With increasing AHI severity, sensitivity significantly reduced ($p = 0.01$) but specificity significantly increased ($p = 0.04$; Figures 7 and S7). One group displayed a perfect score (sensitivity and specificity of 100%) for the detection of both mild and severe OSA in 27 participants with suspected SDB using the MATRx Plus device [61]. This device uses a finger pulse oximeter, a chest-mounted accelerometer, and an abdominal belt with respiratory inductance plethysmography. Authors used a proprietary algorithm, which identifies respiratory events by changes in minute ventilation. Best accuracy scores were derived when an ODI threshold of 3% was applied using the preceding peak method. The preceding peak method

identified respiratory events using the most recent oxygen saturation peak as a reference, rather than using a reference determined by a centred moving window applied to the top 25% of oxygen saturation recordings. The SD102 device was also a high performer with an accuracy of 95% or higher for mild, moderate, and severe OSA in 189 patients referred to a sleep clinic for various reasons [62]. The SD102 is a bedsheet device that contains 99 pressure sensors arranged at 40mm intervals. Changes in pressure are converted into waveforms to analyse the associated breathing pattern.

Oxygen desaturation index

Five studies of oximetry-based devices compared 3% ODI score between a NISMD and PSG, four of which presented mean differences with a measure of variance [44, 57, 60, 63]. Meta-analysis of these four studies showed that, across devices, ODI was overestimated by only 1.26 desaturations per hour (95% CI -1.9 to 4.5). This summary measure did not significantly differ from zero (Figure 8). One study found that the SleepView device significantly underestimated ODI ($p < 0.001$) [55]. One study also compared 3% ODI from a finger oximeter against AHI from PSG [48], as reported in the section above.

Macroarchitectural sleep metrics

Total sleep time

23 studies described accuracy for TST. These included four non-wearable and 20 wearable devices, most commonly WatchPAT (n = 3) and Actiwatch (n = 4) models, and the most common device sensors per study were actigraphy (n = 19), acoustic sensors (n = 7) and plethysmography (n = 8). Three studies did not publish mean differences [44, 59, 64]. Of the 25 devices that did, 84% (21/25) overestimated TST (Figure S8). The summary estimate produced by meta-analysis showed that devices significantly overestimated TST by mean (95% confidence interval) 33.2 (17.6 to 48.8) minutes compared to PSG ($Z = 4.182$, $p < 0.001$, Figure 9). Next, we compared the mean difference in TST between devices that did and did not use actigraphy for TST mean difference and there was no significant difference between these groups either ($t_{df} = 0.335_{24}$, $p = 0.370$). Then, we compared devices that used actigraphy alone against devices that combined additional sensor(s) with actigraphy. We found that combining actigraphy with additional sensor(s) significantly improved TST mean difference versus PSG ($p = 0.002$, Figure 10).

Mean differences ranged from -25.7 minutes (NightOwl [45]) to 87.0 minutes (Withings Pulse [65]) and were ranked in order of performance (Figure S8). Most top performing devices relied on custom algorithms and parameters, individualising classification using time series models to integrate data from surrounding epochs. The best individual performer was Kronowise 3.0 (mean difference -3.73 minutes on a 352-minute mean TST) [66]. This multi-sensor, wrist device uses wrist temperature, light exposure variability, motor activity and wrist position to distinguish sleep and wake. The next best performing device was the Edirol R-4 Pro acoustic device (7.6-minute overestimate [67]). This device uses a bi-directional microphone placed one metre above participants' beds to pick up sounds and a breathing detection system to isolate breathing sounds

from background noises. Eight features (breathing pattern and snore properties) were extracted from audio signal and classified using an algorithm configured as a time series model, with the previous two epochs' data informing the next epoch's classification. Authors also applied an individual decision threshold to account for variation between individual breathing properties using Otsu's method to minimise intra-class variance. Kapella et al customised Actiwatch 2 (a single-sensor accelerometer wrist device) settings to improve mean difference from over 50 minutes with standard settings to 10.9 minutes [68]. Their algorithm classified sleep and wake based on the degree of movement within 2 minutes of a given epoch.

After mean difference, the two next most reported accuracy data were intra-class correlations (ICCs) and correlation coefficients. Four NISMDs had poor agreement (ICC < 0.5; SenseWear Pro Armband, Jawbone Up and Withings Pulse [65], WatchPAT [36]). Actiwatch 2 had good agreement (ICC 0.75-0.90) [69]. The Sleep Profiler – a headband device with frontopolar EEG signals – showed excellent agreement (ICC > 0.90) in two studies scored automatically with manual editing [33] or scored manually [34]. Correlation coefficients (r^2) ranged from 0.418 (Morpheus Ox [44]) to 0.945 (Belun Ring [43]).

Sleep efficiency

15 NISMDs across 12 studies presented accuracy data for SE (Table S3). These included three non-wearable and 12 wearable devices, most commonly Actiwatch (n = 4) models, and the most common device sensors per study were actigraphy (n = 11) and acoustic sensors (n = 4). The summary estimate from meta-analysis showed that devices significantly overestimated SE by

mean 7.6% (95% confidence interval 4.3 to 11.0%, Z = 4.456, $p < 0.001$, Figure 11). As with TST, we also found that devices that combined actigraphy with additional sensor(s) significantly improved SE mean difference versus PSG compared with devices that only used actigraphy ($t_{df} = 2.452_{11}$, $p = 0.016$). Adding sensors to actigraphy improved SE accuracy from 10.0% (SEr 1.7%) to a 1.6% (SEr 2.8) overestimate.

The top performer by mean difference was again Kronowise, which underestimated SE by 0.4% [66] (Figure S8). Other top performers were Readiband (1.0% overestimate; [70]), Edirol R-4 pro (1.5% overestimate; [67]) and Kapella and colleagues' custom settings for Actiwatch 2 (2.6% overestimate; [68]). The Sleep Profiler had good ICC agreement with automatic scoring and manual editing (ICC = 0.86; [33]). All other studies that published ICC showed moderate (ICC 0.5-0.75; [69]) or poor (ICC < 0.5) agreement [36, 65, 69].

Slow wave sleep

Eight studies displayed data for SWS across eleven devices (two non-wearable and nine wearable), including four studies of acoustic and five of actigraphy-based devices. Overall, NISMDs did not measure SWS accurately. Five studies compared mean difference for SWS duration, of which four published a measure of variance. Of these five, four significantly underperformed (Figure 12): Jawbone UP overestimated SWS by 160 minutes [65], Withings Pulse underestimated by 88 minutes [65], Fitbit Charge 2 and Alta HR (combined datasets across two wrist devices) underestimated by 74 minutes [71]. Only the Sleep Profiler performed well, overestimating by only 2.7 minutes with excellent ICC agreement (0.9) and Cohen's kappa

(0.67) with manual scoring [34]). However, 13% of their recordings were discarded prior to scoring because of poor signal quality. The summary measure from meta-analysis across devices showed a mean underestimation of SWS by 16.8 minutes with a wide 95% confidence interval (-93.9 to 60.4 minutes).

Rapid eye movement sleep

Seven studies reported on REM duration or percentage, using one non-wearable and six wearable devices, most-commonly actigraphy, acoustic and/or plethysmography-based devices (n = 4 each). Four of these seven studies displayed mean differences with variance for REM duration versus PSG (Figure 13). Summary estimate from meta-analysis of these four studies showed mean overestimation of REM duration by 3.6 minutes (95% confidence interval -9.1 to 16.1 minutes). Combined data for Fitbit Charge and Alta HR showed mean overestimation by 2.8 minutes across devices [71]. However, this study had large amounts of missing data. Only 52% of sleep recordings had enough quality to allow sleep staging using default, proprietary algorithms, and no further analysis (e.g., correlation) was available for REM duration accuracy. Using a single EEG channel (Fp1-Fp2), the Sleep Profiler overestimated REM duration by only 6.8 minutes with excellent ICC agreement with automatic scoring and manual editing (0.92; [33]), but 10% of recordings were discarded due to poor quality signal. WatchPAT, designed primarily for detection of OSA, had a positive predictive value of 59.1% (technician 1) and 62.2% (technician 2) for REM sleep in patients with COPD [36].

Microarchitectural sleep metrics

We identified only two studies that reported accuracy based on microarchitectural sleep data using the same headband device with EEG - the Sleep Profiler [33, 34]. Lewendoswski et al. used an automated scoring system with manual editing which detected sleep spindles by spikes in alpha and sigma range with simultaneous beta range suppression. SWA was detected in a frequency range of 1-3.5Hz. Although authors assessed night-to-night variability of spindle activity, they did not report accuracy data for or spindle activity or SWA compared to PSG. However, they did use SWA and spindle data to help stage N2 and N3 sleep with best agreement of 80.6% and 75.3%, respectively, across scorers. By comparison, Lucey et al. only used a single frontopolar channel for manual sleep stage scoring and sleep spindle and SWA measurement. Sensitivity was 0.83 for N2 but only 0.29 for N3 across scorers. SWA (1-4.5Hz) across the first and last 20-minute NREM periods were normalised to SWA as a percentage of total SWS. Authors described a significant correlation in SWA between device and PSG for these periods (r^2 and *p* values not given). They also produced Bland-Altman plots of the difference in SWA and reported that there was no significant bias (bias and limits of agreement not given). Finally, this group compared the declines in SWA across the night (as a percentage) and compared the Sleep Profiler against PSG with a paired *t*-test and found this difference was not significant. We excluded studies of several other non-invasive headband [73] and ear EEG [74-76] devices from review because they were conducted on young cohorts (mean age <40 years).

Bias assessment

Assessing for methodological bias using the QUADAS-2 tool (Figure 14) found that 24 (46%) studies had a high risk of bias for the index test on account of classification thresholds not being specified, algorithms not being specified, and/or authors not being blinded to PSG results before analysing index test data. All reference test definitions were standardised across studies but 28 (54%) were scored as having an unclear risk of bias for this category because they did not specify that PSG was scored by a trained professional. 12 (23%) studies had a high risk for study flow and timing because >10% of participants were removed from analysis (e.g., due to technical errors or insufficient sleep time). Since we included both healthy and unhealthy participants to reflect the heterogeneity of individuals susceptible to AD, we did not use the QUADAS-2 applicability questions during our bias assessment. 16 studies (including 13 different devices) had a potential conflict of interest as they were either funded by the company that makes the device, or its authors were salaried by that company (Table 1).

Discussion

This systematic review of validity studies of NISMDs demonstrated mixed accuracy for measuring sleep metrics associated with AD in mid-to-late age adults. Considering the inability of actigraphy to differentiate wakefulness from sleep when an individual is motionless, it was unsurprising that most actigraphy-based devices overestimated TST (and, by extension, SE), and that combining this with additional sensors improved accuracy. However, it was surprising that some devices recorded AHI so accurately yet substantially overestimated TST, which is the denominator for estimation of AHI. For example, one group observed that WatchPAT overestimated AHI by only 0.5 events per hour whilst overestimating TST by mean of 70.4

minutes [58]. This suggests that different algorithms or classification thresholds were used to calculate TST for AHI score and TST alone, but full software information was unavailable. Unfortunately, most algorithms and classification threshold were not open access.

Differences in software are also likely to explain differences between studies, such as for Actiwatch devices, where customising software improved accuracy by as much as 40 minutes [68]. In addition, we observed differences between studies that may be explained by heterogeneity in population characteristics. For example, using the same device and algorithm, one group observed that the non-contact SleepMinder device underestimated AHI by -3.8 events per hour in patients with congestive heart failure with mean age 68 [77], whilst another found the device overestimated AHI by 16.8 events per hour in patients with suspected sleep disorders with mean age 56 [78]. Allowing for this heterogeneity across studies, we did not find that multi-sensor devices outperformed single sensor devices for AHI, where many single-sensor devices performed well including the BresoDX portable acoustic device [42] or the Nemuri scan bedsheet sensor [40]. By comparison, combining additional sensors with actigraphy improved accuracy for TST and SE. This difference may be explained by differences in signal. Single sensor devices recording AHI can detect signal directly associated with apnoea, such as pauses in breathing detected by an acoustic sensor or cessation of chest wall movement detected by a mattress with pressure sensors. By comparison, changes in wrist or arm movement detected by actigraphy are peripherally and indirectly associated with sleep.

Overall, several NISMDs accurately measured the AHI, ODI, and detected OSA, and may therefore be suitable for future longitudinal, naturalistic studies investigating the relationship

between OSA and AD. OSA is associated with cognitive impairments in middle aged adults and the development of AD in older, symptomatic adults [25]. Treating OSA, such as with weight loss strategies or continuous positive airway pressure, is emerging as an important potential preventative strategy for reducing the incidence of AD [79].

In addition to treating OSA, SWS is emerging as a promising preventative target for AD [13, 80). Reduced SWS is associated with cognitive decline and neuropathological changes of AD [18, 23]. However, we found that NISMDs based on plethysmography, actigraphy, pressure sensors or microphones measured SWS inaccurately. These sensors rely heavily on cardiorespiratory signals to infer SWS, an indirect approach also limited by the physiological variability of these signals overnight, though use of smoothing and temporal filtering may help [81]. Future devices may benefit from age-specific adjustments to further improve accuracy, considering the strong additional influences of age on heart and respiratory rate variability during sleep [82]. In the meantime, EEG signals appear necessary for accurate quantification of SWS.

Two studies were included that recorded SWS using the same EEG headband device (the Sleep Profiler; [33, 34]). When all three frontal channels of the Sleep Profiler were used, agreement with PSG was excellent [34]. However, when only a single channel was used, accuracy was poor (sensitivity = 0.294; [33]), in fact worse than that measured by devices relying on cardiorespiratory and motion signals. Authors attributed this low sensitivity to the poor sleep quality of their sample (having only 2.3-5.5% SWS in total), and signal artefact relating to poor electrode contact. Both studies of this headband device excluded large volumes of data due to poor signal quality. Thus, while it appears preferable to detect SWS with EEG signals, it is

harder to ensure EEG signal quality than for more robust cardiorespiratory signals such as heart and respiratory rate.

We also sought to identify validity studies of NISMDs that measured sleep microarchitectural features including SWA and sleep spindles. These neurophysiological hallmarks are particularly pertinent to AD pathology and symptomatology, given that sleep-dependent memory consolidation involves the coordination of hippocampal sharp wave ripples with thalamic spindle activity and the up state of slow waves [13], and the growing evidence that the AD neuropathology disrupts both sleep spindle and SWA [20, 22, 33]. Despite not accurately delineating SWS epochs with a single frontal channel, the Sleep Profiler was able to accurately measure SWA using this channel. Whilst this device was more accurate in measuring SWS duration with three channels than one, it is unclear whether accuracy for SWA would be improved with further channels since this was not tested. With the potential to add additional EEG channels, and have better electrode contact, EEG-based devices show promise as measurement tools for AD research. However, we only found one device that met our search criteria. Whilst there are several other, promising, non-invasive EEG devices available [73-76], they have not yet been tested on older cohorts.

Even if accuracy improves, the topographic distributions of electrodes of EEG-based NISMDs may limit the detection of SWA and spindles. Slow waves and sleep spindles are non-uniformly distributed across the cortex and change their expression with ageing and in AD [83]. In addition, there is significant topographical variability between individuals of similar ages [84]. AD is principally associated with loss of fast spindles [85, 86], which are predominant in parietal

and central areas [89]. NISMDs relying on frontal channels are therefore not well designed to detect spindle changes associated with AD. By contrast, SWA has a predominantly frontal distribution, and reductions in frontal SWA are associated with beta-amyloid deposition [22]. Devices designed to detect frontal SWA may therefore not be appropriate for detection of central and parietal fast spindles, and vice versa, unless that device uses more widespread channels. Topographically specific changes in sleep architecture expression are likely to represent distinct neuropathological profiles [83]. Determining such topographic specificity may be important for detecting and treating focal sleep deficits, and for bidirectional translation between human and animal models of sleep and AD [90], and potentially other diseases, such as depression and schizophrenia. Thus, we are surprised by the lack of validation of NISMDs that can report sleep EEG features from more than one skull location and view this as an area of growth for future deep phenotyping of sleep in AD research. Furthermore, future research should investigate device accuracy in participants with mild cognitive impairment or AD; we did not identify any such data from our search. Accurate, validated assessment of sleep in this population may unlock its potential as an early disease marker and is a pre-requisite for precise assessment of interventions designed to improve disordered sleep.

In addition to issues with signal quality and sensor placement, data on NISMDs are limited by the lack of studies which measure accuracy across more than one night. Considering that one of the clear benefits of non-invasive devices is their ability to capture longitudinal data, it was disappointing to find only two studies that measured accuracy over more than a single night. Both studies aggregated data from two nights [35, 36]. Although we excluded ambulatory PSG-based studies, many of these still only recorded accuracy data from a single night (e.g., [75, 91]).

The dependence on a single night's sleep introduces additional heterogeneity due to ultradian sleep dynamics, stochastic measurement error, and the measurement bias associated with capturing data from one overnight measurement in an unfamiliar setting. Aggregate data from multiple nights would be more clinically meaningful and robust for research purposes. However, whilst there is a need for more aggregate data for sleep macroarchitecture, these data may be less important for sleep microarchitecture. For example, using the Sleep Profiler device, Levendowski et al., showed that recording two nights rather than one did not significantly improve characterisation of sleep spindle activity [34].

In addition to the lack of longitudinal data and data in participants with cognitive impairment, our results may be limited by excluding adults of mean age < 40. Sleep changes in younger adults may also be associated with AD risk [92], although within the included studies there were several participants under the age of 40 including participants aged 18.

This review is further constrained by issues with incomplete data capture. Several studies did not publish mean differences, and the quality of accuracy data reporting was poor for macroarchitectural data. For example, whilst measurement of REM duration appeared accurate across Fitbit and Sleep Profiler devices, only four of the seven available studies (57%) reported a mean difference with a measure of variance. Even within studies that reported mean difference data, large amounts of data were excluded from analysis due to signal quality. In addition, our findings did not account for the comfort or the cost of each device, which is an important factor for experimental design, particularly as AD research extends to Low and Middle-Income Countries. Lastly, whilst we have focussed on objectively measured sleep, subjectively measured

sleep changes are also associated with AD (e.g., [3]). Only one study [69] investigated the differences in accuracy between NISMD and subjective sleep measurement (for TST) versus PSG. If NISMD and subjective sleep measurements are comparably accurate for some macroarchitectural or breathing-related sleep metrics, then there may be less benefit in using NISMDs in future research.

Conclusion

Several devices may be suitable for future longitudinal studies investigating the relationship between sleep and AD risk, especially for OSA. However, whilst some devices may accurately quantify TST, SE, and REM, current devices that do not have EEG do not accurately quantify SWS. We have identified a significant lack of validity studies investigating the accuracy of NISMDs for detecting microarchitectural phenomena relevant to AD in mid-to-late life cohorts. The only device included in our study, which can record SWA and sleep spindles, is limited by issues with signal quality and other available EEG devices have not yet been tested on older cohorts. Our review provides a platform for the design of future device validity studies and research exploring the relationship between sleep and AD risk.

Practice Points
1. A wide range of recent single and multi-sensor devices can accurately measure the apnoea hypopnoea index, oxygen desaturation index, and diagnose OSA in mid-to-late life adults for the purposes of Alzheimer's disease research

2. Some multi-sensor devices can also accurately measure total sleep time and sleep efficiency in mid-to-late life adults
3. Accepting issues with data reporting, some devices may also accurately measure REM duration in mid-to-late life adults
4. Available non-electroencephalography devices do not accurately measure slow wave sleep duration in mid-to-late life adults

Research Agenda

1. Device validation studies should consistently report outcomes including sensitivity, specificity, and mean difference with a measure of variance
2. Independent (i.e., without conflict of interest) device validation studies are needed in mid-to-late life adults and across diverse racial and ethnic populations
3. We propose validation of more electroencephalography-based devices in older cohorts, especially those that record slow wave and spindle activity across several electrode locations
4. Further research is needed to investigate device accuracy across multiple nights to produce more robust and clinically meaningful aggregate sleep data
5. Future studies should investigate device accuracy in populations with mild cognitive impairment and Alzheimer's disease

Funding

SFG, KL and HM are funded by a National Institute for Health Research Academic Clinical Fellowship. JB receives funding from Alzheimer's Research UK (supported by the Margaret

Jost Fellowship and the Don Thoburn Memorial Scholarship), and the David Telling Charitable Trust.

References

[1] Prince M WA, Guerchet M, Ali GC, Wu YT, Prina M. World Alzheimer report 2015—the global impact of dementia: an analysis of prevalence, incidence, cost and trends. London: Alzheimer's Disease International; 2015.

[2] Sabia S, Fayosse A, Dumurgier J, van Hees VT, Paquet C, Sommerlad A, et al. Association of sleep duration in middle and old age with incidence of dementia. *Nat Commun.* 2021; **12**: 2289.

[3] Benedict C, Byberg L, Cedernaes J, Hogenkamp PS, Giedratis V, Kilander L, et al. Self-reported sleep disturbance is associated with Alzheimer's disease risk in men. *Alzheimers Dement.* 2015; **11**: 1090-7.

[4] Gabelle A, Gutierrez LA, Jaussent I, Navucet S, Grasselli C, Bennys K, et al. Excessive Sleepiness and Longer Nighttime in Bed Increase the Risk of Cognitive Decline in Frail Elderly Subjects: The MAPT-Sleep Study. *Front Aging Neurosci*. 2017; **9**: 312.

[5] Pase MP, Himali JJ, Grima NA, Beiser AS, Satizabal CL, Aparicio HJ, et al. Sleep architecture and the risk of incident dementia in the community. *Neurology*. 2017; **89**: 1244-50.

[6] Hahn EA, Wang HX, Andel R, Fratiglioni L. A Change in Sleep Pattern May Predict Alzheimer Disease. *Am J Geriatr Psychiatry*. 2014; **22**: 1262-71.

[7] Bubu OM, Brannick M, Mortimer J, Umasabor-Bubu O, Sebastiao YV, Wen Y, et al. Sleep, Cognitive impairment, and Alzheimer's disease: A Systematic Review and Meta-Analysis. *Sleep*. 2017; **40**: zsw032.

[8] Shi L, Chen SJ, Ma MY, Bao YP, Han Y, Wang YM, et al. Sleep disturbances increase the risk of dementia: A systematic review and meta-analysis. *Sleep Med Rev*. 2018; **40**: 4-16.*

[9] Robbins R, Quan SF, Weaver MD, Bormes G, Barger LK, Czeisler CA. Examining sleep deficiency and disturbance and their risk for incident dementia and all-cause mortality in older adults across 5 years in the United States. *Aging (Albany NY)*. 2021; **13**: 3254-68.

[10] Andre C, Laniepce A, Chetelat G, Rauchs G. Brain changes associated with sleep disruption in cognitively unimpaired older adults: A short review of neuroimaging studies. *Ageing Res Rev*. 2021; **66**: 101252.*

[11] Jack CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 2010; **9**: 119-28.

[12] Stickgold R, Walker MP. Sleep-dependent memory consolidation and reconsolidation. *Sleep Med*. 2007; **8**: 331-43.

[13] Mander BA, Winer JR, Jagust WJ, Walker MP. Sleep: A Novel Mechanistic Pathway, Biomarker, and Treatment Target in the Pathology of Alzheimer's Disease? *Trends Neurosci*. 2016; **39**: 552-66.*****

[14] Xu W, Tan CC, Zou JJ, Cao XP, Tan L. Sleep problems and risk of all-cause cognitive decline or dementia: an updated systematic review and meta-analysis. *J Neurol Neurosurg Psychiatry*. 2020; **91**: 236-44.*****

[15] Ettore E, Bakardjian H, Sole M, Nogueira ML, Habert MO, Gabelle A, et al. Relationships between objectives sleep parameters and brain amyloid load in subjects at risk for Alzheimer's disease: the INSIGHT-preAD Study. *Sleep*. 2019; **42**: zsz137.

[16] Molano JRV, Roe CM, Ju YE. The interaction of sleep and amyloid deposition on cognitive performance. *J Sleep Res*. 2017; **26**: 288-92.

[17] Winer JR, Mander BA, Kumar S, Reed M, Baker SL, Jagust WJ, et al. Sleep Disturbance Forecasts beta-Amyloid Accumulation across Subsequent Years. *Curr Biol*. 2020; **30**: 4291-98.

[18] Ju YE, Ooms SJ, Sutphen C, Macauley SL, Zangrilli MA, Jerome G, et al. Slow wave sleep disruption increases cerebrospinal fluid amyloid-beta levels. *Brain*. 2017; **140**: 2104-11.

[19] Varga AW, Wohlleber ME, Gimenez S, Romero S, Alonso JF, Ducca EL, et al. Reduced Slow-Wave Sleep Is Associated with High Cerebrospinal Fluid A beta 42 Levels in Cognitively Normal Elderly. *Sleep*. 2016; **39**: 2041-8.

[20] Kam K, Parekh A, Sharma RA, Andrade A, Lewin M, Castillo B, et al. Sleep oscillation-specific associations with Alzheimer's disease CSF biomarkers: novel roles for sleep spindles and tau. *Mol Neurodegener*. 2019; **14**: 10.

[21] Winer JR, Mander BA, Helfrich RF, Maass A, Harrison TM, Baker SL, et al. Sleep as a Potential Biomarker of Tau and beta-Amyloid Burden in the Human Brain. *J Neurosci*. 2019; **39**: 6315-24.

[22] Mander BA, Marks SM, Vogel JW, Rao V, Lu B, Saletin JM, et al. beta-amyloid disrupts human NREM slow waves and related hippocampus-dependent memory consolidation. *Nat Neurosci*. 2015; **18**: 1051-7.*

[23] Lucey BP, McCullough A, Landsness EC, Toedebusch CD, McLeland JS, Zaza AM, et al. Reduced non-rapid eye movement sleep is associated with tau pathology in early Alzheimer's disease. *Sci Transl Med*. 2019; **11**: eaau6550.

[24] Scarpa JR, Jiang P, Gao VD, Vitaterna MH, Turek FW, Kasarskis A. NREM delta power and AD-relevant tauopathy are associated with shared cortical gene networks. *Sci Rep*. 2021; **11**: 7797.

[25] Bubu OM, Andrade AG, Umasabor-Bubu OQ, Hogan MM, Turner AD, de Leon MJ, et al. Obstructive sleep apnea, cognition and Alzheimer's disease: A systematic review integrating three decades of multidisciplinary research. *Sleep Med Rev*. 2020; **50**: 101250.*****

[26] De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable Sleep Technology in Clinical and Research Settings. *Med Sci Sports Exerc*. 2019; **51**:1538-57.*

[27] Perez-Pozuelo I, Zhai B, Palotti J, Mall R, Aupetit M, Garcia-Gomez JM, et al. The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ Digit Med.* 2020; **3**:42**.***

[28] Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2020; **396**: 413-46.

[29] McGrath TA, Moher D, McInnes MDF. Steps toward more complete reporting of systematic reviews of diagnostic test accuracy: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA). *Syst Rev*. 2019; **8**: 166.

[30] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016; **5**: 210.

[31] Santiago JA, Potashkin JA. The Impact of Disease Comorbidities in Alzheimer's Disease. *Front Aging Neurosci*. 2021; **13**: 631770.

[32] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011; **155**: 529-36.

[33] Lucey BP, McLel, S. J, Toedebusch CD, Boyd J, Morris JC, et al. Comparison of a single-channel EEG sleep study to polysomnography. *J Sleep Res*. 2016; **25**: 625-35.*

[34] Levendowski DJ, Ferini-Strambi L, Gamaldo C, Cetel M, Rosenberg R, Westbrook PR. The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *J Clin Sleep Med*. 2017; **13**: 791-803.*

[35] Spielmanns M, Bost D, Windisch W, Alter P, Greulich T, Nell C, et al. Measuring Sleep Quality and Efficiency With an Activity Monitoring Device in Comparison to Polysomnography. *J Clin Med Res*. 2019; **11**: 825-33.

[36] Holmedahl NH, Fjeldstad O-M, Engan H, Saxvig IW, Gronli J. Validation of peripheral arterial tonometry as tool for sleep assessment in chronic obstructive pulmonary disease. *Sci Rep*. 2019; **9**: 19392.

[37] Abad J, Munoz-Ferrer A, Cervantes MA, Esquinas C, Marin A, Martinez C, et al. Automatic Video Analysis for Obstructive Sleep Apnea Diagnosis. *Sleep*. 2016; **39**: 1507-15.

[38] Munoz-Ferrer A, Cervantes M-A, Garcia-Olive I, Vicente I, Folgado C, Ruiz-Manzano J, et al. In-Home Diagnosis of Obstructive Sleep Apnea Using Automatic Video Analysis. *Archivos de Bronconeumologia (English Edition).* 2020; **56**: 704-9.

[39] Lu M, Fang F, erson JE, Ma C, Wang Q, Zhan X, et al. Validation of a portable monitoring device for the diagnosis of obstructive sleep apnea: electrocardiogram-based cardiopulmonary coupling. *Sleep Breath*. 2019; **23**: 1371-8.

[40] Kogure T, Kobayashi M, Inoue Y, Okawa T, Nakajima T. Validation of a sheet-shaped body vibrometer for screening of obstructive sleep apnea. *Drug Discov Ther*. 2017; **11**: 126-32.

[41] Andres-Blanco AM, Alvarez D, Crespo A, Arroyo CA, Cerezo H, ez A, et al. Assessment of automated analysis of portable oximetry as a screening test for moderate-to-severe sleep apnea in patients with chronic obstructive pulmonary disease. *PLoS ONE*. 2017; **12**: e0188094.

[42] Alshaer H, Fernie GR, Tseng W-H, Bradley TD. Comparison of in-laboratory and home diagnosis of sleep apnea using a cordless portable acoustic device. *Sleep Med*. 2016; **22**: 91-6.

[43] Gu W, Leung L, Kwok KC, Wu IC, Folz RJ, Chiang AA. Belun Ring Platform: A novel home sleep apnea testing system for assessment of obstructive sleep apnea. *J Clin Sleep Med*. 2020; **16**: 1611-7.

[44] Li Y, Gao H, Ma Y. Evaluation of pulse oximeter derived photoplethysmographic signals for obstructive sleep apnea diagnosis. *Medicine (Baltimore)*. 2017; **96**: e6755.

[45] Massie F, De Almeida DM, Dreesen P, Thijs I, Vranken J, Klerkx S. An evaluation of the Night Owl home sleep apnea testing system. *J Clin Sleep Med*. 2018; **14**: 1791-6.

[46] Meng L, Wu H, Xu H, Guan J, Yi H, Yin S. Validation of a novel sleep-monitoring system for diagnosing obstructive sleep apnea: A comparison with polysomnography. *Exp Ther Med*. 2016; **12**: 2937-41.

[47] Boyd SB, Upender R, Walters AS, Goodpaster RL, Stanley JJ, Wang L, et al. Effective Apnea-Hypopnea Index ("Effective AHI"): A New Measure of Effectiveness for Positive Airway Pressure Therapy. *Sleep*. 2016; **39**: 1961-72.

[48] Pinheiro GdL, Cruz AF, Domingues DM, Genta PR, Drager LF, Strollo PJ, et al. Validation of an Overnight Wireless High-Resolution Oximeter plus Cloud-Based Algorithm for the Diagnosis of Obstructive Sleep Apnea. *Clinics (Sao Paulo)*. 2020; **75**: e2414.

[49] Assefa SZ, Diaz-Abad M, Korotinsky A, Tom SE, Scharf SM. Comparison of a simple obstructive sleep apnea screening device with standard in-laboratory polysomnography. *Sleep Breath*. 2016; **20**: 537-41.

[50] Ioachimescu OC, Dholakia SA, Venkateshiah SB, Fields B, Samarghandi A, Anand N, et al. Improving the performance of peripheral arterial tonometry-based testing for the diagnosis of obstructive sleep apnea. *J Investig Med*. 2020; **68**: 1370-8.

[51] Cho JH, Kim HJ. Validation of ApneaLinkTM Plus for the diagnosis of sleep apnea. *Sleep Breath*. 2017; **21**: 799-807.

[52] Crinion SJ, Boyle P, Russell A, Traynor M, Kent BD, Tiron R, et al. Ambulatory detection of sleep apnea using a non-contact biomotion sensor. *J Sleep Res*. 2020; **29**: e12889.

[53] Kim JW, Kim T, Shin J, Choe G, Lim HJ, Rhee CS, et al. Prediction of Obstructive Sleep Apnea Based on Respiratory Sounds Recorded Between Sleep Onset and Sleep Offset. *Clin Exp Otorhinolaryngol*. 2019; **12**: 72-8.

[54] Tiron R, Lyon G, Kilroy H, Osman A, Kelly N, O'Mahony N, et al. Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology. *J Thorac Dis*. 2020; **12**: 4476-95.

[55] Zou J, Meng L, Liu Y, Xu X, Liu S, Guan J, et al. Evaluation of a 2-channel portable device and a predictive model to screen for obstructive sleep apnea in a laboratory environment. *Respir Care*. 2015; **60**: 356-62.

[56] Ribeiro RC, Mizoguchi EI, Pinto JA, de Godoy LBM, Hirsch LAM, Gomes LM. Accuracy of peripheral arterial tonometry in the diagnosis of obstructive sleep apnea. *Braz J Otorhinolaryngol.* 2015; **81**: 473-8.

[57] Jen R, Orr JE, DeYoung P, Smales E, Malhotra A, Owens RL, et al. Accuracy of WatchPAT for the Diagnosis of Obstructive Sleep Apnea in Patients with Chronic Obstructive Pulmonary Disease. *COPD*. 2020; **17**: 34-9.

[58] Kasai T, Takata Y, Yoshihisa A, Takeishi Y, Chin K, Ando SI, et al. Comparison of the Apnea-Hypopnea Index Determined by a Peripheral Arterial Tonometry-Based Device With That Determined by Polysomnography - Results From a Multicenter Study. *Circ Rep*. 2020; **2**: 674-81.

[59] Pillar G, Berall M, Berry R, Etzioni T, Shrater N, Hwang D, et al. Detecting central sleep apnea in adult patients using WatchPAT-a multicenter validation study. *Sleep Breath*. 2020; **24**: 387-98.

[60] Tondo P, Drigo R, Scioscia G, Ballarin A, Rossi E, Floriani AF, et al. Usefulness of sleep events detection using a wrist worn peripheral arterial tone signal device (WatchPAT) in a population at low risk of obstructive sleep apnea. *J Sleep Res*. 2021: **30**: e13352.

[61] Topor ZL, Grosse J, Mosca EV, Jahromi SAZ, Zhu Y, Bruehlmann S, et al. Validation of a new unattended sleep apnea monitor using two methods for the identification of hypopneas. *J Clin Sleep Med*. 2020; **16**: 695-703.

[62] Miyata S, Otake H, Ando M, Okuda M, Fujishiro H, Iwamoto K, et al. Patient characteristics affecting accurate detection of sleep apnea using a bed sheet-type portable monitor. *Sleep Breath*. 2020; **24**: 783-90.

[63] Smith D, Park J, Hay K, Hoey L, Leong G, Leong M, et al. Use of a limited-channel device for obstructive sleep apnoea diagnosis in a tertiary sleep disorders centre. *Intern Med J*. 2020; **50**: 1109-14.

[64] Tal A, Shinar Z, Shaki D, Codish S, Goldbart A. Validation of contact-free sleep monitoring device with comparison to polysomnography. *J Clin Sleep Med*. 2017; **13**: 517-22.

[65] Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. *PLoS ONE*. 2019; **14**: e0210569.

[66] Madrid-Navarro CJ, Escamilla-Sevilla F, Cuesta FJP, Campos M, Rol MA, Madrid JA, et al. Validation of a device for the ambulatory monitoring of sleep patterns: A pilot study on Parkinson's disease. *Front Neurol*. 2019; **10**: 356.

[67] Dafna E, Tarasiuk A, Zigel Y. Sleep-wake evaluation from whole-night non-contact audio recordings of breathing sounds. *PLoS ONE*. 2015; **10**: e0117382.

[68] Kapella MC, Vispute S, Zhu B, Herdegen JJ. Actigraphy scoring for sleep outcome measures in chronic obstructive pulmonary disease. *Sleep Med*. 2017; **37**: 124-9.

[69] Choi SJ, Kang M, Sung MJ, Joo EY. Discordant sleep parameters among actigraphy, polysomnography, and perceived sleep in patients with sleep-disordered breathing in comparison with patients with chronic insomnia disorder. *Sleep Breath*. 2017; **21**: 837-43.

[70] Dunican IC, Slater JA, Maddison KJ, Eastwood PR, Murray K, Jones MJ, et al. Laboratory and home comparison of wrist-activity monitors and polysomnography in middle-aged adults. *Sleep Biol Rhythms*. 2018; **16**: 85-97.

[71] Moreno-Pino F, Porras-Segovia A, Lopez-Esteban P, Artes A, Baca-Garcia E. Validation of fitbit charge 2 and fitbit alta hr against polysomnography for assessing sleep in adults with obstructive sleep apnea. *J Clin Sleep Med*. 2019; **15**: 1645-53.

[72] Arnal PJ, Debellemaniere E, Ballard ME, Thorey V, Hern, ez AB, et al. The dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*. 2020; **43**: zsaa097.

[73] Mikkelsen KB, Tabar YR, Christensen CB, Kappel SL, Toft HO, Hemmsen MC, et al. Accurate whole-night sleep monitoring with dry-contact ear-EEG. *Sci Rep*. 2019; **9**: 16824.

[74] Nakamura T, Goverdovsky V, ic DP, Morrell MJ. Automatic Sleep Monitoring Using Ear-EEG. *IEEE J Transl Eng Health Med*. 2017; **5**: 2800108.

[75] Tabar YR, Mikkelsen KB, Kidmose P, Rank ML, Hemmsen MC, Otto M, et al. Ear-EEG for sleep assessment: a comparison with actigraphy and PSG. *Sleep Breath*. 2020; **25**: 1693-1705.

[76] Savage HO, Khushaba RN, Zaffaroni A, Colefax M, Farrugia S, Schindhelm K, et al. Development and validation of a novel non-contact monitor of nocturnal respiration for identifying sleep-disordered breathing in patients with heart failure. *ESC Heart Fail*. 2016; **3**: 212-9.

[77] Weinreich G, Terjung S, Wang Y, Werther S, Zaffaroni A, Teschler H. Validation of a non-contact screening device for the combination of sleep-disordered breathing and periodic limb movements in sleep. *Sleep Breath*. 2018; **22**: 131-8.

[78] Dunietz GL, Chervin RD, Burke JF, Conceicao AS, Braley TJ. Obstructive sleep apnea treatment and dementia risk in older adults. *Sleep*. 2021; **44**: zsab076.

[79] Wunderlin M, Zust MA, Feher KD, Kloppel S, Nissen C. The role of slow wave sleep in the development of dementia and its potential for preventative interventions. *Psychiatry Res Neuroimaging.* 2020; **306**: 111178.

[80] Long X, Fonseca P, Aarts RM, Haakma R, Rolink J, Leonhardt S. Detection of nocturnal slow wave sleep based on cardiorespiratory activity in healthy adults. *IEEE J Biomed Health Inform*. 2017: **21**: 123-133.

[81] Schumann AY, Bartsch RP, Penzel T, Ivanov PC, Kantelhardt JW. Aging Effects on Cardiac and Respiratory Dynamics in Healthy Subjects across Sleep Stages. *Sleep*. 2010; **33**: 943-55.

[82] Mander BA. Local Sleep and Alzheimer's Disease Pathophysiology. *Front Neurosci*. 2020; **14**: 525970.

[83] Finelli LA, Achermann P, Borbely AA. Individual "fingerprints" in human sleep EEG topography. *Neuropsychopharmacology*. 2001; **25**: S57-S62.

[84] Westerberg CE, Mander BA, Florczak SM, Weintraub S, Mesulam MM, Zee PC, et al. Concurrent Impairments in Sleep and Memory in Amnestic Mild Cognitive Impairment. *J Int Neuropsychol Soc*. 2012; **18**: 490-500.

[85] Gorgoni M, Lauri G, Truglia I, Cordone S, Sarasso S, Scarpelli S, et al. Parietal Fast Sleep Spindle Density Decrease in Alzheimer's Disease and Amnesic Mild Cognitive Impairment. *Neural Plast*. 2016; **2016**: 8376108.

[86] De Gennaro L, Ferrara M. Sleep spindles: an overview. *Sleep Med Rev*. 2003; **7**: 423-40.

[87] Busche MA, Kekus M, Adelsberger H, Noda T, Forstl H, Nelken I, et al. Rescue of long-range circuit dysfunction in Alzheimer's disease models. *Nature Neurosci*. 2015; **18**: 1623-30.

[88] Mouritzen NJ, Larsen LH, Lauritzen MH, Kjaer TW. Assessing the performance of a commercial multisensory sleep tracker. *PLoS ONE*. 2020; **15**: e0243214.

[89] Muto V, Koshmanova E, Ghaemmaghami P, Jaspar M, Meyer C, Elansary M, et al. Alzheimer's disease genetic risk and sleep phenotypes in healthy young men: association with more slow waves and daytime sleepiness. *Sleep*. 2021; **44**: zsaa137.

[90] Cheung J, Leary EB, Lu H, Zeitzer JM, Mignot E. PSG Validation of minute-to-minute scoring for sleep and wake periods in a consumer wearable device. *PLoS ONE*. 2020; **15**: e0238464.

[91] De Zambotti M, Claudatos S, Inkelis SI, Colrain IM, Baker FC. Evaluation of a Consumer Fitness-Tracking Device to Assess Sleep in Adults. *Chronobiol Int.* 2015; **32**: 1024-28.

[92] Edouard P, Campo D, Bartet P, Yang R-Y, Bruyneel M, Roisman G, et al. Validation of the Withings Sleep Analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome. *J Clin Sleep Med*. 2021; **17**: 1217-27.

[93] Hayano J, Yamamoto H, Nonaka I, Komazawa M, Itao K, Ueda N, et al. Quantitative detection of sleep apnea with wearable watch device. *PLoS ONE*. 2020; **15**: e0237279.

[94] Kahawage P, Jumabhoy R, Hamill K, de Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: In-lab validation against polysomnography. *J Sleep Res.* 2020; **29**: e12931.

[95] Li QY, Berry RB, Goetting MG, Staley B, Soto-Calderon H, Tsai SC, et al. Detection of upper airway status and respiratory events by a current generation positive airway pressure device. *Sleep*. 2015; **38**: 597-605.

[96] Li A, Chen S, Quan SF, Powers LS, Roveda JM. A deep learning-based algorithm for detection of cortical arousal during sleep. *Sleep*. 2020; **43**: zsaa120.

[97] Nilius G, Domanski U, Schroeder M, Franke K-J, Hogrebe A, Margarit L, et al. A randomized controlled trial to validate the Alice PDX ambulatory device. *Nat Sci Sleep*. 2017; **9**: 171-80.

[98] Oliveira MG, Treptow EC, Fukuda C, Nery LE, Valadares RM, Tufik S, et al. Diagnostic accuracy of home-based monitoring system in morbidly obese patients with high risk for sleep apnea. *Obesity Surg*. 2015; **25**: 845-51.

[99] Terjung S, Geldmacher J, Brato S, Werther S, Teschler H, Taube C, et al. Classification of sleep and wake using a novel minimal-contact single-channel device. *Somnologie (Berl)*. 2018; **22**: 144-51.

[100] Ward KL, McArdle N, James A, Bremner AP, Simpson L, Cooper MN, et al. A comprehensive evaluation of a two-channel portable monitor to "rule in" obstructive sleep apnea. *J Clin Sleep Med*. 2015;11(4):433-44.

Figure 1: PRISMA flowchart

Figure 2: Summary of devices

In total, we identified 42 different devices across 52 studies, including ten non-wearable and 32 wearable devices. The number in brackets (n) denotes the number of studies for each device.

ECG = electrocardiography, EEG = electroencephalography.

Figure 3: Study and cohort characteristics

A-C: These frequency histograms display the number of studies according to sample size, % female, and mean BMI for each study.

D: This plot displays the mean age and age range for each study (black diamond) with a summary measure for all studies (large blue diamond). Where an age range was not given by study authors, a reference range is displayed instead. The mean age cut-off for inclusion was 40 years.

BMI = body mass index.

Figure 4: Most devices accurately measured the AHI score

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

AHI = apnoea hypopnoea index, CHF = congestive heart failure, CI = confidence interval, COPD = chronic obstructive pulmonary disease, ECG = electrocardiography, OSA = obstructive sleep apnoea, SD = sleep disorders (mix).

Figure 5: AHI score accuracy by device sensor type

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

AHI = apnoea hypopnoea index, CHF = congestive heart failure, CI = confidence interval, COPD = chronic obstructive pulmonary disease, OSA = obstructive sleep apnoea, SD = sleep disorders (mix).

Figure 6: WatchPAT tended to slight overestimation of AHI score

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

AHI = apnoea hypopnoea index, CI = confidence interval, COPD = chronic obstructive pulmonary disease, OSA = obstructive sleep apnoea.

Figure 7 Sensitivity and specificity for detecting OSA

Higher AHI severity was associated with significantly lower sensitivity but higher specificity. Each study represents an individual data point.

AHI = apnoea hypopnoea index, OSA = obstructive sleep apnoea

Mild = AHI 5-15, Moderate = AHI 15-30, Severe = AHI >30

Figure 8: Most devices accurately measured ODI

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

CI = confidence interval, COPD = chronic obstructive pulmonary disease, ODI = oxygen desaturation index, OSA = obstructive sleep apnoea.

Figure 9: TST was significantly overestimated across devices

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

CI = confidence interval, COPD = chronic obstructive pulmonary disease, OSA = obstructive sleep apnoea, SD = sleep disorders (mix), TST = total sleep time.

Figure 10: Adding additional sensors to actigraphy improved device accuracy for TST

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

CI = confidence interval, COPD = chronic obstructive pulmonary disease, OSA = obstructive sleep apnoea, SD = sleep disorders (mix), TST = total sleep time.

Figure 11: SE was significantly overestimated across devices

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

CI = confidence interval, COPD = chronic obstructive pulmonary disease, OSA = obstructive sleep apnoea, SD = sleep disorders (mix), SE = sleep efficiency.

Figure 12 Most devices inaccurately recorded SWS duration

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG.

CI = confidence interval, OSA = obstructive sleep apnoea, SWS = slow wave sleep.

Figure 13 Devices may record REM duration accurately

Mean and 95% confidence intervals are displayed. Mean difference was calculated as the index device minus PSG. Only four out of seven studies presented mean difference with variance for REM duration or percentage, and two of these studies reported significant issues with data loss due to signal quality.

CI = confidence interval, OSA = obstructive sleep apnoea, REM = rapid eye movement, SD = sleep disorders (mix).

Figure 14 Overall risk of study bias was high

Risk was deemed 'high' if not meeting ≥1 of the signalling criteria, 'unclear' if information not available for ≥1 signalling criteria, and 'low' if meeting all signalling criteria.

QUADAS-2 = Quality assessment of diagnostic accuracy studies-2

Study no.	Author, Year (Reference)	n	Sample description	Mean age (RR or R)	BMI (kg/m^2)	Female (%)	Wearable Y/N (location)	Device name	Device sensors	Sleep metric(s) assessed	Algorithm and/or scoring software description	Funding source
1	Abad et al., 2016 [37]	50	Suspected OSA	53 (RR 26 to 81)	30	?	N (Bedside)	SleepWise	Video camera Microphone	AHI	Open access algorithm published with source paper. Video image integration, pixel comparison, and noise filtering provided binary array for respiratory event detection. Magnitude of respiratory airflow reduction was divided by mean respiratory motion and duration of event.	D

2	Alshaer et al., 2016 [42]	135	Suspected OSA	53 (R 23 to 84)	31	41	Y (Face)	BresoDx	Microphone	AHI	Open access acoustic analysis with two rules. The first detects apnoea by signal flatness (absent breath sounds for minimum 10 seconds), width (minimum gap between breaths) and depth (by baseline amplitude). Hypopnoea was detected by gradual airflow reductions (using the 'falling edge factor'), width and depth.	C

3	Andres-Blanco et al., 2017 [41]	110 & 68	Non-COPD & COPD	55 (RR 20 to 90) & 65 (RR 43 to 86)	29 & 29	31 & 12	Y (Finger)	Wrist Ox2	Oximetry	AHI	Open access regression-based multilayer perceptron artificial neural network as described in source paper.	D
4	Assefa et al., 2016 [49]	56	Suspected OSA	49 (RR 21 to 76)	38	54	Y (Oronasal)	ApneaStrip	Airflow sensor Thermal sensor	AHI	Data scored automatically with proprietary flow-based algorithms.	D
5	Boyd et al., 2016 [47]	28	Severe OSA	51 (RR 30 to 73)	36	25	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	D

6	Cheung et al., 2020 [90]	41	Suspected SDB	42 (R 19 to 72)	25	59	Y (Wrist)	Huami Arc	Actigraphy	TST	Cole-Kripke algorithm with a threshold of 10 units (estimated with training data). Time in bed is assumed to be sleep except when the activity data is of sufficient length and/or amplitude (i.e., above a wake threshold value) to infer that the participant is more likely to be awake than asleep.	D

7	Cho et al., 2017 [51]	149	Suspected OSA	43 (RR 18 to 73)	26	15	Y (Oronasal, finger, chest)	ApneaLink Plus	Pressure sensor Oximetry Airflow sensor	AHI	Analysed by three modes: automatic, automatic against AASM criteria, and manual. Apnoea and hypopnoea defined as ≥80% and ≥50% reduction, respectively, for ≥10 seconds. Proprietary automatic algorithm was generally most accurate.	D
8	Choi et al., 2017 [69]	36 & 30	SDB & Insomnia	50 (RR 39 to 77) & 58	26 & 22	17 & 80	Y (Wrist)	Actiwatch 2	Actigraphy	TST SE	Analysed automatically by proprietary Actiware version 5.70 sleep software.	D

				(RR 21 to 78)								
9	Crinion et al., 2020 [52]	67	Suspected OSA	52 (RR 27 to 77)	31	29	N (Bedside)	SleepMinder	Pressure sensor	AHI	Automatic analysis by proprietary algorithm which required ≥50% reduction in non-contact motion amplitude for ≥10 seconds to score a respiratory event.	E
10	Dafna et al., 2015 [67]	70	Sleep clinic patients	55 (R 19 to 82)	31	39	N (Bedside)	Edirol R-4 pro	Microphone	TST SE	Adaboost algorithm using time-series model to classify 30 second epochs.	D

11	De Zambotti et al., 2015 [91]	28	Healthy and insomnia	50 (RR 42 to 58)	25	100	Y (Wrist)	Jawbone UP	Actigraphy	TST	Sleep-wake scoring by automatic proprietary algorithm	D
12	Dunican et al., 2018 [70]	50	Healthy	57 (R 46 to 73)	27	60	Y (Wrist)	Readiband v3 (1) GT3X+ (2)	Actigraphy	TST SE	Readiband analysed by Readiband Sync proprietary algorithm. GT3X analysed by Actilife 6 software with Cole algorithm.	D
13	Edouard et al., 2021 [92]	118	Suspected OSA	49 (RR 26 to 73)	33	57	N (Mattress)	Withings Sleep Analyzer	Pressure sensor Microphone	AHI TST SE	Proprietary deep-learning algorithm that uses body movement, breathing patterns, cardiac activity, and	B

										SWS	snoring to estimate AHI.	
14	Gruwez et al., 2019 [65]	22	Confirmed OSA	53 (RR 28 to 78)	31	23	Y (Upper arm [1], wrist [2 + 3])	SenseWear Pro (1) Jawbone UP (2) Withings Pulse O2 (3)	Actigraphy	TST SE SWS	Automatic scoring by proprietary algorithms.	D
15	Gu et al., 2020 [43]	50	Healthy	55 (RR 23 to 87)	31	46	Y (Finger)	Belun Ring	Plethysmography Oximetry Actigraphy	AHI TST	Automatic scoring by proprietary cloud based neural networks algorithm.	A + B
16	Hayano et al.,	41	Confirmed SDB	48 (RR 42 to 58)	25	17	Y (Wrist)	E4 wristband	Plethysmography	AHI	Analysed by auto-correlated wave detection with adaptive	F

	2020 [93]										threshold algorithm developed by heart variability signals from ECG.	
17	Holmedahl et al., 2019 [36]	16	COPD	61 (RR 44 to 79)	26	56	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI TST SE SWS REM	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	B
18	Ioachimescu et al., 2020 [50]	500	Suspected OSA	53 (R 24 to 92)	32	20	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	D

19	Jen et al., 2020 [57]	33	COPD	63 (RR 49 to 77)	28	39	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI ODI	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	D
20	Kahawage et al., 2020 [94]	42	Insomnia	49 (R 19 to 82)	?	55	Y (Wrist)	Actiwatch Spectrum Pro (1) Fitbit Alta HR (2)	Actigraphy	TST SE SWS REM	Automatic scoring by proprietary algorithm. 30s epochs.	D
21	Kapella et al., 2017 [68]	50	COPD	63 (RR 47 to 80)	27	30	Y (Wrist)	Actiwatch 2	Actigraphy	TST SE	Actiware 6.0.8. Validated algorithm that codes epochs as sleep or wake by examining activity	D

											counts for two minutes surrounding the epoch.	
22	Kasai et al., 2020 [58]	120	Suspected SDB	58 (RR 35 to 81)	26	15	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI	Automatic scoring by zzzPAT 4.3.62 (proprietary) algorithm. 30s epochs.	A + B
23	Kim et al., 2019 [53]	116	Habitual snorers	50 (RR 18 to 83)	26	33	N (Bedside)	SURP-102	Microphone	AHI	10-fold cross-validation using simple logistic regression. Features extracted with jAudio, a Java- based audio feature extraction software.	D

24	Kogure et al., 2017 [40]	70	Suspected OSA	49 (R 20 to 80)	26	17	N (Bedsheet)	Nemuri scan	Pressure sensor	AHI	Respiratory events scores automatically and defined as ≥30% reduction for ≥10 seconds or consecutive increases in respiratory effort (x5 times baseline). TST algorithm open access - like Cole-Kripke algorithm.	B
25	Levendowski et al., 2017 [34]	47	Suspected SDB or healthy	61 (RR 32 to 90)	27	62	Y (Forehead)	Sleep Profiler	Electroencephalography (Fp1, Fp2, FpZ)	TST SWS REM	Automatic scoring by Sleep Profiler algorithm.	B

									Plethysmography Microphone			
26	Li et al., 2015 [95]	45	Confirmed OSA	53 (RR 26 to 80)	34	36	Y (Facemask positive airway pressure device)	System One REMstar	Airflow sensor	AHI	Analyse-it Software, Ltd. version 2.26.	A + B
27	Li et al., 2020 [96]	43	Participants from the Multi-Ethnic Study of	69 (R 56 to 87)	?	51	Y (Wrist)	Actiwatch Spectrum	Actigraphy	TST SE	Hidden Markov Model (unsupervised & best performer), also tested Actiwatch software & pre-trained supervised UCSD algorithm.	D

			Atherosclerosis									
28	Li et al., 2017 [44]	49	Suspected OSA	44 (RR 18 to 77)	27	22	Y (Finger)	Morpheus Ox	Oximetry Plethysmography	AHI ODI TST	MedCalc version 16.2.0 (MedCalc Software BVBA).	D
29	Lu et al., 2019 [39]	179	Suspected OSA with cardiovascular disease	45 (R 21 to 72)	28	85	Y (Chest)	Single-lead ECG	Electrocardiography	AHI	Scored using open access algorithm. ECG-derived respiratory signal and inter-beat intervals. Coherences and cross-power of these signals Fourier transformed to determine high and low frequency cardiopulmonary	D

											coupling regimes which correlate with different sleep stages.	
30	Lucey et al., 2016 [33]	29	Referrals to sleep clinic or recruited from ageing study	54 (R 25 to 80)	?	41	Y (Forehead)	Sleep Profiler	Electroencephalography (Fp1-Fp2) Plethysmography Microphone	TST SE REM	Manually scored using Sleep Profiler Manual.	D
31	Madrid-Navarro et al., 2019 [66]	70	Referrals to sleep clinic	56 (RR 25 to 86)	29	37	Y (Wrist)	Kronowise 3.0	Thermal sensor Actigraphy Light sensor	TST SE	TAPL algorithm to distinguish sleep and wake. Keywake algorithm to mark sleep epochs	D

											Kronoware 10.0 software Circadianware All implemented on the Kronowizard cloud platform.	
32	Massie et al., 2018 [45]	101	Referrals to sleep clinic	53 (RR 28 to 78)	29	44	Y (Finger)	Night Owl	Actigraphy Plethysmography	AHI TST	Analysed using proprietary Night Owl algorithm. PSG and NightOwl data were synchronized by matching heart rate traces.	B
33	Meng et al., 2016 [46]	131	Suspected OSA	44 (RR 21 to 68)	27	16	Y (Finger) + N	Unnamed system	Pressure sensor Oximetry	AHI	Respiratory event reductions in respiratory movement	D

							(Bedsheet)				amplitude of ≥30% (hypopnoea) or ≥50% (apnoea) for ≥10 seconds. Closed access customized algorithm.	
34	Miyata et al., 2020 [62]	189	Referrals to sleep clinic	56 (RR 20 to 92)	25	30	N (Bedsheet)	SD102	Pressure sensor	AHI	Respiratory event determined by decrease in pressure alteration of >30% of mean respiratory pressures for ≥10 seconds. Proprietary.	D
35	Moreno-Pino et al., 2019 [71]	65	Confirmed OSA	59 (RR 32 to 86)	30	35	Y (Wrist)	Fitbit Charge 2 (1) Fitbit	Actigraphy Plethysmography	TST SWS REM	Both devices scored automatically using proprietary algorithm.	D

								Alta HR (2)				
36	Munoz-Ferrer et al., 2020 [38]	38	Suspected OSA	47 (RR 25 to 69)	31	63	N (Bedside)	SleepWise	Microphone	AHI	Open access algorithm as per Abad et al., 2016 (above).	D
37	Nilius et al., 2017 [97]	85	Suspected OSA	49 (RR 23 to 76)	30	20	Y (Finger, oronasal, thorax, abdomen)	Alice PDx	Oximetry Airflow sensor Thermal sensor Pressure sensor	AHI	Not specified.	D
38	Oliveira et al.,	32	Suspected OSA	49 (RR 28 to 71)	41	44	Y (Finger,	StarDust II	Pressure sensor	AHI	Not specified.	D

	2015 [98]						oronasal, chest)		Airflow sensor Actigraphy Plethysmography			
39	Pillar et al., 2020 [59]	84	Suspected OSA +/- CHF	57 (R 22 to 83)	30	36	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI TST	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	A
40	Pinheiro et al., 2020 [48]	304	Suspected OSA	55 (R 18 to 90)	31	44	Y (Finger)	Oxistar	Oximetry Actigraphy	AHI ODI	Automatic scoring by proprietary algorithm.	B

41	Ribeiro et al., 2015 [56]	30	Suspected OSA	43 (R 24 to 71)	?	33	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI SWS REM	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	D
42	Savage et al., 2016 [76]	47	CHF	68 (R 41 to 84)	29	17	N (Bedside)	SleepMinder	Pressure sensor	AHI	Proprietary algorithm which required ≥50% reduction in non-contact motion amplitude for ≥10 seconds to score a respiratory event.	A
43	Smith et al., 2020 [63]	100	Suspected OSA	55 (RR 22 to 88)	35	51	Y (Oronasal, finger)	Apnea Link	Airflow sensor Oximetry	AHI ODI	Automatic scoring by proprietary software.	D

44	Spielmanns et al., 2019 [35]	26	Healthy	55 (R 27 to 86)	30	35	Y (Wrist)	Polar A300	Actigraphy	TST	Not specified.	D
45	Tal et al., 2017 [64]	43	Suspected SDB	46 (R 18 to 72)	35	21	N (Mattress)	EarlySense	Pressure sensor	TST SWS REM	Not specified.	A
46	Terjung et al., 2018 [99]	49	Suspected SDB	56 (R 23 to 82)	30	70	N (Mattress)	VitaLog	Pressure sensors	TST SE	Modified version of open access algorithm – 30s epochs assigned using modified Gorny algorithm	D
47	Tiron et al., 2020 [54]	120	Healthy	54 (RR 25 to 79)	28	32	N (Bedside)	Firefly	Microphone	AHI	Proprietary Firefly AI/ML algorithms as described in source paper.	B

48	Tondo et al., 2021 [60]	47	Low risk of OSA	52 (RR 24 to 80)	26	38	Y (Wrist, finger, chest)	WatchPAT	Plethysmography Oximetry Microphone Actigraphy	AHI ODI	Automatic scoring by zzzPAT (proprietary) algorithm. 30s epochs.	D
49	Topor et al., 2020 [61]	27	Suspected SDB	55 (R 40 to 71)	?	29	Y (Finger, chest, abdomen)	MATRx Plus	Actigraphy Plethysmography Oximetry Airflow sensor Pressure sensor	AHI	Proprietary algorithm which identifies respiratory events by changes in minute ventilation.	A + B
50	Ward et al., 2015 [100]	104	Suspected OSA	51 (RR 24 to 77)	31	39	Y (Oronasal, finger)	ApneaLink Ox	Pressure sensor Oximetry	AHI	Initially analysed by proprietary Apnea Link Software version 8.00.	A

											Data then manually reviewed after importing ApneaLink signals. Apnoea defined as reduction in airflow of ≥80% for minimum 10 seconds and hypopnoea as ≥50% or ≥30% plus 3% oxygen desaturation for minimum 10 seconds.	
51	Weinreich et al., 2018 [77]	57	Suspected sleep disorders	56 (RR 29 to 84)	30	19	N (Bedside)	SleepMinder	Pressure sensor	AHI	Automatic analysis by proprietary SleepMinder software, which applies bandpass filtering to movement	B

											signals to differentiate body movement from respiration and assign 30s epochs.	
52	Zou et al., 2015 [55]	93	Healthy & suspected OSA	43 (RR 18 to 69)	27	25	Y (Oronasal, finger)	SleepView	Pressure sensor Oximetry	AHI ODI	Automatic analysis by proprietary SleepView software.	D

Table 1: Summary of studies included for review

Where range (R) was not given, a 95% reference range (RR) was calculated using the given standard deviation instead.

AASM = American academy of sleep medicine, AI/ML = artificial intelligence/machine learning, COPD = chronic obstructive pulmonary disease, CHF = congestive heart failure, ECG = electrocardiography, EEG = electroencephalography, OSA = obstructive sleep apnoea, REM = rapid eye movement, R = range, RR = reference range, SDB = sleep disordered breathing, TAPL = thermometry, actigraphy, position, light, TST = total sleep time.

Funding sources:

A = funded by company that helps produce, market, or distribute the device

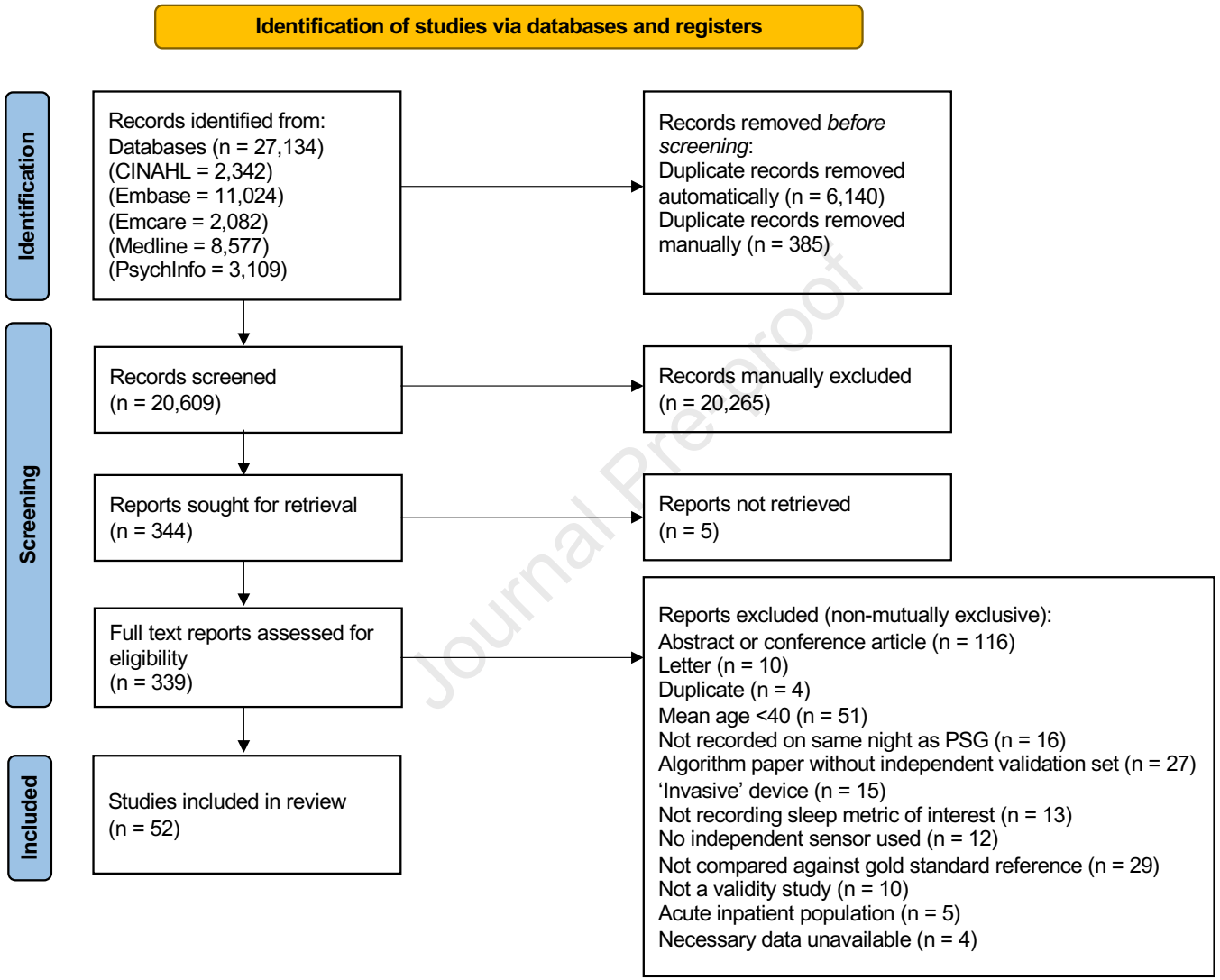
B = one or more authors is a shareholder, sponsored by, or an employee of the company that makes device

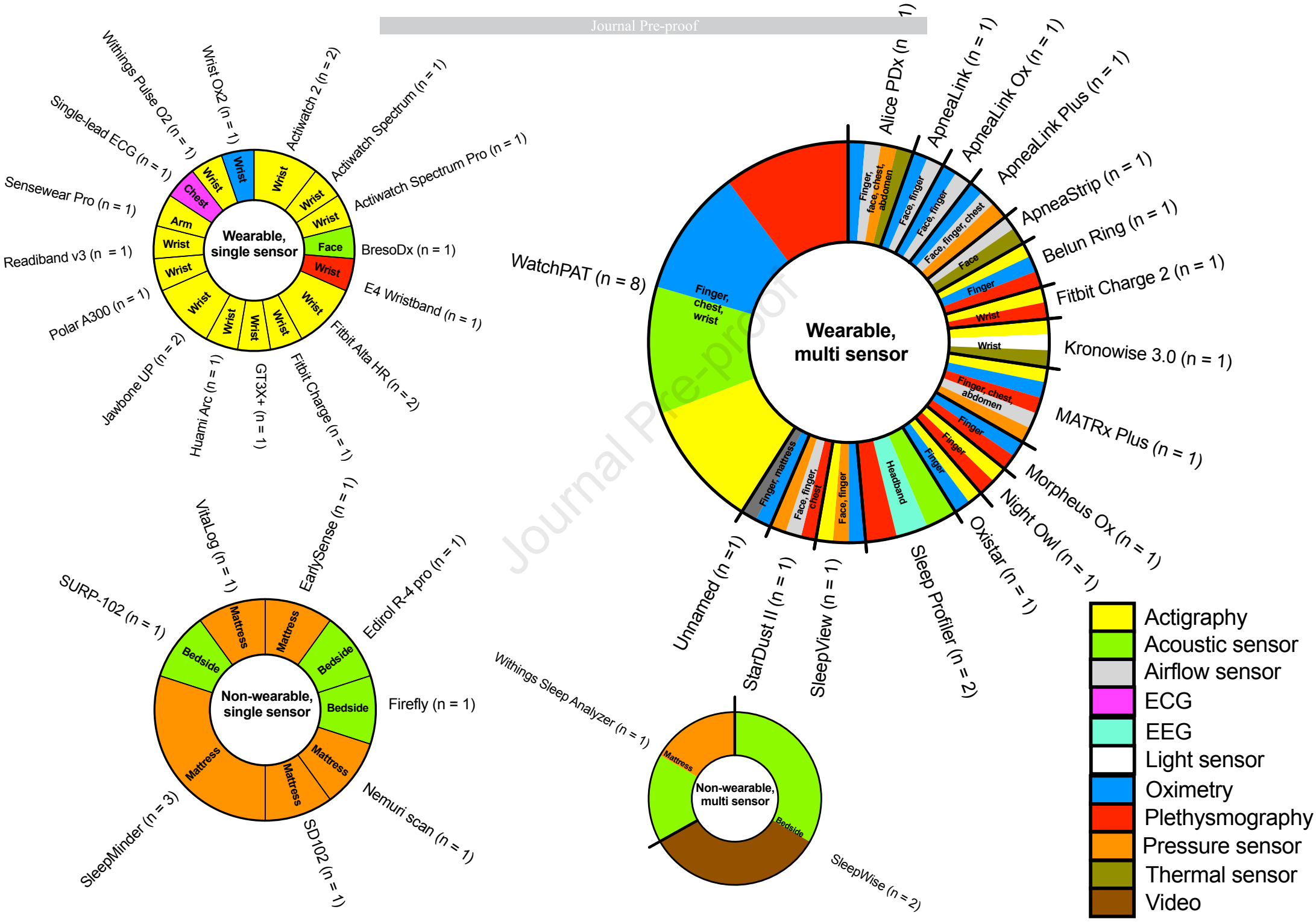
C = authors own company that makes device

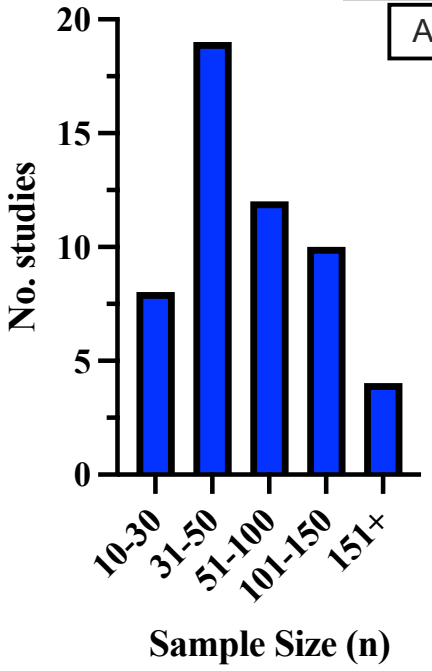
D = no conflicts of interest declared

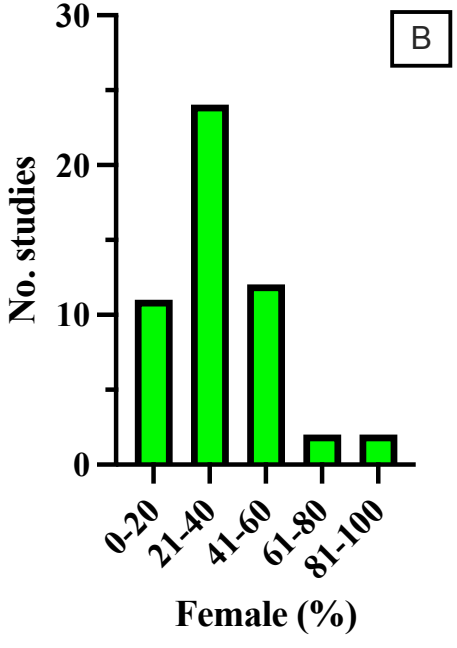
E = not specified

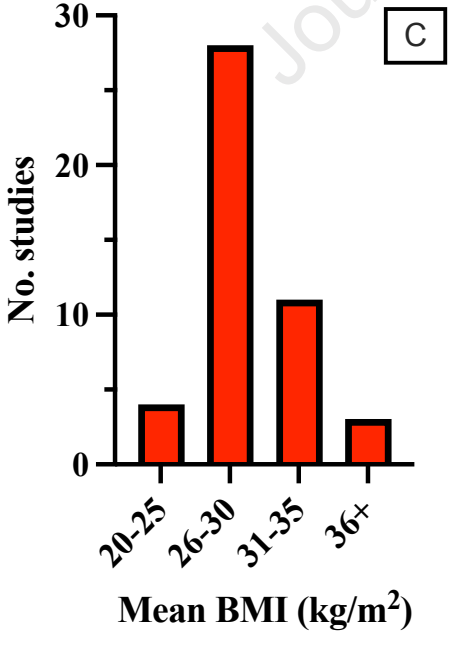
F = other

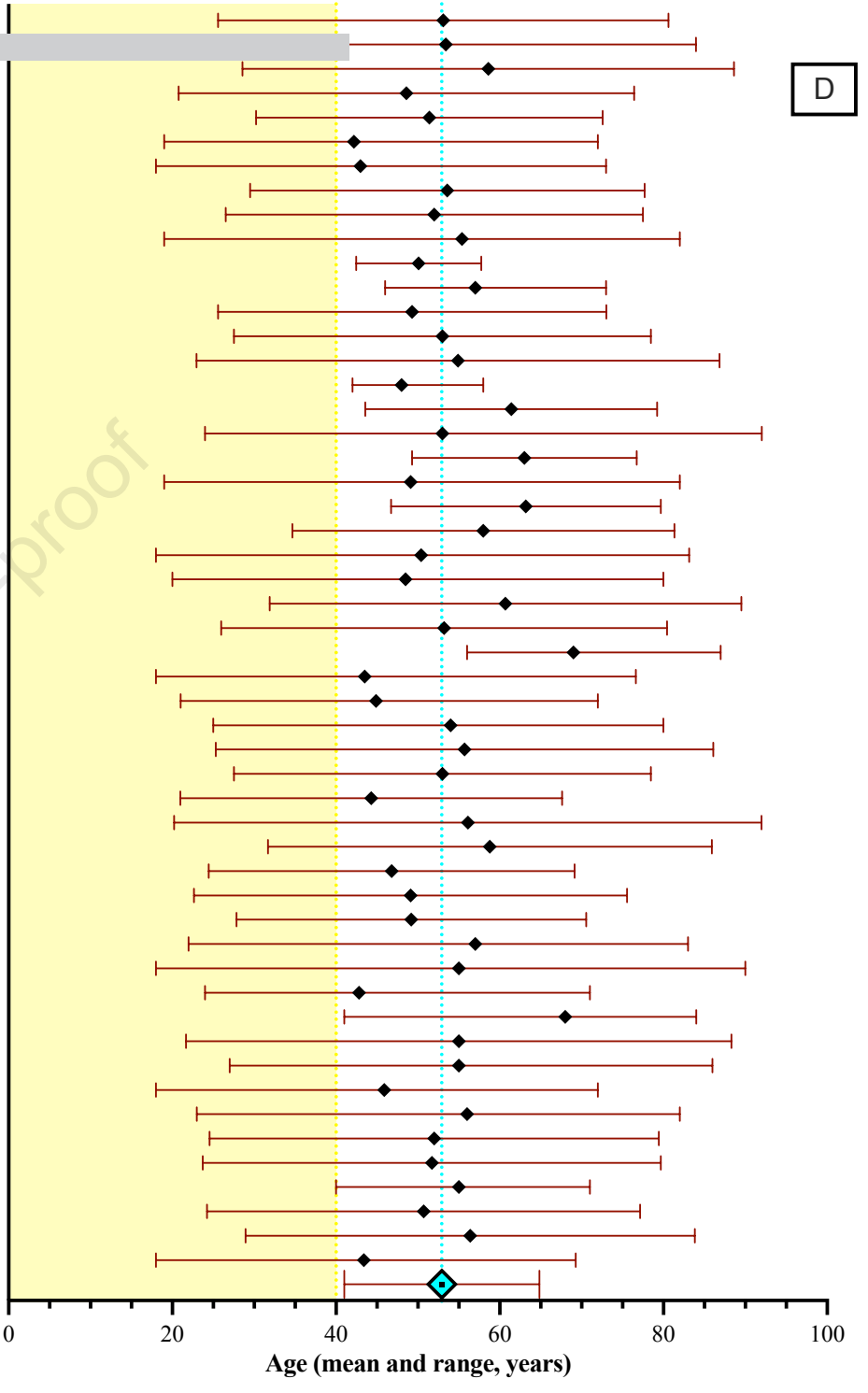


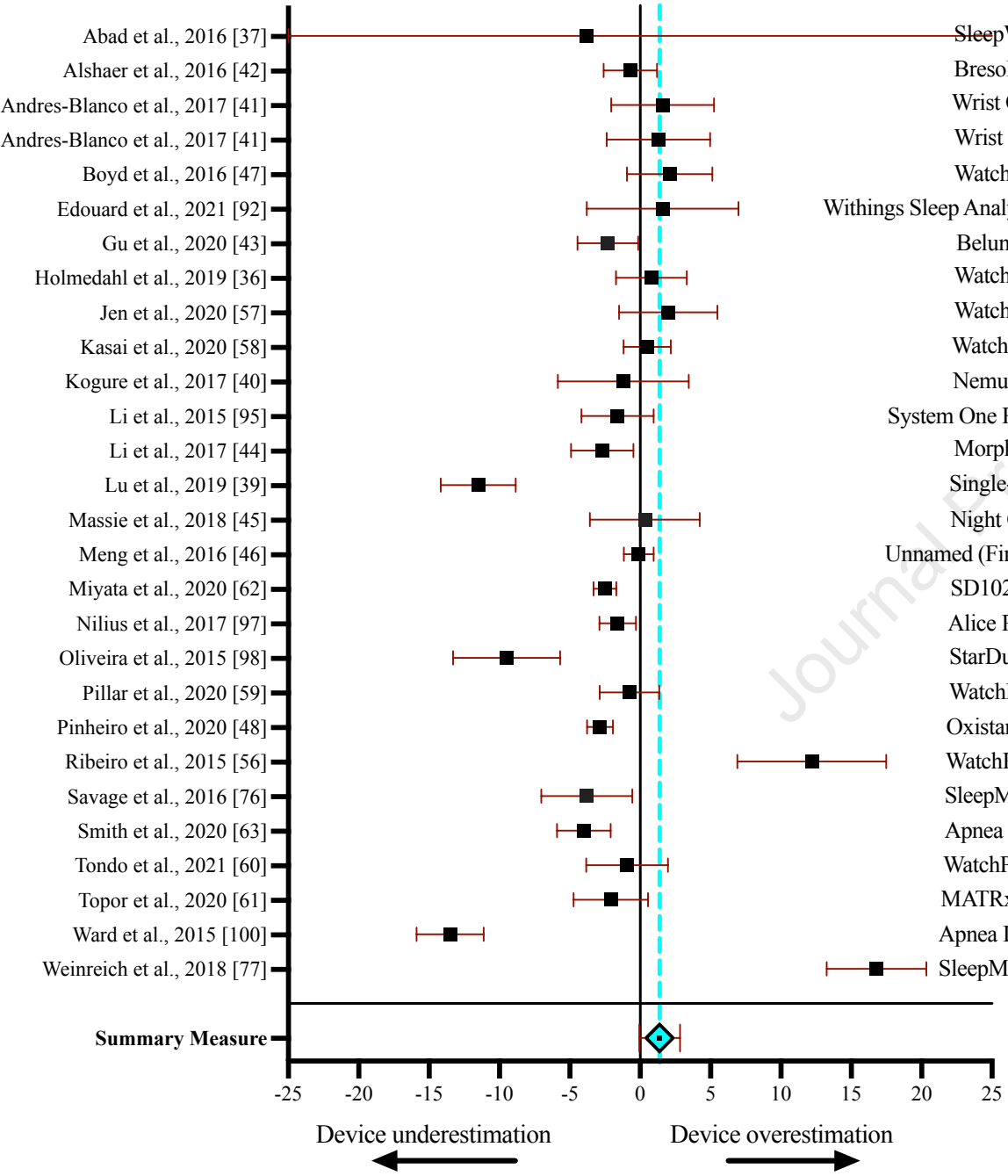






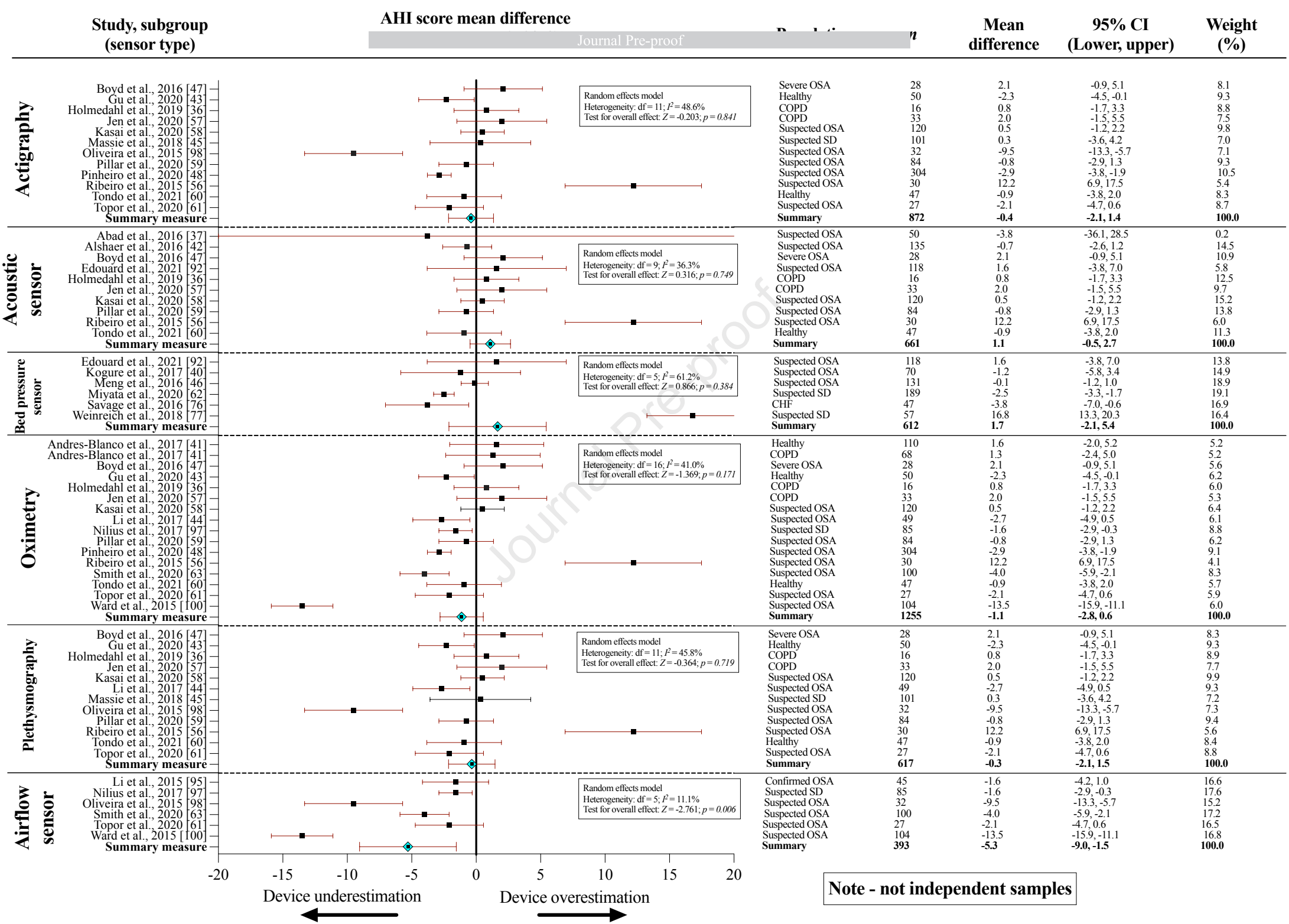


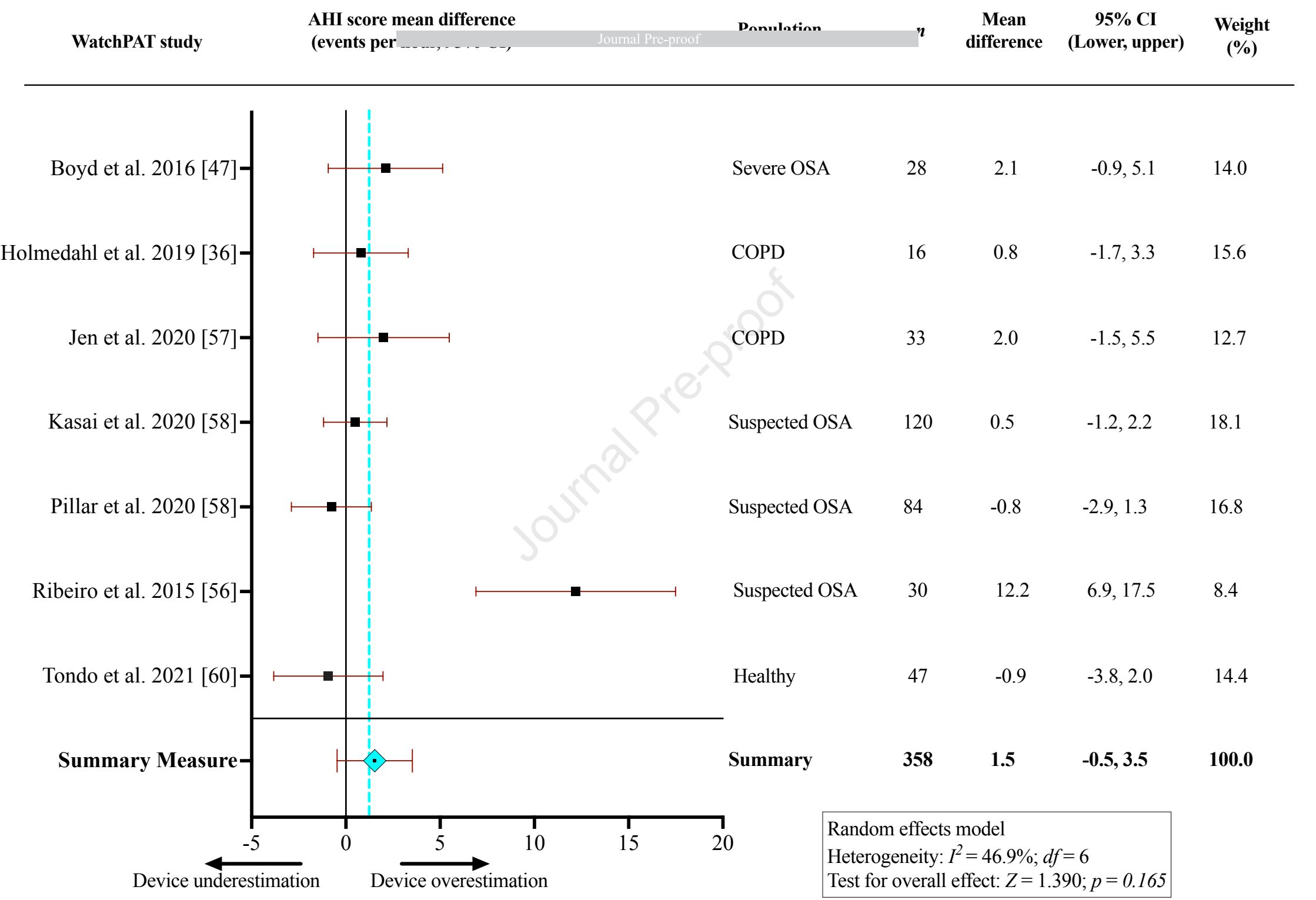


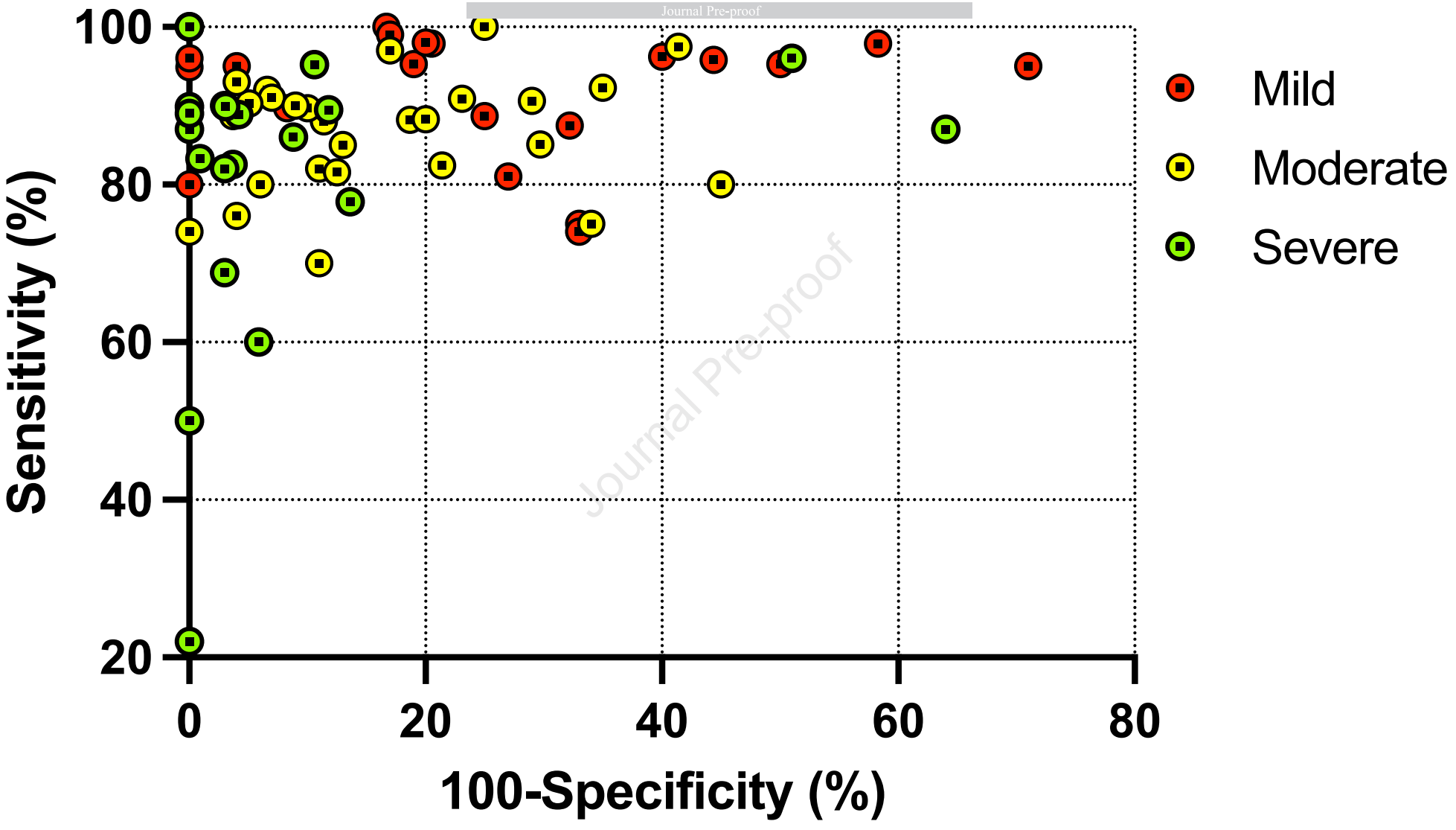


Study	AHI score mean difference (events per hour, 95% CI)	Device name (Location)	Population	*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Abad et al., 2016 [37]		SleepWise (Bedside)	Suspected OSA	50	-3.8	-36.1, 28.5	0.2
Alshaer et al., 2016 [42]		BresoDx (Face)	Suspected OSA	135	-0.7	-2.6, 1.2	4.0
Andres-Blanco et al., 2017 [41]		Wrist Ox2 (Finger)	Healthy	110	1.6	-2.0, 5.2	3.4
Andres-Blanco et al., 2017 [41]		Wrist Ox2 (Finger)	COPD	68	1.3	-2.4, 5.0	3.4
Boyd et al., 2016 [47]		WatchPAT (Multiple)	Severe OSA	28	2.1	-0.9, 5.1	3.6
Edouard et al., 2021 [92]		Withings Sleep Analyzer (Mattress)	Suspected OSA	118	1.6	-3.8, 7.0	2.7
Gu et al., 2020 [43]		Belun Ring (Finger)	Healthy	50	-2.3	-4.5, -0.1	4.0
Holmedahl et al., 2019 [36]		WatchPAT (Multiple)	COPD	16	0.8	-1.7, 3.3	3.8
Jen et al., 2020 [57]		WatchPAT (Multiple)	COPD	33	2.0	-1.5, 5.5	3.5
Kasai et al., 2020 [58]		WatchPAT (Multiple)	Suspected OSA	120	0.5	-1.2, 2.2	4.1
Kogure et al., 2017 [40]		Nemuri Scan (Mattress)	Suspected OSA	70	-1.2	-5.8, 3.4	3.0
Li et al., 2015 [95]		System One REMStar (Face)	Confirmed OSA	45	-1.6	-4.2, 1.0	3.8
Li et al., 2017 [44]		Morpheus Ox (Finger)	Suspected OSA	49	-2.7	-4.9, 0.5	3.9
Lu et al., 2019 [39]		Single-lead ECG (Chest)	Suspected OSA	179	-11.5	-14.2, 8.8	3.8
Massie et al., 2018 [45]		Night Owl (Finger)	Suspected SD	101	0.3	-3.6, 4.2	3.3
Meng et al., 2016 [46]		Unnamed (Finger/Mattress)	Suspected OSA	131	-0.1	-1.2, 1.0	4.2
Miyata et al., 2020 [62]		SD102 (Mattress)	Suspected SD	189	-2.5	-3.3, 1.7	4.3
Nilius et al., 2017 [97]		Alice PDx (Multiple)	Suspected SD	85	-1.6	-2.9, -0.3	4.2
Oliveira et al., 2015 [98]		StarDust II (Multiple)	Suspected OSA	32	-9.5	-13.3, -5.7	3.3
Pillar et al., 2020 [59]		WatchPAT (Multiple)	Suspected OSA	84	-0.8	-2.9, 1.3	4.0
Pinheiro et al., 2020 [48]		Oxistar (Finger)	Suspected OSA	304	-2.9	-3.8, -1.9	4.3
Ribeiro et al., 2015 [56]		WatchPAT (Multiple)	Suspected OSA	30	12.2	6.9, 17.5	2.7
Savage et al., 2016 [76]		SleepMinder (Bedside)	CHF	47	-3.8	-7.0, -0.6	3.6
Smith et al., 2020 [63]		Apnea Link (Multiple)	Suspected OSA	100	-4.0	-5.9, -2.1	4.0
Tondo et al., 2021 [60]		WatchPAT (Multiple)	Healthy	47	-0.9	-3.8, 2.0	3.7
Topor et al., 2020 [61]		MATRx Plus (Multiple)	Suspected OSA	27	-2.1	-4.7, 0.6	3.8
Ward et al., 2015 [100]		Apnea Link Ox (Multiple)	Suspected OSA	104	-13.5	-15.9, -11.1	3.9
Weinreich et al., 2018 [77]		SleepMinder (Bedside)	Suspected SD	57	16.8	13.3, 20.3	3.4
Summary Measure			**Summary**	**2,409**	**1.4**	**-0.1, 2.8**	**100.0**

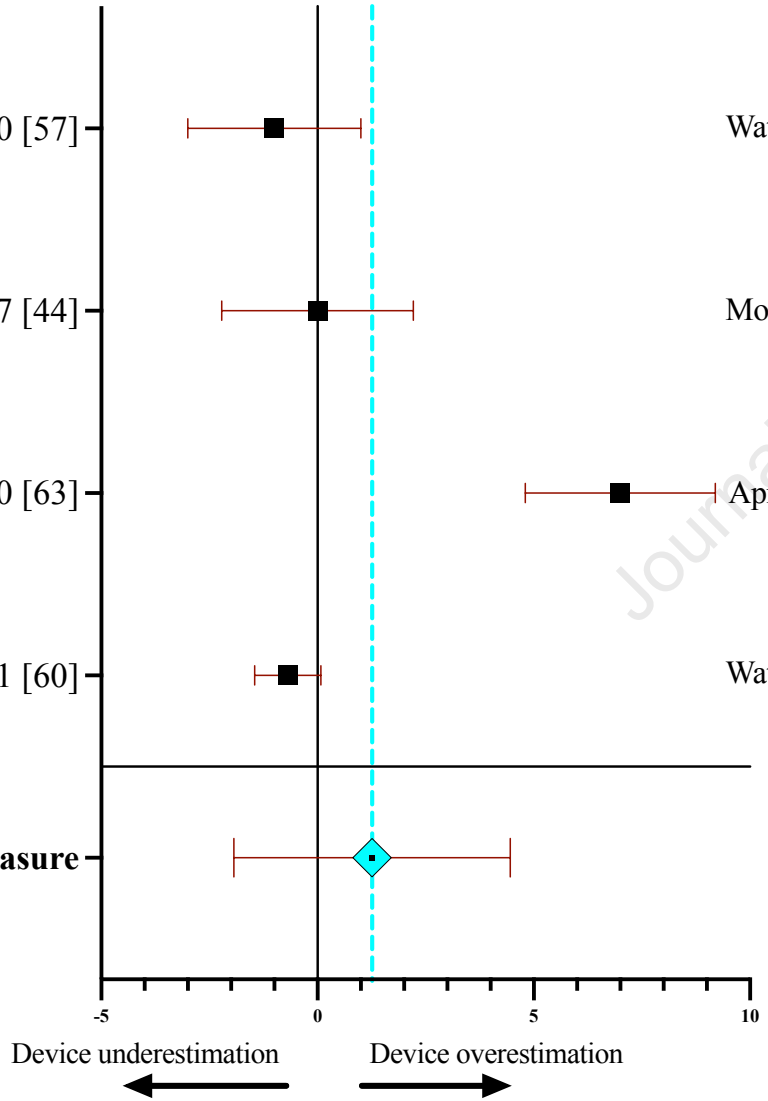
Random effects model
Heterogeneity: $I^2 = 51.3\%$; $df = 27$
Test for overall effect: $Z = 1.872$; $p = 0.061$



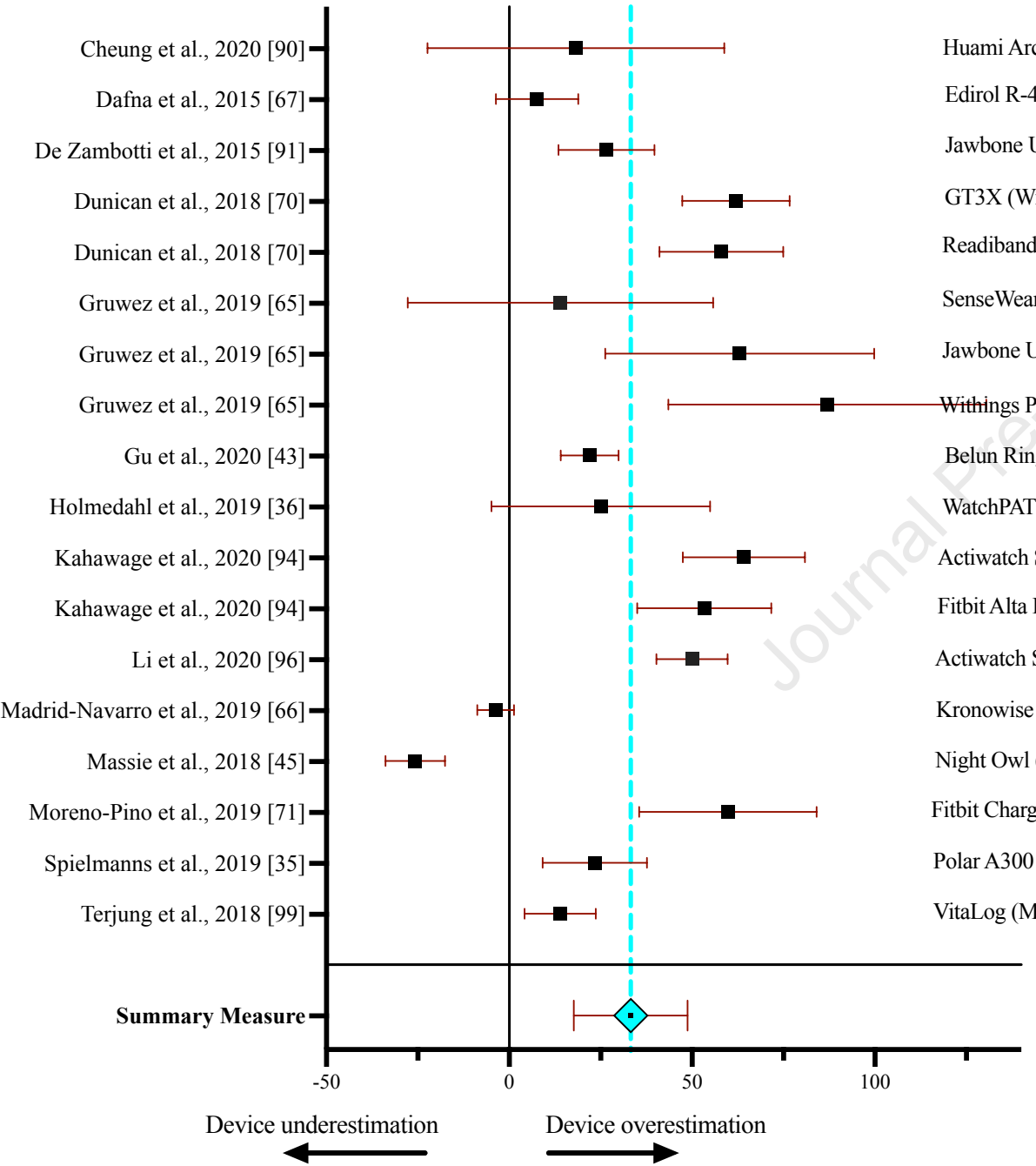




Study	ODI mean difference (number of 3% desaturations per hour, 95% CI)	Device name (Location)	Population	*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Jen et al., 2020 [57]		WatchPAT (Multiple)	COPD	33	-1.0	-3.0, 1.0	24.7
Li et al., 2017 [44]		Morpheus Ox (Finger)	Suspected OSA	49	0.0	-2.2, 2.2	24.2
Smith et al., 2020 [63]		Apnea Link (Multiple)	Suspected OSA	100	7.0	4.8, 9.2	24.2
Tondo et al., 2021 [60]		WatchPAT (Multiple)	Healthy	47	-0.7	-1.5, 0.1	26.9
Summary Measure			**Summary**	**229**	**1.26**	**-1.9, 4.5**	**100.0**

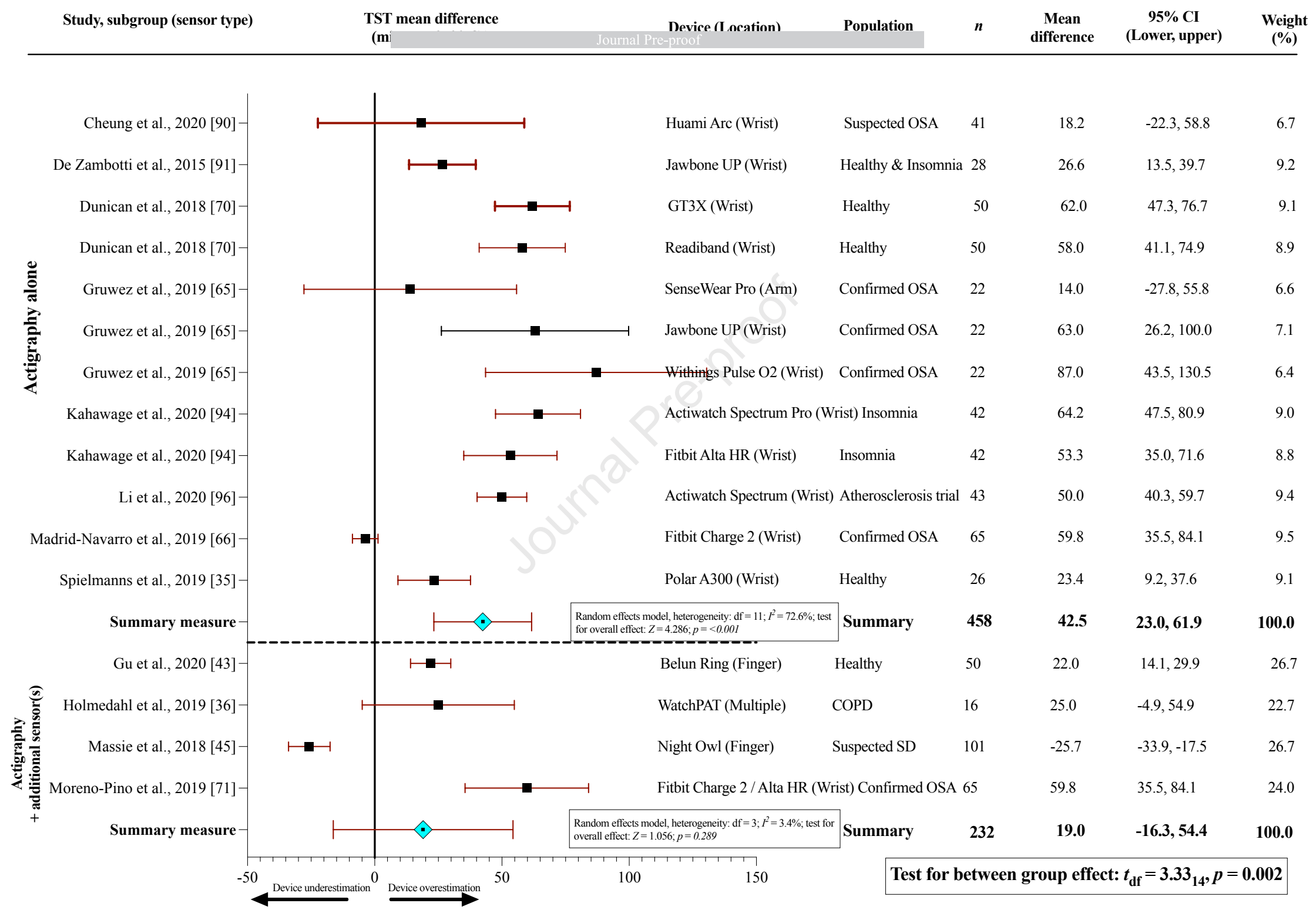


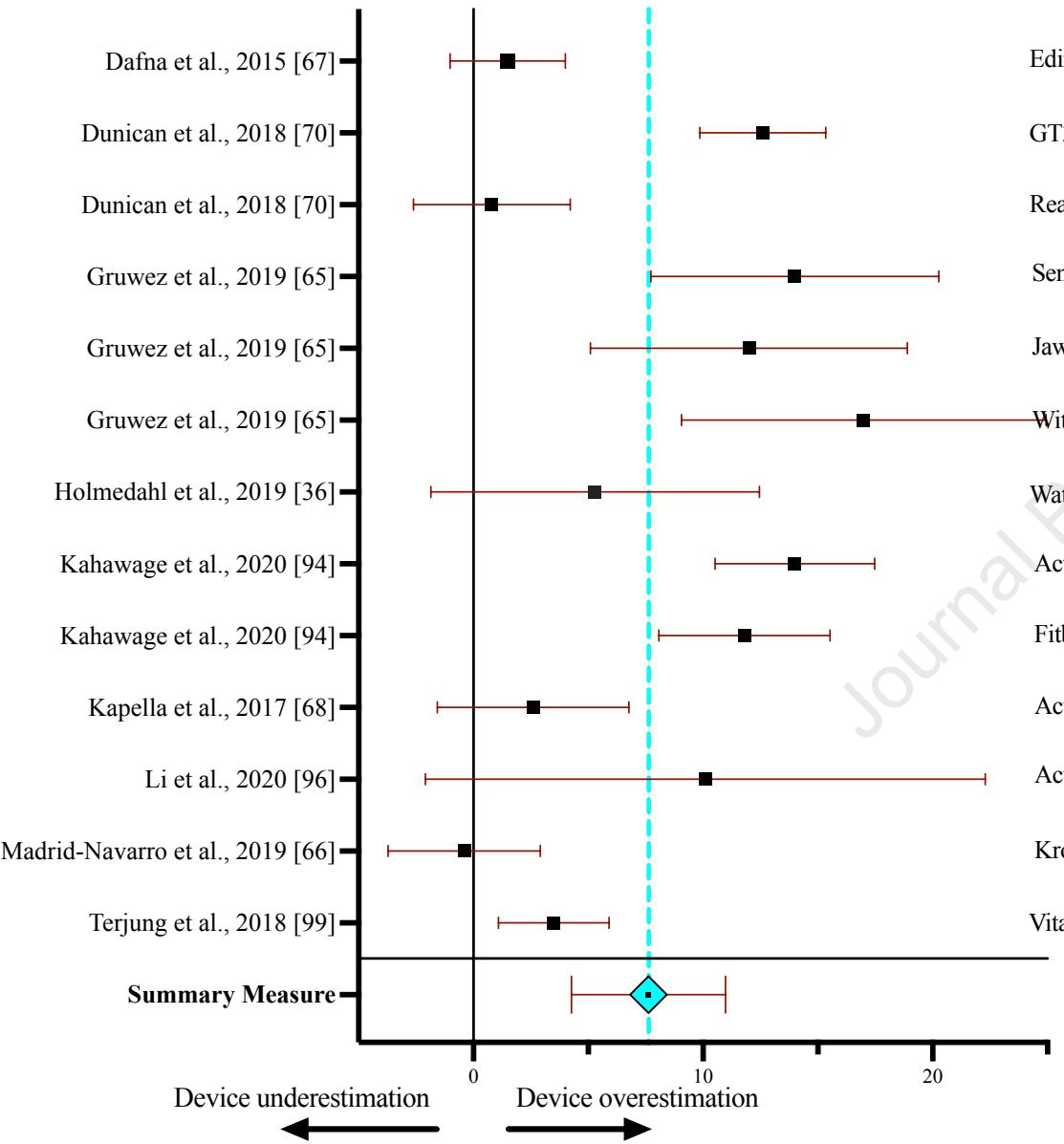
Random effects model
Heterogeneity: $I^2 = 25.2\%$; $df = 3$
Test for overall effect: $Z = 0.774$; $p = 0.441$



Study	TST mean difference (minutes, 95% CI)	Device name (Location)	Population	*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Cheung et al., 2020 [90]		Huami Arc (Wrist)	Suspected OSA	41	18.2	-22.3, 58.8	4.4
Dafna et al., 2015 [67]		Edirol R-4 pro (Bedside)	Confirmed SD	70	7.6	-3.67, 18.9	6.0
De Zambotti et al., 2015 [91]		Jawbone UP (Wrist)	Healthy & Insomnia	28	26.6	13.5, 39.7	6.0
Dunican et al., 2018 [70]		GT3X (Wrist)	Healthy	50	62.0	47.3, 76.7	5.9
Dunican et al., 2018 [70]		Readiband (Wrist)	Healthy	50	58.0	41.1, 74.9	5.8
Gruwez et al., 2019 [65]		SenseWear Pro (Arm)	Confirmed OSA	22	14.0	-27.8, 55.8	4.3
Gruwez et al., 2019 [65]		Jawbone UP (Wrist)	Confirmed OSA	22	63.0	26.2, 100.0	4.6
Gruwez et al., 2019 [65]		Withings Pulse O2 (Wrist)	Confirmed OSA	22	87.0	43.5, 130.5	4.2
Gu et al., 2020 [43]		Belun Ring (Finger)	Healthy	50	22.0	14.1, 29.9	6.1
Holmedahl et al., 2019 [36]		WatchPAT (Multiple)	COPD	16	25.0	-4.9, 54.9	5.1
Kahawage et al., 2020 [94]		Actiwatch Spectrum Pro (Wrist)	Insomnia	42	64.2	47.5, 80.9	5.8
Kahawage et al., 2020 [94]		Fitbit Alta HR (Wrist)	Insomnia	42	53.3	35.0, 71.6	5.7
Li et al., 2020 [96]		Actiwatch Spectrum (Wrist)	Atherosclerosis trial	43	50.0	40.3, 59.7	6.1
Madrid-Navarro et al., 2019 [66]		Kronowise 3.0 (Wrist)	Suspected SD	70	-3.7	-8.8, 1.3	6.2
Massie et al., 2018 [45]		Night Owl (Finger)	Suspected SD	101	-25.7	-33.9, -17.5	6.1
Moreno-Pino et al., 2019 [71]		Fitbit Charge 2 / Alta HR (Wrist)	Confirmed OSA	65	59.8	35.5, 84.1	5.4
Spielmanns et al., 2019 [35]		Polar A300 (Wrist)	Healthy	26	23.4	9.2, 37.6	5.9
Terjung et al., 2018 [99]		VitaLog (Mattress)	Suspected OSA	49	13.9	4.2, 23.6	6.1
Summary Measure		**Summary**		**809**	**33.2**	**17.6, 48.8**	**100.0**

Random effects model
Heterogeneity: $I^2 = 36.2\%$; $df = 17$
Test for overall effect: $Z = 4.182$; $p = <0.001$

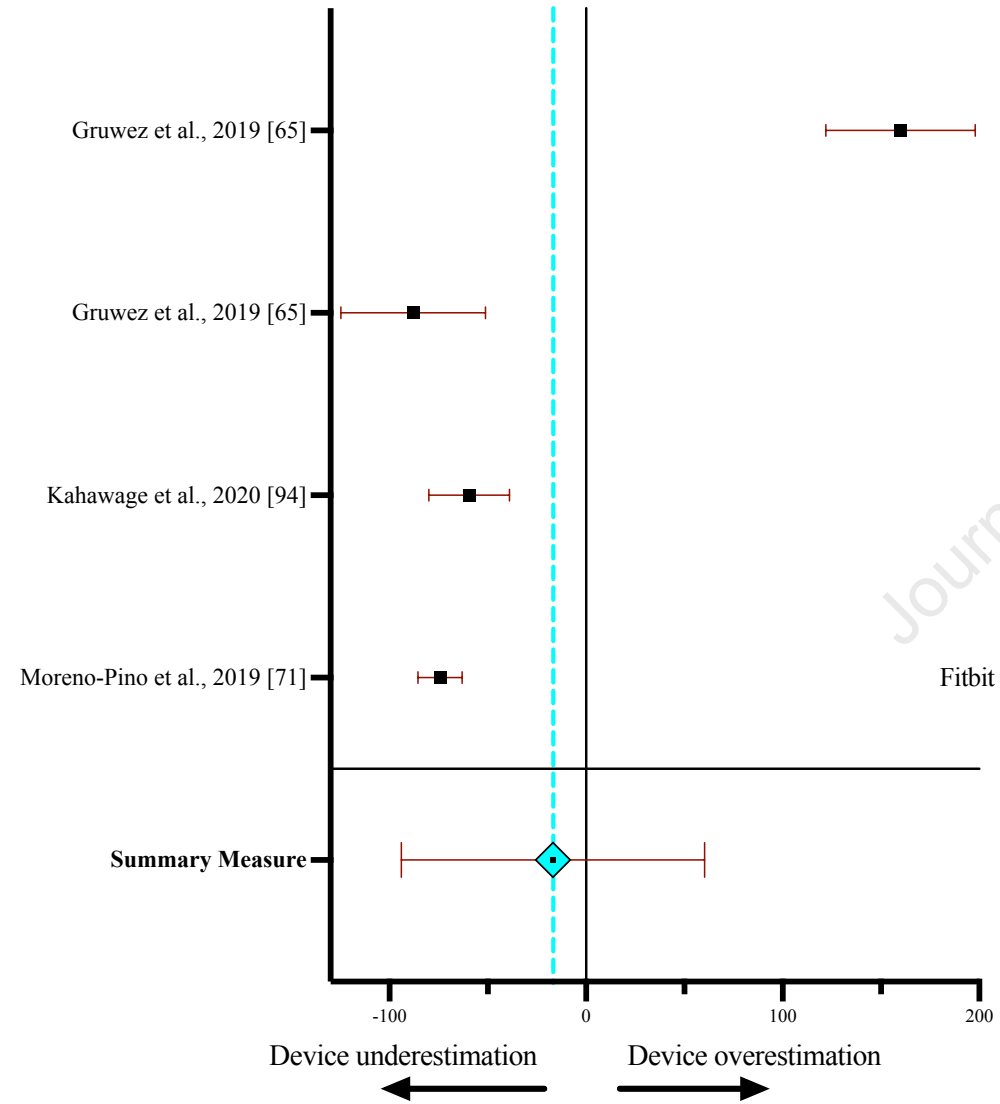




Study	SE mean difference (%, 95	Device name (Location)	Population	*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Dafna et al., 2015 [67]		Edirol R-4 pro (Bedside)	Confirmed SD	70	1.5	-1.0, 4.0	8.9
Dunican et al., 2018 [70]		GT3X (Wrist)	Healthy	50	12.6	9.9, 15.3	8.9
Dunican et al., 2018 [70]		Readiband (Wrist)	Healthy	50	0.8	-2.6, 4.2	8.6
Gruwez et al., 2019 [65]		SenseWear Pro (Arm)	Confirmed OSA	22	14.0	7.7, 20.3	7.1
Gruwez et al., 2019 [65]		Jawbone UP (Wrist)	Confirmed OSA	22	12.0	5.1, 18.9	6.7
Gruwez et al., 2019 [65]		Withings Pulse O2 (Wrist)	Confirmed OSA	22	17.0	9.1, 24.9	6.2
Holmedahl et al., 2019 [36]		WatchPAT (Multiple)	COPD	16	5.3	-1.9, 12.5	6.6
Kahawage et al., 2020 [94]		Actiwatch Spectrum Pro (Wrist)	Insomnia	42	14.0	10.5, 17.5	8.6
Kahawage et al., 2020 [94]		Fitbit Alta HR (Wrist)	Insomnia	42	11.8	8.1, 15.5	8.4
Kapella et al., 2017 [68]		Actiwatch 2 (Wrist)	COPD	50	2.6	-1.6, 6.8	8.2
Li et al., 2020 [96]		Actiwatch Spectrum (Wrist)	Atherosclerosis trial	43	10.1	-2.1, 22.3	4.2
Madrid-Navarro et al., 2019 [66]		Kronowise 3.0 (Wrist)	Suspected SD	70	-0.4	-3.7, 2.9	8.6
Terjung et al., 2018 [99]		VitaLog (Mattress)	Suspected OSA	49	3.5	1.1, 5.9	9.0
			Summary	**548**	**7.6**	**4.3, 11.0**	**100.0**

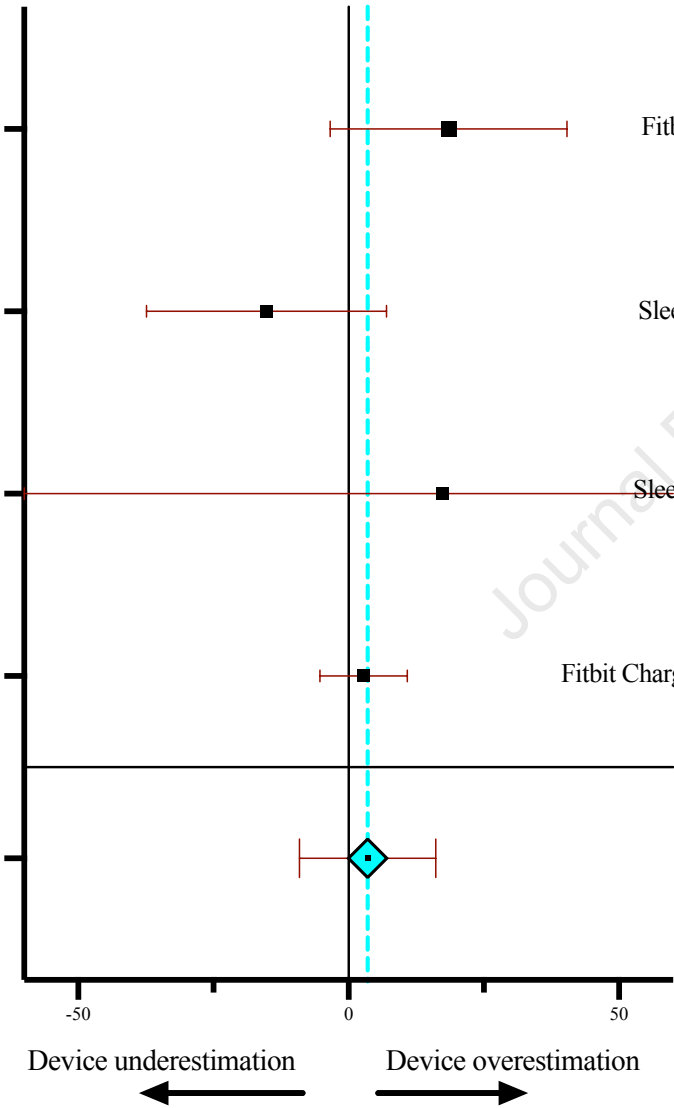
Random effects model
Heterogeneity: $I^2 = 3.9\%$; $df = 12$
Test for overall effect: $Z = 4.456$; $p = <0.001$

Study	SWS duration mean difference (minutes, 95% CI)	Device name (Location)		*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Gruwez et al., 2019 [65]		Jawbone UP (Wrist)	Confirmed OSA	22	160	122.0, 198.0	24.4
Gruwez et al., 2019 [65]		Withings Pulse O2 (Wrist)	Confirmed OSA	22	-88.0	-124.8, -51.2	24.5
Kahawage et al., 2020 [94]		Fitbit Alta HR (Wrist)	Insomnia	42	-59.4	-79.9, -38.9	25.4
Moreno-Pino et al., 2019 [71]		Fitbit Charge 2 / Alta HR (Wrist)	Confirmed OSA	65	-74.2	-85.4, -63.0	25.7
Summary Measure			**Summary**	**151**	**-16.8**	**-93.9, 60.4**	**100.0**

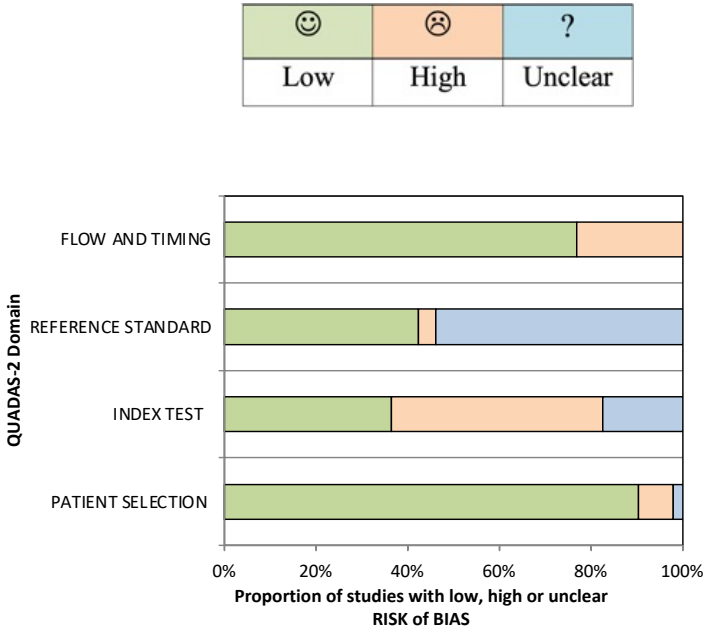


Random effects model
Heterogeneity: $I^2 = 54.3\%$; $df = 3$
Test for overall effect: $Z = -0.426$; $p = 0.667$

Study	REM duration mean difference (minutes, 95% CI)	Device name (Location)	Population	*n*	Mean difference	95% CI (Lower, upper)	Weight (%)
Kahawage et al., 2020 [94]		Fitbit Alta HR (Wrist)	Insomnia	42	18.5	-3.4, 40.4	10.4
Levendowski et al., 2017 [34]		Sleep Profiler (Headband)	Healthy & OSA	47	-15.2	-37.4, 7.0	14.4
Lucey et al., 2016 [33]		Sleep Profiler (Headband)	Suspected SD	29	17.4	-166.4, 201.3	11.5
Moreno-Pino et al., 2019 [71]		Fitbit Charge 2 / Alta HR (Wrist)	Confirmed OSA	65	2.78	-5.3, 10.9	63.8
Summary Measure			**Summary**	**183**	**3.5**	**-9.1, 16.1**	**100.0**



Random effects model
Heterogeneity: $I^2 = 29.2\%$; $df = 3$
Test for overall effect: $Z = 0.545$; $p = 0.582$



Study	Risk of bias			
	Patient selection	Index test	Reference standard	Flow and timing
Abad et al., 2016 [37]	☺	☺	☺	☹
Alshaer et al., 2016 [42]	☺	☺	?	☺
Andres-Blanco et al., 2017 [41]	☺	☹	?	☺
Assefa et al., 2016 [49]	☺	☺	☺	☹
Boyd et al., 2016 [47]	☺	☺	☺	☺
Cheung et al., 2020 [90]	☺	?	?	☺
Cho et al., 2017 [51]	☺	☹	☺	☺
Choi et al., 2017 [69]	☹	☹	?	☺
Crinion et al., 2020 [52]	☺	☺	☺	☺
Dafna et al., 2015 [67]	☺	☺	?	☺
De Zambotti et al., 2015 [91]	☺	☹	?	☺
Dunican et al., 2018 [70]	☺	☺	?	☺
Edouard et al., 2021 [92]	☺	☺	☺	☹
Gruwez et al., 2019 [65]	☺	?	?	☺
Gu et al., 2020 [43]	☹	☺	☺	☺
Hayano et al., 2020 [93]	☺	☹	?	☺
Holmedahl et al., 2019 [36]	☺	?	?	☹
Ioachimescu et al., 2020 [50]	☺	☺	☺	☺
Jen et al., 2020 [57]	☺	?	☺	☺
Kahawage et al., 2020 [94]	☺	☹	?	☹
Kapella et al., 2017 [68]	☹	?	?	☺
Kasai et al., 2020 [58]	☺	☺	☺	☺
Kim et al., 2019 [53]	☺	☹	?	☺
Kogure et al., 2017 [40]	☺	?	?	☺
Levendowski et al., 2017 [34]	☺	?	?	☹
Li et al., 2015 [95]	☺	☺	☺	☺
Li et al., 2020 [96]	☺	?	?	☺
Li et al., 2017 [44]	☺	☺	?	☺
Lu et al., 2019 [39]	☺	☺	?	☺
Lucey et al., 2016 [33]	☺	☹	☹	☺
Madrid-Navarro et al., 2019 [66]	☺	?	?	☺
Massie et al., 2018 [45]	☺	☹	☹	☺
Meng et al., 2016 [46]	☺	☺	☺	☺
Miyata et al., 2020 [62]	☺	☺	☺	☺
Moreno-Pino et al., 2019 [71]	☺	☹	?	☺
Munoz-Ferrer et al., 2020 [38]	☺	☹	?	☺
Nilius et al., 2017 [97]	☺	☺	☺	☺
Oliveira et al., 2015 [98]	☺	☹	?	☺
Pillar et al., 2020 [59]	☺	☹	☺	☺
Pinheiro et al., 2020 [48]	☺	☹	☺	☹
Ribeiro et al., 2015 [56]	☺	☹	☺	☺
Savage et al., 2016 [76]	☺	☹	☺	☹
Smith et al., 2020 [63]	☺	☹	☺	☺
Spielmanns et al., 2019 [35]	☺	☹	?	☹
Tal et al., 2017 [64]	?	☹	?	☹
Terjung et al., 2018 [99]	☺	☺	?	☺
Tiron et al., 2020 [54]	☺	☹	?	☺
Tondo et al., 2021 [60]	☺	☹	☺	☺
Topor et al., 2020 [61]	☺	☹	?	☺
Ward et al., 2015 [100]	☹	☺	?	☹
Weinrich et al., 2018 [77]	☺	☹	☺	☺
Zou et al., 2015 [55]	☺	☹	☺	☹