Medicine & Science IN Sports & Exercise

The Official Journal of the American College of Sports Medicine
www.acsm-msse.org

. . . Published ahead of Print

Wearable Sleep Technology in Clinical and Research Settings

Massimiliano de Zambotti[1], Nicola Cellini[2], Aimee Goldstone[1],
Ian M Colrain[1,3], and Fiona C Baker[1, 4]

[1]Center for Health Sciences, SRI International, Menlo Park, CA, US; [2]Department of General Psychology, University of Padova, Padova, Italy; [3]Melbourne School of Psychological Sciences, University of Melbourne, Parkville, Victoria, Australia; [4]Brain Function Research Group, School of Physiology, University of the Witwatersrand, Johannesburg, South Africa

Accepted for Publication: 7 December 2018

Medicine & Science in Sports & Exercise, Publish Ahead of Print
DOI: 10.1249/MSS.0000000000001947

Wearable Sleep Technology in Clinical and Research Settings

Massimiliano de Zambotti[1], Nicola Cellini[2], Aimee Goldstone[1],
Ian M Colrain[1,3], and Fiona C Baker[1, 4]

[1]Center for Health Sciences, SRI International, Menlo Park, CA, US; [2]Department of General Psychology, University of Padova, Padova, Italy; [3]Melbourne School of Psychological Sciences, University of Melbourne, Parkville, Victoria, Australia; [4]Brain Function Research Group, School of Physiology, University of the Witwatersrand, Johannesburg, South Africa

Corresponding author

Massimiliano de Zambotti, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, 94025 Tel: (650) 859-2714; Fax: (650) 859-2743; E-mail: massimiliano.dezambott@sri.com; maxdeze@gmail.com

Conflict of interest. Authors declare no conflict of interest. MdZ, FB, and IC have received research funding unrelated to this work from Ebb Therapeutics Inc., Fitbit Inc., and International Flavors & Fragrances Inc. **Funding**. This work was supported by the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA; U01 AA021696 to IMC and FCB) and by the National Heart, Lung, and Blood Institute (NHLBI; R01HL139652 to MdZ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. NC was supported by the University of Padua under the STARS Grants program, and the present work was carried out within the scope of the research program "*Dipartimenti di Eccellenza*", which is supported by a grant from MIUR to the Department of General Psychology, University of Padua.

Abstract

The accurate assessment of sleep is critical to better understand and evaluate its role in health and disease. The boom in wearable technology is part of the digital health revolution and is producing many novel, highly sophisticated and relatively inexpensive consumer devices collecting data from multiple sensors and claiming to extract information about users' behaviors, including sleep. These devices are now able to capture different bio-signals for determining, for example, heart rate and its variability, skin conductance, and temperature, in addition to activity. They perform 24/7, generating overwhelmingly large datasets (Big Data), with the potential of offering an unprecedented window on users' health. Unfortunately, little guidance exists within and outside the scientific sleep community for their use, leading to confusion and controversy about their validity and application. The current state-of-the-art review aims to highlight use, validation and utility of consumer wearable sleep-trackers in clinical practice and research. Guidelines for a standardized assessment of device performance is deemed necessary, and several critical factors (proprietary algorithms, device malfunction, firmware updates) need to be considered before using these devices in clinical and sleep research protocols. Ultimately, wearable sleep technology holds promise for advancing understanding of sleep health, however, a careful path forward needs to be navigated, understanding the benefits and pitfalls of this technology as applied in sleep research and clinical sleep medicine.

Keywords. Wearables; Polysomnography, Validation, Actigraphy, Digital health, Sleep

The landscape for wearable sleep-tracking technologies

Wearable sleep-trackers (e.g., wristbands, armbands, smartwatches, headbands, rings, sensor clips) are part of a larger consumer sleep technology (CST) family. CST includes smartphones, in-bed sensors, and contactless sensors, as well as other devices designed to enhance sleep and/or improve sleep behaviors such as neurostimulators, bio-feedback devices, and brainwave entrainment systems.

We consider ‘wearable sleep-trackers’ as those over-the-counter, relatively low-cost devices available without prescription or clinical recommendations. With many originally designed as fitness-trackers, these devices now claim to measure several bio-signals (e.g., heart rate and its variability, skin conductance, temperature), in addition to motion, from which information about behaviors, including sleep, can be extracted. Their accessibility (cloud-based platforms used for data storage and integration), usability (mobile user interfaces), novelty, and affordability has led to their widespread use and contributed to an increased awareness about the importance of sleep in the general population.

Within the research and clinical sleep communities, there is growing recognition of the potential benefits of using wearable sleep trackers. Benefits include the easy accessibility of an incredible and unprecedented amount of information about sleep and other behaviors, collected in peoples’ natural environments for extensive periods. Data can be collected at any time without active engagement from the users (who simply wear a device) and without the need of specialized technicians processing the data (which are usually provided in a summary form, such as total minutes spent asleep). However, despite these potential advantages, a fundamental issue is still unsolved. For many of the devices and associated systems, there are inadequate data available about their validity, accuracy and reliability in measuring the various sleep parameters and other indices, such as those reflecting cardiac function, that they report.

Although new regulatory models such as the Digital Health Software Precertification (Pre-Cert) Program (1) may ultimately affect the consumer wearable space, currently the US Food and Drug Administration does not regulate consumer-level wearables that provide “general wellness” information. There also is no consensus among sleep clinicians and research scientists on how to deal with the wearable boom, and no widely accepted standards as to how to implement the use of these devices in research and clinical sleep settings.

Alarmingly, with little knowledge and understanding of the performance of consumer wearables, the use of these devices is growing exponentially within the scientific field. For example, the Fitabase website (https://www.fitabase.com/research-library/), which keeps track of publications using Fitbit devices in research, lists >650 abstracts and journal papers for the Fitbit devices alone.

The focus of the current state-of-the-art review is on the use and validation of consumer wearable sleep-trackers and an evaluation of their utility in clinical practice and research. For the use and validation of other sleep technologies including mobile platforms for screening and monitoring sleep, the use of wearables in healthcare, please see (2-7).

Comprehensive literature searches were performed across the main electronic databases of PubMed, Google Scholar, Web of Science and PsycINFO for studies published in the English language about use and validation of wearables sleep tracking technology. One or more of the following terms were used: “wearable”, “sleep”, “validation”, “accuracy”, “sensitivity”, “specificity”, “reliability”, “polysomnography”, “comparison”, “fitness-tracker”, “sleep-tracker”, “actigraphy”, “commercial device”, “Fitbit”, “Jawbone”, “Misfit”, “Basis”, “Withings”, “ŌURA”. Full-text manuscripts were reviewed for relevance. Studies evaluating device performance were included only if they used 1) standard polysomnography (PSG) as the main reference for comparison, and 2) showed “acceptable standards” for methodological rigor, including adequate statistics and methods for PSG – device comparison (e.g., Bland-Altman method and/or epoch-by-epoch comparison).

Objective measurement of sleep: Polysomnography and actigraphy

PSG is the gold standard method to assess sleep and is the main reference for device validation. PSG is a comprehensive measure of sleep, based on the simultaneous recording of cortical (electroencephalogram [EEG]), submental muscle (electromyogram), and electroocular activity via the standardized positioning (international 10/20 EEG system) of scalp surface electrodes (8). As part of the PSG assessment, a number of additional physiological signals (e.g., electrocardiogram [ECG], respiration, leg movements, nasal pressure, oxygen desaturation and body position) are routinely assessed and help to characterize the complex nature of sleep and potential presence of sleep disorders. Following standardized visual rules based on the American Academy of Sleep Medicine (AASM) recommendations (8), sleep is manually scored in 30-s intervals by visual identification of specific phasic (e.g., arousals, K-complexes, spindles) and tonic (e.g., percentage of slow wave sleep within an epoch) features from the multiple EEG and physiological channels to assign each epoch as either: wake, N1, N2, N3 or REM sleep. PSG is usually confined to sleep laboratory research and clinical settings as it requires specialized equipment (a dedicated PSG acquisition system) and expertise (professionally trained personnel) for recording, scoring and interpreting PSG data. Although portable ambulatory PSG systems exist, the use of PSG is too expensive and impractical to be feasible for measuring sleep for prolonged periods outside of research studies.

The accepted alternative to PSG for non-laboratory settings is actigraphy. Actigraphy devices (mainly wrist-worn devices) rely on an accelerometer to measure patterns of activity (motion) and estimate sleep/wake states accepting the simple assumption that motion implies wake, and no-motion implies sleep. Due to their small size, comfort and waterproof properties, actigraphy devices are designed to be worn 24/7 and thus are suitable for prolonged recordings in non-laboratory settings. The device's accelerometer detects the occurrence and degree of motion in multiple directions (e.g., 3-axis), which is converted into a digital signal to derive an activity count. Then, depending on the sleep-wake threshold of the algorithm, an epoch is determined as wake if its activity count exceeds the threshold, or sleep if it is below the threshold. Data can be stored at different rates, which contributes to how long a device can store continuous data. Owing to limitations in data storage, the majority of the literature using actigraphy is based on 1 min resolution for data collection. Algorithms used by actigraphy are either provided by the manufacturer (e.g., Philips Respironics, Inc. Bend, OR) or publicly available (e.g., Cole–Kripke and Sadeh algorithms), and have been validated against PSG in healthy and clinical populations, on infants through the elderly (see 9, 10).

Although the majority of studies report high sensitivity (ability to detect true sleep) and accuracy (overall ability to detect true wake and sleep), actigraphy is inherently impaired in detecting true wake (specificity) as it is unable to identify motionless wake. For studies that have included healthy participants, specificity ranged from 26.9% to 77%, (11-20), while others that have included a variety of patient groups report specificity values ranging between 32.5% and 80% (21-23). Although many studies report specificity less than 50%, this finding is often minimized or overlooked, and actigraphy is accepted as providing an accurate estimate of PSG. Studies that have assessed the accuracy of actigraphy (in the classification of PSG sleep and wake epochs) using the different sensitivity thresholds of the Philips Respironics algorithms (11, 15, 17, 21, 23-25), as well as publicly available algorithms (15, 19, 26), have consistently shown that there is a trade-off between sensitivity and specificity. For example, for Philips Respironics algorithms, the "low" threshold requires smaller activity counts to deem an epoch as wake, therefore increasing specificity but at the cost of sensitivity. Conversely, the “medium” threshold increases sensitivity at the cost of specificity, due to the greater activity count threshold required for wake. Whether researchers should aim for high overall accuracy and sensitivity and acknowledge that sleep is overestimated, or whether they should instead aim to more accurately detect wake at the cost of sleep is still an open question, and is probably best decided based on the object of the investigation. For example, if the aim of a study is to determine changes in the amount of sleep disruption following a sleep treatment, it would be better to prioritize high accuracy in wake detection. Differently, if the purpose of a study is to evaluate changes in time spent asleep across adolescence, an algorithm prioritizing accuracy in sleep detection would be preferred. Furthermore, although studies have validated particular devices and algorithms against PSG and have reported that some algorithms are more accurate than others (15, 19, 26), the differences between devices, algorithms, participant groups and study designs makes it very difficult to draw firm conclusions across studies as to which device and algorithm is best. In addition, studies have reported specific device × algorithm interactions (19) and threshold × group interactions (23), further complicating the conclusions that can be drawn between studies and populations.

Although actigraphy has a number of advantages, there are limitations to consider. It is less costly than a PSG system, however, clinical devices are often upwards of $1000 each, which remains a limiting factor, particularly when sleep needs to be recorded on large datasets in populations like adolescents who may be reluctant to wear a research-grade device. Furthermore, although actigraphy does not require an “expert” to manually score sleep records or monitor recordings overnight, an experienced staff member with expertise in sleep analysis is still required to identify any issues with the actigram, such as artefacts or missing data. Additionally, although there are alternative algorithms which are publicly available, they are not integrated into existing software, and require expertise to conduct further post-hoc analysis. Even when algorithms have been shown to be less affected by wake (e.g., regression algorithms (17)) they have not been widely evaluated or adopted and researchers often apply settings recommended by the manufacturer (e.g., "medium" sensitivity threshold), despite them not necessarily being appropriate for their sample. Thus, there is still no consensus on specific recommendations for different patient groups, devices and algorithm thresholds for actigraphy.

Among the several limitations and the immobility of the actigraphy field (27), probably the cost of actigraphy and the requirement of technical staff and time for processing the data are among the main factors leading researchers and clinicians to consider consumer wearables as an alternative solution to easily collect sleep data in non-laboratory settings.

Consumer wearable sleep trackers

The availability and easy use of wearable sleep trackers contrasts with their hidden complexity, frequently leading to an erroneous adoption of these devices, and misleading interpretation of their outcomes.

In the following sections, we aim to summarize the advances made in the sleep wearable consumer market, the published validation studies, and the main factors and challenges to consider before using a consumer wearable sleep tracker in clinical and research settings.

These aspects should be taken as a starting point for researchers and clinicians to initiate a discussion about clarification and standardization for evaluating the accuracy and reliability of wearable sleep trackers. The conditions for which these new tools should be accepted and used in clinical and research settings need to be determined. Here, we propose initial guidelines to evaluate consumer wearable sleep technology.

It is important to recognize that consumer wearables are commercial devices designed for general consumers and are not specifically developed for clinical or research purposes. The algorithms used by these devices are proprietary and no raw data (direct sensor reading before any algorithms' implementation) are currently available. Also, wearable companies can change their algorithms without notice, an important aspect to consider when using a device over a certain period of time, and particularly for longitudinal studies. Although the number of validation studies is growing, validation clearly moves at a slower pace than the wearable industry, which keeps introducing new devices every year. Thus, evidence for the validation of a specific device model may be available when that model is no longer produced.

Lastly, it is important to understand that the second generation of multisensory consumer sleep trackers is fundamentally different from the first motion-based generation of consumer wearables (and actigraphy). The use of multiple sensors should theoretically overcome some of the challenges in detecting sleep and wake patterns, as discussed next. However, there are no direct comparisons – at least in the public domain - between motion-based and multisensory consumer sleep trackers, and their theoretical advantages over the previous generation remain to be empirically proven.

Advances made in sleep wearable technology: Toward a multisensory approach for sleep detection

The first generation of consumer sleep wearables (e.g., Jawbone UP, Fitbit Tracker "original", Fitbit Ultra, Fitbit Flex, Misfit Shine), similarly to standard actigraphy, extracted motion-based features from a built-in accelerometer-type sensor to measure wake and sleep. As for standard actigraphy, the limitation is that people can lie in bed awake for prolonged periods without moving, and in that case, the algorithm would misclassify wake epochs as sleep. For this reason, the first generation of consumer sleep wearables were limited in detecting wake. Also, despite attempts to differentiate sleep stages using motion-based pattern classification algorithms (see 28), these devices are limited to the binary detection of sleep and wake. Based on this intrinsic limitation, it is unlikely that further improvements in the levels of accuracy in sleep measurement (wake/sleep and sleep stage classification) will be achieved with motion-only based devices.

More intriguing is the new generation of wearables. The technological advances in sensor technology including miniaturization, low power consumption, low cost, connectivity and functionality of bio-sensors, allow new-generation wearables to continuously record a broad range of bio-signals (see (5, 29), for a review about methods and measurements of relevant wearable digital parameters) using, for example, skin temperature and optical photoplethysmography (PPG) sensors in addition to motion sensors that may advance sleep stage classification (30, 31).

Analysis of beat-to-beat cardiac information extracted from peripheral sensors such as PPG, can offer a valid approximation of ECG-derived heart rate variability [HRV; beat-to-beat variations in heart rate], a reliable indicator of cardiac autonomic nervous system (ANS) function, at least under conditions of minimal movement such as during sleep (see 32). For example, our group tested the accuracy of a multisensory sleep wearable (Fitbit Charge HR) against gold standard ECG in measuring heart rate during sleep in healthy sleepers, and we found an average ECG-PPG discrepancy for heart rate of <1 bpm (33). The comparison was based on min-by-min averages of HR across the night since beat-to-beat PPG data is currently inaccessible from consumer wearables, and thus, beat-to-beat accuracy levels are still unknown. Also, it is unknown whether the level of accuracy we found in healthy sleepers can be maintained in patients with sleep disorders (34), as well as during wake-time activities when the accuracy of wearable-based HR data is more questionable (35).

The main rationale underlying attempts to stage sleep (e.g., “light [PSG N1+N2]”, “deep [PSG N3]” and REM) in addition to the dichotomous distinction between sleep and wake states, relying in part on derived HRV data, is based on the concept of central nervous system (CNS) and ANS coupling (see 36). Sleep is not merely reflected by changes in cortical EEG activity but is characterized by changes in several other bio-systems including the functioning of the ANS, which regulates the majority of the organism’s internal functions (e.g., myocardial function, circulation, digestion) and mediates an individual’s responses to environmental challenges. ANS measures fluctuate across the night under homeostatic and circadian influences, and these fluctuations, particularly those reflecting vagal function (e.g., high frequency HRV), are tightly coupled with fluctuations in CNS EEG indices (e.g., activity in the slow delta EEG frequency band) (36).

A growing body of evidence indicates that wake and sleep stage classification could benefit by combining motion data and autonomic features (e.g., heart rate, HRV indices) (see 31, 37, 38-40). It remains unclear whether other recorded bio-signals (e.g., skin temperature, skin conductance (41)) will advance sleep staging in the future. However, at this juncture, the correspondence of these bio-signals with sleep-related EEG features and PSG stages is less evident, and future research is warranted to determine whether their addition could improve wake-sleep classification.

Our group provided promising results for the first validation studies of the new generation of multisensory wearables for PSG stage classification in healthy individuals, with reasonable differentiation of “light sleep” (PSG N1+N2) and REM sleep, although classification of slow wave sleep and wake were less consistent (42, 43) (see Table 1). Also, these multisensory wearables still had relatively low specificity in detecting wake.

There could be several reasons for this failure, among which, the most likely seems to be that attempts to classify sleep stages using multisensors is still in the early stages. As reviewed in (36), sleep is characterized by a sophisticated range of phasic, coordinated cortico-cardiac oscillations, reflecting the complexity of the dynamic communication between central and periphery. To leverage this complexity to achieve new improvements in sleep staging and sleep-wake classification, the wearable industry may benefit from input from domain experts within the sleep science and other fields (e.g., Network Physiology (44)) investigating the characterization and dynamic interactions of multiple aspects of central and peripheral systems which underlie the generation of different physiological states (sleep/wake, ‘light’, ‘deep’ and REM sleep).

We should also acknowledge that these devices are facing the challenge of performing 4 choices (wake, “light”, “deep”, and REM sleep) compared to the simplest dichotomous choice between sleep and wake, impacting their ability to discriminate between sleep and wake. Further, for validation studies relative to PSG, any automatic sleep scoring algorithm is referenced to manually-scored epochs of sleep. The AASM manual scoring system for PSG has high inter- and intra-scorer variability (45, 46), challenging the notion of stability of the gold-standard reference method, although a 10% of disagreement between scorers in the 5-choices (wake, N1, N2, N3, REM sleep) for PSG sleep staging is tolerated.

Finally, the influence of factors like demographics (e.g., age, sex) and environmental conditions (e.g., stress exposure, evening medication or alcohol use, environmental temperature) on the multiple signals recorded by these devices (e.g., HR and its variability) (see Section 5.1), and thus their capability in accurately staging sleep, should not be underestimated.

Validation of sleep wearables

Results of validation studies

New wearable devices and algorithms are introduced on the market every year. Due to the dynamic field, and the slow pace of scientific validations, it is challenging to provide an overall picture for the accuracy of wearable sleep trackers. Table 1 summarizes studies in chronological order that have examined the performance of wearables against gold standard PSG. Fitbit (33, 42, 47-54) and Jawbone (28, 52, 55-58) sleep trackers are among the wearables more frequently tested against PSG. In some studies, both consumer-based wearable devices and standard actigraphy were simultaneously used, together with PSG. In this review, we did not consider any direct comparison between wearable devices and standard actigraphy or sleep logs (see Section 5.2), which are summarized elsewhere (2).

It is important to realize that what we call "validation studies" are actually "second-step validations" whereby post-processed signals (e.g., heart rate) (see 33) and derived behaviors (e.g., sleep) are compared against gold standard methods; any comparison based on raw data is not available due to the black box nature of these devices. These limitations cannot be easily overcome. For details about algorithm validation and sensor validation see (6).

Despite several differences existing among studies, participants usually wore the wearable sleep trackers (and standard actigraphy) on the wrist of the non-dominant hand, for a fixed time, from lights-off to lights-on. The majority of studies were conducted in the laboratory and only a few studies have been conducted in free-living conditions (49, 51, 53, 57, 59). The latter point needs to be carefully considered since performance may differ at home relative to controlled in-lab conditions. Data from the first-generation motion-based wearables were usually manually extracted in a 1-min resolution and then matched with the resolution of PSG, or vice versa. In contrast, recent studies have been able to directly compare PSG and device epochs with a 30-s resolution, the same resolution used for PSG sleep stage classification.

To date, there are no accepted standard rules or regulations on how to evaluate and interpret the performance of commercial wearable sleep trackers and there is a wide range of validation measures used between studies. Overall, wearables show high sensitivity (above 90%) in detecting sleep but lower specificity in detecting wake, which is reflected in a general overestimation of PSG total sleep time (TST) and underestimation of wake after sleep onset (WASO), a performance that is in line with the majority of actigraphy literature (10). In studies that used both a consumer-wearable and clinical actigraph, compared to PSG, in the same participants, this pattern was still evident (47, 50-52, 55, 58, 59). Studies assessing the performance (accuracy in wake and sleep stage classification) of the second generation multisensory wearable devices in healthy participants, indicated a relatively higher performance in classifying PSG N1+N2 ("light sleep") (42, 43, 54, 58) and PSG REM sleep (60-75% agreements) (42, 43, 54, 58), compared to a relative lower performance for PSG wake and N3 sleep classification (42, 43, 54, 58). A relatively poorer performance for REM detection was found in one study testing a multisensory device in patients with hypersomnolence and mix sleep disturbances (58) (see Table 1).

Impact of nocturnal wake periods and age on device performance

Several studies have shown that greater sleep disruption (i.e., increased wake intrusions during a sleep period) exacerbates PSG-device biases, for actigraphy (see 17, for an example) as well as consumer wearables. In an adult sample of midlife women wearing Jawbone UP over two PSG nights, the PSG-device discrepancies in detecting WASO as well as TST were greater on the night with the higher amount of PSG WASO (56). Similarly, in a sample of adolescents wearing

ŌURA rings, we found that the PSG-device discrepancy in assessing WASO depended on the amount of PSG wakefulness (43). In several other studies, the relations between PSG-device discrepancies and alterations in PSG sleep were not directly tested but observed as a qualitative interpretation of the Bland-Altman plots (see section below for details about Bland-Altman plots). Similarly, the presence of sleep disorders, possibly driven by increases in the amount of PSG sleep disruption, may also affect devices performance. However, few studies directly tested device performance in patients with sleep disorders (see Section 7 and Table 1 for details), reporting mixed results, probably due to the use of different wearables and sample characteristics.

Factors other than sleep disruption also affect device performance. For example, some evidence suggests that performance may vary as a function of age, particularly in children and adolescents. When testing a sample of sixty-five healthy adolescents, our group showed that with increasing age, the performance of Jawbone UP significantly shifted from underestimating to overestimating TST and SE, and from overestimating to underestimating SOL and WASO (28). Similar results were provided by Toon et al. (55), who tested Jawbone UP against PSG in groups of preschool children, primary school children, and adolescents. In contrast, age, body mass index and sex did not affect device performance when testing a novel multisensory wearable (the first version of the ŌURA ring) in forty-one healthy adolescents (43). Therefore, it remains unclear if age, particularly across different developmental groups, affects the performance of motion-based wearables only. More research aimed to understand the factors accounting for variations in device performance across age is needed.

Detecting naps with wearable devices

Since consumer wearables may be worn around the clock, they have the potential of being used to track sleep outside of the nocturnal period. To our knowledge, few studies have assessed the performance of consumer sleep trackers in measuring daytime naps. Cook et al. (58) investigated the capability of the Jawbone UP3 to correctly identify the number of sleep-onset REM periods (SOREMPs) during a multiple sleep latency test in patients with hypersomnolence/mixed sleep disturbances, while Sargent et al. (60) tested the capability of Fitbit Charge HR in detecting daytime naps in athletes. Both studies showed strong limitations of these devices in automatic daytime sleep assessment. These limitations could be due to specific algorithm requirements for a minimum duration of sleep to allow sleep classification which are, so far, unknown to the users. For example, currently https://help.fitbit.com/ reports that "*Naps at least an hour in length will be automatically detected by your device and stored in your sleep history*", and in another help section states that "*Your device needs at least 3 hours of sleep data to estimate your sleep stages, so you won't see sleep stages for shorter naps*". Also, the poor performance in detecting naps may be due to the low specificity of wearables (including actigraphy) in distinguishing sleep from quiet wakefulness. Daytime sleep is common in pediatric and older adult populations as well in some sleep disorders (e.g., narcolepsy) or shift-workers, and frequently overlooked compared to night-time sleep (see 61). The ability to automatically track day-time sleep (even < 1h) is extremely important. Wearable companies should provide clear guidelines about the daytime sleep tracking capability of their devices, including whether and how the daytime sleep periods are merged with nighttime sleep (e.g., a 30 minutes nap plus a 6 h nocturnal sleep is displayed as a total of 6 h and 30 minutes of sleep) or showed as two separate sleeping periods. Future studies need to investigate the ability of wearables not only to assess nighttime sleep, but all sleeping periods during 24h.

Detecting sleep onset and offset with wearable devices

Another frequently overlooked aspect of wearables, is the ability of a device to accurately assess the onset and offset (morning awakening) of sleep. This is particularly important given that the timing of sleep onset and offset directly affect the determination of the sleep duration and its derived measures. Sleep onset is PSG-defined as the first epoch of any sleep stage, according to AASM criteria (8). In contrast, standard actigraphy determines sleep onset based on immobility time thresholds (see 62) within "rest intervals" determined by sleep diaries checked off-line by expert scorers. Event markers, used by individuals pressing a button on the device, and information about light exposure from embedded light sensors may also be available on some actigraphy models and used to determine lights-off and lights-on times.

The new generation consumer sleep-trackers use proprietary algorithms to automatically determine bedtime. Thus, lights-off and lights-on are determined without asking any active engagement from users (the off-line adjustment of these intervals is still available for some devices). However, commercial devices, like actigraphy, are limited in reliable determination of lights-off times, making it challenging to determine sleep onset latency without supplementary information from users about their self-reported lights-out times. Pesonen and Kuula (59) investigated the accuracy of a consumer device in determining onset and offset of sleep in children and adolescents, compared to PSG in an at-home setting. In that study, there were no significant differences in the onset and offset of sleep as derived by the Polar A370 sleep tracker compared to those determined by PSG. However, in the group of adolescents, although the mean differences were not significant, the standard deviation of the differences for the sleep onset estimation was quite wide (38 min) suggesting high variability in device performance for sleep onset time between individuals. Similarly, in healthy young adults, Liang and Martell (53) found that most of the time (68%) there was a positive delay (between 0 and 20 min) in sleep onset estimation from Fitbit Charge 2 compared to a single channel PSG at home, whereas in 24% of the cases, the delay was >20 min. Further research is needed to address the accuracy of consumer devices in determining timing for onset and offset of sleep (as well as the timing of REM onset, and the onset and offset of NREM-REM sleep cycles), particularly in populations in which sleep timings are altered (e.g., delayed sleep phase syndrome).

Testing and understanding the performance of a consumer wearable sleep tracker

To aid comparison across studies, it would be beneficial to use standard means of testing validity. Figure 1 outlines our recommended steps for evaluating the performance (validity) of a wearable against PSG, and these steps are further discussed here.

When validating a sleep device, controlled in-laboratory PSG should be the reference. However, given the barriers and limitations of in-laboratory PSG (e.g., cost, time, artificial setting) and the need for evaluation of wearable devices in more naturalistic settings (where wearables are used), the utilization of validated unattended ambulatory PSG (Type II, comprehensive portable PSG) is also appropriate. This is particularly true in evaluating the performance of wearable devices in convenient populations in which ambulatory PSG is routinely used, like in the evaluation and management of sleep-related breathing disorders (63). One of the main challenges in the PSG-device comparison in at-home environments is the accurate selection of the time windows for comparison, particularly the bed-time (lights-off) which is usually determined by participants' self-reported data. Careful instructions for logging lights-off and lights-on times for both night-time and day-time sleep may partially overcome the limitation. Any direct comparison between wearables and standard actigraphy for device validation should be avoided. In fact, this may result in inconclusive and misleading outcomes. When both wearables and standard actigraphy are used in conjunction with PSG, both devices should be compared directly with PSG and data outcomes interpreted accordingly. It is also important to consider that the current PSG scoring system (64) is similar to the one introduced almost 50 years ago (see 65), which relies on the discrete arbitrary and visual determination of sleep composition. Given that, we believe that PSG records used in study validation should always be double scored (two independent scorers) to avoid potential rater-specific biases in the outcomes. A high (usually >90%) inter-scorer agreement (or *inter-rater reliability*) should be set.

Synchronization

In validation studies, the first step is to guarantee an accurate **PSG-device synchronization**. Although most wearable devices do not disclose specific timing about how sleep parameters or epoch-by-epoch staging is calculated (e.g., server clock, device clock), synchronization is critical, particularly when performing epoch-by-epoch (EBE) analysis. We recently showed the impact of PSG-device synchronization misalignments on PSG-device discrepancies (42).

At a minimum, synchronization of the computer times where PSG and the wearable devices are running should be performed; however, this procedure does not guarantee an accurate PSG-device synchronization given that the precise onset/offset of the automatic device sleep staging algorithm is unknown. In our lab, it is common practice to start the PSG recording (time 0) at a rounded time (e.g., 22:32:00).

Direct comparison between PSG and wearable outcomes

Comparing PSG outcomes and PSG-equivalent sleep outcomes provided by the device via statistical tools is the first step in assessing the reliability of any sleep tracker. **Within-subject tests** (e.g., *t*-tests, repeated measure ANOVAs) compare the mean and standard deviation (SD) of several outcomes of the devices versus PSG. This step is fundamental to interpret potential significance in overestimating/underestimating PSG outcomes by the device, forming the basis to interpret Bland-Altman biases (see below). However, these analyses do not account for the heterogeneity of the participants' behavior, i.e., high variability in their behaviors, such as some subjects having very high and other subjects very low amounts of WASO. The latter issues can be overcome using mixed-effects models which can account for both the average population behavior and the natural heterogeneity of participant outcomes (66).

Concordance and agreement between PSG and wearables

The **Bland-Altman plot** is the most important tool to assess concordance between instruments and should be used to evaluate the overall performance of a device, by plotting the PSG-device discrepancies (y-axis) against the PSG values (x-axis), for each parameter of interest (the most common are TST, WASO, time spent in N1, N2, N3, REM sleep). In the original Bland-Altman plots, mean differences between devices are plotted on the x-axis (67), but since PSG is the accepted gold standard method for sleep assessment, a more conservative approach using PSG as a reference is recommended. While the Bland-Altman plots allow a visual (qualitative) assessment of both agreement and heteroscedasticity (i.e., whether there is an increase error as a function of the magnitude of the measured value), quantitative indices such as mean differences (or biases), SD and ±95%CI of the biases, lower and upper limits of agreement (mean difference ±1.96*SD) and ±95%CI of the agreement limits should be reported. A significant direct comparison test and a positive bias indicates that the device underestimated the observed PSG sleep outcome, whereas a significant direct comparison test and a negative bias indicates that the device overestimated the PSG sleep measure.

There is a general tendency to overemphasize the magnitude of the biases and underestimate the width of the agreement limits. However, it should be kept in mind that even if the biases are not significant, the performance of a device cannot be considered good when the discrepancies are "quite wide". A common practice is to report the number or percentage of participants falling outside the Bland-Altman agreement limits, which emphasizes potential large discrepancies between the PSG and the device. Still, this metric is dependent and needs to be interpreted by considering the distribution of the PSG-device discrepancies, which vary greatly across studies. Unfortunately, we are still relying on a case-by-case interpretation of the results based on our expertise and best judgement, more than on standardize performance quality metrics.

As shown in Table 1, a common metric used to investigate performance of a device is "*a-priori set clinically satisfactory ranges*" (see 15, 28, 33, 43, 48, 51, 52, 55, 59), i.e. fixed thresholds (usually, ≤ 30-min PSG-device difference for TST and WASO, and ≤ 5% difference for SE) to determine whether a bias is clinically significant or not. However, use of these fixed thresholds has limitations. We believe the rationale behind these proposed ranges, leading back to the frequently cited study of Werner et al. (68), remains unclear. Further clarification is required before advocating the use of the current "*a-priori set clinically satisfactory ranges*", and careful interpretation of these measures is needed.

Sometimes it is necessary to adjust the PSG-device bias if it is not constant across the range of measurement and shows significant heteroscedasticity. For example, logarithmic transformation of the values, calculating the ratio, or the percentage difference, instead of the absolute difference, can be done (see 69, 70). Finally, simple **regression tests** should be used to explore potential systematic dependency of PSG-device discrepancies in sleep outcomes on the amount of PSG sleep disruption and demographic factors possibly affecting motion patterns and/or other biological domains used by the proprietary scoring algorithms (see 28, 43).

Although frequently used in the literature, Pearson's correlations between PSG and device outcomes are misleading and should be avoided in evaluating and interpreting PSG-device agreements in measuring sleep outcomes. Indeed, simply correlating PSG and device sleep outcomes assesses the extent to which two measures covary, and not whether they are close together (see 71). For example, if the sleep tracker systematically reports a sleep onset latency two times longer than the PSG, the correlation coefficient would be 1 (perfect correlation), whereas in reality, the sleep tracker is not providing a valid measure of sleep onset.

A more appropriate approach is the use of **intraclass correlation (ICC)** which allows quantification of the PSG-device agreement for sleep outcomes. Following Cicchetti’s guidelines for interpreting ICC reliability coefficients (see 71), clinical significance is stated as “poor” for coefficients of less than 0.40, “fair” for coefficients lying between 0.40 and 0.59, “good” for coefficients lying between 0.60 and 0.74, and “excellent” for coefficients between 0.75 and 1.00. However, although some authors consider a device as “valid” based on ICC outcomes (51), there is still no consensus as to what are the minimum requirements for considering a device “valid” (72).

Accuracy of a device

Epoch-by-epoch (EBE) analysis is the preferred approach to assess the accuracy of a device. EBE should be performed in a 30-s resolution to evaluate *sensitivity* (proportion of PSG epochs correctly identified as “sleep” by a wearable device, see Figure 1) and *specificity* (proportion of PSG epochs correctly identified as “wake” by the device) of a device. When appropriate, the accuracy in detecting PSG sleep stages should be evaluated as the proportion of PSG epochs of a specific PSG sleep stage correctly identified by the device. A clarification on EBE terminology is needed. Currently, we believe that the terms “sensitivity” and “specificity” (widely used in the actigraphy literature) should be used when referring to the ability of a device to correctly classify PSG sleep and wake epochs. When evaluating the PSG-device concordance in the EBE sleep stages classification ("light", "deep" and REM sleep), we suggest wording the outcomes as: "*agreement for*" (e.g., *the EBE agreement for REM sleep is 0.60, reflecting the fact that 60% of the PSG REM sleep epochs are correctly classified as REM sleep by the device*). In our opinion, usage of standardized terminology will prevent confusion and misinterpretation of outcomes from validation studies.

EBE overall accuracy (proportion of PSG epochs correctly identified as "sleep" and "wake" by a wearable device) is frequently reported when evaluating a device performance. However, this measure is misleading due to a strong bias toward the extremely high sensitivity of most devices and the consequent tendency of evaluating the performance of a device based on its "accuracy". The relationship between sensitivity and specificity can also be visually assessed using the Receiver Operating Characteristic (ROC) curves, which provide a visual and quantitative measure of the accuracy of the device (see 73).

EBE analysis should be performed for each individual and the outcomes should be provided as mean, SD and ±95%CI of the mean. The determination of PSG-equivalent epochs of specific sleep stages from a device is not always straightforward (e.g., PSG N1 and N2 sleep may be represented as "light sleep" (see 42, 43)), but this information can be available directly from device manufacturers.

A common issue in performing EBE analysis is that wearables devices do not always provide 30-s sleep scoring data, which should be the ideal recording time to match with standard PSG scoring (8). Thus, different strategies have been adopted to match PSG and device epochs. A common strategy is to convert 30-s PSG epochs into 1-min epochs as W-W = W, W-S or S-W = W, and S – S = S (33, 48, 55, 56). Others (47, 50, 51, 58), split the device 1-min epochs into two equal 30-s epochs to match the PSG 30-s epochs resolution. Results of these procedures can overinflate the amount of PSG wake. For example, as little as 16 s of PSG wake (e.g., alpha rhythm more than 50% of the epoch over the occipital region according to AASM rule for wake) can result in 1 min of wake.

Another measure that can be derived from EBE analysis is the Cohen's kappa coefficient, which is an index of interrater reliability that reflects the percentage of measurement agreement (in this case, the sleep/wake scoring) of two methods not due to chance. However, since during sleeping periods the proportion of sleep epochs is generally higher than the wake epochs, it is possible to fall into ‘‘the first paradox of kappa statistic’’ (74), that occurs when two measures have a high agreement but a low kappa. A way to correct this bias is to calculate a prevalence- and bias-adjusted kappa (PABAK), which weights the number of sleep and wake epochs (75).

A full representation of the EBE analysis is the **error matrix** (or **confusion matrix**). The error matrix allows assessment of the device performance in classifying PSG wake and sleep (as well as stages of sleep) epochs via a cross-tabular representation of the PSG-device epoch-by-epoch classifications. The advantage is to obtain a more complete picture providing not only the proportion of PSG epochs correctly classified by the device but also the source of the potential misclassification (see Figure 1, and (42)). For a better reading of the confusion matrix we previously calculated mean, SD and ±95%CI of the proportion of agreement between PSG epochs and predicted (device) epochs (42).

Other strategies have been proposed to capture the PSG-device accuracy accounting for sleep timing, sleep stage distribution and cycles across the night (see also Table 2). For example, in one of the first validation studies for wearable sleep trackers, Montgomery-Downs et al. (47) calculated EBE sensitivity separately for wake before and after sleep onset (an approach that may be useful when performing EBE analysis outside the controlled laboratory settings in which lights-off and lights-on time cannot always be accurately obtained). Authors also calculated EBE sensitivity separately for PSG N1, N2, N3 and REM sleep, and in epochs containing arousals.

Our group, recently introduced a PSG-device comparison based on the ability of the device to correctly identify PSG NREM-REM cycles across the night (42).

Reliability for sleep assessment

Less emphasis has been placed on assessment of device reliability (see Section 5.2.5), which has been measured using within-subject analyses (e.g., paired *t*-tests) in the only two studies assessing intra-device reliability (a person wearing multiple devices simultaneously) (47, 48). Also, most validation studies have been based on single-night in-lab recordings due to several pragmatic and logistic reasons (e.g., easy to control and implement, cost-effective, validation study nested into other research protocols).

Factors to consider when choosing a wearable sleep tracker

A critical requirement for using a wearable in research is to have access to the data. Most wearable companies have some form of access to an Application Programming Interface (API) and software development kit (SDKs), which allows post-processed data access and integration, developing applications and services (e.g., https://dev.fitbit.com/; https://jawbone.com/up/developer; https://build.misfit.com/; https://developer.health.nokia.com/api). Some companies also have cloud services or web dashboards which allow to directly export summary data in easy-to-read files (e.g., *.csv, *.xls), ready for analysis. An initial bridge between research and industry is offered by third party research services, usually requiring a subscription, like Fitabase (Small Steps Labs LLC.; https://www.fitabase.com/; supporting Fitbit devices and, more recently, Garmin devices) which allows access to more technical information, assistance with setting up projects, and pre-processed (but not raw) data at different time resolutions.

Other factors to consider if choosing a wearable in research or for clinical purposes are shown in Figure 2. Reliability should be a major point of consideration given that these devices may be particularly useful for long-term recording in non-laboratory settings, i.e. in epidemiological studies. In the following paragraphs we will highlight some important reliability issues (see Sections 6.1 and 6.2).

It is also critical to consider the sample being studied. Demographics and other characteristics of the sample may impact device performance (see 28, 43, 55). If a specific device shows a certain performance in an adult sample, one cannot assume that it will have the same performance in children or adolescents. The same is also true for sleep disorders, meaning that one cannot assume that a device validated in a healthy population will show the same performance in individuals with sleep disorders (see Section 7). Some consumer wearables offer different sensitivity settings (e.g., "*normal*" or "*sensitive*" mode). The "*normal*" setting is usually indicated for most users, whereas indication for using the "*sensitive*" setting implies its use in the presence of sleep disturbances. However, no clear indication for using different sensitivity settings are provided by wearables manufactures. As summarized in Table 1, the few studies comparing different algorithm sensitivities in Fitbit devices (48, 50, 51) indicated overall a poorer performance of the devices used in "sensitive" mode.

Device position may also affect the accuracy of a device, particularly for the new generation of multisensory sleep trackers. Other than the effects of position on the pattern of motion, other bio-signals may be directly or indirectly affected by the position or incorrect position of a device (e.g., PPG signals depend on how accurately blood flow is detected, skin conductance is affected by sweating) (see 43). This is particularly important when considering using the device in free-living condition, when technicians may be unavailable, and participants need to self-apply the device.

Inter-device reliability

Inter-device reliability can be taken to mean that several devices used in the same conditions can provide the same outcome. An at-home study based on three participants wearing two Fitbit "original" devices overnight showed high reliability of these devices (percentage of EBE agreement of 96.5%, 99.1% and 97.6%) (47). Similar results were reported by Meltzer and colleagues (48), who examined intra-device reliability in 7 subjects wearing 2 Fitbit Ultra devices on the same wrist. Nevertheless, the authors prudently suggested that a device should not be switched with another device in the middle of a research protocol. Inter-device reliability is often overlooked and deserves further attention.

Device malfunctioning and other issues

A common issue with wearable trackers is data loss. In one study (48), 19% of the Fitbit Ultra data (12 participants) were not recorded due to technical issues. Of note, in the same study 14% of the data recorded with both the Actiwatch Spectrum and the AMI Motionlogger were unusable for technical issues (48). Other studies reported 4.3% of unusable sleep data (2 recordings) for Fitbit Charge 2 (42), and 12.5% (7 devices) for Fitbit Alta HR (54). Sargent and colleagues (60) reported 10 missing recording (out of 60) from Fitbit Charge HR due to an error in transcription (unclear whether this was a human or a device error). Mantua and colleagues (49) testing several devices against PSG, reported that data from 25% of Fitbit Flex (10 devices), 10% of Basis Health (2014 edition), 37.5% of Misfit Shine and 10% of Withings Pulse O2 devices could not be used (either for user errors, gross mis-estimation or other miscellaneous reasons). Of note, in the same study authors reported that 12.5% of the data from Actiwatch Spectrum were unusable (1 device for gross mis-estimation and 4 for malfunctions). More recently, Kang and colleagues (55) reported only 2% of the data lost with the Fitbit Flex and 5%

with the Actiwatch 2. Toon et al. (55) reported unusable data from 4% of the Actiwatch 2 and 13% of Jawbone UP devices. Missing data were due to participant behaviors (e.g., child taking off the UP during the night) or device malfunctions (e.g., actigraphy recording ceased due to battery malfunction). In another study (58), 17.5% of the data recorded with the Jawbone UP3 were unusable due to unspecified malfunctions.

Gruwez et al. (57) reported missing data from 14% of the Withings Pulse 02,7% of the Jawbone UP MOVE, and 5% of the SenseWear Pro Armband recordings. In another study with 20 participants wearing the SenseWear Pro3 Armband the authors were able to use data from all but one recordings (76). Interestingly, the same armband showed high reliability even when recording several nights of sleep (77). In contrast, Lillehei and colleagues (78) using Fitbit One over 5 consecutive nights reported about 86% of missing data. Baroni et al. (79) showed a similar picture, with only 14% of the Fitbit Flex devices used in their study able to collect six or seven nights of sleep, and 35% of them failed to record any nights of sleep.

Overall, these studies show mixed results. Considering that the main advantages of wearables is to collect data for several days, future studies are warranted to provide further data on the long-term reliability of wearables. A detailed report for reliability should include not only the number of recordings/device failure, but also information about the source of unusable data (e.g., due to mechanical failure, human factors, software issues).

It is important to remember that wearable companies adopt different decision criteria as to whether to provide a data outcome. For example, Fitbit Inc states that "*The Fitbit system does not return sleep stages under various conditions. These include cases where the heartbeat signal (and hence the heart rate variability) is not cleanly detected throughout the night, if the total sleep duration is less than three hours, or if the battery runs out of power during the sleeping period*". These criteria are based on different factors including a test of the integrity and amount of data they collect, which is not accessible to us. Thus, even when a sleep outcome is provided, we do not know specifically how much "reliable" information is used to provide that value.

The potential role of sleep wearables in clinical sleep disorders, intervention delivery and patient monitoring

Although the gold standard to evaluate the presence of sleep disorders is PSG, actigraphy has been commonly used in clinical practice to provide additional characterization of individuals with sleep disorders and to assess their treatment response (see 80). Nevertheless, so far only a few motion-based (first generation) consumer wearables have been tested in patients with clinical sleep disorders.

Two studies targeted children and adolescents with sleep disordered breathing (SDB). Meltzer et al. (48) showed that discrepancies between PSG and Fitbit Ultra changed as a function of SDB status and device sensitivity settings ("*normal*" or "*sensitive*"). Specifically, the study showed that despite Fitbit Ultra "*normal*" setting overestimated PSG TST and underestimated PSG WASO in both children with or without OSA, the PSG-device discrepancies were greater in mild OSA and further exacerbated in children with moderate/severe OSA. The authors also reported that most of the participants were outside the a priori-set "clinically satisfactory ranges" (i.e., TST <30 min and SE < 5%; see above for concerns about the use of these agreement limits). A reverse pattern was observed for the "*sensitive"* setting, characterized by greater PSG-device discrepancies in the no OSA category (TST underestimation and WASO overestimation), which progressively lessened in mild OSA and moderate/severe OSA categories (see Table 1 for details). Toon et al. (55) tested the Jawbone UP and showed no differences in PSG-Jawbone UP discrepancies in estimating TST, WASO, or SE as a function of SDB severity (i.e., primary snoring, mild or moderate-severe OSA). Moreover, the authors observed from the Bland-Altman plots that the Jawbone UP sleep outcomes were more consistent with PSG measures than were Actiwatch 2-PSG outcomes. Nevertheless, similar to Metzer et al. (48), the majority of the participants fell outside a priori-set “clinically satisfactory ranges”. The authors indicated that, on the one hand, the Jawbone UP should be used as a diagnostic tool with caution; on the other hand, they observed that the Jawbone UP performance was, overall, similar to the Actiwatch 2.

Few studies have evaluated device performances in individuals with insomnia. Kang et at. (51) reported an overall good performance of the Fitbit Flex in the “*normal”* mode for good sleepers (no significant PSG-device differences for SOL, WASO, and SE, fair to excellent ICCs, and the majority of the participants fell inside the “satisfactory clinical agreement limits”). However, the Fitbit Flex showed more difficulties to assess sleep in the insomnia group. Specifically, the Fitbit Flex significantly overestimated PSG TST, SE and underestimate WASO in the insomnia group. Moreover, only 39.4% of the sample fell within the a priori-set “clinical agreement range”. Again, as in Meltzer et al. (48), the “*sensitive”* mode showed a different, and less reliable pattern than the “*normal”* mode. Despite claims that the “*sensitive”* setting should be used in the presence of sleep disturbances, probably due to an algorithm that maximizes specificity (i.e., wake detection) at the detriment of sensitivity (i.e., sleep detection), these validation studies suggest that the “*sensitive*” setting is less reliable than the “*normal”* setting even in the presence of sleep disorders. Differently from Kang et at. (51), our group failed to find any difference in PSG-Jawbone UP discrepancies between women with and without insomnia disorder (56). The different wearables and sample used prevent any study comparison.

Two recent studies by Cook and colleagues tested the performance of the Jawbone UP3 (58) and the Fitbit Alta HR (54) against PSG and standard actigraphy (AW-2, only tested against the Jawbone UP3) in patients with different type of central disorders of hypersomnolence (including narcolepsy) and other sleep disorders tested at night and during multiple sleep latency tests

(MSLT). The Jawbone UP3 overestimated TST and SE, and underestimated WASO and SOL compared to PSG, but showed a similar performance to the AW-2. It also showed a good sensitivity (0.97) and a low specificity (0.39) and low agreement for single stage scoring, in particular for REM sleep (0.30). The Fitbit Alta HR provided similar results, with overestimation of TST and SE, compared to PSG. However, while sensitivity was similar to the Jawbone UP3 (0.96), specificity was slightly better (0.58), and in general showed a higher agreement for the discrimination of light, deep, and REM sleep (see Table 1). Of note, both devices failed to detect any SOREMPs during the MSLT. Authors concluded that the Jawbone UP3 and the Fitbit Alta HR cannot substitute the standard PSG to assess sleep in central disorders of hypersomnolence.

To our knowledge no studies have validated any consumer wearable trackers for circadian rhythm disorders. Indeed, these conditions are less common that insomnia or OSA. However, considering that actigraphy is a recommended tool for the diagnosis of circadian disorders (see 80), the lack of study with this clinical condition is somewhat surprising and future studies with wearables need to fill this gap. At-home PSG could be a viable approach for addressing validation within this patient population. However, due to the challenges in the longitudinal use of ambulatory PSG systems, a more reasonable approach would involve the assessment of cross-sectional PSG-device biases in individuals with altered sleep-wake times. Advancements should also be made to not only consider PSG-device validation of classical outcomes (time spent asleep/awake and in different sleep stages) but also consider major indices such as sleep onset and wake-up times used to assess circadian alterations (e.g., delayed/advanced/irregular sleep-wake phases, jet lag). Reliable determination of sleep onset is challenging with current wearables and advancements in algorithms or, possibly, the addition of other sensors to enhance the detection of sleep onset would be valuable (see section 5.1 "*Detecting sleep onset and offset with wearable devices*").

Currently, there is insufficient evidence to consider consumer wearables as a potential stand-alone diagnostic tool for sleep disorders.

An important concern of the general enthusiasm around the concept of “quantify self” is evident in the growing tendency for people to self-diagnose and even change their sleep habits based on the interpretation of unregulated information of their consumer sleep trackers. For example, people may try to stay longer in bed if their wearable device does not show a ‘magic number’ of 8 hours slept. Sleep feedback could be particularly problematic in those suffering from insomnia, who may exacerbate their anxiety and worry about sleep if their trackers display “poor sleep” performance. On the other hand, inaccurate feedback of “good sleep” may prevent or delay individuals from looking for professional help. We are also facing the situation in which patients are asking their physicians to evaluate their wearable sleep graphics. This use of potentially inaccurate information about sleep may not only alter the individuals’ perception of sleep, but challenge the clinician’s evaluation of their sleep pattern and potential treatments (see 81, 82). However, some guidelines are now available for clinicians on how to deal with CST data in clinical settings, as provided by the AASM (83).

Nevertheless, if regulated, consumer wearable sleep-trackers may still be useful in clinical settings to provide additional information about patients’ sleep-wake patterns (e.g., assess regularities and abnormalities in individuals’ sleep schedules), and monitor treatment responses and recovery. In this framework, a few studies have combined consumer sleep-trackers and smartphone Apps to provide different type of interventions (e.g., internet-based cognitive-behavioral therapy) (84-87) or to assess the effect of interventions on the sleep pattern (88, 89) with mixed results. Sleep trackers may be useful to monitor patient’s compliance to a particular sleep intervention such as sleep restriction. In general, clinicians should be aware of the risk that patients start to trust their tracker outcomes more than their physician’s clinical judgment.

Sleep trackers, if sufficiently validated, may potentially be useful to screen for sleep disorders in the future. So far, to our knowledge, only smartphone applications using phone and additional external sensors to extract and combine multiple features (position, audio, oxygen saturation) have been used to screen for sleep apnea (90, 91), with some promising results. Similarly, sleep trackers, in particular the second generation of multisensory sleep trackers, may help to screen for potential sleep disorders in order to increase the number of individuals who can ask for a clinical evaluation. However, although wearable technology has been used to assess sleep quality in OSA patients, no currently available consumer wearable devices are suitable for diagnosing OSA. Guidance from the Centers for Medicare and Medicaid Services indicates four types of equipment for diagnosis of OSA: 1) in-laboratory PSG (Type I); 2) in-home PSG (Type II); 3) in-home measures of respiratory effort, airflow, cardiac data and blood oxygen saturation (Type III); 4) in-home measures of blood oxygen saturation and airflow (Type IV) (92). It is also the position of the AASM that care should be taken in the interpretation of the results of at-home sleep apnea testing, with raw data that should be reviewed and interpreted by a board-certified sleep medicine physician (93). In a recent position statement (83) the AASM in reference to CST (which includes sleep wearables devices) clearly stated that "*CSTs cannot be utilized for the diagnosis and/or treatment of sleep disorders at this time*".

Limitations, barriers and future direction for the use of wearable sleep trackers

There is a lack of incentives from both the scientific community and industry (which frequently relies on their own internal non-peer-reviewed tests) to perform dedicated scientific validation of sleep-tracking wearables. Thus, the existing validation studies are frequently initiated by the curiosity of isolated researchers or research groups, moved by the need to find affordable, accurate, and reliable alternatives to the expensive medical grade devices for measuring sleep in natural contexts. Further studies are needed to validate wearable devices in different populations and conditions, particularly in individuals with sleep disorders, in whom studies are few. Recently, the National Institute of Health (NIH) recognized the potential of wearables for biosensing applications and the need to fill the gap between validation and use of wearables within the scientific field. NIH promoted several initiatives within the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs, and other funding opportunities to promote the development (e.g., wearable devices to monitor blood alcohol levels and identify biomarkers of drug addiction relapse in real time, identifying physiologic changes with old age) and validation of wearable devices for health measurement and intervention delivery (e.g., wearables to improve diagnosis and early treatment in minority and health disparity populations).

The consumer wearable market is extremely crowded, and the wearable industry is struggling with market differentiation. For the scientific sleep community, the necessity of opening the “black-box” wearable devices is important for raw data access and standardization, but raw data access and cloud services do not come free. Within this scenario, it is unclear if a line of consumer products and platforms more focused on the needs of researchers and clinicians would fit the consumer wearable companies’ business model. On the other hand, it is still unclear if the consumer wearables devices will maintain the advantage over standard actigraphy in the recording of multiple bio-signals and related assessment of sleep staging. In fact, within the medical space, new actions by actigraphy companies may be taken (e.g., moving to a multisensory approach and still offering validated algorithms based on multiple channels of information) (27). In addition, it is still unclear what the limit of the level of performance is for these early-stage non-EEG consumer wearable devices, and whether further advancement and integration of peripheral information will be able to more accurately approximate EEG-defined sleep staging. Also, the role of EEG consumer wearable devices within the sleep and circadian fields is still unclear. The Zeo headband (Zeo, Inc.), which was the first product of its kind, showed promising results in sleep measurement when compared to gold-standard lab-grade PSG (94-96). After its failure (the company went out of business in 2013), other EEG-based wearable headbands (e.g., Muse, Dreem, Neuroon) populated the market, and have shown promise in detecting sleep stages in clinical and non-clinical populations (97, 98). However, they are taking a different path from Zeo, more toward sleep-hacking (e.g., neuromodulation, brain entrainment for sleep enhancement) than sleep-tracking per se. The use of in-ear EEG (EEG recorded within the ear canal) is also of interest, but it is still in its infancy (99). It is unlikely that, in the immediate future, these EEG-like devices will make the same impact as multisensory non-EEG wearables in the field of sleep and circadian science due to their greater invasiveness and relatively higher costs, limited use (they cannot be easily worn 24/7), and the challenges in recording good quality EEG signals in uncontrolled, non-laboratory conditions.

The sleep community still should clearly state what are their specific minimal requirements (e.g., raw data access, algorithm standardization, validation steps) for accepting and potentially introducing a consumer wearable sleep tracker in research and clinical sleep settings. This should be the first step to opening a discussion with industry (Table 3).

Time is critical because consumer sleep wearables are increasingly used in observational and interventional studies (see 2), and are already implemented in corporate wellness programs (84, 100, 101). Also, consumer sleep trackers are a core part of the growing area of the Internet of Things (102) and Big Data for eHealth and mHealth (103) applications, an unstoppable digital health revolution. Press releases and reports from wearable companies, based on analysis of billions of wearable data of unproven accuracy, are also growing in popularity. The ability to map sleep in entire countries, breaking down sleep data by regions, in association with major historical events, investigate sex- and age-differences in sleep patterns in large populations (see 104, for example), is of value. But, without knowing the performance of these devices, and without a scientific approach to the Big Data, any interpretation of these results could be misleading. Having large amounts of data from these devices does not necessary reduce the within- and between-subject variability. On the contrary, it may amplify the inaccuracy of the results they provide, particularly when the discrepancy between PSG and device (bias) for the measure of interest differs from zero or when the discrepancy varies as a function of the PSG measure (i.e. the bias is not constant). In the new generation of multisensory wearables, the consistency of the algorithm used in measuring sleep is further challenged by the fact that the relationship between cardiac features and sleep stages may vary as a function of different factors such as age, sex and even by geographical area where participants live (105). Of concern is the growing perception from the public that population-based sleep data as provided by wearables companies (obtained by a specific sub-sample of the general population - the wearables users) are the new normative sleep data. In addition, people may be making changes to their sleep behaviors based on their wearable outcomes and frequently non-scientific validated "tips" for sleeping better (direct or indirect claims made by most wearable companies). Similarly, there could be situations in which people do not take actions when they should, due to potentially false feedbacks from their wearable device (e.g., they may truly have severe sleep disruption or altered sleep patterns, but their device is telling them that their sleep is good).

Another factor to consider in using consumer wearables, particularly concerning a potential role in precision medicine, is their accuracy at the level of the individual. The translation of group-average results of validation studies to the individual is challenging due to several factors such as variability in demographics, sleep and daytime habits which may affect the performance of a device. An open-access data repository of de-identified PSG and wearable data, including demographics and other information collected from validation studies, may ultimately allow correction for key factors affecting device performance. For example, the characterization of the relation (function) between demographics (e.g., age and sex) and PSG-device biases on a group level (see 28), could be used to adjust device outcomes at the individual level.

Conclusion

Sleep is fundamental for health (106). About one-third of the population is struggling with their sleep, a number that is estimated to increase. In our 24/7 sleepless society, sleep wearables may have a key role to better characterize and understand sleep and, within the framework of precision medicine, to ultimately improve health, safety and well-being for individuals and society. Collection of continuous data, day and night, could also lead to better understanding of links between sleep and daytime behaviors such as exercise.

Wearable sleep trackers are being increasingly adopted by both the general public and sleep researchers and clinicians. The second generation of multisensory sleep trackers opens a path for greater accuracy in measuring sleep, as compared to the motion-based approach to sleep/wake assessment. However, the proven theoretical advantage of the multisensory approach to sleep staging needs further empirical validation. Currently, these devices should be used cautiously, and interpretation of their outcomes should be carefully considered to avoid generating large inaccurate datasets leading to potential misleading scientific conclusions, assessment of sleep disturbances, and therapeutic decisions.

Further work is needed to investigate the potential use and performance, pros and cons, and limitations of these novel sleep trackers, particularly in sleep disorder populations. Keeping in mind the differential and overlapping motivations of various end-users' groups (e.g., research and clinical sleep community, wearable industry, consumers), partnership with industry is beneficial to combine excellence and speed in technological advancement from industry and advanced psychophysiological knowledge and scientific rigor from sleep science.

Acknowledgments

The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of the present study do not constitute endorsement by the American College of Sports Medicine (ACSM).

Conflict of interest. Authors declare no conflict of interest. MdZ, FB, and IC have received research funding unrelated to this work from Ebb Therapeutics Inc., Fitbit Inc., and International Flavors & Fragrances Inc.

Funding. This work was supported by the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA; U01 AA021696 to IMC and FCB) and by the National Heart, Lung, and Blood Institute (NHLBI; R01HL139652 to MdZ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. NC was supported by the University of Padua under the STARS Grants program, and the present work was carried out within the scope of the research program "*Dipartimenti di Eccellenza*", which is supported by a grant from MIUR to the Department of General Psychology, University of Padua.

References

1. Food and Drug Administration (FDA). *Fostering Medical Innovation: A Plan for Digital Health Devices; Software Precertification Pilot Program. Avaliable at: https://www.fda.gov/MedicalDevices/DigitalHealth/UCM567265*. Web site access: October 4, 2018.

2. Baron KG, Duffecy J, Berendsen MA, Cheung IN, Lattie E, Manalo NC. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. *Sleep Med Rev*. 2018;40:151-9.

3. Lorenz C, Williams A. Sleep apps: what role do they play in clinical medicine? *Curr Opin Pulm Med*. 2017;23(6):512-6.

4. Shelgikar A, Anderson P, Stephens M. Sleep Tracking, Wearable Technology, and Opportunities for Research and Clinical Care. *Chest*. 2016;150(3):732-43.

5. Roomkham S, Lovell D, Cheung J, Perrin D. Promises and Challenges in the Use of Consumer-grade Devices for Sleep Monitoring. *IEEE Reviews in Biomedical Engineering*. 2018;11:53-67.

6. Bianchi M. Sleep devices: wearables and nearables, informational and interventional, consumer and clinical. *Metabolism: clinical and experimental*. 2018;84:99-108.

7. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. *Per Med*. 2018;15(5):429-48.

8. Iber C. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine; 2007.

9. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;15(4):259-67.

10. Van de Water A, Holmes A, Hurley D. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography-a systematic review. *J Sleep Res*. 2011;20(1 Pt 2):183-200.

11. Belanger ME, Bernier A, Paquet J, Simard V, Carrier J. Validating actigraphy as a measure of sleep for preschool children. *J Clin Sleep Med*. 2013;9(7):701-6.

12. de Souza L, Benedito-Silva AA, Pires ML, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. *Sleep*. 2003;26(1):81-5.

13. Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiol Behav*. 2001;72(1-2):21-8.

14. Jean-Louis G, Kripke DF, Mason WJ, Elliott JA, Youngstedt SD. Sleep estimation from wrist movement quantified by different actigraphic modalities. *J Neurosci Methods*. 2001;105(2):185-91.

15. Meltzer L, Walsh C, Traylor J, Westin A. Direct comparison of two new actigraphs and polysomnography in children and adolescents. *Sleep*. 2012;35(1):159-66.

16. Meltzer LJ, Wong P, Biggs SN et al. Validation of Actigraphy in Middle Childhood. *Sleep*. 2016;39(6):1219-24.

17. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. *Sleep*. 2007;30(10):1362-9.

18. Pollak CP, Tryon WW, Nagaraja H, Dzwonczyk R. How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep*. 2001;24(8):957-65.

19. Quante M, Kaplan ER, Cailler M et al. Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. *Nat Sci Sleep*. 2018;10:13-20.

20. Shin M, Swan P, Chow CM. The validity of Actiwatch2 and SenseWear armband compared against polysomnography at different ambient temperature conditions. *Sleep science (Sao Paulo, Brazil)*. 2015;8(1):9-15.

21. Kushida CA, Chang A, Gadkary C, Guilleminault C, Carrillo O, Dement WC. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Med*. 2001;2(5):389-96.

22. Marino M, Li Y, Rueschman MN et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. 2013;36(11):1747-55.

23. Ward TM, Lentz M, Kieckhefer GM, Landis CA. Polysomnography and actigraphy concordance in juvenile idiopathic arthritis, asthma and healthy children. *J Sleep Res*. 2012;21(1):113-21.

24. Farabi SS, Quinn L, Carley DW. Validity of Actigraphy in Measurement of Sleep in Young Adults With Type 1 Diabetes. *J Clin Sleep Med*. 2017;13(5):669-74.

25. Cellini N, Buman M, McDevitt E, Ricker A, Mednick S. Direct comparison of two actigraphy devices with polysomnographically recorded naps in healthy young adults. *Chronobiol Int*. 2013;30(5):691-8.

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

27. Goldstone A, Baker F, de Zambotti M. Actigraphy in the digital health revolution: still asleep? *Sleep*. 2018;41(9).

28. de Zambotti M, Baker F, Colrain I. Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep*. 2015;38(9):1461-8.

29. Jeong IC, Bychkov D, Searson P. Wearable Devices for Precision Medicine and Health State Monitoring. *IEEE Transactions on Biomedical Engineering*. 2018.

30. Beattie Z, Oyang Y, Statan A et al. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol Meas*. 2017;38(11):1968-79.

31. Fonseca P, Weysen T, Goelema M et al. Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults. *Sleep*. 2017;40(7).

32. Schäfer A, Vagedes J. How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *Int J Cardiol*. 2013;166(1):15-29.

33. de Zambotti M, Baker F, Willoughby A et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav*. 2016;158:143-9.

34. Khandoker A, Karmakar C, Palaniswami M. Comparison of pulse rate variability with heart rate variability during obstructive sleep apnea. *Med Eng Phys*. 2011;33(2):204-9.

35. Benedetto S, Caldato C, Bazzan E, Greenwood D, Pensabene V, Actis P. Assessment of the Fitbit Charge 2 for monitoring heart rate. *PloS one*. 2018;13(2):e0192691.

36. de Zambotti M, Trinder J, Silvani A, Colrain I, Baker FC. Dynamic coupling between the central and autonomic nervous systems during sleep: a review. *Neurosci Biobehav Rev*. 2018;90:84-103.

37. Fonseca P, Long X, Radha M, Haakma R, Aarts R, Rolink J. Sleep stage classification with ECG and respiratory effort. *Physiol Meas*. 2015;36(10):2027-40.

38. Willemen T, Van Deun D, Verhaert V et al. An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification. *IEEE J Biomed Health Inform*. 2014;18(2):661-9.

39. Beattie Z, Pantelopoulos A, Ghoreyshi A, Oyang Y, Statan A, Heneghan C. Estimation of sleep stages using cardiac and accelerometer data from a wrist-worn device. *Sleep*. 2017;40(Suppl 1):A26.

40. Aktaruzzaman M, Rivolta M, Karmacharya R et al. Performance comparison between wrist and chest actigraphy in combination with heart rate variability for sleep classification. *Comput Biol Med*. 2017;89:212-21.

41. Herlan A, Ottenbacher J, Schneider J, Riemann D, Feige B. Electrodermal activity patterns in sleep stages and their utility for sleep versus wake classification. *J Sleep Res*. 2018:e12694.

42. de Zambotti M, Goldstone A, Claudatos S, Colrain I, Baker F. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int*. 2018;35(4):465-76.

43. de Zambotti M, Rosas L, Colrain IM, Baker FC. The Sleep of the Ring: Comparison of the ŌURA Sleep Tracker Against Polysomnography. *Behav Sleep Med*. 2017;21:1-15.

44. Bartsch RP, Liu KK, Bashan A, Ivanov PC. Network Physiology: How Organ Systems Dynamically Interact. *PLoS ONE*. 2015;10(11):e0142143.

45. Younes M, Raneri J, Hanly P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. *J Clin Sleep Med*. 12(6):885-94.

46. Younes M, Kuna S, Pack A et al. J Clin Sleep Med. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*. 2018;14(2):205-13.

47. Montgomery-Downs H, Insana S, Bond J. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;16(3):913-7.

48. Meltzer L, Hiruma L, Avis K, Montgomery-Downs H, Valentin J. Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents. *Sleep*. 2015;38(8):1323-30.

49. Mantua J, Gravel N, Spencer RM. Reliability of Sleep Measures from Four Personal Health Monitoring Devices Compared to Research-Based Actigraphy and Polysomnography. *Sensors*. 2016;16(5):646.

50. Cook J, Prairie M, Plante D. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *J Affect Disord*. 2017;217:299--305.

51. Kang S-G, Kang JM, Ko K-P, Park S-C, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res*. 2017;97:38-44.

52. Maskevich S, Jumabhoy R, Dao P, Stout J, Drummond S. Pilot Validation of Ambulatory Activity Monitors for Sleep Measurement in Huntington's Disease Gene Carriers. *J Huntingtons Dis*. 2017;6(3):249-53.

53. Liang Z, Martell MAC. Validity of Consumer Activity Wristbands and Wearable EEG for Measuring Overall Sleep Parameters and Sleep Structure in Free-Living Conditions. *Journal of Healthcare Informatics Research*. 2018;2:152–78.

54. Cook J, Eftekari S, Dallmann E, Sippy M, Plante D. Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography. *J Sleep Res*. in press:e12789.

55. Toon E, Davey M, Hollis S, Nixon G, Horne R, Biggs S. Comparison of Commercial Wrist-Based and Smartphone Accelerometers, Actigraphy, and PSG in a Clinical Cohort of Children and Adolescents. *J Clin Sleep Med*. 2015;12(3):343-50.

56. de Zambotti M, Claudatos S, Inkelis S, Colrain I, Baker F. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chrobiol Int*. 2015;37(2):1024-8.

57. Gruwez A, Libert W, Ameye L, Bruyneel M. Reliability of commercially available sleep and activity trackers with manual switch-to-sleep mode activation in free-living healthy individuals. *Int J Med Inform*. 2017;102:87-92.

58. Cook JD, Prairie ML, Plante DT. Ability of the Multisensory Jawbone UP3 to Quantify and Classify Sleep in Patients With Suspected Central Disorders of Hypersomnolence: A Comparison Against Polysomnography and Actigraphy. *J Clin Sleep Med*. 2018;14(5):841-8.

59. Pesonen A, Kuula L. The Validity of a New Consumer-Targeted Wrist Device in Sleep Measurement: An Overnight Comparison Against Polysomnography in Children and Adolescents. *J Clin Sleep Med*. 2018;14(4):585-91.

60. Sargent C, Lastella M, Romyn G, Versey N, Miller D, Roach G. How well does a commercially available wearable device measure sleep in young athletes? *Chronobiol Int*. 2018;35(6):754-8.

61. Lambrechtse P, Ziesenitz V, Cohen A, van den Anker J, Bos E. How reliable are commercially available trackers in detecting daytime sleep. *Br J Clin Pharmacol*. 2018;84(3):605-6.

62. Meltzer LJ, Walsh CM, Peightal AA. Comparison of actigraphy immobility rules with polysomnographic sleep onset latency in children and adolescents. *Sleep Breath*. 2015;19(4):1415-23.

63. Corral-Peñafiel J, Pepin J, Barbe F. Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome. *Eur Respir Rev*. 2013;22(129):312-24.

64. Berry R, Brooks R, Gamaldo C et al. AASM Scoring Manual Updates for 2017 (Version 2.4). *J Clin Sleep Med*. 2017;13(5):665-6.

65. Kales A, Rechtschaffen A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Bethesda, MD; 1968.

66. Baayen R. Mixed-effects models. In: AC Cohn, C Fougeron, MK Huffman editors. *The Oxford handbook of laboratory phonology*: Oxford University Press; 2012, pp. 668-77.

67. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-10.

68. Werner H, Molinari L, Guyer C, Jenni O. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Arch Pediatr Adolesc Med*. 2008;162(4):350-8.

69. Ludbrook J. Confidence in Altman-Bland plots: a critical review of the method of differences. *Clinical and experimental pharmacology & physiology*. 2010;37(2):143-9.

70. Bland J, Altman D. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-60.

71. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*. 1994;6(4):284-90.

72. Atkinson G, Nevill A. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217-38.

73. Metz C, Herman B, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med*. 1998;17(9):1033-53.

74. Feinstein A, Cicchetti D. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543-9.

75. Byrt T, Bishop J, Carlin J. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-9.

76. Roane B, Van Reen E, Hart C, Wing R, Carskadon M. Estimating sleep from multisensory armband measurements: validity and reliability in teens. *Journal of sleep research*. 2015;24(6):714-21.

77. Ridgers N, Hnatiuk J, Vincent G, Timperio A, Barnett L, Salmon J. How many days of monitoring are needed to reliably assess SenseWear Armband outcomes in primary school-aged children? *J Sci Med Sport*. 2016;19(12):999-1003.

78. Lillehei AS, Halcón LL, Savik K, Reis R. Effect of Inhaled Lavender and Sleep Hygiene on Self-Reported Sleep Issues: A Randomized Controlled Trial. *J Altern Complement Med*. 2015;21(7):430-8.

79. Baroni A, Bruzzese J, Di Bartolo C, Shatkin J. Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. *Sleep Breath*. 2016;20(2):853-4.

80. Morgenthaler T, Alessi C, Friedman L et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. *Sleep*. 2007;30(4):519-29.

81. Van den Bulck J. Sleep apps and the quantified self: blessing or curse? *J Sleep Res*. 2015;24(2):121-3.

82. Baron K, Abbott S, Jao N, Manalo N, Mullen R. Orthosomnia: Are Some Patients Taking the Quantified Self Too Far? *J Clin Sleep Med*. 2017;13(2):351-4.

83. Khosla S, Deak MC, Gault D et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *J Clin Sleep Med*. 2018;14(05):877-80.

84. Crowley O, Pugliese L, Kachnowski S. The Impact of Wearable Device Enabled Health Initiative on Physical Activity and Sleep. *Cureus*. 2016;8(10):e825.

85. Kang S, Kang J, Cho S et al. Cognitive Behavioral Therapy Using a Mobile Application Synchronizable With Wearable Devices for Insomnia Treatment: A Pilot Study. *J Clin Sleep Med*. 2017;13(4):633-40.

86. Melton BF, Buman MP, Vogel RL, Harris BS, Bigham LE. Wearable devices to improve physical activity and sleep: A randomized controlled trial of college-aged African American women. *Journal of Black Studies*. 2016;47(6):610-25.

87. Luik AI, Machado PF, Espie CA. Delivering digital cognitive behavioral therapy for insomnia at scale: does using a wearable device to estimate sleep influence therapy? *npj Digital Medicine*. 2018;1(1):3.

88. Dunican I, Martin D, Halson S et al. The Effects of the Removal of Electronic Devices for 48 Hours on Sleep in Elite Judo Athletes. *J Strength Cond Res*. 2017;31(10):2832-9.

89. Rondanelli M, Opizzi A, Monteferrario F, Antoniello N, Manni R, Klersy C. The effect of melatonin, magnesium, and zinc on primary insomnia in long-term care facility

residents in Italy: a double-blind, placebo-controlled clinical trial. *J Am Geriatr Soc*. 2011;59(1):82-90.

90. Behar J, Roebuck A, Shahid M et al. SleepAp: an automated obstructive sleep apnoea screening application for smartphones. *IEEE journal of biomedical and health informatics*. 2015;19(1):325-31.

91. Daly J, Roebuck A, Palmius N et al. SleepCare: obstructive sleep apnoea screening using mobile health technology. In: *Proceedings of the Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*. 2014. p. 1-4.

92. Centers for Medicare and Medicaid Services. Decision memo for sleep testing for obstructive sleep apnea (OSA)(CAG-00405N). *Medicare Coverge Database. Baltimore, MD: CMS*. 2009.

93. Rosen IM, Kirsch DB, Chervin RD et al. Clinical use of a home sleep apnea test: an American Academy of Sleep Medicine position statement. *J Clin Sleep Med*. 13(10):1205-7.

94. Shambroom J, Fábregas S, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. *J Sleep Res*. 2012;21(2):221-30.

95. Tonetti L, Cellini N, de Zambotti M et al. Polysomnographic validation of a wireless dry headband technology for sleep monitoring in healthy young adults. *Physiol Behav*. 2013;118:185-8.

96. Cellini N, McDevitt E, Ricker A, Rowe K, Mednick S. Validation of an automated wireless system for sleep monitoring during daytime naps. *Behav Sleep Med*. 2015;13(2):157-68.

97. Debellemaniere E, Chambon S, Pinaud C et al. Performance of an Ambulatory Dry-EEG Device for Auditory Closed-Loop Stimulation of Sleep Slow Oscillations in the Home Environment. *Front Hum Neurosci*. 2018;12:88.

98. Chambon S, Galtier M, Arnal P, Wainrib G, Gramfort A. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Trans Neural Syst Rehabil Eng*. 2018;26(4):758-69.

99. Goverdovsky V, von Rosenberg W, Nakamura T et al. Hearables: Multimodal physiological in-ear sensing. *Scientific Reports*. 2017;7.

100. Abraham J, White K. Tracking The Changing Landscape Of Corporate Wellness Companies. *Health affairs (Project Hope)*. 2017;36(2):222-8.

101. Henning A, van de Ven K. 'Counting your steps': The use of wearable technology to promote employees' health and wellbeing. In: Elsevier; 2017.

102. Swan M. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*. 2012;1(3):217-53.

103. Firouzi F, Rahmani AM, Mankodiya K et al. Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics. *Future Generation Computer Systems*. 2018;78(2):583-6.

104. Walch O, Cochran A, Forger D. A global quantification of" normal" sleep schedules using smartphone data. *Sci Adv*. 2016;2(5):e1501705.

105. Hill CM, Bucks RS, Cellini N et al. Cardiac autonomic activity during sleep in high altitude resident children compared to lowland residents. *Sleep*. 2018.

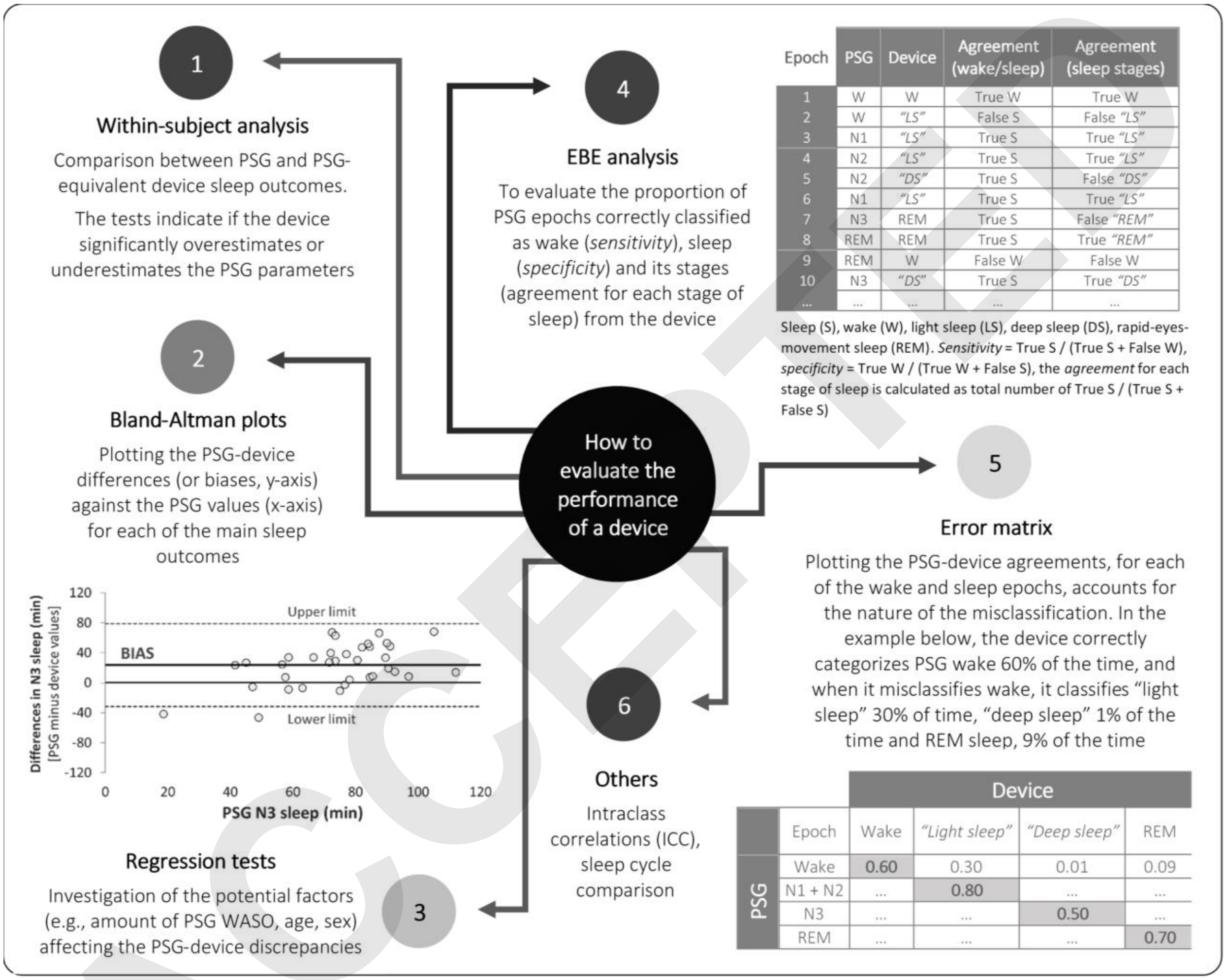
106. Grandner M. Sleep, Health, and Society. *Sleep Med Clin*. 2017;12(1):1-22.

Figure Captions

Figure 1 Recommendations for the analysis and evaluation of the performance of a consumer wearable sleep tracker against polysomnography (PSG). EBE, epoch-by-epoch; WASO, wake after sleep onset

Figure 2 Critical factors to consider when evaluating the potential use of a consumer wearable sleep trackers in research and clinical sleep settings

Figure 1


How to evaluate the performance of a device

1

Within-subject analysis

Comparison between PSG and PSG-equivalent device sleep outcomes.

The tests indicate if the device significantly overestimates or underestimates the PSG parameters

2

Bland-Altman plots

Plotting the PSG-device differences (or biases, y-axis) against the PSG values (x-axis) for each of the main sleep outcomes

Differences in N3 sleep (min)
[PSG minus device values]
120
80
40
0
-40
-80
-120
Upper limit
BIAS
Lower limit
0
20
40
60
80
100
120
PSG N3 sleep (min)

3

Regression tests

Investigation of the potential factors (e.g., amount of PSG WASO, age, sex) affecting the PSG-device discrepancies

4

EBE analysis

To evaluate the proportion of PSG epochs correctly classified as wake (*sensitivity*), sleep (*specificity*) and its stages (agreement for each stage of sleep) from the device

Epoch	PSG	Device	Agreement (wake/sleep)	Agreement (sleep stages)
1	W	W	True W	True W
2	W	*"LS"*	False S	False *"LS"*
3	N1	*"LS"*	True S	True *"LS"*
4	N2	*"LS"*	True S	True *"LS"*
5	N2	*"DS"*	True S	False *"DS"*
6	N1	*"LS"*	True S	True *"LS"*
7	N3	REM	True S	False *"REM"*
8	REM	REM	True S	True *"REM"*
9	REM	W	False W	False W
10	N3	*"DS"*	True S	True *"DS"*
...	...	...	...	...

Sleep (S), wake (W), light sleep (LS), deep sleep (DS), rapid-eyes-movement sleep (REM). *Sensitivity* = True S / (True S + False W), *specificity* = True W / (True W + False S), the *agreement* for each stage of sleep is calculated as total number of True S / (True S + False S)

5

Error matrix

Plotting the PSG-device agreements, for each of the wake and sleep epochs, accounts for the nature of the misclassification. In the example below, the device correctly categorizes PSG wake 60% of the time, and when it misclassifies wake, it classifies "light sleep" 30% of time, "deep sleep" 1% of the time and REM sleep, 9% of the time

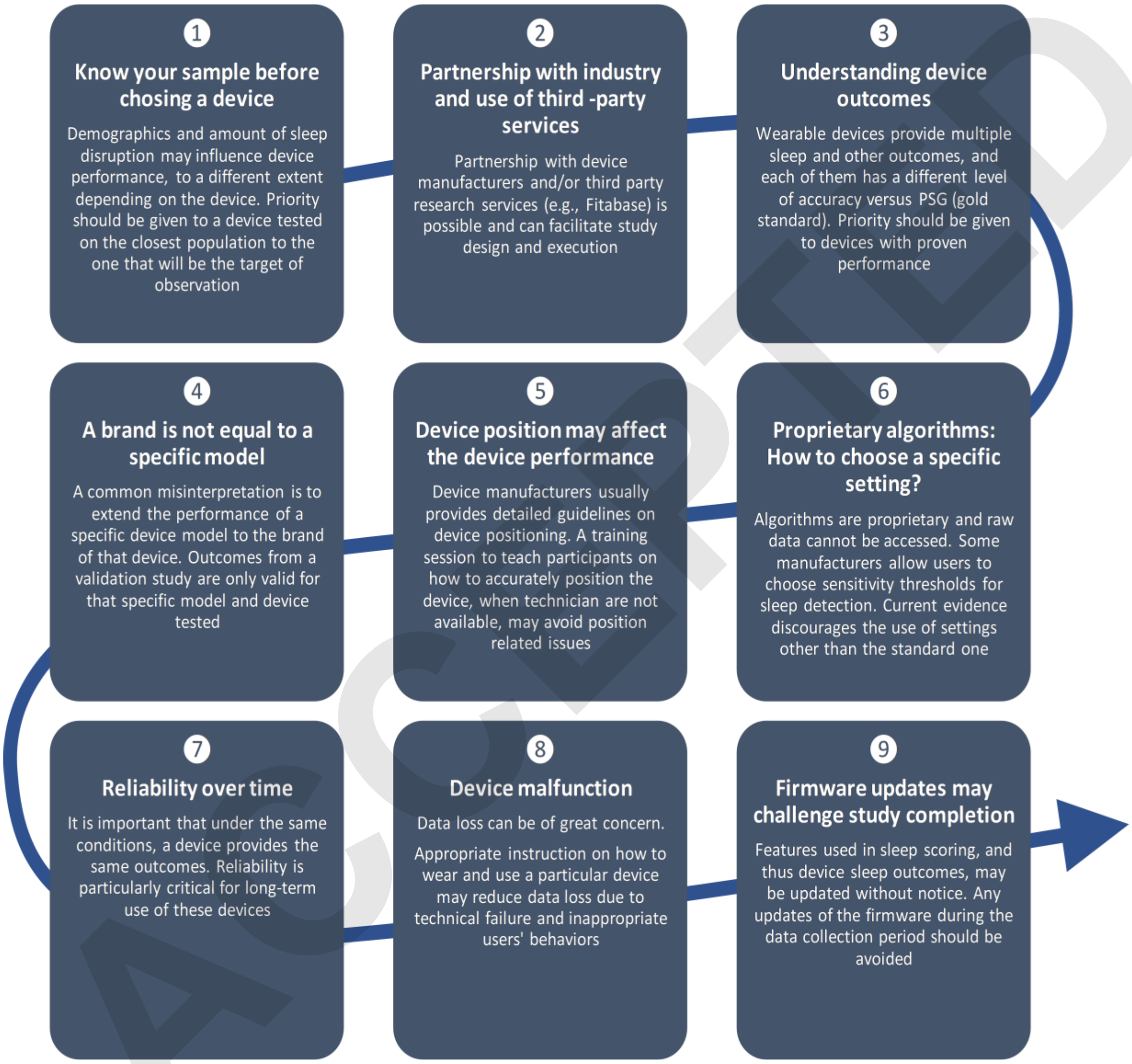
		Device			
	Epoch	Wake	*"Light sleep"*	*"Deep sleep"*	REM
PSG	Wake	0.60	0.30	0.01	0.09
	N1 + N2	...	0.80	...	...
	N3	...	...	0.50	...
	REM	...	...	...	0.70

6

Others

Intraclass correlations (ICC), sleep cycle comparison

Figure 2



1

Know your sample before chosing a device

Demographics and amount of sleep disruption may influence device performance, to a different extent depending on the device. Priority should be given to a device tested on the closest population to the one that will be the target of observation

2

Partnership with industry and use of third -party services

Partnership with device manufacturers and/or third party research services (e.g., Fitabase) is possible and can facilitate study design and execution

3

Understanding device outcomes

Wearable devices provide multiple sleep and other outcomes, and each of them has a different level of accuracy versus PSG (gold standard). Priority should be given to devices with proven performance

4

A brand is not equal to a specific model

A common misinterpretation is to extend the performance of a specific device model to the brand of that device. Outcomes from a validation study are only valid for that specific model and device tested

5

Device position may affect the device performance

Device manufactures usually provides detailed guidelines on device positioning. A training session to teach participants on how to accurately position the device, when technician are not available, may avoid position related issues

6

Proprietary algorithms: How to choose a specific setting?

Algorithms are proprietary and raw data cannot be accessed. Some manufacturers allow users to choose sensitivity thresholds for sleep detection. Current evidence discourages the use of settings other than the standard one

7

Reliability over time

It is important that under the same conditions, a device provides the same outcomes. Reliability is particularly critical for long-term use of these devices

8

Device malfunction

Data loss can be of great concern.

Appropriate instruction on how to wear and use a particular device may reduce data loss due to technical failure and inappropriate users' behaviors

9

Firmware updates may challenge study completion

Features used in sleep scoring, and thus device sleep outcomes, may be updated without notice. Any updates of the firmware during the data collection period should be avoided

Table 1 Peer-reviewed journal articles evaluating the performance of wearable sleep trackers against standard polysomnography (PSG). Results about comparisons between overnight summary sleep outcomes from wearables (or actigraphy, when available) and PSG (PSG-device biases) are reported. When available, results from epoch-by-epoch (EBE) analysis are reported. Sample characteristic, type of devices, and amount of PSG sleep disruption are also provided for each study to allow better interpretation of the study results. When not specified, wearables data were collected using the default (normal) setting, and PSG records were scored in 30-s epochs.

							PSG- device biases		**EBE analysis**	
Study	**Authors**	**Sample characteristics**	**Age Range** (or mean and SD)	**Standard Actigraphy type**	**Wearable Device type**	**PSG SE** (group mean)	**Standard Actigraphy** (mean, and SD of the biases when available)	**Wearable Device** (mean, and SD of the biases when available)	**Standard Actigraphy** (group mean)	**Wearable Device** (group mean)
2012 *In-lab*	Montgomery-Downs et al. (47)	24 healthy adults (10 female)	19 – 41 y	Actiwatch-64 (Minimitter, Inc.)	Fitbit “original” (Fitbit, Inc.)	< 85 %	Overestimated PSG SE (9.3 ± 9.7%) and TST (43.0 ± 46.6min)	Overestimated PSG SE (14.5 ± 10.7%) and TST (67.1 ± 51.3 min)	Sensitivity: 0.96 Specificity: 0.39	Sensitivity: 0.98 Specificity: 0.20
2015 *In-lab*	Meltzer et al. (48)	63 children (32 female). 23% of the sample had 1.5 ≥ AHI ≤ 5 (mild OSA), and 16% of the sample had AHI > 5 (moderate OSA). The analyses were conducted on 49 children due to several device failures	3 – 17 y	AMI Motionlogger (Ambulatory Monitoring, Inc.) or Actiwatch Spectrum (Phillips Respironics) (analyses were conducted on sub-groups of 12 children for devices)	Fitbit Ultra (Fitbit, Inc.) using both “normal” and “sensitive” settings	< 85 %	No significant PSG-actigraphy biases	Overestimated PSG TST by 41 min and SE by 8%, and underestimated WASO by 32 min, using the “normal” setting. Discrepancies > 100 min for TST and WASO and > 20% using the “sensitive” setting. With increasing AHI (as well as with increasing in developmental age), the mean PSG-device discrepancies increased using the “normal” and decreased using the “sensitive” settings	No direct comparison with PSG was performed	Sensitivity of 0.87 for the “normal”, and of 0.70 for the “sensitive’ setting. Specificity of 0.52 for the “normal”, and of 0.70 for the “sensitive” setting

2015 *In-lab*	Toon et al. (55)	78 children (27 female). 41% of the sample had 1 > RDI ≤ 5 (mild OSA), 28% of the sample had RDI > 5 (moderate OSA), 6% had PLMI > 5, 31% had a diagnosis of primary snoring. In addition, 51% of the sample had other comorbidities (e.g., chronic inflammation, behavioral disorders) and 29% were under medication (e.g., methylphenidate)	3 – 18 y	Actiwatch 2 (Phillips Respironics)	Jawbone UP	< 85 %	Underestimated PSG SOL by an average of 21 min	No significant PSG-device biases. However, Jawbone UP underestimated PSG SOL in those participants with primary snoring (mean difference of 9.7 min). Also, biases for TST and SE changes from underestimating to overestimating, across developmental age. Differently, the bias for WASO changed from overestimating to underestimating, across developmental age	Sensitivity: 0.93 Specificity: 0.63	Sensitivity: 0.92 Specificity: 0.66
2015 *In-lab*	de Zambotti et al. (28)	65 healthy adolescents (28 female)	12 – 22 y	-	Jawbone UP (Jawbone Inc.)	> 90 %	-	Overestimated PSG TST (10.0 ± 20.5 min) and SE (1.9 ± 4.2 %), and underestimated WASO (10.6 ± 14.7 min)	-	-
2015 *In-lab*	de Zambotti et al. (56)	28 midlife women (12 of them meeting the DSM-IV criteria for insomnia)	44 – 60 y	-	Jawbone UP (Jawbone Inc.)	< 85 %	-	Overestimated PSG TST (26.6 ± 35.3 min) and SOL (5.2 ± 9.6 min), and underestimated WASO (31.2 ± 32.3 min). No differences in device performance according to disease status	-	Sensitivity: 0.96 Specificity: 0.37 No differences in device performance according to disease status
2016 *At-home*	Mantua et al. (49)	40 healthy adults (19 female)	18 – 30 y	Actiwatch Spectrum (Phillips Respironics)	Basis Health (Intel, Corp.), Fitbit Flex (Fitbit, Inc.) Misfit Shine (Misfit, Inc.),	< 85 %	No significant PSG-actigraphy biases for TST and SE	Overestimation of PSG TST for both Misfit Shine (~ 75 min for the bias) and Withings Pulse O2 (~ 12 min for the bias), which also overestimated PSG SE with a bias > 5 %; Basis Health underestimated SE with a bias > 10 %. We decided not to report	-	-

					Withings Pulse 02 (Withings, Inc.)			results for sleep staging due to the unusual aggregation of PSG N3 + REM, considered as “deep” sleep		
2016 *In-lab*	de Zambotti et al. (33)	32 healthy adolescents (15 female)	12 – 21 y	-	Fitbit Charge HR (Fitbit, Inc.)	> 90 %	-	Overestimated PSG TST (8.0 ± 21.0 min) and SE (1.8 ± 4.5 %), and underestimated PSG WASO (5.6 ± 14.3 min)	-	Sensitivity: 0.97 Specificity: 0.42
2017 *In-lab*	Cook et al. (50)	21 unmedicated adults (17 female) with DSM-IV major depressive disorder	26.5 ± 4.6 y	Actiwatch 2 (Phillips Respironics)	Fitbit Flex (Fitbit, Inc.) using both “normal” and “sensitive” settings	< 85 %	Overestimated PSG TST (by 40.6 min) and SE (by 7.0 %), and underestimated SOL (by 13.5 min) and WASO (by 27.1 min)	Overestimated PSG TST (by an average of 46.0 min) and SE (by an average of 8.1 %), and underestimated WASO (by an average of 44.0 min) in the “normal” setting. Wide PSG-device biases (> 60 min for TST and WASO, and > 15 % for SE) for the “sensitive” setting	Sensitivity: 0.97 Specificity: 0.31	Sensitivity of 0.98 for the “normal”, and of 0.78 for the “sensitive” setting. Specificity of 0.35 for the “normal”, and of 0.80 for the “sensitive” setting
2017 *At-home*	Kang et a. (51)	33 drug-free individuals with (19 female) and 17 without (11 female) DSM-5 insomnia disorder	18 – 60 y	Actiwatch 2 (Phillips Respironics)	Fitbit Flex (Fitbit, Inc.) using both “normal” and “sensitive” settings	< 85 % in insomniac > 90 % in controls	Underestimated PSG TST (by an average of 17.8 min) and SE (by an average of 4.8 %) in controls. Underestimated PSG WASO (by an average of 21.6 min) in individuals with insomnia	Overestimated PSG TST (by an average of 6.5 min), using the “normal” setting in controls. Overestimated PSG TST (by an average of 32.9 min) and SE (by an average of 7.9% by 30.5 min), and underestimated WASO (by an average of), using the “normal” setting in individuals with insomnia. No data were provided for the “sensitive” setting	Sensitivity: 0.95 in controls and 0.96 in insomniacs Specificity: 0.61 in controls and 0.45 in insomniacs	Sensitivity of 0.97 (0.97 in insomniacs) for the “normal” and of 0.65 (0.64 in insomniacs) for the “sensitive” setting, in controls. Specificity of 0.36 (0.36 in insomniacs) for the “normal” and of 0.82 (0.89 in insomniacs) for the “sensitive” setting, in controls
2017 *In-lab*	Maskevich et al. (52)	7 participants (6 female) carrying Huntington’s gene, with disease	54.1 ± 6.4 y	Actiwatch Spectrum Pro (Phillips	Jawbone UP2 (Jawbone Inc.), and	Not provided. Sleep	Overestimated PST TST (by 74.0 ± 54.4 min)	Both Jawbone UP2 and Fitbit One overestimated PSG TST (by > 60 min) and SE (by > 15 %), and	Sensitivity: 0.97 Specificity: 0.31	Sensitivity of 0.99 for both Jawbone UP2 and Fitbit One.

		severity ranging from presymptomatic (N = 4) to early symptomatic (N = 3)		Respironics)	Fitbit One (Fitbit, Inc.)	was scored in 1 min epochs	and SE (by 14.8 ± 11.0 %)	underestimated WASO (by > 30 min)		Specificity of 0.34 for Jawbone UP2 and of 0.27 for Fitbit One
2017 *At-home*	Gruwez et al. (57)[a]	15 healthy adults Demographics unclear for the final sample analyzed	18 – 40 y	SenseWear Pro (BodyMedia, Inc.)	Jawbone UP MOVE (Jawbone Inc.), and Withings Pulse 02 (Withings Inc.)	> 90 %	No significant PSG-actigraphy biases	Withings Pulse 02 overestimated PSG TST (by an average of 33 min), TIB (by an average of 16 min), and SE (by an average of 5 %). No significant biases were found for Jawbone UP MOVE. We decided to do not report results for sleep staging due to the unclear classification of “light” and “deep” sleep from the device manufacturers.	-	-
2017 *In-lab*	de Zambotti et al. (43)	42 healthy adolescents (13 female)	14 – 22 y	-	ŌURA ring (Ouraring, Inc.)	> 90 %	-	Underestimated PSG N3 (19.6 ± 41.2 min) and overestimated REM (-17.2 ± 50.2 min)	-	Sensitivity of 0.96, and specificity of 0.48. Agreements for N1+N2 of 0.65, for N3 of 0.51, and for REM of 0.61
2018 *In-lab*	de Zambotti et al. (42)	44 healthy adults (26 female). Separate analyses on 9 with PSG evidences of PLMS > 15/h	19 – 61 y	-	Fitbit Charge 2 (Fitbit Inc.)	> 85 % in both groups	-	Overestimated PSG TST (9 ± 24 min) and N1 + N2 (34 ± 34 min), and underestimated SOL (4 ± 9 min) and N3 (24 ± 28 min), in the main group. Underestimated PSG N3 (28 ± 35 min) in the PLMS group	-	Sensitivity of 0.96 for the main, and 0.95 for the PLMS group. Specificity of 0.61 for the main, and 0.62 for the PLMS group. Agreement for N1+N2 of 0.81 for the main, and 0.78 for the PLMS group.

										Agreement for N3 of 0.49 for the main, and 0.36 for the PLMS group. Agreement for REM of 0.74 for the main, and 0.62 for the PLMS group
2018 *In-lab*	Sargent et at. (60)[b]	12 healthy elite athletes (sex not specified)	18.3 ± 1.0 y	-	Fitbit Charge HR (Fitbit, Inc.)	Not provided	-	No significant PSG-device biases for TST for Fitbit TST obtained in “automatic mode” for night-time sleep, as well as when bed timing was manually adjusted to match the bedtime opportunities for both night-time and day-time sleep (see notes below about the protocol). However, PSG-device discrepancies in TST in “automatic mode” was > 240 min in 4 participants. Lack of details on how data were obtained and analysis performed (please refer to the original study (60))	-	-
2018 *At-home*	Pesonen and Kuula (59)[c]	17 healthy children (9 female) and 17 healthy adolescents (8 female)	9 – 13 y children, 14 – 20 y adolescents	Actiwatch 2 (Phillips Respironics)	Polar A370™ (Polar Electro, Inc.)	> 95 %	Overestimated PSG WASO in children (by 20.9 min) and adolescents (by 14.3 min). Underestimated PSG TST in children (by 43.6 min)	Overestimated PSG WASO in children (by 24.4 min) and adolescents (by 12.5 min). Underestimated PSG TST in children (by 28.9 min) and adolescents (by 20.6 min)	Sensitivity: 0.93 in children and 0.93 in adolescents Specificity: 0.68 in children and 0.58 in adolescents	Sensitivity: 0.93 in children and 0.91 in adolescents Specificity: 0.77 in children and 0.83 in adolescents

							and adolescents (by 26.8 min).			
2018 *At-home*	Liang and Martell (53)[d]	25 healthy young adults (10 women)	24.8 ± 4.4 y	-	Fitbit Charge 2 (Fitbit, Inc.)	> 85 %	-	Underestimated PSG TST (by 12.3 min), SOL (by 11.1 min), N1 + N2 sleep (by 42.4 min), REM sleep (by 11.6 min), and % N3 sleep (by 10.2 %). Overestimated PSG WASO (by 24.5 min) and % of WASO (by 6,5 %), % of N1 + N2 sleep (by 13.8 %), % of REM sleep (by 4.6 %), N3 sleep (by 39.8 min)	-	-
2018 *In-lab*	Cook et al. (58)	43 clinical adult patients (29 females): 3 with a diagnosis of narcolepsy, 13 with idiopathic hypersomnia, 17 with idiopathic hypersomnia not otherwise specified, 6 with mild obstructive sleep apnea, and 4 with hypersomnolence related to another condition	33.3 ± 1.0 y	Actiwatch 2 (Phillips Respironics)	Jawbone UP3 (Jawbone Inc.)	> 85 %	Overestimated PSG TST (by 43.9 min) and SE (by 7.5%). Underestimated PSG SOL (by 12.9 min) and WASO (by 33.9 min).	Overestimated PSG TST (by 39.6 min) and SE (by 6.8%). Underestimated PSG SOL (by 5.1 min) and WASO (by 34.3 min)	Sensitivity: 0.97 Specificity: 0.31	Sensitivity: 0.97 Specificity: 0.39 Agreements for N1+N2 of 0.60, for N3 of 0.49, and for REM of 0.30
2018 *In-lab*	Cook et al. (54)	49 adult patients (46 females) with suspected central disorders of hypersomnolence: 14 with idiopathic hypersomnia, 19 with idiopathic hypersomnia not otherwise specified/unspecifi	30.3 ± 9.8 y	-	Fitbit Alta HR (Fitbit, Inc.)	> 85 %	-	Overestimation of PSG TST (by 11.6 min), SE (by 2%) and N3 sleep (by 18.2 min)	-	Sensitivity: 0.96 Specificity: 0.58 Agreements for N1+N2 of 0.73, for N3 of 0.67, and for REM of 0.74 *No significant differences in sensitivity, specificity and*

		ed, 6 with narcolepsy and 10 with mixed diagnoses								*agreements for sleep stages among sub-groups*

[a], unclear is how the time in bed (**TIB**) for the at-home PSG assessment was determined; [b], the experimental design of the study included 3 fixed night-time (10pm-7am; 11pm-7am; 12am-7am) and 2 fixed day-time (2pm-4pm; 3pm-4pm) “sleep opportunities” over three days. Analyses were based on averaged periods for night-time and day-time; [c], authors reported using a pre-product Polar fitness tracker corresponding to the commercially available Polar A370; authors calculated SE as TST/time between sleep onset and offset*100. Thus, the percentage does not account for the wake time between lights-off and sleep onset; [d], the comparison between Fitbit Charge 2 and PSG is based on a PSG portable clinical 1-channel EEG device (sleep stages were automatically analyzed in 30-s epochs and visually checked); **AHI**, apnea-hypopnea index (average number of apnea and hypopnea episodes per hour of sleep); **PLMI**, periodic limb movement index; **RDI**, respiratory disturbance index (average number of apnea and hypopnea episodes, and respiratory event-related arousals per hour of sleep); **REM**, rapid-eye-movement; **SOL**, sleep onset latency; **SE**, sleep efficiency; **TST**, total sleep time; **WASO**, wake after sleep onset

Table 2 An example of a qualitative approach to evaluate device performance.

An alternative way to evaluate the performance of a device is whether it adequately (compared to PSG) captures a significant literature effect (e.g., a group difference in sleep architecture between healthy individuals and those with a sleep disorder, sleep recovery after cognitive-behavioral treatments in insomnia sufferers, sleep alterations following acute stress-inducing experimental manipulation). For example, similarly to PSG, we previously showed that a multisensory sleep tracker (the ŌURA ring) was able to significantly detect the age-related decline in N3 sleep in an adolescence sample. This finding is encouraging given that the device showed its greatest limitation in PSG N3 classification (51% agreement in detecting PSG N3 sleep) (43).

Table 3 Assuming that the proprietary algorithms used by consumer wearables will remain proprietary, what else can the wearable industry do to facilitate the use of consumer wearable sleep-trackers in clinical and sleep research settings?

Open access to raw data	Allows application of publicly available algorithms to wearable raw accelerometer data (and/or plethysmography derived IBIs) obtaining a standardized sleep stage classification
Allow the choice of a specific version of the proprietary algorithm used for sleep classification when exporting/extracting sleep data	Allows consistency for data collection within a study period, by avoiding uncontrollable algorithm updates that may affect sleep parameter calculations Also allows researchers to choose a specific wearable device model using a specific algorithm with proven validation
Have a separate line of products more aligned with research and clinical needs	Would remove many concerns of using an uncontrolled consumer product for research and clinical sleep assessment
Increase partnership with sleep research and clinical centers	Allows access to domain expertise in basic sleep science and clinical sleep disorders, which can lead to consistent use of accepted terminology, and insight into the meaning and value of Big Data