SLEEP, 2024, 47, 1–31
https://doi.org/10.1093/sleep/zsad325
Advance access publication 27 December 2023
Review

Review

State of the science and recommendations for using wearable technology in sleep and circadian research

Massimiliano de Zambotti[1,2,*,†], Cathy Goldstein[3,†], Jesse Cook[4], Luca Menghini[5], Marco Altini[6], Philip Cheng[7], Rebecca Robillard[8,9]

[1]Center for Health Sciences, SRI International, Menlo Park, CA, USA,
[2]Lisa Health Inc., Oakland, CA, USA,
[3]Sleep Disorders Center, Department of Neurology, University of Michigan-Ann Arbor, Ann Arbor, MI, USA,
[4]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA,
[5]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy,
[6]Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands,
[7]Sleep Disorders and Research Center, Henry Ford Health, Detroit, MI, USA,
[8]School of Psychology, University of Ottawa, Ottawa, ON, Canada, and
[9]Canadian Sleep Research Consortium, Canada

*Corresponding author. Massimiliano de Zambotti, Center for Health Sciences, SRI International, 333 Ravenswood Ave, 94025, Menlo Park, CA, USA.Email: massimiliano.dezambotti@sri.com
†These authors contributed equally to this work.

Abstract

Wearable sleep-tracking technology is of growing use in the sleep and circadian fields, including for applications across other disciplines, inclusive of a variety of disease states. Patients increasingly present sleep data derived from their wearable devices to their providers and the ever-increasing availability of commercial devices and new-generation research/clinical tools has led to the wide adoption of wearables in research, which has become even more relevant given the discontinuation of the Philips Respironics Actiwatch. Standards for evaluating the performance of wearable sleep-tracking devices have been introduced and the available evidence suggests that consumer-grade devices exceed the performance of traditional actigraphy in assessing sleep as defined by polysomnogram. However, clear limitations exist, for example, the misclassification of wakefulness during the sleep period, problems with sleep tracking outside of the main sleep bout or nighttime period, artifacts, and unclear translation of performance to individuals with certain characteristics or comorbidities. This is of particular relevance when person-specific factors (like skin color or obesity) negatively impact sensor performance with the potential downstream impact of augmenting already existing healthcare disparities. However, wearable sleep-tracking technology holds great promise for our field, given features distinct from traditional actigraphy such as measurement of autonomic parameters, estimation of circadian features, and the potential to integrate other self-reported, objective, and passively recorded health indicators. Scientists face numerous decision points and barriers when incorporating traditional actigraphy, consumer-grade multi-sensor devices, or contemporary research/clinical-grade sleep trackers into their research. Considerations include wearable device capabilities and performance, target population and goals of the study, wearable device outputs and availability of raw and aggregate data, and data extraction, processing, and analysis. Given the difficulties in the implementation and utilization of wearable sleep-tracking technology in real-world research and clinical settings, the following State of the Science review requested by the Sleep Research Society aims to address the following questions. What data can wearable sleep-tracking devices provide? How accurate are these data? What should be taken into account when incorporating wearable sleep-tracking devices into research? These outstanding questions and surrounding considerations motivated this work, outlining practical recommendations for using wearable technology in sleep and circadian research.

Key words: wearables; clinical trials research; behavioral sleep medicine; circadian rhythms; actigraphy; sleep tracking

Submitted for publication: July 17, 2023; **Revised:** December 21, 2023

- Performance evaluation studies assessing the measurement qualities of wearable sleep trackers are increasingly published and promoted by the community. The currently available landscape for performance evaluation studies of wearable sleep-tracking technology is biased, i.e., most evidence is derived from study cohorts of young healthy adults under controlled laboratory conditions with little representation of minorities and individuals with chronic illnesses or sleep disorders. There is often no or minimal regard for other physiologic or external factors that may challenge the performance of these devices in real-world applications as intended. Here, we provide recommendations on how to interpret and contextualize studies evaluating the performance of wearable sleep-tracking technology and provide considerations when integrating wearable sleep-tracking in research.
- Automatically processed outputs from *contemporary research/clinical-grade devices* should not be exempt from rigorous performance evaluation due to the "*research/clinical-grade*" status of the device.
- While widely used and perceived as easy to use by most, *consumer-grade devices* hold hidden complexity (e.g., data access, fees, privacy, and security) driven by the fact that these devices are largely designed for consumers and not for research/clinical use.
- Data preprocessing steps may be required to properly use and interpret large-scale wearable sleep-tracking technology data.
- Careful interpretation of study results based on wearable sleep-tracking technology data is needed.

Why do we need recommendations for the use of wearable sleep-tracking technology?

Over recent years, the capabilities and modalities to measure sleep in free-living conditions have drastically changed. From the consumer space, several new devices with sleep-tracking capabilities have been introduced to the market (e.g., wearables, nearables, and in-bed sensors). Moving beyond traditional sleep/wake assessments, many of these devices now measure sleep-related physiology (e.g., breathing rate, skin temperature, and cardiac autonomic indices during sleep) and generate proxies of sleep stages and cardiopulmonary sleep-related events (e.g., O2 desaturation). They offer opportunities for continuous, unobtrusive, and large-scale sleep monitoring in the individual's typical sleeping environment. Though the appropriateness of their use remains a matter of debate, these largely unregulated devices (mainly wearables) are increasingly adopted in sleep research and investigations across other scientific disciplines. Patients frequently bring wearable acquired data to their healthcare provider with variable uptake. The characteristics of a consumer product (e.g., user-centered design and functionalities and maintenance of manufacturer intellectual property), the lack of clarity surrounding the algorithms applied to the acquired physiological signals and the ability of the output to approximate gold-standard measures, and the lack of recommendations regarding appropriate research and clinical use are of concern.

On the other hand, the field of research and clinical-grade ambulatory sleep monitoring is rapidly evolving with resultant challenges for end users. For instance, Philips Respironics recently announced the discontinuation of the Actiwatch, the most widely used and well-accepted US Food and Drug Administration (FDA) cleared wrist-worn actigraph. In parallel, a new class of research/clinical-grade devices has emerged, featuring multiple sensors, machine learning-based sleep stages classifications, wireless communication protocols, and cloud services similar to commercial sleep trackers while providing the ability to access raw signal from their devices and maintaining transparency around algorithm development and software updates, similar to traditional actigraphy.

Sleep researchers and clinicians now face the difficult task of selecting accurate and reliable alternative sleep-tracking devices that are acceptable to study participants and patients without the necessary foundational knowledge to make an informed decision. The Sleep Research Society (SRS) received requests from its members and others in the sleep field for recommendations and support regarding the utilization of wearable sleep-tracking devices for sleep and circadian research in this rapidly evolving technological landscape. In response to these requests and to promote the informed and proper use of sleep-tracking technology, the SRS recruited a panel of experts to provide state of science and recommendations for using wearable technology in sleep and circadian research. The goals of this manuscript are to provide answers regarding the data, accuracy, appropriate selection, and implementation of wearable sleep-tracking devices. The panel members were carefully selected to ensure comprehensive coverage of expertise in various aspects related to the utilization of sleep wearable technology. This included individuals with deep insights into both research and clinical applications, as well as those well-versed in the domains of sleep and circadian rhythms. Moreover, we incorporated perspectives from both academia and industry to offer a well-rounded view. Our panel also boasted a diverse array of competencies, encompassing the rationale and practical applications of wearable technology, the intricate hardware and software components that underpin these devices, and the nuances of processing data collected through wearables. Additionally, we considered the essential statistical factors crucial in the analysis of wearable data. Throughout this collaborative effort, the panel convened for several online meetings. These sessions served as a platform for discussions, debates, and collective decision-making. The culmination of these interactions resulted in the crafting and endorsing the final manuscript. In addition, the SRS reviewed and endorses the major findings of this manuscript.

Here, we cover wearable sleep-tracking technology only due to its wide adoption, with particular emphasis on consumer-grade devices. We specifically refer to devices physically donned by an individual that monitor non-electroencephalographic (EEG) signals (i.e., motion, pulse, temperature) to provide an estimate of sleep-related parameters. For instance, these include the widely used wrist-worn smartwatches and, most recently, rings. Therefore, EEG devices including ambulatory polysomnography (PSG) and consumer-grade headbands as well as devices that use photoplethysmography (PPG) for the main purpose of monitoring oxygen saturation and quantification of other respiratory and cardiovascular parameters, home sleep apnea tests, and other devices used for cardiopulmonary or seizure monitoring, are not covered here.

We would like to emphasize that it is not the intention of the panel to endorse or discredit any product. However, given the nature of this work, it was deemed necessary that examples of commercial entities and specific devices be provided to the reader. The examples were chosen based on different criteria including market shares and popularity of products, as well as specific relevant features offered by a device and/or company.

Introduction to wearable sleep-tracking technology

A comprehensive understanding of sleep and its interactions with health and disease requires that sleep is measured accurately, reliably, efficiently, and longitudinally. PSG is recognized as the gold-standard measurement for sleep. Many worthy characteristics substantiate PSG as the gold standard. However, inherent limitations (e.g., cost, time, specialized personnel resources, dependence on manual scoring with imperfect inter-rater reliability, to name a few) reduce the scalability of PSG sleep assessments and prevent longitudinal and ecologically valid (i.e., in the everyday environments of the sleeper) sleep measurement. Over the years, different solutions have been proposed to measure sleep outside the laboratory. Among them, actigraphy, i.e., wearable devices with an embedded accelerometer, gained popularity.

Although somewhat arbitrary, to help readers navigate the sleep wearable ecosystem, we have classified wearable devices into three main groups: *Traditional research/clinical-grade actigraphy*, *consumer-grade devices*, and *contemporary research/clinical-grade devices*. While there is some overlap, these classes broadly differ in key characteristics (Table 1). It is worth mentioning that the line between consumer-grade and research/clinical-grade devices has been blurred, with a substantial gray area between wellness and health tools. For example, it is now common to see certain functions that have been cleared by the US FDA as "software as a medical device" (e.g., atrial fibrillation detection [1]) applied to devices that are not cleared by the FDA.

Key milestones [2–19] for the wearable sleep-tracking technology field are provided in Figure 1.

Table 1. Key characteristics across different classes of wearable devices

	Traditional research/clinical-grade actigraphy	Consumer-grade devices	Contemporary research/clinical-grade devices
Example(s)	Philips Respironics Actiwatch2 (discontinued), Ambulatory Monitoring, Inc., Motionlogger, Axivity AX3-6	Fitbit's devices, OURA's rings, Samsung Galaxy Watches	GeneActive, Empatica EmbracePlus, Belun ring, SleepImage, Oxitone
Sensor(s)	3-Axis accelerometer + additional sensors (e.g., ambient light, event marker, temperature, etc.)	3-Axis accelerometer and PPG + additional sensors (e.g., skin temperature, skin conductance, ambient light, barometer, gyroscope, GPS, ECG, etc.)	3-Axis accelerometer and PPG + additional sensors (e.g., skin temperature, skin conductance, ambient light, etc.)
Feedback(s)	No	Audio/visual (e.g., display, embedded speaker), haptic (e.g., vibration)	Audio/visual (e.g., display, embedded speaker), haptic (e.g., vibration)
External cloud service(s)	No	Yes	Yes
API/SDK	Certain devices only (e.g., AX3-6)	Yes, to some extent	Yes, to some extent
Raw data access	Yes (mostly)	No, for most of the cases	Yes
Communication protocol(s)	Wire (e.g., USB)	Wireless	Wireless
Embedded algorithm(s)	Mostly, open-access (e.g., Cole–Kripke and Sadeh) algorithms to classify sleep/wake pattern	Proprietary algorithms to classify sleep, activity, and other aspects of health/wellness	Open-access and proprietary algorithms to classify sleep, activity, and other aspects of health/wellness
Target audience	Mainly, researchers and clinicians	Mainly, consumer	Mainly, researchers and clinicians
App-based interface	No	Yes	Yes
Capability	24/7 sleep/wake and activity assessments	24/7 sleep/wake and stages, activity, and other physiological data (e.g., cardiorespiratory function) assessments.	24/7 sleep/wake and sleep stages, activity, and other physiological data (e.g., cardiorespiratory function) assessments.
Battery life	Typically, 30 + days	2–10 days, largely varying depending on the sensors' configuration settings.	2–30 + days, largely varying depending on the sensors' configuration settings.

This classification (arbitrary, defined by the expert panel) aims to help the readers understand important differences and considerations when choosing between research/clinical-grade devices and consumer-grade devices. We would also like to recognize that traditional actigraphy evolved. Nowadays, research/clinical-focused wearables leverage similar technological advantages brought by "the consumer wearable revolution". We wish to underscore that it is not our intention to either endorse or discredit any product. Nonetheless, due to the nature of this work, we found it necessary to offer readers examples of commercial entities and specific devices. These examples were selected based on various criteria, including market presence, product popularity, and the specific relevant features offered by a device or company. PPG, photoplethysmography; ECG, electrocardiography; GPS, global positioning system

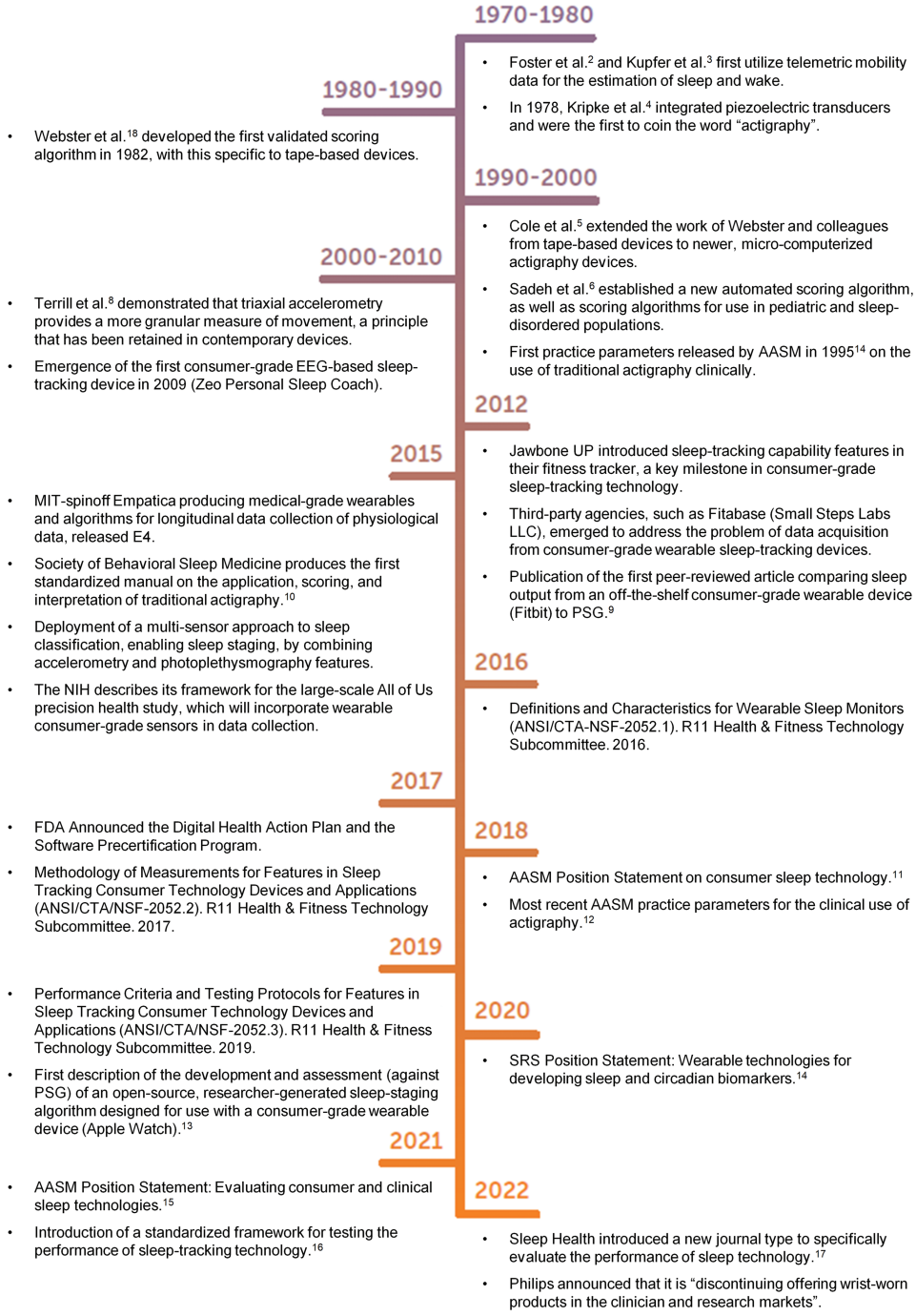

Figure 1. Timeline for innovation, technological advancements, regulation, and standardization initiative.

Traditional research/clinical-grade actigraphy

Traditional actigraphy has served a longstanding and important role as a cost-efficient and unobtrusive accepted alternative to PSG for objectively estimating sleep longitudinally, both for research and clinical purposes [20]. Fundamentally, actigraphy strictly relies upon changes in gravitational acceleration detected through piezoelectric sensors to assess movement, with "validated" and publicly accessible algorithms employed to translate movement into estimates of sleep and wake states. Although actigraphy can be collected from multiple body locations (e.g., ankle, legs, waist), wrist actigraphy is the predominant approach, with devices usually worn on the nondominant wrist [21].

Traditional actigraphy has notable limitations. For example, it applies a simple dichotomous classification approach (sleep vs. wake), which prevents a more comprehensive description of sleep architecture. In addition, traditional actigraphy has consistently been shown to demonstrate a poor ability to accurately detect wake episodes during the attempted sleep period, which is generally thought to be an inherent limitation of the strict reliance on accelerometry and the absence of other biological signals. Ultimately, this results in the device often misclassifying motionless wake as sleep. As such, performance abilities of traditional actigraphy will vary greatly based upon the behavior of the user, as well as the proportion of wake during a sleep period. Moreover,

no empirically determined, standardized setting configurations exist, despite research highlighting that device performance is significantly influenced by acquisition settings, with the out-of-the-box settings potentially resulting in notable bias within certain populations (e.g., among persons experiencing certain sleep disorders, such as insomnia [22], hypersomnia, and movement-related sleep disorders, as well as other conditions including neurodegenerative movement disorders) [23]. It is worth mentioning that traditional actigraphy can be relatively expensive (upwards of $1000 per device), requires patients/participants to return the device for data download, and relies upon trained personnel for the cleaning and interpretation of the data. Consequently, although less intrusive than PSG, these inherent limitations reduce the scale and scope of actigraphic sleep measurement.

Box 1 shows a historical perspective of traditional research/clinical-grade actigraphy.

Consumer-grade devices

The landscape of available tools for objectively and conveniently measuring sleep in free-living conditions began to change with the emergence of commercially available sleep-tracking products at the end of the 2000s. In 2009, the Zeo Personal Sleep Coach took the scene with a novel headband tracking brain activity and movement during sleep. Data were summarized on a bedside display or smartphone application. Yet, it was the arrival of the Jawbone UP in 2011 that introduced a motion-based sleep-tracking feature as part of a fitness-tracking band. Although these two technologies are currently out of business, this marked the beginning of the boom of consumer-grade sleep-tracking technology across the past decade. Box 2 shows a historical perspective of consumer-grade wearables.

Currently, consumer-grade sleep-tracking devices vary notably in approach (i.e., contact-based/wearable or contactless/nearable), hardware (i.e., wrist-worn wearables vs. ring wearables vs. EEG headbands; variation in incorporated sensors), and software (i.e., scoring algorithms) capabilities, and performance [21, 24]. Most of these consumer-grade devices do not exclusively target sleep but offer a broad range of features (e.g., tracking activity, heart rate [HR]). As such, sleep has different degrees of centrality/importance among different devices, depending on the use cases the device company focuses on, though many feature sleep as the primary output. Initially, the original purpose of wearables revolved around the quantified self to promote wellness, now, many devices cross the fine line between wellness and medical products. Most of these devices collect clinical parameters (e.g., oxygen saturation, electrocardiogram [ECG]) and include FDA-cleared features like atrial fibrillation detection.

The most common sensors embedded in consumer-grade wearable devices are accelerometry and PPG. However, different devices use a broad range of additional sensors (e.g., skin conductance and temperature) enabling the measurement of sleep together with many other markers of health with a single tool. The integration of PPG in consumer-grade wearables was transformative, as PPG measured pulse and pulse rate variability (a proxy for HR variability [HRV]) and allowed for the estimation of sleep stages given the characteristic HR and HRV signatures marking each stage. PPG pulse rate variability also provides a surrogate marker for the autonomic nervous system function.

Wearables usually have a companion mobile app, and frequently a web interface, where more details are accessible. Most wearables utilize wireless connectivity and have different modules to provide user feedback (audio/visual display, haptic sensor). Wearable data are typically processed both in the dedicated firmware (e.g., HR) and in the cloud (e.g., sleep data). Battery life varies widely across devices and is dependent on the sensors used and settings.

Box 1. Traditional research/clinical-grade actigraphy: a historical perspective.

The origin of using movement data to estimate sleep and wake dates to Foster et al. [2] and Kupfer et al. [3] in the early 1970s. However, it was not until 1978 that Kripke and colleagues coined the term "actigraphy," while also advancing the technology to utilize piezoelectric transducers [4]. The next two decades brought major evolutions largely in the form of algorithmic advancements (e.g., validated automated scoring) that greatly improved measurement accuracy relative to polysomnography (PSG), and enhanced the overall utility of actigraphy as a sleep measurement tool in free-living conditions. Specifically, Webster et al. developed the first automated scoring algorithm in 1982 [18], but this was specific to tape-based devices. The work of Cole et al. [5] further advanced actigraphy by establishing a validated scoring algorithm for micro-computerized actigraphic devices [20]. Subsequently, Sadeh et al. [6] developed a new approach for the automatic scoring of data collected with micro-computerized actigraphic devices, while also establishing the first algorithms for use in pediatric and sleep-disordered populations.

Actigraphy progression largely stalled until 2010, when Terrill et al. [8] demonstrated that triaxial accelerometry, improved the capture of more types of movements. Subsequent generations of traditional actigraphy largely adopted this hardware change by implementing microelectromechanical systems that yielded relatively prolonged recordings of triaxial accelerometry in the raw form and at high resolution. Over the recent years, advances in the field of actigraphy have been largely focused on improving algorithms through the integration of machine learning sleep and wake scoring algorithms and new or evolved hardware [26].

Actigraphy notably grew in popularity as a research tool across the 1990s, with a multitude of investigations emerging that employed the technology in free-living conditions, prompting the American Academy of Sleep Medicine (AASM) to release the first practice parameters for the use of actigraphy clinically in 1995 [7]. In 2003, Ancoli-Israel et al. [37] produced a review paper for the AASM on the role of actigraphy in the measurement of sleep and circadian rhythms, which informed an updated clinical practice parameters [19]. Over the past two decades, the AASM has produced updated versions of the clinical practice parameters, grounded in research findings derived from actigraphic studies, with the most recent published in 2018 [12]. Yet, an important advancement in the application of actigraphy for both research and clinical purposes came from the efforts of Ancoli-Israel et al. [10] in 2015 on behalf of the Society of Behavioral Sleep Medicine, which resulted in a standardized manual for applying, scoring, and interpreting actigraphy. These recommendations further solidified actigraphy's role as the primary tool for objectively measuring sleep and circadian rhythms in free-living conditions.

Box 2. Consumer-grade wearables: a historical perspective.

The earliest generations of consumer-grade wearable sleep trackers, including devices such as the Jawbone UP and Fitbit Flex, were designed for the primary purpose of activity tracking and relied solely on a single-sensor accelerometer; therefore, sleep tracking was a supplementary feature. These devices generally showed poor performance in differentiating sleep from wake relative to polysomnography (PSG), and worse performance characteristics than traditional actigraphy.

Initially, only summary data were available from single-sensor consumer-grade wearable devices until third-party companies (e.g., Fitabase [launched in 2012; Small Steps Labs LLC]) provided researchers with more granular data at a cost. However, raw data was and remains largely unavailable. Over the coming years, newer device generations demonstrated improved performance characteristics, despite no apparent changes in hardware, which suggested improved scoring algorithms. The integration of photoplethysmography (PPG), affording the ability to capture heart rate, was a notable advancement for consumer-grade wearables as the estimate of heart rate variability allowed for sleep staging (i.e., light, deep, and rapid-eye-movement sleep). The Jawbone UP3 (2015) was among the first consumer-grade devices that included both accelerometry and PPG sensors [158]. Adding PPG greatly improved sleep quantification estimations (evident when both contextualizing historical performance statistics and empirically evaluated) [105], but the first generation of multi-sensor devices still demonstrated poor sleep classification abilities.

Over the recent years, countless multi-sensor consumer-grade wearables have been developed, with these either being newer generations of older models or novel devices entering the marketplace. Sleep-staging classifiers have markedly improved during this time, correctly identifying sleep stages about 50%–70% of the time [60, 159]. Moreover, some multi-sensor devices also measure blood oxygen levels (SpO2) which may have clinical relevance (e.g., in screening for sleep-disordered breathing). At this juncture, in well-designed, highly controlled in-laboratory protocols, modern multi-sensor consumer-grade wearable devices are either comparable or superior to traditional actigraphy in sleep measurements. In addition to superior performance, consumer-grade wearable sleep trackers have the advantage of a significantly lower cost than traditional actigraphy, with many devices being available for less than 100–200 USD.

The prominent use of consumer-grade wearable sleep trackers among the general population is also notably attractive, given the capacity to expand the scope and scale of population-level sleep research. With improved machine learning algorithms and other innovations related to data acquisition, processing, storage, and exchange, the utility and potential promise of these devices to objectively measure sleep in free-living conditions continues to progress.

We would like to also warn the readers that despite the apparent simplicity, appropriate use of consumer-grade sleep-tracking devices for research/clinical purposes is quite challenging and holds hidden complexity. As the name implies, consumer-grade devices are designed for consumers. This confers advantages, which include optimized user interface and user experience designs and functionalities increasing the ease of operation. However, given that the primary purpose is sleep tracking to inform the consumer, there are significant limitations for research and clinical use cases, including undisclosed (black-box) algorithms, (typically) inaccessible raw data, privacy & security concerns, lack of control over software updates, unknown consistency and reliability of hardware components, and uncertain long-term availability of a device and model. While these shortcomings may be overcome in the future, at the current time, the use of consumer-grade sleep-tracking devices for research and clinical purposes requires acceptance of these limitations.

We would like to point the readers to the following resources for further details, pros and cons, capability, and use of consumer-grade sleep-tracking wearable technology [21, 24, 25].

Contemporary research/clinical-grade devices

The traditional actigraphy field is evolving and there is a growing availability of "contemporary" research/clinical-grade wearable devices. These devices share characteristics of both traditional clinical/research-grade actigraphy (e.g., raw data availability, disclosure of sleep-classifier algorithms) and consumer-grade devices (e.g., wireless communication protocols, cloud services, raw accelerometry data storage as opposed to activity counts, integration of PPG sensors). Some identify the main distinction between traditional and contemporary accelerometer devices as the ability to access raw acceleration data prior to a reduction in activity counts [26]. While many of these devices maintain an exclusive focus on sleep tracking, most of them are positioned as multipurpose health tools. Of note, most contemporary research/clinical-grade wearable devices and associated services are available under subscription, which can pose certain challenges to research and clinical work. However, they do provide flexibility by allowing the user the capacity to select or even develop their own sleep classification algorithm.

What data can wearable sleep-tracking devices provide?

Wearable technology, and particularly multi-sensor devices, currently provide a broad range of measurements including "raw" sensor data and aggregate data, continuously over 24/7 periods. Largely (but not always), consumer-grade devices restrict raw data access. Beyond concerns related to intellectual property, this is often also due to the limited memory capacity and intended use of these devices. Most of these aggregate measurements are generated by either open-access or proprietary undisclosed algorithms, with the latter mainly applying to consumer-grade devices. The granularity and accuracy of both raw data and aggregate measures largely vary across devices (i.e., not all devices and not all the data have the same level of accuracy when compared to the gold standard).

A detailed review of the main signals and measures used in wearable sleep technology is provided elsewhere [24, 27, 28]. Here, we provide an overview of the main types of data provided by wearable technologies. For each type of data, we briefly describe the corresponding measures and provide practical warnings for using them in research/clinical work.

"Raw" Data

Here, we will use the term "raw data," to refer to the actual signal values recorded by the device sensors at a specific sampling frequency. Whereas they are often confused with pre-processed

high-granularity aggregate indicators (e.g., 30 s sleep staging), raw data typically allow for a more reproducible computation of the key measures used in research and clinical work [29]. Raw data are more commonly (but not always) available from research/clinical-grade devices, whereas most consumer-grade trackers only provide aggregate measurements. Table 2 shows an overview of the types of data that are commonly provided by wearable devices.

Aggregate indicators

Sleep measures

Sleep aggregated indicators refer to the classical overnight sleep parameters (e.g., total sleep time (TST), sleep efficiency (SE), time spent in different stages of sleep) as defined by American Academy of Sleep Medicine (AASM) standards [32]. Table 3 shows an overview of the common aggregated sleep measures provided by wearable devices.

Proxies of circadian measures

The human circadian system is a critical input into a broad range of behaviors, some of which can be captured with actigraphy data. As such, these actigraphy-based factors can sometimes be used to model outputs and/or inputs into the circadian system. These variables are admittedly imperfect because human behaviors are driven by multiple factors, including social and other behavioral factors that can be misalignment with physiology; however, evidence suggests that these measures remain informative. For example, bedtime is robustly correlated with dim light melatonin onset in healthy individuals (e.g., nonshift workers who are well-entrained) despite being influenced by multiple physiological and socio-behavioral factors [34–36].

A few considerations are important when considering the use of wearable devices for inferences regarding the circadian system. Most devices do not provide such outputs by default; instead, either high-granularity actigraphy data (i.e., steps, activity count, and/or light data) or the sleep measures derived from actigraphy data need to be processed further through additional software.

Cosinor analysis has long been applied to actigraphy to estimate acrophase, mesor, period, and amplitude of the rest-activity rhythm. A cosine curve with a period at or near 24 hours is fit to the data by the least-squares method and from that, acrophase (time to peak activity), amplitude (peak to nadir difference), and mean of the curve can be identified [37]. However, because circadian rhythms are sometimes non-sinusoidal, alternatives such as transformation of the cosine curve or other nonparametric methods often result in a better fit to activity data [37, 38]. The pseudo-F statistic captures the extent that an individual's sleep–wake activity conforms to the extended cosine model and is a marker of the strength of the circadian rhythm [37, 38]. However, given the prevalence of non-sinusoidal patterns in rest-activity rhythms, nonparametric approaches may provide a better proxy for circadian rhythms derived from actigraphy [39].

The use of models of the human circadian system (e.g., the Kronauer model [40] and its subsequent variations [41–43], or the Hannay Model [44]) has emergent evidence for producing accurate estimates of circadian phase, particularly when there could be misalignment between behaviors (e.g., sleep–wake timing) and physiology. These methods are usually compared against gold standard circadian variables that are collected under highly controlled laboratory settings, such as dim light melatonin onset. Importantly, the use of these approaches with wearable devices may rely on activity as the only available input into these light-based models, though evidence across multiple populations has supported the robustness of this method [45, 46], perhaps due to the limitations of measuring light on the wrist [47]. Additionally, these methods typically require longer data collection than sleep indicators, especially if there is significant day-to-day variability in activity and light exposure patterns. In a sample of fixed night shift workers with significant circadian disruption, two weeks of actigraphy data were able to produce estimates that had strong agreement with dim light melatonin onset [46]. Notably, these predictions are comparable to other behavioral proxies (e.g., bedtime) in healthy and entrained individuals, but are two to four times more accurate in populations with significant circadian misalignment (i.e., night shift workers) [45, 46].

Table 4 shows an overview of the common sleep timing and circadian proxy parameters that can be derived from wearable device collected data.

Activity

Table 5 shows an overview of the common activity data provided by wearable devices.

Respiratory

Table 6 shows an overview of the common respiratory data provided by wearable devices.

Cardiovascular

Most devices use continuous PPG data to process pulse rate (simply referred to as HR) and pulse rate variability (simply referred to as HRV). Single-time measurement of HR and HRV is also possible via the ECG technique, which some devices offer by instructing the user to touch an electrode placed on the device with their free hand. Care should be taken when deciding which sensor to use if HR and HRV data are of interest, due to several issues characterizing the PPG technique and the processing techniques applied to the PPG signal. Table 7 shows an overview of the common cardiac data provided by wearable devices.

How do wearable sleep-tracking devices perform?

How to interpret a performance evaluation ("validation") study

Performance evaluation studies are conducted to determine the capacity of wearable sleep-tracking technologies to accurately measure the indicators of interest. In this context, the term "performance evaluation" has been recommended instead of "validation" to account for the fast-paced continuous update of device features, which prevents establishing an absolute level of validity for a given device [17].

Briefly, the performance of a device is evaluated by simultaneously measuring sleep in the same individuals with both the device of interest and a reference method (usually gold standard, manually scored PSG). The device output is then compared with the reference output recorded during the same time interval at epoch-by-epoch and/or overnight summary levels, which provide distinct information regarding the performance of the device [16]. While performance evaluation can be potentially applied to any raw signal provided by the device (e.g., PPG, HR, temperature, acceleration), most studies focus on aggregated sleep indicators (e.g., TST) and high-granularity sleep/wake and sleep stage classification (e.g., 30- or 60-s epoch-by-epoch classification).

Here, we outline the main aspects to consider when reading a performance evaluation study and when interpreting the related

Table 2. "Raw" data from wearables

"Raw" data	Warning
Accelerometry. Most wearables available on the market include one or more accelerometry sensors. Whereas single-axis accelerometers are commonly employed by traditional research/clinical-grade devices, consumer-grade and contemporary research/clinical-grade devices typically use triaxial sensors (i.e., measuring acceleration on the *x*-, *y*- and *z*-axes). Note that raw accelerometer data typically comes at a relatively high sampling frequency (e.g., 20–100 Hz, meaning that 20–100 values per second are provided for each axis). Accelerometry data is often submitted to a first processing stage within the device to compute a set of features or aggregated variables, such as activity counts, steps, time spent at different intensities, and energy expenditure (e.g., calories). When raw accelerometry data is available, it can be processed using low-pass filters to isolate the gravity component, and potentially body position (depending on the location of the sensor), as well as bandpass filters to isolate the motion component (i.e., sensor acceleration net to the gravity component) [30].	*Access to raw accelerometry data is rarely provided, despite a recent industry trend towards improving data sharing (e.g., raw accelerometry data is now available for certain consumer-grade devices through an application programming interface/software development kit).* *Caution should be paid when classifying sleep or computing aggregated measures from raw acceleration with legacy algorithms, as many in the literature are adapted to signal from a single-axis accelerometer.*
Photoplethysmography (PPG). PPG uses light reflection or transmission to capture changes in blood volume during a cardiac cycle. Although PPG values are only indirect estimates of the actual changes in cardiac activity, being often expressed with arbitrary measurement units, such raw PPG data are also representative of physiological (e.g., vasomotor, blood pressure) activity and processes that cannot be captured by other signals. A few contemporary research/clinical-grade devices provide raw PPG signal values, enabling researchers and clinicians to apply standardized and more reproducible procedures for computing aggregate estimates of heart rate (HR) and heart rate variability (HRV) (e.g., via pulse peaks detection), rather than relying on the scores automatically computed by the device. Additionally, raw data might be used for the development of new algorithms and applications (e.g., PPG-based arrhythmia detection). PPG signal is highly susceptible to artifacts [24] and typically, can only be considered reliable under conditions of no movement when it comes to HRV analysis.	*Raw PPG signal is rarely accessible. Not having access to the PPG signal results in the inability to apply customized algorithms to analyze the PPG waveform, and failure to identify problems with signal quality or other potential issues that require the visual inspection of raw data, as opposed to the resulting aggregate indicators (e.g., heart rate or HRV).*
Pulse rate (second resolution). While not considered as true 'raw' data, some wearables provide high-granularity pulse rate (i.e., the PPG equivalent for HR) data, up to second resolution. This is not to be confused with inter-beat intervals (IBIs) or PP intervals (i.e., peak-to-peak intervals extracted from PPG), discussed below (Section 3.2.5).	*The pulse rate from PPG is frequently referred to as HR. It is important to consider that these are not the original values recorded by device sensors, but pre-processed data that have been elaborated based on predefined parameters.* *With the second resolution, pulse rate (or HR) HRV cannot be calculated.*
Peak-to-peak intervals (PP-intervals; beat-to-beat resolution). PP intervals refer to the actual time intervals between consecutive pulses (heartbeats), and as such, cannot be reported at a fixed sampling rate (e.g., when HR is 60 bpm, there will be 60 IBIs in one minute). PP intervals are required and usually used to process HR and HRV measures in a certain time window (Section 3.2.5). Most wearable sensors use a 5-min window to compute HR and HRV during sleep and do not provide PP-intervals, as they might not be transmitted from the wearable to the app or software, to save battery power and bandwidth. In these cases, IBIs are processed directly in firmware (i.e., on the device, outside the control of the researcher), and possible artifacts are also discarded at this stage if artifact removal is present. In certain cases, PP-intervals might come with a signal quality estimate generated by the wearable manufacturer, to enable the researcher to assess the likelihood of having collected high-quality data.	*Some wearables may refer to PP intervals (or IBIs) as pulse rate or HR on a beat-to-beat resolution. When no signal quality is reported, artifact removal should be used to clean the IBIs time series from potential issues, as PPG is particularly prone to artifacts. While conceptually the same as electrocardiography (ECG)-derived RR intervals, PP-intervals, are not necessarily equivalent in the context of certain applications, such as HRV analysis. In particular, outside resting conditions in healthy individuals, changes in blood pressure might result in inconsistencies between HRV derived from PPG and ECG, despite the fact that both RR and PP intervals are generally called IBIs* [31].
Electrodermal activity (EDA). EDA sensors (frequently named galvanic skin response or GSR sensors) are of growing popularity in wearable sensing technology. GSR data are used in modeling stress and mood, but also in sleep classification models. Specifically, the tonic (level) and phasic component (responses) of skin conductance are recorded by passing a weak constant voltage between two electrodes placed on the skin surface, and by applying low-pass filters (e.g., 5 Hz) to the resulting conductance values. In turn, such raw measurements (rarely accessible from consumer-grade devices) can be used to adjust sleep/wake and sleep stage classifications.	*Raw EDA is rarely accessible, with similar implications to those highlighted for acceleration and BVP. When analyzing EDA, it is necessary to consider both its thermoregulatory (e.g., sweating, hot flashes) and non-thermoregulatory origins (e.g., cognitively/emotionally induced arousal) depending on the application. It is important to note that EDA assessed at various body locations (e.g., finger, palm of the hand, wrist) may yield diverse outcomes and exhibit distinct relationships with the target processes of interest, such as stress.*
Temperature. Some devices can also estimate variations in "core" and peripheral hemodynamic status based on measurements of skin temperature (i.e., the temperature recorded from the skin surface, oscillating around 32°C and 35°C within the 0.01–2.00 Hz range). Surface thermistors and infrared thermopiles are often used to collect such raw measurements, which are rarely provided by contemporary consumer-grade and research/clinical-grade devices.	*Wearables may not report the temperature in absolute units but as relative changes with respect to a person's previous day or night average, making single time point measurements of limited utility. Additionally, during the day, data from wearables is often confounded by environmental temperature, poor contact between the sensor and the skin, and other issues that might cause the data to be less reliable.*

performance metrics. The two key analyses to focus on are the Bland–Altman plots, which quantify the discrepancies between the device and the reference method in measuring the overnight aggregated measure of interest, and the epoch-by-epoch comparison between the two methods.

Figure 2 shows two Bland–Altman plots [55]. In both plots, the x-axis represents the size of the measurement, that is the range of values over which any measurement can lie, based on the considered sample. This can be quantified either by computing the average between device and reference measurements [55] or, as

Table 3. Aggregate sleep indicators

Sleep indicator	Warning
High-granularity sleep/wake and sleep stage classification: While sometimes erroneously considered "raw data," high-resolution (e.g., 30-s or 1-min intervals) aggregate sleep staging is commonly provided by wearables. This consists of the output of the classification models used by these devices based on the available features (e.g., acceleration, heart rate, temperature). Two main types of outputs are usually provided by wearables: sleep stages (light, deep, and rapid-eye-movement [REM] sleep) and dichotomous sleep/wake classification.	*Although the availability of epoch-by-epoch sleep classifications allows for a more reproducible computation of aggregate sleep indicators, the algorithms are often undisclosed, and sometimes they are provided with a low sampling rate (e.g., 1-min or 5-min epochs, rather than standard 30-s epochs).*
Bedtime and wake-up time: Bedtimes and wake-up times are by definition subjectively reported behavioral indicators reflecting the time (hh:mm:ss) a person chooses to start trying to fall asleep and conclude their attempts at sleep, respectively. The time elapsed between bedtime and wake-up (time in bed, see below), is necessary for computing key standard indicators such as sleep onset latency, wake after sleep onset, total sleep time, and sleep efficiency. Although most devices provide automatic detection of the beginning and end of each "sleep period," such automatically determined times are usually an approximation based on sleep onset and the last sleep-to-wake transition. Due to the limited evidence on the accuracy of these automatic classifications and the tendency for individuals to flank the time in bed period with nonmoving, resting wakefulness, they should be used with caution. Some devices also allow manual setting/adjusting of the period attempting sleep (e.g., by manually initializing and ending a sleep period or by post hoc adjusting these intervals). This might be the preferred choice for research and clinical applications.	*Bedtime and wake-up time should only be used when self-reported or manually signaled by the users (e.g., by pressing a dedicated event-marker button, or by directly inputting the dedicated device app). When this is not possible, sleep diaries based on third-party applications, personal digital assistants, or paper-and-pencil questionnaires should be used instead.* *It is worth noting that event markers, as well as retrospective reports of bedtime and wake-up times, can also pose challenges, including potential issues like recall biases and difficulties in consistently and accurately pressing the marker.*
Time in bed (TIB): TIB is defined as the time between bedtime and wake-up time and thus is subjectively determined and affected by the same considerations reported above for bedtime and wake-up time. Particularly, most consumer-grade as well as standard actigraphy devices that provide automatic estimates of TIB are actually outputting "sleep period" durations based on motion/activity thresholds, identifying the first and the last epochs classified as sleep.	*TIB should be only referred to and used when bedtime and wake-up time are self-reported or manually signaled by the users. TIB is not the equivalent of the "sleep period" (typically reported from consumer-grade devices).*
Sleep onset (SO): SO refers to the time (hh:mm:ss) at which the person falls asleep (e.g., first wake-to-sleep transition) from a behavioral and physiological perspective (e.g., changes in motion, temperature, heart rate). In consumer-grade wearables, it is usually coincident with the onset of the "sleep period," which is the first epoch classified as sleep.	*SO operationalization should be confirmed by evaluating high-granularity (e.g., 30 s or 1 min) sleep classification data, to be identified as the timestamp corresponding to the first transition from wake to sleep. When included in data analysis, it can be operationalized as a continuous variable consisting of the time lag (minutes) from midnight (as a conventional arbitrary choice).*
Sleep onset latency (SOL): SOL is the time interval (minutes) from the subjectively reported time at which the person starts trying to fall asleep (i.e., bedtime) to the objectively determined SO time. As highlighted above, both information are necessary to reliably measure SOL.	*No device can provide SOL without a measure of the subjective determination of bedtime.*
"Sleep period" duration: Based on the above considerations, "sleep period" duration is the nonstandard time interval (minutes) between SO and the last sleep-to-wake transition, as automatically determined by the device. Although not included in AASM standards [32], we argue that this indicator can provide useful information for most applications, provided that researchers and clinicians are aware of its differences from TIB.	*As mentioned above, the "sleep period" is not equivalent to TIB. Some devices output multiple sleep periods at night, following proprietary logic. This segmentation may be incorrect. Procedures have been implemented to account for this potential issue [33].*
Total sleep time (TST): TST is the total time (minutes) classified as sleep within a TIB or a "sleep period" interval. It is among the most widely reported parameters in performance evaluation studies, and one of the most widely used in both empirical investigations and clinical interventions. Among wearable outputs, TST demonstrates the most consistent definition, greatest "accuracy," and least variation in accuracy across different devices.	*The TST is mainly dependent on the TIB/ "sleep periods." Please refer to those indicators for warnings.*
Wake after the sleep onset (WASO): WASO is the total wake time (minutes) within a TIB (or a "sleep period").	*Some devices allow specifying "sensitivity" thresholds/ consideration of "small awakenings." These settings affect the total amount of wake and should be carefully chosen based on the population under observation.* *The assessment of Wake After Sleep Onset (WASO) stands out as one of the primary limitations associated with actigraphy-based wearable sleep trackers, especially when considering their use in clinical populations with anticipated sleep disruptions.*

Table 3. Continued

Sleep indicator	Warning
Time spent in "light," 'deep', REM sleep: Sleep stages are provided for multi-sensor devices with sensor configuration including accelerometry and photoplethysmography. The algorithms used to stage sleep from consumer devices are largely unknown. However, the known changes in heart rate and heart rate variability that characterize sleep stages likely inform the sleep-staging algorithms of sleep-tracking devices (though the actual features utilized remain proprietary). Some devices use additional features from other sensors (e.g., skin temperature) as well as non-physiological features to model the within-night distribution of stages. For example, other features may be included based on the neurophysiology of sleep (e.g., circadian regulation of stage REM sleep) that make assumptions about the user (e.g., "normal" circadian entrainment), which may result in bias and inaccuracies in edge cases. "Light sleep" is usually equivalent to polysomnography (PSG) N1 + N2 sleep, while "deep sleep" is considered the equivalent of PSG N3 sleep.	*There are instances in which sleep stages are not provided by a device that typically provides them. These might include, for instance, excessively short sleep periods and low battery (e.g., causing the deactivation of some sensors or switching to reduced/intermittent sampling). In some cases, when sleep stages are not provided, a device still provides sleep/wake classifications or intermediate states (e.g., "restless" sleep). In these cases, it is currently advisable to use wake/sleep dichotomization only and treat intermediate classifications as wake.* *Sleep stage classification requires at least accelerometer and photoplethysmography (PPG) data. The use of an accelerometer sensor only is not sufficient to provide 4-level staging.*

Table 4. Aggregate circadian proxies

Circadian proxies	Warning
Sleep midpoint: The time between sleep onset and end of sleep.	*It is not typically provided as the default sleep output. It can be calculated using high-granularity sleep/wake and/or stages data.*
Rest-activity rhythms: Components of rest-activity rhythms often include amplitude (a proxy for the strength of the rhythm), mesor (mean level of activity), and acrophase (peak activity). These can be estimated with cosinor analysis or with nonparametric methods.	*These typically require additional computations beyond what is typically available with standard actigraphy software.*
Core body temperature minimum (CBT_{min}) and/or *melatonin secretion onset estimates:* Both CBTmin and melatonin secretion have been established as valid and reliable markers of SCN activity that also have physiological relevance to sleep and other functions.	*These indicators are estimated from higher-order analysis of actigraphy data by mathematical models of the circadian system* (e.g., *www.predictDLMO.com*).

previously recommended for sleep-tracking performance evaluation [16], by simply considering the reference measurements (as exemplified in the figure). In contrast, the y-axis represents the differences between the device and the reference method. In most cases, the differences are computed as device–reference, implying that values above zero indicate overestimations, whereas values below zero indicate underestimations. In other cases, studies may report reference–device, implying an opposite interpretation regarding device overestimation/underestimation. Importantly, both variables are expressed in the original measurement units (e.g., minutes for TST, % for SE), making the plot and the related metrics easily interpretable.

Figure 2A depicts the situation where the differences are evenly distributed over the size of the measurement. In other words, the estimated bias, that is the mean difference (device–reference) across participants/nights, is predicted to keep the same value regardless of the measurement value (uniform bias). In such cases, it is possible to determine whether the bias is overall significantly higher (overestimation) or lower than zero (underestimation) based on statistical testing. In our example, the mean difference in total sleep time is 6.75 min, indicating a tendency to slightly overestimate total sleep time compared to the reference method. However, since the 95% confidence intervals (CI) around the bias include zero (i.e., 95% CI = [–6.19% to 19.69%] min) it can be concluded that, on average, device-based measurements do not significantly differ from reference-based measurements.

Figure 2B shows a different scenario where the bias is not uniform but rather proportional to the size of the measurement. Specifically, it shows a negative proportional bias that only approaches zero for higher sleep efficiency measurements, whereas the device tends to overestimate sleep efficiency for measurements lower than 85%. In such cases, the bias can no longer be generally evaluated as significant vs. nonsignificant, because it strictly depends on the size of measurement, with important implications for certain device applications. For example, such a device might be unsuitable for a clinical trial aiming at evaluating an intervention to improve sleep efficiency. Considering a subject with a pre-intervention sleep efficiency lower than 85% and a post-intervention value of 95%, we cannot determine whether such a difference is due to the intervention or rather to the change in measurement error.

In addition to the mean bias, quantifying the systematic measurement error implied by the device, both plots also show the 95% limits of agreement (LOAs). The 95% LOAs quantifies the random variability of the differences around the bias, or the limits within which most differences are predicted to lie [55]. Both bias and LOAs are critical to evaluate device performance. Indeed, a device showing an average difference close to zero but very large LOAs might not be useful for some applications as it might provide very inaccurate measurements for some subjects. In Figure 2, LOAs are represented by gray solid lines. Similar to the bias, LOAs are usually plotted with their 95% confidence intervals (CI) (gray dashed lines) and they can be either uniform (i.e., parallel to the bias line, as in Figure 2A) or proportional to the size of measurement (i.e., narrower or wider LOAs for higher compared to lower measurement values), a condition termed heteroscedasticity. For instance, Figure 2B shows a device that tends to return more consistent and less randomly varying differences for higher SE measurements (negative heteroscedasticity), with similar implications to those considered above for proportional biases.

In summary, reading and correctly interpreting a Bland–Altman plot requires considering several aspects of how the plot

Table 5. Aggregate activity indicators

Activity indicator	Warning
Activity counts and steps: Most traditional sensors use accelerometers to derive activity counts, a unitless measure representative of motion. Activity counts are then used as the independent variable in the linear regression model developed to predict energy expenditure [48]. The typical limitations of these approaches are the following; the accuracy of the monitor is highly dependent on the activities used to develop the model and a single model does not fit all possible activities [49]. Additionally, activity counts might not be detected when motion is decoupled from the sensor location, e.g., when cycling while wearing a sensor at the hip or wrist. Consumer wearables have moved away from activity counts and provide more intuitive metrics, such as steps. Similar to activity counts, steps are estimated differently by each manufacturer, and therefore not directly comparable. Additionally, sensor location will impact the derived steps, e.g., more steps might be detected when a participant is moving their hands if the sensor is a watch, wristband, or ring, while other physical activities might be underestimated (e.g., cycling).	*Activity counts and steps suffer from similar limitations,* i.e., *the inability of a single sensor to capture a variety of body movements, often leading to underestimations of movement for certain activities* (e.g., *cycling, rowing, or activities with limited full body motion) and overestimation of movement for other activities* (e.g., *activities with a high level of hand or arm movement when using wrist-based sensors or rings*). *Steps across different devices may not be comparable.*
Energy expenditure: Energy expenditure estimates are normally derived from accelerometer data or combining accelerometer and heart rate (HR) data. Accelerometers exploit the relationship between motion and calories burned. However, the limitations just discussed, still apply (e.g., the sensor might be placed in a location where motion is decoupled from energy expenditure, such as the wrist or hand during cycling). HR data could be used in these cases to exploit the relationship between oxygen uptake and HR and estimate energy expenditure more accurately during exercise. However, HR monitors typically provide low accuracy for energy expenditure estimation during sedentary behavior [50], given that HR is affected by many other factors (e.g., stress and emotions). Additionally, the relationship between oxygen uptake and HR is highly individual and would require individual calibration for optimal accuracy [51, 52]. Photoplethysmography (PPG)-HR estimate is per se problematic, particularly during activities, with inaccuracy due to the several sources of artifacts on the PPG signal (Table 1). Thus, despite the theoretical advantage of using HR data to estimate energy expenditure, estimates derived from watches, wristbands, or rings might be impacted by compounded errors due to the potential inaccuracy of PPG-derived HR, an important predictor of the energy expenditure model.	*Energy expenditure estimations suffer from different limitations based on sensor location and types of signals used for the estimate* (e.g., *accelerometer only or combined accelerometer and HR*). *In most circumstances, energy expenditure estimates are of poor accuracy and have not been validated in different populations.*

Table 6. Aggregate respiratory indicators

Respiratory indicator	Warning
Breathing rate: Breathing rate is typically not measured directly by consumer wearables but is usually estimated from the photoplethysmography (PPG) signal based on respiratory sinus arrhythmia. As such, breathing rate is basically another pulse rate variability feature, looking at changes over a longer timeframe, with respect to the typical peak-to-peak features used in heart rate variability (HRV) analysis. Breathing rate estimation suffers from the issues reported in **Tables 2** and **7** in terms of PPG data quality, i.e., it can only be assessed reliably when there is limited or no motion.	*Limited validations of breathing rate algorithms are available in the scientific literature, making it difficult to trust the output provided.*
Oxygen saturation: Oxygen saturation (SpO2) is estimated by the ratio of the pulsatile and slow-varying component of the PPG signal.	*PPG sensors utilized by wearable sleep-tracking technologies are distinct from medical-grade pulse oximeters given reflection/reflectance mode, green spectrum light, and location of placement; therefore, blood oxygen saturation measurements from such devices should be interpreted with caution.*
Apnea-hypopnea index (AHI): The AHI is the number of apneas and hypopneas per hour of sleep with apneas and hypopneas identified by reductions in airflow measured by the nasal–oral thermistor and nasal pressure transducer. The AHI is used for objective confirmation of obstructive sleep apnea and has been adopted as a quantification of severity. Wearable-derived AHI is modeled from PPG signal [53, 54].	*Studies that assess the ability of wearable AHI to approximate PSG AHI or distinguish between individuals with and without obstructive sleep apnea are limited. Furthermore, the limitations noted above for raw data and aggregate respiratory indicators derived from wearables can introduce inaccuracy and bias to wearable predicted AHI values.*

was generated (e.g., the meaning of *x*- and *y*-axes, accounting for proportional biases and heteroscedasticity). Whereas the bias is probably the most immediately interpretable information, its interpretation without considering its trend over the size of measurement and the dispersion of the differences around it (LOAs) might be misleading.

Figure 3 shows examples of error matrices (or confusion matrices), which are the main output of epoch-by-epoch analyses. Error matrices are cross-tabular representations of the total number or proportion of epochs classified by the device and the reference method in each of two (e.g., sleep vs. wake) or more categories (e.g., sleep stages). An error matrix can be obtained either by summing the total number of epochs in each classification category across participants/nights (absolute error matrix, shown in Figure 3A) or by dividing each value by the corresponding marginal frequency (highlighted in gray) and then averaging such proportions across

Table 7. Aggregate cardiac indicators.

Cardiac indicator	Warning
Heart rate (HR) or pulse rate (PR): HR is the frequency of heartbeats, resulting from heart contractions each minute, and it is commonly measured with an electrocardiogram. Most wearables measure pulse rate, i.e., the equivalent of HR when using photoplethysmography (PPG) technology. Pulse rate and HR can be equivalent under certain circumstances [160]. The HR signal, when measured at rest (e.g., during sleep), can be considered highly reliable and less artifact-prone with respect to HR variability (HRV). However, not only the accuracy but also the timing—or protocol—of the measurement should be assessed. Some consumer-grade devices do not disclose or mix data collected during the day and during the night to provide a measure of "resting HR," an approach that might be suboptimal, as not only the physiological response but also the participant's behavior, will impact the data. Other devices use an approach that the authors would deem more accurate when it comes to assessing resting physiology, i.e., averaging the data for the entire night (e.g., Whoop or Oura). By taking a full average, changes due to the circadian rhythm and sleep stages will not add additional variability to the data, which is the case when using sporadic measurements or shorter time windows. During exercise, pulse rate tends to be of poor accuracy, for a number of reasons ranging from sensor positioning, fit, artifacts, and low signal-to-noise ratio [161]. Thus, pulse rate data should mostly be used when measured at rest, while heart rate data (e.g., from a Polar chest strap [162]) should be used for exercise measurements.	*HR and PR measurements from wearables tend to be accurate when measured in motionless conditions, such as sleep. The same cannot be said of measurements during movement or exercise. Moreover, the timing of the measurement matters for data analysis and interpretation, whereas some devices tend to capture data in different ways (e.g., sporadically, using day and night data, using only night data).* *High-granularity HR data can also be accessible by some devices (e.g., second or minute resolution) and these data can be combined with sleep-staging data (30-s resolution) for more accurate averaging of HR across the time windows of interest.*
HRV or pulse rate variability (PRV): HRV results from a number of factors, including parasympathetic modulation of heart rhythm, the baroreflex, mechanical stimuli, and hormones. When measured at rest, it is often used as a marker of parasympathetic modulation of heart rhythm in the context of the stress response (e.g., a stressor would cause a reduction in parasympathetic activity, reflected in a lower HRV, typically [163]). When it comes to measuring HRV, we have additional complexities with respect to HR. First, what wearables measure, using optical signals (PPG), is not HRV but PRV. While HR and pulse rate are equivalent, HRV (i.e., the electrical activity of the heart) and PRV (i.e., changes in pulse rate measured at the periphery, either the wrist, finger, or ear, normally) are not always equivalent. Measurements taken at rest in healthy subjects show an almost perfect correlation between HRV and PRV [160, 164], but this is not the case during exercise or in other populations [165] as pulse transit time and blood pressure might impact PRV and HRV differently. Secondly, the sampling window needs to be considered, similar to what we have discussed for HR. In this context, methods relying only on a few minutes of the night have shown to be unreliable (e.g., Whoop up to version 3 and Apple Watch, all versions [166]), while a better approach is to use the average of the entire night. Some wearables might provide researchers with additional tools able to extract not only PRV features, but also PP intervals (peak-to-peak intervals derived from the PPG signal), and therefore allow the researchers to compute additional features. Ideally, a wearable able to provide 5-min resolution HRV samples across the entire night or for at least 4–5 h during the night should be used in order to properly assess resting physiology during sleep. Sensors able to provide this type of data should be favored against sensors able to provide only sporadic measurements during the night or fewer data samples. Among the many HRV features that can be computed from RR or PP intervals, time domain features, and in particular rMSSD, should be favored when possible, due to their standardization and clear physiological interpretation.	*Consumer-grade wearables report PRV as a surrogate of HRV, which can be considered a valid alternative under certain conditions (i.e., measurements in healthy participants at rest). The protocol and sampling strategy used should be carefully analyzed, as some consumer wearables provide sporadic data points that are typically vulnerable to noise and artifacts. Importantly, a single artifact over a 5-min window can dramatically change HRV and PRV [167]. Artifacts can derive from noise in the signal, e.g., motion disrupting the PPG waveform, or actual cardiac abnormalities, such as ectopic beats or forms of arrhythmia. In these cases, most wearables will not report any issues as signal quality metrics are not provided. Without access to an electrocardiogram, issues due to sensor malfunctioning, movement, or heart rhythm abnormalities are indistinguishable, and therefore care should be taken to assess the likelihood of such issues in the study population of interest.*
Proprietary biomarkers without scientific or clinical comparator or relevance: Some consumer-grade devices provide aggregated metrics (e.g., global indices of sleep disturbances, general stress, and recovery/readiness scores) of unknown operationalization and questionable utility.	*Metrics lacking standard definition and/or operationalization are of limited utility and should not be used in clinical and research studies.*

participants/nights (proportional error matrix, shown in Figure 3B). Whereas the former provides an overall idea of classification performance, the latter is more informative of device classification metrics such as sensitivity (i.e., the proportion of reference-based epochs in a target stage/condition that are correctly classified by the device) and specificity (i.e., the proportion of reference-based epochs different than a target stage/condition that are correctly classified by the device), while accounting for individual variability. For instance, Figure 3B shows a device with an average sensitivity to "light" sleep of 79%, while showing that the remaining 4% of light sleep epochs are misclassified as wake (5%), "deep sleep" (6%), and rapid-eye-movement (REM) sleep (10%).

Epoch-by-epoch analysis can be considered as a more in-depth accuracy check than Bland–Altman plots. Whereas Bland–Altman plots are the first thing to look at, informing on-device performance at a macroscopic level (overnight aggregate indicators), epoch-by-epoch analysis zooms in at a microscopic level (single epoch) to inform whether the device is actually doing what it claims to do, namely whether it accurately classifies epochs of sleep/wake. Finally, it should be noted that such analyses are usually conducted with a single-night research design, only providing a snapshot of device accuracy and not considering its precision over multiple measurements. Future studies should use multi-night designs to cover such a lack of knowledge.

Detailed information and procedural steps in evaluating the performance of wearable sleep technology in terms of Bland–Altman, epoch-by-epoch, and other analyses are provided elsewhere [14, 16, 25]. Particularly, a standardized analytical framework to evaluate the performance of wearable sleep trackers, including related open-source R-based codes and functions, has been recently published to facilitate studies that evaluate the sleep-tracking capabilities of wearable sleep technologies [16]. The same pipeline has been recently extended and implemented in Python [56].

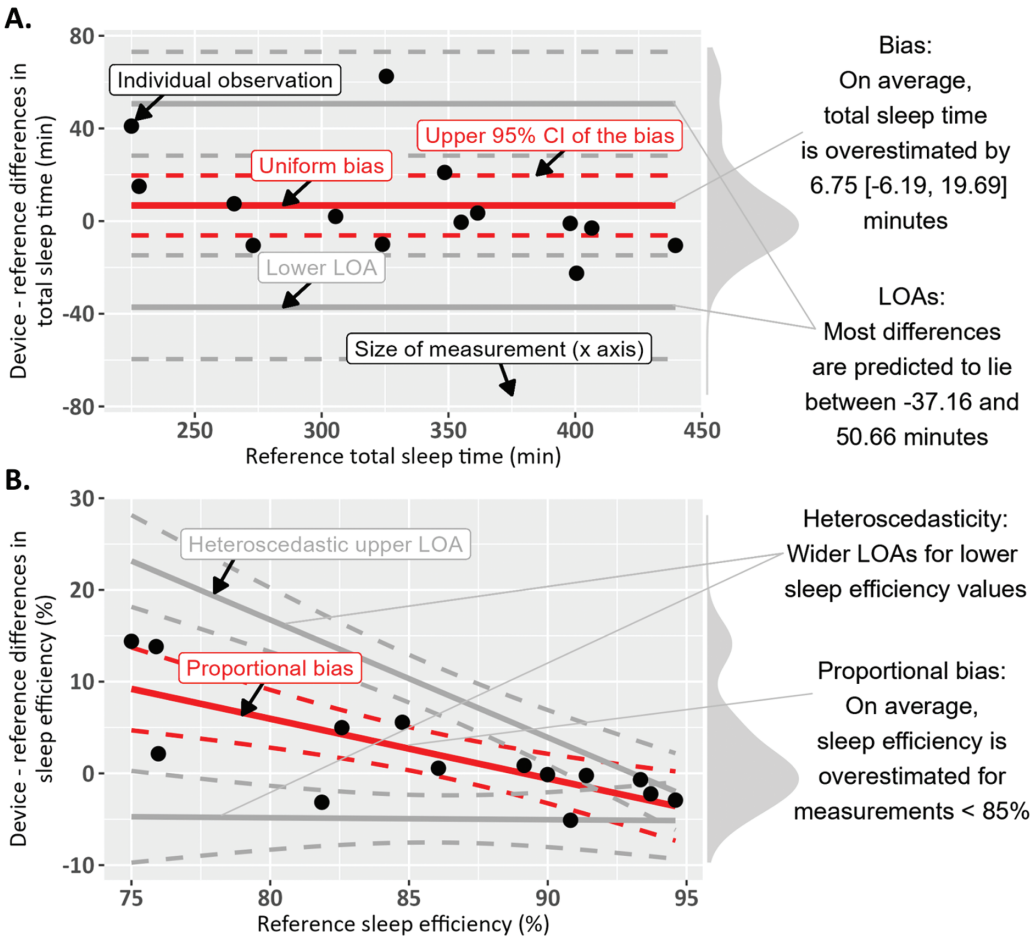

Figure 2. Examples of Bland–Altman plots. CI = confidence intervals; LOA = limit of agreement.

Overview of wearable devices performance

It is our position that any tools used in research/clinical settings require rigorous evaluation, i.e. comparison of sleep outputs compared to gold-standard PSG. As an overall practical warning, there is an erroneous assumption that because a device is considered a research/clinical tool and "validation" studies are available, the data provided are "good." For instance, traditional actigraphy has long been considered the gold standard alternative to PSG in non-laboratory settings, despite the low ability to correctly classify wake, which rarely exceeds 50%. In addition, we should not assume that the performance of contemporary research/clinical-grade devices is "good" just because their intended uses are research and clinical applications. These devices should go through the same rigorous evaluation and receive the same level of scrutiny that is applied to consumer-grade devices.

It is essential to acknowledge the increasing prevalence of industry-sponsored performance evaluation studies, paralleling inherent challenges, and the limited feasibility of conducting unbiased, independent assessments within academic settings. While acknowledging the good faith of an investigator, to strike a harmonious balance between dependent and independent performance evaluations, it is imperative to allocate research grants that specifically support impartial third-party evaluations. One effective approach could involve establishing supplementary grant mechanisms designed to facilitate and promote independent assessments.

While a review of the vast number of performance evaluation studies is outside the scope of this manuscript, we here provide a high-level overview of the current knowledge regarding the performance of sleep-tracking wearable technology.

Importantly, within the realm of sleep and circadian research, we lack a standardized operational framework for defining the threshold of what would be deemed "adequate" to validate and endorse the use of a device. Additionally, despite the availability of best practices in the evaluation and reporting of wearable sleep-tracking device performance [14, 16, 25, 56], some studies fail to adhere to these recommendations; therefore, the following should be interpreted with caution.

The sections that follow will discuss the capacity of wearable device-derived aggregate sleep parameters (as defined in Table 3) to approximate the equivalent PSG values. High-granularity sleep/wake and sleep stage classification performance is reported if epoch-by-epoch comparison was undertaken and summary sleep indicator (e.g., TST) accuracy is cited if Bland–Altman plots (and associated values) were generated in the study. Studies that compare wearable device-acquired non-sleep parameters (e.g., pulse oximetry) to corresponding gold-standard measures have utilized heterogeneous study protocols and statistical reporting; therefore, the available performance measures are cited.

Sleep–wake differentiation and sleep staging

Traditional research/clinical-grade actigraphy

The performance of standard actigraphy has been repeatedly tested against reference standards (PSG) across different samples and conditions. Despite being currently referred to as the accepted alternative to PSG in a nonlaboratory setting, standard

A.

	Device wake	Device light	Device deep	Device REM	Reference total
Reference wake	806	467	29	71	1373
Reference light	272	**3969**	359	501	5101
Reference deep	31	**1059**	885	13	1988
Reference REM	30	445	49	898	1422
Device total	1139	5940	1322	1483	9884

3,969 light sleep epochs were correctly classified as light sleep

1,059 deep sleep epochs were misclassified as light sleep

B.

	Device wake	Device light	Device deep	Device REM
Reference wake	0.61 (0.16) [0.52, 0.69]	0.31 (0.18) [0.22, 0.41]	0.02 (0.03) [0.01, 0.03]	0.06 (0.11) [0, 0.1]
Reference light	0.05 (0.03) [0.03, 0.07]	**0.79 (0.11) [0.73, 0.85]**	0.06 (0.07) [0.03, 0.09]	0.1 (0.07) [0.06, 0.13]
Reference deep	0.02 (0.02) [0, 0.03]	**0.52 (0.25) [0.39, 0.65]**	0.45 (0.25) [0.32, 0.59]	0.01 (0.02) [0, 0.02]
Reference REM	0.02 (0.02) [0, 0.03]	0.24 (0.28) [0.09, 0.37]	0.03 (0.04) [0, 0.04]	0.71 (0.31) [0.57, 0.88]

On average, 79% of light sleep epochs were correctly classified as light sleep

On average 52% of deep sleep epochs were misclassified as light sleep

Figure 3. Examples of absolute (A) and proportional error matrix (B) computed from epoch-by-epoch sleep staging comparison between a device and a reference method. In (B), proportions are reported as mean (standard deviation) (95% confidence intervals). REM = rapid-eye-movement.

actigraphy performance shows a profound limitation in wake assessment (referred to as low specificity in the binary sleep/wake classification). The ability to correctly classify epochs of nocturnal wake rarely exceeds 50% (at a level of sensitivity, i.e., the ability to correctly classify sleep epochs, >90%) with performance degrading as a function of the amount of wake time at night. Thus, the more wakefulness during the attempted sleep period, the less accurate the device is. The problem manifests in devices that can significantly underestimate wakefulness, potentially by hours. This limitation restricts their utility within the demographic that could derive the greatest benefit from precise sleep assessment, namely individuals afflicted with sleep disorders [57]. The low specificity of actigraphy is primarily due to the misclassification of motionless wake as sleep [37, 58].

Of note, to mitigate actigraphy's problem of low specificity for wake, one group used 70 h of PSG data to enrich their training dataset with wake epochs such that wake and sleep epochs were equivalent. This technique improved classification to a more optimal trade-off between sensitivity (89%) and specificity (80%) though specificity deteriorated during testing on an independent validation set [59].

Consumer-grade devices

The ability of consumer-grade sleep-tracking devices to approximate sleep as defined by PSG was previously considered inferior to FDA-cleared actigraphy, primarily due to the lack of high-quality evidence [11]. However, a growing body of literature has revealed that multi-sensor consumer-grade, sleep-tracking devices can differentiate sleep from wake similar to or better than research/clinical-grade actigraphy.

Epoch-by-epoch analyses reveal that overall, when consumer-grade devices (and their associated native algorithms at the time of the study) are compared to in-laboratory scored PSG, sensitivity for sleep is usually greater than 90%, while the sensitivity for wakefulness is relatively lower and more variable (ranging from 20% to 70%). See [20, 25, 60–66], for reference.

Studies in adults that report mean bias between PSG and consumer-grade device-derived summary sleep indicators typically demonstrate that devices overestimate PSG TST (by up to more than an hour), though some investigations demonstrated no significant bias, or an underestimation [60, 64, 66–68]. Device-derived sleep onset latency (SOL) measurements display either no significant difference from PSG SOL or mean discrepancies (both overestimates and underestimates) of less than 15 min. PSG wake after sleep onset (WASO) is typically underestimated by consumer-grade devices (by up to an hour), though some investigations demonstrated no significant bias or an overestimation of WASO [60, 64, 66–68]. In line with the observed misclassification tendencies, the discrepancy between PSG and consumer-grade device-measured SE displays greater magnitude differences in the overestimation (by up to 20%) than the underestimation direction [25, 60, 66]. These findings are distinct from those in the pediatric and adolescent population, where PSG TST may be underestimated and WASO overestimated by consumer-grade device output [61–63], though this was not always the case [69]. For both children and adults, proportional biases have been reported across indices of both sleep duration and disturbances, generally pointing toward a greater inaccuracy of the devices on nights with more disturbed sleep, a well-known recognized limitation in the actigraphy literature [57].

Epoch-by-epoch assessment of sleep staging demonstrates widely ranging accuracies, about 50%–90% for light sleep (PSG N1 + N2), and about 30%–80% for deep (PSG N3) and REM sleep derived from consumer-grade devices compared to PSG. Overall, devices have better and less variable performance for light sleep and REM sleep, compared to deep sleep. Currently, there are no clear performance trends for biases and proportionality of biases for sleep staging [20, 25, 60–66].

Importantly, while not always directly evaluated [66], there is a consistent trend indicating an overall improvement in performance over time for consumer-grade devices.

Contemporary research/clinical-grade devices

Modern research/clinical-grade devices provide access to raw acceleration as opposed to automatic data reduction to activity counts [70]. Therefore, instead of using algorithms that use activity counts as an input, sleep–wake estimates can be derived from these devices by using existing software such as GGIR (https://www.accelting.com/ggir-software/), an open-source R-package to process multi-day raw accelerometer data (m/s^2). Given the flexibility in analyzing data derived from contemporary research/clinical-grade devices, interpretation of performance evaluation studies that compare the output of these devices to polysomnogram must take into account the classifier and data cleaning methodologies used (e.g., sleep diary time in bed (TIB) versus algorithm determined sleep period time).

For example, wrist-worn raw acceleration analyzed with GGIR to differentiate between sleep and wake bounded by TIB determined with a sleep diary, demonstrated sensitivity of 91% and specificity of 45% in the epoch-by-epoch analysis [71]. In that study, TST was overestimated by about 30 min. No assessment of other sleep parameters or appraisal of proportional bias was available. When the same classifier (GGIR) was used without a sleep diary, instead employing an algorithm for sleep period detection, sleep onset, wake time, and sleep duration (measured at the right wrist) were not significantly different between the device and PSG in a sleep disorders population. However, sleep duration (left wrist) and SE were overestimated by 30 min and 9%, respectively. With the use of the same methodology, but in a healthy population, sleep onset was underestimated by 20 min, but no other significant biases in sleep parameters were observed. In both groups, visual inspection of Bland–Altman plots (for sleep duration) suggested a negative proportional bias, that is at lower sleep durations, the error was larger (particularly when the device was worn on the left wrist). However, this was not formally tested. Sensitivity to detect sleep was 92% and 93%, in clinic-based and healthy sleepers, respectively and specificity was not reported [71].

A more recent investigation went a step further, applying the GGIR sleep classifier (both with and without sleep diary) and the Cole–Kripke algorithm (with the Tudor-Locke algorithm to measure sleep period time) to determine sleep parameters from a contemporary research/clinical-grade device and then compared outputs to PSG. GGIR analysis, with and without a sleep diary, overestimated PSG TST by 31 and 26 min, respectively. This overestimation was even greater (47 min) with use of the Cole–Kripke sleep–wake classifier within bounds set by the Tudor-Locke algorithm. Agreement between wearable device and PSG TST was poor (intraclass correlation coefficient [ICC] = .27 to .44) across all three methods of accelerometry analysis. Longer SOL and wakefulness after sleep onset were associated with a greater overestimation of PSG TST. Sensitivity and specificity were not reported [72].

Despite differential methods, these findings are consistent with the tendency of motion-based sleep tracking to misclassify nonmoving wakefulness as sleep.

Machine learning analysis of data acquired with contemporary research/clinical-grade devices provides a novel methodology that may be superior to original algorithms to quantify sleep from raw acceleration data [73] and may even identify off-wrist time [74]. When the original van Hees sleep-staging algorithm and the random forest analysis of raw acceleration were compared again in a large data set of corecorded accelerometry and PSG, the epoch-by-epoch analysis revealed that the original method had better sensitivity but lower specificity than the machine learning algorithm (sensitivity = 84% vs. 78%, specificity = 48% vs. 56%) [26]. Mean biases between polysomnogram and device-derived sleep parameters were not provided.

Some multi-sensor contemporary research-grade/clinical devices have a nonsleep tracking primary indication (i.e., detection of obstructive sleep apnea (OSA), seizure detection). For these multipurpose tools, few publications are available describing their sleep-tracking capabilities compared to gold-standard. For example, the Belun ring, which records both accelerometry and PPG, is FDA-cleared for the detection of OSA but also provides TST estimates that are highly correlated with PSG ($r = .95$) [53]. Conversely, when compared to polysomnogram, cardiopulmonary coupling employed by the Sleep Image device showed a kappa value suggesting weak agreement with conventional sleep staging (44%) but estimated a measure of cortical arousal, the cyclic alternating pattern, with a kappa value of 77%, which falls into the range of substantial agreement [75]. The Empatica E4 (as well as the new model EmbracePlus), contains PPG, electrodermal activity, and an infrared thermophile in addition to a triaxial accelerometer; however, the investigation assessing its sleep-tracking capabilities compared to PSG used only the actigraphy-based sleep algorithm that demonstrated 96%–97% sensitivity and 39%–40% specificity [76].

Circadian measures

The utilization of wearables to track circadian measures is still a relatively new development, and as such information regarding its performance is still emerging. Except for a few recent studies, much of the extant literature has only tested performance in healthy adults with minimal sleep and circadian disruption. In studies with healthy adults, the mean absolute error (the difference between the measured value and the "true" value) has typically fallen within 2 h [77, 78]. One validation in a sample of fixed night shift workers with significant circadian disruption found a mean absolute error of 2.9 h, suggesting that these methods could be valuable in populations of clinical interest and relevance [46].

Despite being more novel, this is a rapidly growing area where we are likely to see continued improvements in performance. One promising mechanism may be the integration of multiple data sources (e.g., HR, temperature) and analytics as methods to improve the performance of wearables in tracking circadian measures. Further research regarding the accuracy required for interventions is also needed.

Respiratory parameters

The inclusion of PPG sensors in consumer-available, wearable sleep-tracking technologies allows for the computation of respiratory rate (RR) and apnea-hypopnea index (AHI, reflecting the number of apneas and hypopneas per hour of sleep) during sleep (Table 6). More recently, blood oxygen saturation measurement has become available [79]. Studies on the accuracy of these parameters compared to the gold standard are available, but more limited in number than assessments of sleep-staging performance.

Respiratory rate

Venous fluctuations during respiration result in low-frequency alternating current oscillations superimposed on the direct current baseline signal of PPG, allowing for quantification of RR [80, 81]. One investigation compared the RR measured by wearable sleep-tracking technology to respiratory inductance PPG during PSG and reported bias and precision error of 1.8% and 6.7%, respectively [82]. Similarly, another investigation demonstrated a high correlation ($r > .90$) between consumer-grade wearable and PSG-derived RR [83]. Notably, consumer-grade wearable tracking of RR was leveraged during the COVID-19 pandemic and demonstrates the potential public health ramifications of these devices during times of scarce resources [84].

Blood oxygen saturation

When tested in healthy individuals and those with chronic lung disease, the Apple Watch Series 6 displayed a strong correlation with fingertip pulse oximetry ($r = .81$) [85]. Skin color, wrist circumference, and the presence of wrist hair were not predictors of differences in readings from the devices [85]. A systematic review of oxygen saturation derived from the Apple Watch Series 6 reported LOAs from ±2.7%–5.9%, though outliers of 15% were reported [86]. Setting permissible error at 3% below or above readings from an ear lobe pulse oximeter, the Garmin Forerunner 245 had a 50% error rate that increased to 80% at altitude [87]. Importantly, at lower oxygen saturations, a greater error was observed [87].

However, a great deal of uncertainty surrounds the accuracy of even medical-grade PPG in measuring blood oxygen saturation. If an FDA-cleared pulse oximeter reads 90%, true blood oxygen saturation, measured by arterial blood gas (ABG), is between 86% and 94% (https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication; Accessed June 15, 2023). Therefore, ABG is considered the true gold standard for assessing blood oxygen saturation. At high altitude, compared to ABG, the Garmin Fenix 5X Plus displayed an ICC of .55 (ICC < .50 and ICC > .80 are considered poor and good reliability, respectively). The mean absolute percent error was 9.8%, and the mean oxygen saturation difference was 7% [88].

The differences observed when comparing oxygen saturation measurement from wearable sleep-tracking technologies to medical-grade methods are not only statistically significant but clinically significant. Therefore, this output should not be relied upon for medical decision-making [1].

AHI

Data from PPG and accelerometry have also been modeled to estimate the AHI, allowing for the identification of the presence and severity of OSA from wearable sleep-tracking technologies.

A deep learning algorithm applied to wrist-worn PPG and accelerometer approximated PSG AHI well, with a weighted Cohen's kappa = .51 and stratified individuals into OSA severity classes (area under the receiver operating characteristic curve of .84, .86, and .85 for mild, moderate, severe OSA, respectively) [54].

More recently, SpO2 data has been included as input to AHI estimation algorithms. For example, the use of both cardiac and oximetry measurements from the PPG (along with accelerometer) of a consumer-grade device allowed algorithm prediction of OSA (AHI ≥ 5) with accuracy, sensitivity, and specificity of 81.1%, 76.5%, and 100% [89]. A ring device that includes an FDA-cleared oximeter utilizes a neural network applied to ring-acquired oximetry, pulse rate, HRV, accelerometer, and PPG waveform data to predict TST and AHI [90]. The comparison of ring-estimated AHI correlated highly with AHI collected during PSG or home sleep apnea testing. Accuracy, sensitivity, and specificity in categorizing individuals with AHI ≥ 15 were .81 [95% CI, 0.70% to 0.89%], .93 [95% CI 0.77% to 0.99%], and .74 [95% CI, 0.59% to 0.85%], respectively [90]. Given the relevance of OSA for daytime function, management of certain chronic conditions, and risk of incident disease (particularly cardiovascular) [91], the lack of high-quality evidence regarding the ability of consumer-grade wearables to detect OSA prevents their use as a diagnostic tool.

Oxygen desaturation index (ODI) can represent an estimate of the AHI and ring-derived ODI demonstrated good correlation ($r = .91$) and close agreement with PSG AHI allowing for classification of OSA with a sensitivity of 87% and specificity of 83% [92].

The above examples demonstrate the capacity of consumer-available, wearable sleep-tracking technologies to estimate respiratory parameters, which makes these devices potential clinical tools. However, caution must be used, particularly when interpreting oxygen saturation and derived parameters.

HR and HRV measures

Less is known about the performance of wearables in measuring HR and HRV during sleep (e.g., nocturnal HR and HRV measures). Overall, there is a general consensus that the performance of PPG-based devices in the assessment of HR and HRV (usually ECG) is higher during sleep compared to during wake. This is largely driven by the low level of motion occurring during sleep, with motion being a major confounder in PPG sensor readings [1, 24].

In an interesting recent work from Miller et al. [93], the authors evaluated sleep HR and HRV measures from different wearable devices and compared them to gold-standard ECG. In that study, the Apple Watch S6 overestimated HR by an average of .5 beats per minute, had a mean bias of 1.5 beats per minute, and an intraclass correlation of .96. Similar performances were shown by Polar Vantage V, OURA Gen 2, and Whoop 3.0 devices, while Somfit devices showed the poorest performance. These outcomes are not too dissimilar from those provided by others and by earlier studies evaluating previous-generation wearable devices [62, 69, 94, 95]. Thus, HR estimates from wearable devices, particularly when averaged across the entire or extensive sleep periods, seem to have reasonable accuracy. On the other hand, except for Whoop 3.0, Miller et al. [93] showed less convincing data supporting the accuracy of PPG-based wearable technology in HRV (root mean square of successive differences between normal heartbeats) estimation, an area requiring further exploration.

Biases in data derived from performance evaluation studies and limitations in the real-world use of wearable sleep-tracking technology

While several performance evaluation studies do exist and more are upcoming, the available data supporting the performance of wearable devices are biased toward specific conditions, instrumentation, and patient populations. To maintain rigor and reproducibility, performance assessments compare the sleep output of wearable devices to manually scored in-laboratory PSG in carefully controlled study protocols; however, for ease and cost containment, these studies typically span only a single night and often use a limited convenience sample (participants of an ongoing study not specifically designed to test the performance of a wearable device, e.g. adults presenting to the clinical sleep laboratory for suspected sleep apnea at an academic center). It is

important to also consider that, as previously outlined by others [24], several factors can affect device performance (e.g., individual characteristics, environmental conditions), particularly for devices using features from multi-sensor signals for their sleep classification algorithms. For example, PPG readings are well known to be impacted by user characteristics (e.g., skin tone and thickness, hair, and ability to securely wear the device). Gender, tattoos, body temperature, RR, pressure, motion, and ambient light can also impact PPG recording and therefore, can introduce errors into sleep parameters derived from PPG signal [80, 96]. The dependency of EEG-based sleep stages on cardiac autonomic responses can also be affected by individual characteristics, disease conditions, and medication use (for example [97]). Therefore, published data on wearable performance may not directly translate to individuals of different races and ethnicities or patients with heterogeneous sleep and medical conditions across a variety of ages.

Additionally, there is limited evidence of sleep-tracking capabilities during sleep outside the main sleep bout (e.g., daytime sleep and naps) and the reliability of sleep estimates over numerous days and physiological conditions (e.g., alcohol use, acute illness) remains unclear. Furthermore, performance metrics are specific to use with the manual designation of TIB and may not be translated to different sensor hardware, firmware, and software. It is also important to consider that the availability of new consumer-grade devices and models outpace the traditional route to scientifically evaluate them. Thus, performance data for a specific device and model is available when a device may be no longer available.

The following sections discuss some considerations when translating laboratory-cited performance to real-life use of consumer-grade sleep-tracking devices.

In-lab evaluation versus free-living use, and TIB estimation

Most of the literature that compares a device's performance against reference standards is derived from highly controlled, in-laboratory studies as opposed to the free-living condition, which is the intended use of these devices. For example, the utilization of single-night PSG with enforced TIB (specified lights out, lights on) under the supervision of trained personnel limits the translation of the cited performance metrics to the real world.

Calculated summary metrics such as TST, SOL, WASO, and SE (Table 3 for definitions) are contingent not only on the device and associated algorithm's ability to differentiate sleep from wake but also on TIB gated by bedtime and rise time. During the laboratory comparison of a consumer sleep-tracking device to PSG, bedtime, rise time, and TIB attempting sleep are clearly delineated; however, in the free-living environment, sleep period time (as opposed to TIB; Table 3) is used and determined automatically by the device. If incorrect (e.g., sleep period start is designated by the device when an individual is watching TV on the couch and not actively attempting sleep), the accuracy of summary sleep metrics would be expected to suffer even if the sleep–wake classifier displays high sensitivity and specificity compared to PSG. However, minimal literature is available to support or refute this theory.

One study of the WHOOP device tested performance (compared to PSG) when the device used manual input versus automatically detected TIB [98]. WHOOP sleep parameter estimates with automatically detected TIB significantly underestimated polysomnogram light (–9 min) and deep sleep (–16 min) without statistically significant biases between device and polysomnogram-measured TST, wake, and stage REM sleep. WHOOP sleep parameter estimates with manually designated TIB significantly underestimated polysomnogram wake time (17 min) without statistically significant biases between the device and polysomnogram-measured TST, light, deep, and stage REM sleep. Additionally, when sleep classifiers were applied to the same contemporary research/clinical-grade wearable device, both with and without the use of sleep diaries, agreement with polysomnogram TST was poor (ICC = .27 to .44) regardless of the analysis method [72].

In an attempt to eliminate the need for sleep diaries, a heuristic algorithm was developed. The algorithm detected the "sleep period time-window" (which begins with sleep onset and ends with the final awakening) with accelerometry data in the absence of a sleep diary with a mean difference in duration of 2 min compared to PSG in healthy sleepers [71]. Other algorithms to estimate TIB from accelerometry data alone also show promise in maintaining completely passive sleep tracking by consumer wearable devices [99].

Research protocols can include a self-report measure of TIB (e.g., digital sleep diary) and investigators may use this subjective information to clean data before the calculation of summary sleep metrics. However, the inclusion of bedtime and rise time may introduce bias from the participant, which may vary depending on different factors such as the presence or absence of sleep disorders [100]. Additionally, data cleaning itself could introduce bias and inaccuracies given the multitude of decision points in this process.

Proprietary scores or metrics such as "restless sleep", "sleep disturbances", or "readiness" are often output from wearable sleep-tracking devices and do not have a comparable clinical and scientific measure. Therefore, the accuracy and relevance of such scores remain unclear.

Changes and updates in device types, models, and algorithms

Rapid changes in device models, firmware and software of consumer-available, wearable sleep-tracking technologies may impede the translation of laboratory performance metrics to the field. Although actigraphy software does not remain static either, FDA-cleared actigraphs typically disclose sleep-classifier algorithms and require the user to manually update the software version, providing transparency. Conversely, consumer-grade device manufacturers use black-box algorithms based on artificial intelligence to stage sleep and do not typically provide details regarding algorithm development, testing, or the potential changes in performance that may arise from algorithm updates during use in the field. Conversely, machine learning can confer significant advantages in sleep tracking, for example, adapting to an individual's data to improve performance. This capability was leveraged to compare "generalized" and "personalized" sleep-staging algorithms. The "personalized" sleep classifier was superior to the generalized algorithm with sensitivity of 94% and 93% and specificity of 70% and 83%, respectively [65].

Studies that describe the development and testing of investigator-initiated sleep-staging algorithms adapted to off-the-shelf consumer-grade devices are limited (given manufacturer restrictions on providing raw data to investigators) but available. Scientific teams have developed algorithms (typically utilizing machine learning) that can categorize sleep from signals derived from the Fitbit [101], Apple Watch [13, 102–104], Oura ring [102, 105], Amazfit [106], and the Microsoft Band [107]. This type of work holds significant promise to increase the rigor, reproducibility,

and transparency of research with consumer-available, wearable sleep-tracking technologies given the disclosure of demographic characteristics in training datasets, algorithm features, and other details of algorithm development. To further increase reproducibility and transparency, some study teams have even made code for their algorithms open source [13, 103, 104, 106].

Device-agnostic open-access algorithms for processing wearable "raw data" and classifying sleep and other variables of interest are upcoming. Additionally, code to translate raw acceleration data into activity counts [29, 108] may be beneficial in harmonizing data sets where sleep was recorded with traditional research/clinical-grade devices with those that utilize contemporary research/clinical-grade devices as well as consumer-grade devices given the current accessibility of Apple Watch and Fitbit acceleration data. Notably, large public datasets (e.g., Multi-Ethnic Study of Atherosclerosis) of co-recorded actigraphy and PSG have been leveraged for both algorithm training and testing with the rationale that together, actigraphy and PSG sleep measurement modalities derive similar parameters (activity counts and PPG pulse/PRV from the pulse oximeter) that are recorded in a single consumer-grade device alongside an annotated gold-standard [13, 109, 110]. Though an excellent use of existing data, the ability to translate algorithms developed with traditional methods to consumer-grade devices remains uncertain [111] and rigorous, reproducible development and testing of sleep-classifier algorithms at scale are expected to benefit from large, public datasets composed of raw acceleration signal and PPG data acquired from consumer-grade devices co-recorded with PSG in heterogeneous patient populations. Independent testing data sets (not just hold-out data from the training data set) should be used to ensure generalizability. With appropriate disclosure of characteristics of the dataset, open-source code, and performance reporting, researchers could build a library of algorithms appropriate for use in different patient groups.

Study setting and assessment conditions

Device placement

Reflective PPG on the wrist may be a problem when placed by the patient/subject given increased movement artifact that was not present in the performance assessment study when the device was appropriately secured by a lab staff member.

Specific to oxygen saturation readings, in addition to reflective/reflectance technology (as opposed to transmissive/absorptive used by medical-grade oximetry) the location of consumer-available, wearable sleep-tracking technologies may be problematic [112]. Medical-grade PPG is placed on physical locations with dense vascular beds (i.e. fingertip, ear lobe) while consumer-available, wearable sleep-tracking technologies utilize sensors at the dorsum of the wrist or finger. These areas are more prone to movement between the sensor and anatomical location, less vascular, contain hair, and are known to result in less accurate oximetry readings [80, 112]. Additionally, acceleration data may also be impacted by placement, given the observation that proximal versus distal actigraphy placement on the wrist resulted in sleep parameter discrepancies [29].

These multiple sources of error that could arise due to real-world device placement may reduce the performance cited in studies where device placement is consistent.

Sleep bouts outside the main sleep period (daytime/naps)

Short sleep bouts detected outside the main (typically nocturnal) "sleep period" pose a variety of challenges in the translation of laboratory-measured performance. Firstly, the wearable device algorithm has to recognize that sleep is being attempted during the day or outside the usual time (i.e., TIB period must be appropriately designated during an atypical interval). Additionally, daytime sleep or sleep outside the main sleep bout is often different in sleep stage breakdown and is likely to have increased WASO. Such inherent differences in sleep, and the predilection for wearable sleep trackers that use movement to misclassify nonmoving wakefulness as sleep, can render summary measures less accurate than during the main sleep bout. Indeed, this has been demonstrated with the use of traditional actigraphy [113, 114].

The ability of consumer-grade wearable devices to capture daytime sleep has been evaluated both in the sleep lab and at home. In a 3-day laboratory study, compared to PSG, the FitBit correctly identified only 6 of 20 daytime naps (24 of 30 nighttime sleep periods correctly detected) [115]. Using sleep logs as the reference standard in individuals self-selecting their sleep–wake schedules at home, the percent of missed daytime sleep episodes were noted for the following consumer-grade sleep-tracking devices: Fatigue Science: 3.6%; Fitbit: 4.8%; Oura: 6.0%; Polar: 37.3%. Missed episodes were most likely to occur when the daytime TIB period was short, demonstrating the limited capacity for consumer-grade sleep-tracking devices to track naps.

Daytime sleep may also occur in the context of shift work; though consumer-grade wearables have been assessed in shift workers, captured sleep was during the night [94]. Therefore, the ability of consumer-grade devices to provide sleep estimates in individuals who work shifts or have circadian rhythm sleep–wake disorders remains unknown.

Single versus repeated nights of recording

The intended use of consumer-grade wearable sleep-tracking devices is in the home environment over days, weeks, months, and beyond. However, studies comparing the output of consumer-grade wearable sleep-tracking devices to PSG take place during a single night of recording; therefore, the reliability of cited performance metrics over multiple nights remains unclear.

However, emerging work has addressed this issue by utilizing home PSG/EEG over multiple (3–14) nights [66, 116, 117]. Of note, a week of co-recorded multichannel dry EEG (embedded in a headband) and 4 consumer-grade wearable devices revealed that device output aggregate parameters dependent on sleep–wake differentiation (TIB, TST, SE, sleep latency, and WASO) approximated ground-truth (here, EEG defined sleep) better on nights with higher SE [117]. These findings are not surprising given the high sensitivity and low specificity of these devices. Sleep staging (light, deep, REM) was highly variable across nights [117]. These findings highlight the potential discrepancies between laboratory and real-world performance.

Bed partners, children, and pets

While sleep during performance assessment studies takes place with the participant sleeping alone in the bed of the sleep laboratory, sleep at home may take place in a bed shared with another human or pet. Although unknown, theoretically, this could impact the movement-based assessment of sleep.

Notably, in an actigraphic study of bed sharing with dogs, a significant positive relationship between human and dog movement over sleep periods was found, with dogs influencing human movement more than humans influencing dog movement. Dog movement tripled the likelihood of the human transitioning from a nonmoving state to a moving state [118].

Sample demographics and characteristics

Age

Sleep-staging performance of consumer-grade wearable devices does not necessarily generalize to age groups outside the population the assessment took place in, as changes in physiology

across the lifespan may impact accelerometer and PPG-acquired measurements during sleep [96, 119]. Performance assessments of consumer-grade sleep-tracking devices have largely been in adults. However, a few investigations have evaluated performance in children (as young as 3 years of age) and adolescents [62, 120]. Recent publications cite higher specificity, the proportion of true wake correctly identified by the device and algorithm, in adolescents than typically observed in adults (68% and 88%–89% for the Fitbit Charge 3 and Oura ring, respectively) [63, 121]. As expected, the performance of multi-sensor devices in estimating sleep in children and adolescents exceeds that of early models of consumer sleep-tracking devices that only measure movement [122]. Sleep-staging performance (light, deep, REM) varies widely as in adults [63, 121].

Developmental and age-relevant contextual factors should be acknowledged when using wearable technology to track sleep in children and adolescents. Sleep patterns and behaviors undergo a significant transition from childhood to adulthood, driven by biological changes. In childhood, homeostatic sleep drive is greater and reflected by longer sleep duration with a greater proportion of slow wave sleep. Multiple sleep bouts (naps) and sleep outside the child's own bed could also influence the performance of wearable sleep-tracking devices. As children transition into adolescence, their circadian rhythms delay and their homeostatic sleep drive decreases, leading to later bedtimes and a preference for later wake times. However, early school start times can result in sleep deprivation when superimposed on these physiological changes. Based on biological changes, the recommended sleep durations for age groups are as follows: 10–13 h for ages 3–5 years, 9–11 h for ages 6–13 years, 8–10 h for ages 14–17 years, 7–9 h for ages 18–64 years, and 7–8 h for > 65 years [123].

Across adolescence, sleep macro- and micro-structure undergo profound changes thought reflecting brain maturation processes, including a 40% reduction in slow wave activity (or N3 sleep), which is also reflected by a steep decline in Delta power (.3 to 4 Hz) [124]. In the trajectory from childhood to older adulthood, stage REM and N3 sleep proportions decline sharply with modest increases in N1 and N2 and greater increases in WASO [125]. However, various factors such as work demands, stress, social factors (e.g., parental influences on children's sleep schedules), and reproductive stage (e.g., menopause share uniqueness in the characterization of sleep with menopause core symptoms, i.e., hot flashes, directly disrupting women's sleep [126]) and lifestyle choices can affect the sleep patterns. Overall, the evolution of sleep across the lifespan reflects the complex interplay of biological, psychological, and environmental factors. Understanding these changes is crucial when considering the use of wearable sleep tracking in specific age populations.

The complexity of sleep assessment is further compounded by developmental and sex-specific variations in cardiac function, a critical input to wearable devices in sleep staging. These variations interact intricately with a myriad of age-dependent and independent biopsychosocial factors, such as activity levels and lifestyle choices. Research conducted by de Zambotti et al. [127] has shed light on some of these intricacies. They observed significant age-related declines in HR among boys during adolescence, while no such trend was observed in girls (increasing male-female HR difference of ~2.4 beats per minute each year). Results were partially explained by age- and sex-dependent changes in the pattern of activity. In the same study, within-night trajectories in cardiac function exhibited sex-divergent patterns, with boys experiencing more pronounced increases in HRV compared to girls. Moreover, evidence like cardiac dependencies to sleep stage transitions used by wearables to differentiate sleep stages, including non-rapid-eye movement sleep HR differences, were notably more prominent in girls (~3.9 beats per minute) than in boys (~2.4 beats per minute). These intricate nuances represent just a fraction of the multifaceted factors that can influence the performance of wearable sleep monitoring devices.

It is also noteworthy that most consumer-grade sleep-tracking devices are designed for adults; therefore, significant differences in wrist (and finger) circumference and length without proportional changes to the device may result in sensor positioning disparate from that observed when an adult uses the device; however, as an example, the Fitbit Charge HR is designed to fit wrists that are 5.4 to 8.7 inches in diameter (inclusive of the average wrist size of children as young as 3 years of age) [62]. For very young children and infants, device placement is often on the ankle or calf [128]. Additionally, though not definitively known, algorithm training of the sleep classifiers used by consumer-grade devices likely takes place in adults and may not translate to the motion and cardiac physiology observed in pediatric sleep, though the stronger parasympathetic activity during sleep in children may result in improved sleep classification performance [120]. Furthermore, there are specific pediatric scoring rules from 2 months post-term through 18 years of age [129], which require different annotations of the gold standard used for training of algorithms used to classify sleep children and adolescents. Finally, children are more likely to nap and may sleep in a moving context (e.g., stroller) further augmenting potential inaccuracies of wearable sleep tracking [128]. Therefore, though performance assessments in the pediatric population are available, caution should be taken in children.

Further evaluation of the performance of consumer-grade sleep-tracking devices in pediatric as well as elderly populations and, if indicated, the ability to select different sleep-staging algorithms based on age may enhance the accuracy of such devices.

Skin tone and skin thickness

Through melanin's absorption of light, decreased signal intensity is observed when photoplethysmography is recorded in individuals with dark skin tones. As noted previously, given superior stability in the context of motion, green light is typically used by PPG sensors in consumer-grade devices and is more vulnerable to melanin absorption than red and infrared light [112]. Additionally, increasing skin thickness, which is directly correlated with body mass index (BMI), also dampens PPG signal [96]. One investigation demonstrated that together, increased BMI and skin tone caused signal loss of up to 61.2% in consumer-grade wearables [130].

When the accuracy of direct PPG measurements is reduced, noise is introduced into outputs from consumer-grade wearable devices that utilize PPG as an input (e.g. sleep and HRV parameters) [1, 24, 131]. Therefore, the performance of consumer-grade sleep-tracking devices cited from studies in nonobese, light-skinned individuals may be superior to what is observed when the same device is used in individuals with obesity or dark skin color. The differential accuracy of consumer-grade wearables has raised the concern that digital health solutions may reinforce existing disparities in care [132].

Sleep disorders

Outside of traditional research/clinical-grade actigraphs, only a few modern, consumer-grade wearable sleep-tracking devices have been evaluated within an adult sleep-disordered population. To our knowledge, performance evaluations have been conducted in conditions including insomnia disorder [133–135], OSA [136,

137], central disorders of hypersomnolence [138], and frequency of leg movements during sleep [67]. These performance studies rarely include or are compared to a reference group of normal sleepers, which makes it challenging to determine performance alterations specific to a distinct sleep disorder.

In their evaluation of a consumer-grade sleep tracker and PSG, Kang and colleagues [133] identified a significant divergence in performance abilities for a consumer-grade sleep tracker when applied in good sleepers versus patients experiencing insomnia [133]. Acceptable agreement between a consumer-grade sleep tracker and PSG was significantly more common in good sleepers (82.4% displayed acceptable agreement) than in patients experiencing insomnia (39.4% displayed acceptable agreement). While poor performance of wearable devices in clinical populations with sleep disorders may be expected and driven by a possible dependency of device performance on sleep continuity characteristics (e.g., reduced accuracy with increasing WASO and awakenings and decreasing SE), this requires further confirmation.

Unlocking the full potential of these devices, both for research and clinical purposes, requires increasing the empirical attention toward rigorous evaluations of wearable sleep-tracking devices in sleep disorder populations. Presently, there are still major gaps in knowledge on how these devices will perform in response to notably atypical sleep. For example, it is unclear whether consumer-grade devices can capture sleep onset rapid-eye-movement sleep episodes, which is a distinguishing sleep characteristic of narcolepsy. As such, there is a major need for future evaluations performed over samples including a diverse collection of sleep disorders, as well as good sleepers, to better clarify sleep disorder-specific device abilities.

Medical conditions

Use of consumer-grade wearable sleep-tracking devices in individuals with medical conditions that impact HR, HRV, or motion during sleep (e.g., atrial fibrillation, patients with pacemakers, spinal cord injury patients) may render sleep outputs that do not approximate PSG parameters as expected based on performance assessments in healthy individuals. For example, individuals with pacemakers may display fluctuations in the P–P interval (derived from PPG recording) that are not accompanied by R–R interval changes [20]; this may interfere with sleep staging based on pulse rate variability.

Considerations when integrating wearable devices into research

Study protocol design and device setup

Depending on the type of device selected for the study, different steps might be necessary for properly setting up a study. In this section, we outline several examples of critical steps to be taken for successfully implemented wearables use in clinical and research studies.

When designing a study, it is critical to consider the device placement. For example, the form factor of certain devices (e.g., watches or wristbands) might be able to accommodate most people thanks to adjustable straps. In these cases, as per the manufacturer's instruction, it might be preferable to wear the strap a few centimeters further away from the wrist, to improve signal quality. On the other hand, rings, which are becoming increasingly common, might or might not provide adjustable hardware, and therefore selecting devices with an adequate fit for each user becomes key. The accurate positioning of devices is even more critical for outcomes relying on sensor contact (e.g., PPG-based).

It is important to consider that, depending on the intended use, devices may be worn 24/7 (e.g., for circadian measures), or during nighttime only. The variable of interest may also dictate charging behaviors. For example, users interested only in sleep metrics typically do not need to wear the device during the day, thus allowing the devices to charge to full capacity during the day. However, users interested in data across the 24-h and circadian parameters may need to be more strategic about charging behaviors, including the use of multiple shorter charging windows where full charge is not achieved. That said, we know that sleep occurs within the greater context of circadian rhythms, and therefore is impacted by daytime behaviors such as light exposure, activity levels, and daytime naps.

Once the optimal fit or exact positioning of the sensor has been established, each wearable typically comes with an app, usually requiring signing up and setting using some login credential or software. Particularly for devices relying on cloud services, it is best practice to avoid using identifiable information in device settings. At current, we are not aware of devices using demographics in their sleep classification algorithms. Thus, it is unlikely that the absence of this information (or utilizing artificial, "dummy" demographics) would alter the device's behavior and performance. Related to that, it is also recommended to use ad hoc created login accounts and avoid participants setting up and using their own login credentials (however, this step cannot guarantee complete anonymization). It is also important to recognize that devices have different privacy policies, and some devices require the users to select and give approval to use/sense certain data. Thus, it is important to be sure that certain device metrics are enabled. If the user does not grant access to those metrics via the companion app, these metrics will be not available.

The companion device's app typically allows some level of configurability. For example, self-adjusting bedtime and wake-up time are among the relevant available features provided by some devices (e.g., Fitbit). While the device aims to automatically detect these times, lack of motion (e.g., watching a movie or reading from an e-reader in bed) might cause misdetection, and therefore bedtime might need manual adjusting. If manual adjusting cannot be performed (e.g., the software does not provide this functionality), it might be useful to supplement the wearable with, for example, electronic diaries so that bedtime can be annotated correctly. It is worth mentioning that in-app manually adjusting the bedtime and wake-up time by the user may not be equivalent to the user directly self-reporting bedtime and wake-up time with diaries.

In summary, care should be taken to ensure optimal fit and sensor positioning, and clear instructions should be provided to the study participants regarding wearing time and potentially in terms of manual adjustments of the collected data. The same sensors, positioning, and adjustments should be used consistently for the entire duration of the study.

A further critical point of device setup is the temporal synchronization between the wearable device and other data sources, for those cases where multiple sources are used (e.g., wearable and diary, wearable and environmental sensor, multiple wearables). Initial device configuration and subsequent synchronization with the dedicated app are usually the procedures through which the device's internal clock is set up, determining the timestamp associated with each data point. Making sure that such temporal coordinates are synchronized with those recorded by other data sources is advised to facilitate subsequent data processing and to reduce sources of bias in the data collection. Another consideration is the time zone across devices, which could differ and

cause challenges with circadian data. Sudden changes in clock time during travel or daylight savings also pose an additional challenge for circadian rhythm tracking.

It is important to consider that consumer-grade devices come with rich in-app and on-device audiovisual and haptic (e.g., vibration) feedback on users' biobehavioral data, as well as other engagement features. Typically, feedback is provided within a gamification framework (e.g., rewarding/celebrating improved sleep time compared to the last week). It is critical to recognize that feedback may directly and/or indirectly affect the behavior of individuals. Thus, it is critical that when implementing the use of wearables in clinical and research studies, devices and app feedback are minimized. Not all devices allow turning off all feedback and notifications. In that case, different solutions can supplement the lack of customizability of apps and devices. Solutions include the design of specific instructions to participants, as well as the use of physical equipment from covering devices' screens (e.g., from the simple use of black tape to sophisticated customized 3D printed covers).

Critical information to collect from wearable sleep-tracking devices

Some wearable information is critical for study reproducibility and evaluation of potential confounders. Consumer products are frequently updated, leading to inconsistencies in the data reported by the various algorithms used before and after an update. Hardware changes refer to upgrades of the sensor, e.g., when a new version is released. Typically, these do not directly impact a study unless the previous version of the sensor is not sold anymore, and the researchers require the purchasing of additional units. Software changes might concern updated algorithms, e.g., a new sleep stage detection model. Firmware changes are typically associated with low-level features, e.g., improving artifact removal for beat-to-beat intervals used for HR and HRV analysis, which in turn are used for sleep stage estimation or the introduction of a new set of accelerometer features. When the study is not longitudinal in nature, hardware, firmware, and software versions should be annotated as the derived results might not be reproducible with different versions. In addition to that, when the study is longitudinal, measures should be taken to avoid hardware, firmware, and software updates. For example, auto-updates for firmware and software should be disabled, while hardware should not be changed for a newer version.

User adherence, behavior, and data

Standard actigraphy devices require manual initialization (data collection configuration and activate a device in data collection mode) with data retrieved upon protocol completion (e.g., after a couple of weeks). Thus, data collection failures were realized at the end of the protocol.

Conversely, most current wearable devices, including the new generation of clinical/research tools, rely on cloud services which allow the study team to view the state of data collection in real-time, when a participant synchronizes their data and what has been synched at any moment in time.

Data access and preprocessing requirements

Probably the most critical and overlooked aspect of using wearable devices, and particularly consumer-grade devices, is accessing the data. Different wearable classes (Table 1) and devices enable different modalities for gathering the data. Accessing the data outputted by the device is often neither standardized nor possible to automate in a signal recording pipeline. The different options available in terms of accessing the data might therefore be one of the most important decision points when selecting a wearable device. We cover the most common ways to access wearable data here.

Importantly, wearable data outputs are not always ready to use immediately. Once collected and retrieved, some level of preprocessing of the data may be required. This is more relevant for consumer-grade devices. The following discussion outlines data preprocessing and provides examples.

Accessing wearable data

The ability to access the data collected by a wearable device is paramount for any study. Therefore, a device must be chosen with a clear understanding of what data can be accessed and how. There are typically four different modalities by which wearables data can be accessed or exported. Traditional actigraphy generally relies on wired communication (e.g., USB) to configure the device for data acquisition, as well as retrieval of data collected over the measured period that is stored within the device's internal memory. Inherently, this requires participants to acquire the device from the researcher at the beginning of the assessment window and return the device at the end of the assessment window. Although this procedure aids in the degree of control over data acquisition, it is relatively cumbersome in the context of current technological advances, such as cloud-based services. Furthermore, the process doesn't provide insight into the participant's adherence to device use during data collection, which could result in poor data quality and/or yield upon device return. Similarly, researchers are unable to view day-to-day sleep and circadian-related outcomes during data collection, with this information only available through post hoc review after the device has been returned and the data downloaded. Conversely, most consumer-grade devices come with a mobile app where bedtime, wake-up time, sleep stages, and possibly other parameters are reported daily. It is always an option to manually record the parameters outputted by the app. However, this relatively rudimentary solution confers low reproducibility and is more subject to mistakes. Moreover, it is not scalable, i.e., it might be impractical with large samples and/or long assessment durations. In these cases, it would be advisable to use a wearable that provides access to the data with more standardized techniques, namely through centralized data export (e.g., via cloud services) or to automate the data processing pipeline (e.g., via the use of a Software Development Kit [SDK] or Application Programming Interface [API]).

Cloud services typically consist of online platforms where the researcher/clinician can log in and manage the data recorded from each study participant/patient. For some consumer-grade wearables, cloud services are provided as part of commercial services enabling professionals (e.g., sports team coaches, and clinical consultants) to manage the data from multiple clients. However, in these cases, data export might not be present or be limited in terms of time resolution or data streams, whereas some platforms associated with consumer-grade devices provide high-granularity data (e.g., data export via *.csv files of minute-by-minute data). It is paramount to check the minimum output data resolution before choosing a wearable device. Also, because the features and functionalities of centralized data export from consumer products might change quite frequently, we advise the study team to contact and maintain a relationship with the wearable manufacturer to ensure that access to the data will be provided for the duration of the study.

Finally, when available, an SDK or API would allow automation of the signal processing or data collection pipeline, extracting potentially higher resolution data with custom-made software. APIs and SDKs often also come with additional costs and limitations on usage, which should be analyzed and compared between vendors.

Preprocessing time series data derived from wearable sleep-tracking devices

To obtain accurate, reproducible, and high-quality time series wearable data (e.g., day-to-day TST estimates) it is critical to implement some preprocessing of the device outputs. Data preprocessing includes all the procedures applied to the data outputted by the device for obtaining the final time series on each measure to be used in subsequent analyses or clinical reports.

While most research/clinical-grade devices are optimized to provide (almost) ready-to-use outputs, data preprocessing is particularly relevant for consumer-grade devices where the operationalization of aggregate indicators (e.g., TST, WASO, mean HR) is proprietary and undisclosed. For example, wearable companies have different rules for classifying a "sleep period" as daytime or nighttime sleep, determining whether data quality is sufficient for accepting vs. discarding a measurement and whether a unified sleep episode versus several shorter sleep periods is provided. Thus, we recommend some data preprocessing procedures to be implemented before analyzing and reporting the collected data. Importantly, data preprocessing cannot be easily standardized due to its dependency on the specific data type and format outputted by each device (e.g., raw data vs. aggregate indicators). Therefore, there are many degrees of freedom for researchers and clinicians, with important implications for the study results. In all cases, it is paramount to transparently report and justify such preprocessing steps and decisions.

At the highest level of reproducibility, we recommend relying on raw data or, if unavailable, starting with the data exported at the maximal possible resolution (i.e., highest granularity). This minimizes the automatic data aggregation procedures programmed by wearable companies (often undisclosed) while increasing the transparency on how aggregate indicators are computed. For instance, when a device provides both epoch-by-epoch and nightly aggregate measurements, it is advisable to rely on the former and apply standard AASM definitions for computing sleep measures. Open-access tools have been recently made available in both R and Python for standardizing the computation of sleep parameters from epoch-by-epoch data in a replicable and AASM-compliant manner [16, 56]. Similarly, when possible, relying on raw biological signal data (e.g., PPG signal) rather than using automatically aggregated indicators (e.g., hourly mean HR) is advisable for better transparency and reproducibility of all signal preprocessing steps (e.g., data filtering, inter-beat intervals detection).

At a lower, but still acceptable, level of reproducibility accounting for more pragmatic arguments (e.g., need to optimize data preprocessing for clinical applications, lack of time and personnel resources), we still recommend a few key steps are considered to avoid misleading results, namely: temporal synchronization, explicit definition of "sleep periods," data quality check, and visual representation.

First, it is important to verify and synchronize the temporal information associated with each measurement. Temporal synchronization across data points and data types is critical to ascertain that the time series is correctly encoded without gaps, double recordings, and other possible issues due to lack of temporal alignment (e.g., data recorded in different time zones, switches between daylight saving and standard time). Second, it is always advisable to check and, if necessary, adjust the device definition of "sleep period." Most devices aggregate sleep indicators over time windows automatically identified between the first and the last sleep epoch (Table 3). Verifying and making explicit whether this or alternative "sleep period" operationalizations were used is critical for interpreting certain sleep measures (e.g., SOL, WASO, SE). Moreover, some devices might accept excessively short "sleep periods" (e.g., a few minutes) or record multiple periods within the same night. In these cases, it is advisable to apply some procedures for discarding and/or combining data points based on the adopted definition.

For instance, Menghini et al. [63] analyzed 2 months of Fitbit Charge 3 data in a sample of adolescents. To reach a more accurate sampling of nocturnal sleep, the authors arbitrarily defined "nocturnal sleep periods" as sequences of 180+ min of sleep starting between 6 PM and 6 AM. Then, based on that definition, the authors introduced a set of ad hoc data filtering rules that resulted in the removal of about 14% of cases (e.g., diurnal periods starting before 6 PM, nocturnal periods shorter than 180 min). Moreover, the authors realized that multiple distinct "sleep periods" were outputted by the device on some nights, which they combined into a single "sleep period" by considering the time in between as WASO. This approach guaranteed a strict correspondence between the phenomenon focused by the study hypotheses (i.e., nocturnal sleep and sleep disruption) and the data used to generate the study results, in addition to providing better reproducibility of the study procedures.

A third key preprocessing step is the inspection of data quality to filter unreliable observations. For instance, compliance information such as wearing/off-wrist time can be used to exclude observations characterized by excessive data loss. Alternatively, the same information might be considered as a statistical control in the following analyses [139]. Similarly, participants without a minimum amount of accepted data might be discarded from the following analyses. For example, recent studies highlighted that at least five-to-seven nights are necessary to capture reliable individual differences in TST and other indicators [140, 141].

Finally, a visual representation of the time series obtained for each measure is highly recommended to get a more general overview of these and further potential issues. Substantial changes in aggregate measurements or data quality indices over time might highlight algorithm/firmware updates, technical problems with some sensors, or changes in participant use of the device. In such cases, it is advisable to separately analyze the identified subsets of data or to include time as a covariate in the following analyses. For instance, Menghini et al. [63] filtered all temporally isolated data points that were recorded 10 + days or after the remaining measurement.

Device wear time information

Few devices provide information about individual compliance with the ambulatory assessment, mainly based on estimates of the amount of time the person wore the device (e.g., time or proportion of time for on- and off-wrist).

When wearing time is not provided, wearing time can be estimated by integrating multiple signals such as acceleration, temperature, and external light. For example, some devices (e.g., Empatica Embrace Plus) use on-skin detection algorithms to automatically identify and filter the data recorded while not

being used as intended, reducing the likelihood of including unreliable observations. Some devices and applications also provide signal-specific indicators of the recording reliability or data quality, for instance, based on the number of artifacts or missing data points. Both data quality and compliance information can be useful for data cleaning or for evaluating the impact of data quality on the study results. Depending on the device, data quality information can be outputted with several formats, ranging from simple and intuitive qualitative indicators (e.g., "poor" vs. "acceptable/optimal") to more detailed information on the exact amount of data that was automatically removed or corrected.

It is important to know that some authors developed ad hoc processes and best practices to evaluate wearing time and accepting wearable data points. For example, Wing and colleagues [139] considered the availability of minute-level HR data to establish if the device was worn. The authors excluded minutes without HR, HR exceeding thresholds for unlikely/aphysiologic values, and/or HR values that were part of strings of consecutive repeated or missed HR values. They used thresholds of wearing time to exclude days not considered to represent "typical activity."

Integration with experience sampling methods

The integration of wearable with experience sampling methods (ESMs) is of growing interest. ESMs refer to the repeated assessment of current psychological states, experiences, and activities in real-time and free-living conditions [142]. This is done through standardized and short sets of questions pushed to the recipient for response at predefined time points (e.g., every hour, every day at 9 PM) or contingently to specific events or contexts (e.g., bedtime, location). The synergetic use of wearables and ESMs allows for designing ecological momentary assessment and intervention protocols, which are increasingly used for both research and clinical applications in the sleep and circadian field (e.g. [33, 143]).

Usually, there is a need to rely on multiple data sources (e.g., third-party surveying applications) to supplement wearable assessment with ESMs. For this purpose, third-party mobile and/or web-based surveying applications are commonly used (e.g., RedCap, Qualtrics). Open-access options are also available (e.g., m-path, Sensus). Some platforms can extract wearable data while allowing for the use of ESMs (e.g., ilumivu). Interestingly, some devices (e.g., Fitbit Sense + Fitabase) allow self-report data collection via dedicated mobile applications or directly through the device interface. Some features enable the scheduling of customizable messages, and questions, with reminders that can be prompted at predefined times directly on the device screen. Further building on ESM, wearable devices are a powerful input to just-in-time-adapitve interventions [144], personalized interventions delivered in real-time in response to changes in an individual's internal and contextual state.

Privacy, security, and ethical concerns

Although it is not our intent to extensively cover these topics, we would like to highlight some important concerns related to privacy, security, and ethics with the use of sleep wearables, as well as some practical actions that can be taken to mitigate some of the risks in research and clinical work.

While several devices are sold and used internationally, privacy rules and regulations differ across countries. In the United States, the "patchwork" nature of federal and state laws and regulations surrounding the collection and use of health-related data is rather complex [145]. Furthermore, data collected outside of healthcare services (e.g. data under the control of the device manufacturer) may not be regulated by the Health Information Portability and Accountability Act. In many instances, laws enable companies to use data collected by health devices without the user's consent and it is impossible for the user to know for what purposes their data could potentially be used and with whom it could be shared. The sleep-tracking device pool is composed of a mixture of devices, some regulated by the FDA as class II medical devices and many that are considered wellness devices. Similarly in Europe, although the General Data Protection Regulation covers several areas of privacy, data security, and consumer rights, the differentiation of lifestyle health from medical information remains unclear. Overall, similar data collected with different classes of wearables may be regulated differently [146]. Yet, even if some sleep wearables are marketed for "wellness tracking," they are likely to be used for health-related motives and may have specific features that are FDA-cleared (e.g., atrial fibrillation detection, oximetry).

While the venue of wireless communications in sleep wearables creates exciting new possibilities, it also opens the door to privacy threats. In remote health tracking, wireless communications have been identified as the venue via which most attacks take place and this is complicated by the numerous stakeholders involved at different stages of communication processes [147]. A large array of privacy attacks have been identified (e.g. man in the middle, battery drain attacks) and several countermeasures can be implemented in the development phase [148–151]. Some preventative strategies that can be implemented by clinicians and researchers are to refrain from inserting any personally identifiable information in device settings and to evaluate issues related to specific populations like minors. For example, in the US, the "Children's Online Privacy and Protection Act" places "parents in control over what information is collected from their young children online." Wearables leveraging cloud-based data transfers result in the exposure of data to wearable companies and network/internet service providers. In such cases, researchers, and clinicians lose control over several aspects of data management for their participants or patients. Informed consent processes therefore need to transparently disclose that data is shared with third parties.

From an ethical standpoint, sleep wearables also call for considerations about safety and accessibility. Safety concerns have been raised about the medicalization of everyday sleeping experiences since some sleep metrics are of questionable relevance outside of the context of sleep pathologies [152]. For instance, low SE in healthy sleepers who may enjoy spending extended TIB without fueling negative conditioning may be inconsequential. Daily tracking of sleep metrics without proper guidance may lead to pathologizing and hypernudging [153, 154]. This aligns with increasing concerns about orthosomnia, where wearable users become overly perfectionistic and preoccupied with improving their sleep [155]. Special clinical attention may be required to avoid inappropriate coping mechanisms and minimize stress and mental health risks.

There are multiple characteristics of wearable sleep-tracking devices that may increase existing health disparities. As previously discussed, PPG may be less accurate in individuals with dark skin tones. Therefore, direct measurements and aggregate parameters calculated from PPG-derived quantities may also be less accurate in dark-skinned individuals [132]. Additionally, algorithms that estimate health-related metrics from wearable signals often use machine learning. Machine learning algorithms

perform optimally when presented with data that is similar to the data that trained the algorithm. Under-represented groups are likely to be under-represented in training data sets; particularly if inclusion in the data set requires access to wearable devices and medical care. Therefore, machine learning algorithms to derive sleep and other health metrics may underperform in minority populations [156]. Collectively, these shortcomings may result in an understanding of sleep that is incomplete in certain groups and interventions poorly adapted to specific populations [132]. Mitigation of these biases should be a priority in promoting rational, safe wearable technology use in health care.

When wearable data is integrated into clinical care, providers also need to manage their patients' expectations regarding data collection and sporadic, triggered viewing versus continuous monitoring of sleep data. Wearables offer great potential for increasing access to care, provided that equity policies are put in place. As the scientific evidence builds up to support potential clinical applications, there will be a need to advocate for better integration within healthcare systems including IT infrastructure, patient and staff education, and appropriate reimbursement mechanisms for provider and staff time spent viewing longitudinal sleep data derived from wearable sleep-tracking devices. One of the financial aspects of wearable use that is becoming challenging to reconcile with the structure of both patient care and research is business models based on frequent hardware updates and continuous subscription fees. Another key aspect of access pertains to digital literacy and users' difficulties in interpreting sleep data [157]. Importantly, the relevant outcomes related to the use of wearables in clinical sleep medicine are unclear. Though the use of consumer-grade wearable devices as a surrogate for traditional research/clinical-grade actigraphy could allow more providers to follow the recommendations set forth by the AASM and would provide measurement-based care, it remains unknown whether incorporation of longitudinal sleep data would actually improve care or result in untoward consequences (e.g. overreliance on wearable data and neglect of self-report symptoms). Overall, more work is needed to develop optimal means of guiding the use of sleep wearables in clinical populations, and awareness of privacy, security, and ethical implications of wearable use is paramount for sleep healthcare and research.

How to select a wearable sleep-tracking device?

In the attempt to guide the reader to make an informed choice, the following questions should be considered as a checklist to help to select the right device, to serve the specific research/clinical study needs (Figure 4).

Avoiding misleading conclusions: a call for cautious interpretation of data derived from wearable sleep-tracking devices

The magnitude and extent of wearable use and generated scientific outcomes/discoveries from wearables exponentially increased over the past few years, with a forecasted uptrend. Thus, we believe it is our responsibility to call for a cautious interpretation of wearable data.

When interpreting study outcomes based on data collected using wearable sensors, caution must be taken for several reasons. First, while each sensor provides several parameters, many of them described in this manuscript, not all parameters are equally trustworthy. A simple distinction we can make is between what is estimated, and what is measured. Regardless of the transparency of the wearable manufacturer in terms of algorithmic details, estimates tend to involve larger errors. Examples of estimates are sleep stages or recovery scores. Even when something is more directly measured, we still need to establish the measurement error and evaluate the sensor's accuracy with respect to the gold standard (e.g., HR derived from PPG with respect to HR derived from ECG), but we typically have fewer steps in the pipeline. Looking at available published literature for different parameters might be indicative of which parameters are measured with relatively high accuracy, and which parameters are estimated with higher error. For example, parameters for which multiple sensors provide similar values, tend to be more reliable. Parameters, for which different sensors provide dramatically different results, are likely not something that should be trusted yet.

Another important aspect to consider, especially for estimates (e.g., sleep stages), is the population from which data was collected for algorithm development. Each algorithm is typically trained on a number of participants from studies that were run internally by the wearable manufacturer or in a convenience sample (e.g., clinical sleep lab), and may not represent the population of interest of the study that the wearable sensors will be deployed. When our study population differs from the population used to train the algorithms, we might have a larger error for the parameters of interest. Unfortunately, this is a common scenario as specific clinical populations are hardly ever considered when developing algorithms, especially when consumer products target a different market. As noted previously, other characteristics of the population of interest (e.g., skin tone) might also impact certain parameters (e.g., blood oxygenation) with augmented differences between sensor locations (e.g., wrist or finger).

It is also important to consider that some indices like HRV measures strongly require contextualization and are highly dependent on the context in which they are measured (e.g. rest versus exercise), calling for extra caution in their use and interpretation [1].

Conclusion

Wearable sleep technologies, and particularly, ubiquitous and highly accepted consumer-grade sleep-tracking devices hold significant promise in furthering our understanding of normal and disordered sleep and its role in health and disease. This manuscript outlined the benefits and limitations that wearables confer in sleep research and provided recommendations to promote rigorous and reproducible scientific inquiry. Understanding the derivation of sleep and other parameters from wearable acquired signal, the potential for artifact, the performance of the device against the gold-standard measurement (and limitations in translating the cited performance to real-world use), and the post-collection data extraction and processing techniques are critical when conducting research with wearable sleep-tracking technologies. The research study's population of interest and scientific question(s) will drive device selection and implementation. Even if the considerations for device selection and use are rigorously taken into account, inherent characteristics of wearables have the potential to widen health disparities and produce spurious results (e.g., reduced performance of PPG sensors on individuals with dark skin tones and increased skin thickness, wearable's performance can be contingent on the extent of sleep disruption). Therefore, cautious, informed use is crucial. Ultimately,

	Checklist - selecting the right device for a research/clinica study
☐	Does the device provide the output measure(s) of interest with the required granularity? The reader should review what a device can and cannot provide and pay attention to the way the outputs of interest are defined and operationalized, and the time resolution of the outputs (see Section 3).
☐	Can I access the raw data from the device? If interested in "raw data" access, particular attention should be paid for raw data availability by checking that the device provides "true" raw data. It is also an option to directly contact the manufacturer of a device to obtain any missing information and/or confirm data availability.
☐	Do I have financial constraints? While consumer devices are a viable choice when financial restrictions are in place, the reader should consider that there are several hidden costs in using consumer-grade wearable technology (e.g., subscription, third-party services cost for data access, data science expertise to process wearable outcomes, etc.) (see Section 5.4).
☐	How do I access and manage the device data? The reader should carefully review the different options to access wearable data and account for the technical expertise and labor force needed to extract the data and harmonize wearable data with other digitally collected study data (see Section 5.4).
☐	How do I process and analyze the data? There are several available open-access tools to process and analyze wearable data, including codes based on validated methods developed by the scientific community. Many of these tools are listed here https://www.researchsleep.ca/data-processing-tools
☐	Do I plan to perform data collection in conditions of limited mobile data connection / Wi-Fi connection? It is important to consider that devices have limited data storage without synching, and mobile data connection / Wi-Fi may be necessary to run the study.
☐	Is use of the selected device feasible in the population of interest and study design (e.g., can the participants adhere to necessary procedural components, does my study require a specific device's battery life? Is the form factor of a device suitable for my target population?)?.
☐	Has the performance of the device been assessed in similar samples, states, and conditions of the planned study? Different devices have different performance across different sample and condition. Performance evaluation studies should be carefully interpreted in light of study limitation and intended use of a device. See Section 4.
☐	Is the study cross-sectional or longitudinal? What is the timeframe for data collection? Availability of a specific device model and potential updates in the algorithm(s) used should be factors to consider. See Section 5.
☐	Does the device meet the requirements for privacy and security (e.g., data storage, HIPAA-compliace)? Any other related privacy and ethical concerns? See Section 6.

Figure 4. Checklist to help the reader select the right device for a research/clinical study.

the unique properties of multi-sensor consumer-grade wearable sleep trackers will provide a window into sleep beyond what is provided by traditional actigraphy given the ability to co-record autonomic parameters, estimate circadian features, and the potential to integrate other self-reported, objective, and passively recorded health indicators. Combined with experiential sampling methods and just-in-time adaptive interventions, wearables will contribute to the personalization of sleep medicine and wellness.

Disclosure Statement

Financial Disclosure: none.

Non-financial Disclosure: MdZ has received research funding unrelated to this work from Noctrix Health Inc. and Verily Life Science LLC. MdZ is a co-founder and Chief Scientific Officer at Lisa Health Inc. and has ownership of shares in Lisa Health Inc. MA is an advisor at Ôura Health Oy Ltd. CG is on the medical advisory boards of Huxley Medical Inc. and eviCore Healthcare. CG is a 5% inventor of a circadian mobile application licensed to Arcascope LLC. JC serves as a consultant to Cerno Health and Somni, while also having equity in Somni, and previously served as a consultant to Bodymatter Inc. The Sleep Research Society (SRS) board reviewed and endorses the major findings in the final version of this document.

Data Availability Statement

No new data were generated or analyzed in support of this research.

References

1. Petek BJ, Al-Alusi MA, Moulson N, *et al*. Consumer wearable health and fitness technology in cardiovascular medicine: JACC state-of-the-art review. *J Am Coll Cardiol*. 2023;**82**(3):245–264. doi:10.1016/j.jacc.2023.04.054
2. Foster F, Kupfer D, Weiss G, Lipponen V, McPartland R, Delgado J. Mobility recording and cycle research in neuropsychiatry. *J Interdiscip Cycle Res*. 1972;**3**(1):61–72. doi:10.1080/09291017209359298
3. Kupfer DJ, Detre TP, Foster G, Tucker GJ, Delgado J. The application of Delgado's telemetric mobility recorder for human studies. *Behav Biol*. 1972;**7**(4):585–590. doi:10.1016/s0091-6773(72)80220-7
4. Kripke DF, Mullaney D, Messin S, Wyborney VG. Wrist actigraphic measures of sleep and rhythms. *Electroencephalogr Clin Neurophysiol*. 1978;**44**(5):674–676. doi:10.1016/0013-4694(78)90133-5
5. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep*. 1992;**15**(5):461–469. doi:10.1093/sleep/15.5.461
6. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep*. 1994;**17**(3):201–207. doi:10.1093/sleep/17.3.201
7. American Sleep Disorders Association. Practice parameters for the use of actigraphy in the clinical assessment of sleep disorders American Sleep Disorders Association. *Sleep*. 1995;**18**(4):285–287. doi:10.1093/sleep/18.4.285
8. Terrill PI, Mason DG, Wilson SJ. *Development of a Continuous Multisite Accelerometry System for Studying Movements During Sleep*. IEEE; 2010: 6150-6153
9. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath*. Sep 2012;**16**(3):913–917. doi:10.1007/s11`5-011-0585-y
10. Ancoli-Israel S, Martin JL, Blackwell T, *et al*. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav Sleep Med*. 2015;**13**(sup1):S4–S38. doi:10.1080/15402002.2015.1046356
11. Khosla S, Deak MC, Gault D, *et al*.; American Academy of Sleep Medicine Board of Directors. Consumer sleep technology: an American Academy of sleep medicine position statement. *J Clin Sleep Med*. 2018;**14**(5):877–880. doi:10.5664/jcsm.7128
12. Smith MT, McCrae CS, Cheung J, *et al*. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of sleep medicine clinical practice guideline. *J Clin Sleep Med*. 2018;**14**(7):1231–1237. doi:10.5664/jcsm.7230
13. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*. 2019;**42**(12). doi:10.1093/sleep/zsz180
14. Depner CM, Cheng PC, Devine JK, *et al*. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;**43**(2). doi:10.1093/sleep/zsz254
15. Schutte-Rodin S, Deak MC, Khosla S, *et al*. Evaluating consumer and clinical sleep technologies: an American Academy of Sleep Medicine update. *J Clin Sleep Med*. 2021;**17**(11):2275–2282. doi:10.5664/jcsm.9580
16. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2021;**44**(2):zsaa170. doi:10.1093/sleep/zsaa170
17. de Zambotti M, Menghini L, Grandner MA, *et al*. Rigorous performance evaluation (previously, "validation") for informed use of new technologies for sleep health measurement. *Sleep Health*. 2022;**8**(3):263–269. doi:10.1016/j.sleh.2022.02.006
18. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. *Sleep*. 1982;**5**(4):389–399. doi:10.1093/sleep/5.4.389
19. Littner M, Kushida CA, Anderson WM, *et al*.; Standards of Practice Committee of the American Academy of Sleep Medicine. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. *Sleep*. 2003;**26**(3):337–341. doi:10.1093/sleep/26.3.337
20. Lujan MR, Perez-Pozuelo I, Grandner MA. Past, present, and future of multisensory wearable technology to monitor sleep and circadian rhythms. *Front Digit Health*. 2021;**3**:721919. doi:10.3389/fdgth.2021.721919
21. Cook J, Castelan A, Cheng P. Measuring sleep in the bedroom environment. In: Kushida CA, ed. *Encyclopedia of Sleep and Circadian Rhythms*. 2nd ed. Oxford: Academic Press; 2023: 16–29.
22. Te Lindert BH, van der Meijden WP, Wassing R, *et al*. Optimizing actigraphic estimates of polysomnographic sleep features in insomnia disorder. *Sleep*. 2020;**43**(11). doi:10.1093/sleep/zsaa090
23. Cook JD, Eftekari SC, Leavitt LA, Prairie ML, Plante DT. Optimizing actigraphic estimation of sleep duration in suspected idiopathic hypersomnia. *J Clin Sleep Med*. 2019;**15**(4):597–602. doi:10.5664/jcsm.7722
24. de Zambotti M, Cellini N, Menghini L, Sarlo M, Baker FC. Sensors capabilities, performance, and use of consumer sleep technology. *Sleep Med Clin*. 2020;**15**(1):1–30. doi:10.1016/j.jsmc.2019.11.003
25. de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*. 2019;**51**(7):1538–1557. doi:10.1249/MSS.0000000000001947
26. Patterson MR, Nunes AA, Gerstel D, *et al*. 40 years of actigraphy in sleep medicine and current state of the art algorithms. *NPJ Digital Med*. 2023;**6**(1):51. doi:10.1038/s41746-023-00802-1
27. Kwon S, Kim H, Yeo W-H. Recent advances in wearable sensors and portable electronics for sleep monitoring. *Iscience*. 2021;**24**(5):102461. doi:10.1016/j.isci.2021.102461
28. Rentz LE, Ulman HK, Galster SM. Deconstructing commercial wearable technology: contributions toward accurate and free-living monitoring of sleep. *Sensors (Basel)*. 2021;**21**(15):5071. doi:10.3390/s21155071
29. Te Lindert BH, Van Someren EJ. Sleep estimates using microelectromechanical systems (MEMS). *Sleep*. 2013;**36**(5):781–789. doi:10.5665/sleep.2648
30. Altini M, Penders J, Amft O. *Energy Expenditure Estimation Using Wearable Sensors: A New Methodology for Activity-Specific Models*. WH '12: Wireless Health 2012, San Diego, California, Oct. 23–25, Association for Computing Machinery, New York, NY, 2012:1–8
31. 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. doi:10.1016/j.ijcard.2012.03.119
32. Berry RB, Quan SF, Abreu AR, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine; 2020

33. Menghini L, Yuksel D, Prouty D, Baker FC, King C, de Zambotti M. Wearable and mobile technology to characterize daily patterns of sleep, stress, presleep worry, and mood in adolescent insomnia. *Sleep Health*. 2023;**9**(1):108–116. doi:10.1016/j.sleh.2022.11.006
34. Crowley SJ, Acebo C, Fallone G, Carskadon MA. Estimating dim light melatonin onset (DLMO) phase in adolescents using summer or school-year sleep/wake schedules. *Sleep*. 2006;**29**(12):1632–1641. doi:10.1093/sleep/29.12.1632
35. Kantermann T, Sung H, Burgess HJ. Comparing the Morningness-Eveningness Questionnaire and Munich ChronoType Questionnaire to the dim light melatonin onset. *J Biol Rhythms*. 2015;**30**(5):449–453. doi:10.1177/0748730415597520
36. Reiter AM, Sargent C, Roach GD. Finding DLMO: estimating dim light melatonin onset from sleep markers derived from questionnaires, diaries and actigraphy. *Chronobiol Int*. 2020;**37**(9-10):1412–1424. doi:10.1080/07420528.2020.1809443
37. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;**26**(3):342–392. doi:10.1093/sleep/26.3.342
38. Marler MR, Gehrman P, Martin JL, Ancoli-Israel S. The sigmoidally transformed cosine curve: a mathematical model for circadian rhythms with symmetric non-sinusoidal shapes. *Stat Med*. 2006;**25**(22):3893–3904. doi:10.1002/sim.2466
39. Gonçalves B, Adamowicz T, Louzada FM, Moreno CR, Araujo JF. A fresh look at the use of nonparametric analysis in actimetry. *Sleep Med Rev*. 2015;**20**:84–91. doi:10.1016/j.smrv.2014.06.002
40. Kronauer RE, Forger DB, Jewett ME. Quantifying human circadian pacemaker response to brief, extended, and repeated light stimuli over the phototopic range. *J Biol Rhythms*. 1999;**14**(6):500–515. doi:10.1177/074873099129001073
41. Forger DB, Jewett ME, Kronauer RE. A simpler model of the human circadian pacemaker. *J Biol Rhythms*. 1999;**14**(6):533–538. doi:10.1177/074873099129000867
42. Jewett ME, Forger DB, Kronauer RE. Revised limit cycle oscillator model of human circadian pacemaker. *J Biol Rhythms*. 1999;**14**(6):493–499. doi:10.1177/074873049901400608
43. Hilaire MAS, Klerman EB, Khalsa SBS, Wright KP Jr, Czeisler CA, Kronauer RE. Addition of a non-photic component to a light-based mathematical model of the human circadian pacemaker. *J Theor Biol*. 2007;**247**(4):583–599. doi:10.1016/j.jtbi.2007.04.001
44. Hannay KM, Forger DB, Booth V. Macroscopic models for networks of coupled biological oscillators. *Sci Adv*. 2018;**4**(8):e1701047. doi:10.1126/sciadv.1701047
45. Huang Y, Mayer C, Cheng P, *et al*. Predicting circadian phase across populations: a comparison of mathematical models and wearable devices. *Sleep*. 2021;**44**(10). doi:10.1093/sleep/zsab126
46. Cheng P, Walch O, Huang Y, *et al*. Predicting circadian misalignment with wearable technology: validation of wrist-worn actigraphy and photometry in night shift workers. *Sleep*. 2021;**44**(2). doi:10.1093/sleep/zsaa180
47. Figueiro M, Hamner R, Bierman A, Rea M. Comparisons of three practical field devices used to measure personal light exposures and activity levels. *Light Res Technol*. 2013;**45**(4):421–434. doi:10.1177/1477153512450453
48. Crouter SE, Kuffel E, Haas JD, Frongillo EA, Bassett DR Jr. A refined 2-regression model for the actigraph accelerometer. *Med Sci Sports Exerc*. 2010;**42**(5):1029–1037. doi:10.1249/MSS.0b013e3181c37458
49. Altini M, Penders J, Vullers R, Amft O. Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning. *IEEE J Biomed Health Inf*. 2014;**19**(1):219–226. doi:10.1109/JBHI.2014.2313039
50. Ceesay SM, Prentice AM, Day KC, *et al*. The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry. *Br J Nutr*. 1989;**61**(2):175–186. doi:10.1079/bjn19890107
51. Brage S, Brage N, Franks PW, *et al*. Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *J Appl Physiol*. 2004; **96**(1):343–351. doi:10.1152/japplphysiol.00703.2003
52. Altini M, Penders J, Vullers R, Amft O. Personalizing energy expenditure estimation using physiological signals normalization during activities of daily living. *Physiol Meas*. 2014;**35**(9):1797–1811. doi:10.1088/0967-3334/35/9/1797
53. Gu W, Leung L, Kwok KC, Wu I-C, Folz RJ, Chiang AA. Belun ring platform: a novel home sleep apnea testing system for assessment of obstructive sleep apnea. *J Clin Sleep Med*. 2020;**16**(9):1611–1617. doi:10.5664/jcsm.8592
54. Papini GB, Fonseca P, van Gilst MM, Bergmans JW, Vullings R, Overeem S. Wearable monitoring of sleep-disordered breathing: estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography. *Sci Rep*. 2020;**10**(1):13512. doi:10.1038/s41598-020-69935-7
55. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;**1**(8476):307–310. doi:10.1016/S0140-6736(86)90837-8
56. Benedetti D, Menghini L, Vallat R, *et al*. Call to action: an open-source pipeline for standardized performance evaluation of sleep-tracking technology. *Sleep*. 2023;**46**(2). doi:10.1093/sleep/zsac304
57. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;**15**(4):259–267. doi:10.1016/j.smrv.2010.10.001
58. Martin JL, Hakim AD. Wrist actigraphy. *Chest*. 2011;**139**(6):1514–1527. doi:10.1378/chest.10-1872
59. Ryser F, Gassert R, Werth E, Lambercy O. A novel method to increase specificity of sleep-wake classifiers based on wrist-worn actigraphy. *Chronobiol Int*. 2023;**40**(5):557–568. doi:10.1080/07420528.2023.2188096
60. Chinoy ED, Cuellar JA, Huwa KE, *et al*. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*. 2021;**44**(5). doi:10.1093/sleep/zsaa291
61. Lee XK, Chee NI, Ong JL, *et al*. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *J Clin Sleep Med*. 2019;**15**(9):1337–1346. doi:10.5664/jcsm.7932
62. Godino JG, Wing D, de Zambotti M, *et al*. Performance of a commercial multi-sensor wearable (Fitbit Charge HR) in measuring physical activity and sleep in healthy children. *PLoS One*. 2020;**15**(9):e0237719. doi:10.1371/journal.pone.0237719
63. Menghini L, Yuksel D, Goldstone A, Baker FC, de Zambotti M. Performance of fitbit charge 3 against polysomnography in measuring sleep in adolescent boys and girls. *Chronobiol Int*. 2021;**38**(7):1010–1022. doi:10.1080/07420528.2021.1903481
64. Lim SE, Kim HS, Lee SW, Bae K-H, Baek YH. Validation of fitbit inspire 2TM against polysomnography in adults considering adaptation for use. *Nat Sci Sleep*. 2023; **15**:59-67. doi:10.2147/NSS.S391802; 15:67
65. Grandner MA, Bromberg Z, Hadley A, *et al*. Performance of a multisensor smart ring to evaluate sleep: in-lab and home-based evaluation of generalized and personalized algorithms. *Sleep*. 2023;**46**(1). doi:10.1093/sleep/zsac152
66. Ghorbani S, Golkashani HA, Chee NI, *et al*. Multi-night at-home evaluation of improved sleep detection and classification with a

memory-enhanced consumer sleep tracker. *Nat Sci Sleep*. 2022; **14**:645–660. doi:10.2147/NSS.S359789
67. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of fitbit charge 2 compared with polysomnography in adults article. *Chronobiol Int*. 2018;**35**(4):465–476. doi :10.1080/07420528.2017.1413578
68. de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults article. *Chronobiol Int*. 2015;**32**(7):1024–1028. doi:10.3109/07420528.2015.1054395
69. de Zambotti M, Baker FC, Willoughby AR, *et al*. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav*. 2016;**158**:143–149. doi:10.1016/j.physbeh.2016.03.006
70. Plekhanova T, Rowlands AV, Yates T, *et al*. Equivalency of sleep estimates: comparison of three research-grade accelerometers. *J Meas Phys Behav*. 2020;**3**(4):294–303. doi:10.1123/jmpb.2019-0047
71. van Hees VT, Sabia S, Jones SE, *et al*. Estimating sleep parameters using an accelerometer without sleep diary. *Sci Rep*. 2018;**8**(1):12975. doi:10.1038/s41598-018-31266-z
72. Sansom K, Reynolds A, McVeigh J, *et al*. Estimating sleep duration: performance of open-source processing of actigraphy compared to in-laboratory polysomnography in the community. *Sleep Adv*. 2023;**4**(1):zpad028. doi:10.1093/sleepadvances/zpad028
73. Van Hees VT, Sabia S, Anderson KN, *et al*. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS One*. 2015;**10**(11):e0142533. doi:10.1371/journal.pone.0142533
74. Sundararajan K, Georgievska S, Te Lindert BH, *et al*. Sleep classification from wrist-worn accelerometer data using random forests. *Sci Rep*. 2021;**11**(1):24
75. Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. *Sleep*. 2005;**28**(9):1151–1161. doi:10.1093/sleep/28.9.1151
76. Regalia G, Gerboni G, Migliorini M, *et al*. Sleep assessment by means of a wrist actigraphy-based algorithm: agreement with polysomnography in an ambulatory study on older adults. *Chronobiol Int*. 2021;**38**(3):400–414. doi:10.1080/07420528.2020.1835942
77. Phillips AJK, Clerx WM, O'Brien CS, *et al*. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Sci Rep*. Jun 12 2017;**7**(1):3216. doi:10.1038/s41598-017-03171-4
78. Stone JE, Aubert XL, Maass H, *et al*. Application of a limit-cycle oscillator model for prediction of circadian phase in rotating night shift workers. *Sci Rep*. 2019;**9**(1):11032. doi:10.1038/s41598-019-47290-6
79. Zhang Z, Khatami R. Can we trust the oxygen saturation measured by consumer smartwatches? *Lancet Respir Medicine*. 2022;**10**(5):e47–e48. doi:10.1016/S2213-2600(22)00103-5
80. Ryals S, Chiang A, Schutte-Rodin S, *et al*. Photoplethysmography—New applications for an old technology: a sleep technology review. *J Clin Sleep Med*. 2023;**19**(1):189–195. doi:10.5664/jcsm.10300
81. Charlton PH, Kyriacou PA, Mant J, Marozas V, Chowienczyk P, Alastruey J. Wearable photoplethysmography for cardiovascular monitoring. *Proc IEEE*. 2022;**110**(3):355–381. doi:10.1109/jproc.2022.3149785
82. Berryhill S, Morton CJ, Dean A, *et al*. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. *J Clin Sleep Med*. 2020;**16**(5):775–783. doi:10.5664/jcsm.8356
83. Budig M, Stoohs R, Keiner M. Validity of two consumer multisport activity tracker and one accelerometer against polysomnography for measuring sleep parameters and vital data in a laboratory setting in sleep patients. *Sensors (Basel)*. 2022;**22**(23):9540. doi:10.3390/s22239540
84. Natarajan A, Su H-W, Heneghan C, Blunt L, O'Connor C, Niehaus L. Measurement of respiratory rate using wearable devices and applications to COVID-19 detection. *NPJ Digital Med*. 2021;**4**(1):136. doi:10.1038/s41746-021-00493-6
85. Pipek LZ, Nascimento RFV, Acencio MMP, Teixeira LR. Comparison of SpO2 and heart rate values on Apple Watch and conventional commercial oximeters devices in patients with lung disease. *Sci Rep*. 2021;**11**(1):18901. doi:10.1038/s41598-021-98453-3
86. Windisch P, Schröder C, Förster R, Cihoric N, Zwahlen DR, Windisch PY. Accuracy of the apple watch oxygen saturation measurement in adults: a systematic review. *Cureus*. 2023;**15**(2):e35355. doi:10.7759/cureus.35355
87. Hermand E, Coll C, Richalet J-P, Lhuissier FJ. Accuracy and reliability of pulse O2 saturation measured by a wrist-worn oximeter. *Int J Sports Med*. 2021;**42**(14):1268–1273. doi:10.1055/a-1337-2790
88. Schiefer LM, Treff G, Treff F, *et al*. Validity of peripheral oxygen saturation measurements with the Garmin Fēnix® 5X plus wearable device at 4559 m. *Sensors (Basel)*. 2021;**21**(19):6363. doi:10.3390/s21196363
89. Chen Y, Wang W, Guo Y, Zhang H, Chen Y, Xie L. A single-center validation of the accuracy of a photoplethysmography-based smartwatch for screening obstructive sleep apnea. *Nat Sci Sleep*. 2021; **13**:1533–1544. doi:10.2147/NSS.S323286
90. Yeh E, Wong E, Tsai C-W, *et al*. Detection of obstructive sleep apnea using Belun Sleep Platform wearable with neural network-based algorithm and its combined use with STOP-Bang questionnaire. *PLoS One*. 2021;**16**(10):e0258040. doi:10.1371/journal.pone.0258040
91. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. *JAMA*. 2020;**323**(14):1389–1400. doi:10.1001/jama.2020.3514
92. Zhao R, Xue J, Zhang X, *et al*. Comparison of ring pulse oximetry using reflective photoplethysmography and PSG in the detection of OSA in Chinese adults: a pilot study. *Nat Sci Sleep*. 2022; **14**:1427–1436. doi:10.2147/NSS.S367400
93. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. *Sensors (Basel)*. 2022;**22**:6317. doi:10.3390/s22166317
94. Stucky B, Clark I, Azza Y, *et al*. Validation of Fitbit charge 2 sleep and heart rate estimates against polysomnographic measures in shift workers: naturalistic study. *J Med Internet Res*. 2021;**23**(10):e26476. doi:10.2196/26476
95. Cao R, Azimi I, Sarhaddi F, *et al*. Accuracy assessment of oura ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: comprehensive analysis. *J Med Internet Res*. 2022;**24**(1):e27487. doi:10.2196/27487
96. Fine J, Branan KL, Rodriguez AJ, *et al*. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors*. 2021;**11**(4):126. doi:10.3390/bios11040126
97. de Zambotti M, Trinder J, Silvani A, Colrain IM, Baker FC. Dynamic coupling between the central and autonomic nervous systems during sleep: a review. *Neurosci Biobehav Rev*. 2018;**90**:84–103. doi:10.1016/j.neubiorev.2018.03.027

98. Miller DJ, Roach GD, Lastella M, *et al.* A validation study of a commercial wearable device to automatically detect and estimate sleep. *Biosensors*. 2021;**11**(6):185. doi:10.3390/bios11060185
99. Skovgaard EL, Pedersen J, Møller NC, Grøntved A, Brønd JC. Manual annotation of time in bed using free-living recordings of accelerometry data. *Sensors (Basel)*. 2021;**21**(24):8442. doi:10.3390/s21248442
100. Bianchi MT, Thomas RJ, Westover MB. An open request to epidemiologists: please stop querying self-reported sleep duration. *Sleep Med*. 2017;**35**:92–93. doi:10.1016/j.sleep.2017.02.001
101. 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):1968–1979. doi:10.1088/1361-6579/aa9047
102. Roberts DM, Schade MM, Mathew GM, Gartenberg D, Buxton OM. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep*. 2020;**43**(7). doi:10.1093/sleep/zsaa045
103. Perez-Pozuelo I, Posa M, Spathis D, *et al.* Detecting sleep outside the clinic using wearable heart rate devices. *Sci Rep*. 2022;**12**(1):7956. doi:10.1038/s41598-022-11792-7
104. Zhai B, Guan Y, Catt M, Plötz TU. Ubi-SleepNet: advanced multimodal fusion techniques for three-stage sleep classification using ubiquitous sensing. *Proc ACM Interactive Mob Wearable Ubiquitous Technol*. 2021;**5**(4):1–33. doi:10.1145/3494961
105. Altini M, Kinnunen H. The promise of sleep: a multi-sensor approach for accurate sleep stage detection using the oura ring. *Sensors (Basel)*. Jun 23 2021;**21**(13):4302. doi:10.3390/s21134302
106. Olsen M, Zeitzer JM, Richardson RN, *et al.* A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography. *IEEE Trans Biomed Eng*. 2022;**70**(1):228–237. doi:10.1109/TBME.2022.3187945
107. Zhang X, Kou W, Eric I, *et al.* Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. *Comput Biol Med*. 2018;**103**:71–81. doi:10.1016/j.compbiomed.2018.10.010
108. Neishabouri A, Nguyen J, Samuelsson J, *et al.* Quantification of acceleration as activity counts in ActiGraph wearable. *Sci Rep*. 2022;**12**(1):11958. doi:10.1038/s41598-022-16003-x
109. Fonseca P, Weysen T, Goelema MS, *et al.* Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*. 2017;**40**(7). doi:10.1093/sleep/zsx097
110. Song T-A, Chowdhury SR, Malekzadeh M, *et al.* AI-driven sleep staging from actigraphy and heart rate. *PLoS One*. 2023;**18**(5):e0285703. doi:10.1371/journal.pone.0285703
111. Van Gilst M, Wulterkens B, Fonseca P, *et al.* Direct application of an ECG-based sleep staging algorithm on reflective photoplethysmography data decreases performance. *BMC Res Notes*. 2020;**13**:513. doi:10.1186/s13104-020-05355-0
112. Charlton PH, Marozas V. Wearable photoplethysmography devices. *Photoplethysmography*. Elsevier; 2022: 401-439
113. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. *Sleep*. 2007;**30**(10):1362–1369. doi:10.1093/sleep/30.10.1362
114. Chinoy ED, Cuellar JA, Jameson JT, Markwald RR. Daytime sleep-tracking performance of four commercial wearable devices during unrestricted home sleep. *Nature and Science of Sleep*. 2023;**15**:151–164. doi:10.2147/NSS.S395732
115. Sargent C, Lastella M, Romyn G, Versey N, Miller DJ, Roach GD. How well does a commercially available wearable device measure sleep in young athletes? *Chronobiol Int*. 2018;**35**(6):754–758. doi:10.1080/07420528.2018.1466800
116. Svensson T, Chung U-I, Tokuno S, Nakamura M, Svensson AK. A validation study of a consumer wearable sleep tracker compared to a portable EEG system in naturalistic conditions. *J Psychosom Res*. 2019;**126**:109822. doi:10.1016/j.jpsychores.2019.109822
117. Chinoy ED, Cuellar JA, Jameson JT, Markwald RR. Performance of four commercial wearable sleep-tracking devices tested under unrestricted conditions at home in healthy young adults. *Nat Sci Sleep*. 2022;**14**:493–516. doi:10.2147/NSS.S348795
118. Hoffman CL, Browne M, Smith BP. Human-animal co-sleeping: an actigraphy-based assessment of dogs' impacts on women's nighttime movements. *Animals*. 2020;**10**(2):278. doi:10.3390/ani10020278
119. Allen J, Murray A. Age-related changes in the characteristics of the photoplethysmographic pulse shape at various body sites. *Physiol Meas*. 2003;**24**(2):297–307. doi:10.1088/0967-3334/24/2/306
120. Wulterkens BM, Fonseca P, Hermans LW, *et al.* It is all in the wrist: wearable sleep staging in a clinical population versus reference polysomnography. *Nat Sci Sleep*. 2021; **13**:885–897. doi:10.2147/NSS.S306808
121. Chee N, Ghorbani S, Golkashani HA, Leong RLF, Ong JL, Chee MWL. Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. *Nat Sci Sleep*. 2021;**13**:177–190. doi:10.2147/NSS.S286070
122. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*. 2015;**38**(8):1323–1330. doi:10.5665/sleep.4918
123. Hirshkowitz M, Whiton K, Albert SM, *et al.* National sleep foundation's sleep time duration recommendations: methodology and results summary. *Sleep Health*. 2015;**1**(1):40–43. doi:10.1016/j.sleh.2014.12.010
124. Baker FC, Willoughby AR, Massimiliano Z, *et al.* Age-related differences in sleep architecture and electroencephalogram in adolescents in the national consortium on alcohol and neurodevelopment in adolescence sample. *Sleep*. 2016;**39**(7):1429–1439. doi:10.5665/sleep.5978
125. Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*. 2004;**27**(7):1255–1273. doi:10.1093/sleep/27.7.1255
126. de Zambotti M, Colrain IM, Javitz HS, Baker FC. Magnitude of the impact of hot flashes on sleep in perimenopausal women. *Fertil Steril*. 2014;**102**(6):1708–15.e1. doi:10.1016/j.fertnstert.2014.08.016
127. de Zambotti M, Javitz H, Franzen PL, *et al.* Sex-and age-dependent differences in autonomic nervous system functioning in adolescents. *J Adolesc Health*. 2018;**62**(2):184–190. doi:10.1016/j.jadohealth.2017.09.010
128. Schoch SF, Kurth S, Werner H. Actigraphy in sleep research with infants and young children: current practices and future benefits of standardized reporting. *J Sleep Res*. 2021;**30**(3):e13134. doi:10.1111/jsr.13134
129. Troester M, Quan S, Berry R. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darian: American Academy of Sleep Mecine; 2023
130. Boonya-Ananta T, Rodriguez AJ, Du Le V, Ramella-Roman JC. Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal. *Biomed Opt Express*. 2021;**12**(12):7445–7457. doi:10.1364/BOE.439893

131. Menghini L, Gianfranchi E, Cellini N, Patron E, Tagliabue M, Sarlo M. Stressing the accuracy: wrist-worn wearable sensor validation over different conditions. *Psychophysiology*. 2019;**56**(11):e13441. doi:10.1111/psyp.13441
132. Colvonen PJ, DeYoung PN, Bosompra N-OA, Owens RL. *Limiting Racial Disparities and Bias for Wearable Devices in Health Science Research*. US: Oxford University Press; 2020: zsaa159
133. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res*. Jun 2017;**97**:38–44. doi:10.1016/j.jpsychores.2017.03.009
134. Kahawage P, Jumabhoy R, Hamill K, de Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in insomnia disorder I: in-lab validation against polysomnography. *J Sleep Res*. 2020;**29**(1):e12931. doi:10.1111/jsr.12931
135. Hamill K, Jumabhoy R, Kahawage P, de Zambotti M, Walters EM, Drummond SPA. Validity, potential clinical utility and comparison of a consumer activity tracker and a research-grade activity tracker in insomnia disorder II: outside the laboratory. *J Sleep Res*. 2020;**29**(1):e12944. doi:10.1111/jsr.12944
136. Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. *PLoS One*. 2019;**14**(1):e0210569. doi:10.1371/journal.pone.0210569
137. Moreno-Pino F, Porras-Segovia A, Lopez-Esteban P, Artes A, Baca-Garcia E. Validation of fitbit charge 2 and fitbit alta HR against polysomnography for assessing sleep in adults with obstructive sleep apnea. *J Clin Sleep Med*. 2019;**15**(11):1645–1653. doi:10.5664/jcsm.8032
138. Cook JD, Plante DT. Wearable technology as a tool for sleep-wake estimation in central disorders of hypersomnolence. *Curr Sleep Med Rep*. 2019;**5**:193–200. doi:10.1007/s40675-019-00156-9
139. Wing D, Godino JG, Baker FC, *et al.* Recommendations for identifying valid wear for consumer-level wrist-worn activity trackers and acceptability of extended device deployment in children. *Sensors (Basel)*. 2022;**22**(23):9189. doi:10.3390/s22239189
140. Aili K, Åström-Paulsson S, Stoetzer U, Svartengren M, Hillert L. Reliability of actigraphy and subjective sleep measurements in adults: the design of sleep assessments. *J Clin Sleep Med*. 2017;**13**(1):39–47. doi:10.5664/jcsm.6384
141. Lau T, Ong JL, Ng BK, *et al.* Minimum number of nights for reliable estimation of habitual sleep using a consumer sleep tracker. *Sleep Adv*. 2022;**3**(1):zpac026. doi:10.1093/sleepadvances/zpac026
142. Myin-Germeys I, Kuppens P. The open handbook of experience sampling methodology. *The Open Handbook of Experience Sampling Methodology*; Chicago, IL, USA: Independently Publisher; 2021: 1–311
143. Castilla D, Navarro-Haro MV, Suso-Ribera C, Díaz-García A, Zaragoza I, García-Palacios A. Ecological momentary intervention to enhance emotion regulation in healthcare workers via smartphone: a randomized controlled trial protocol. *BMC Psychiatry*. 2022;**22**(1):164. doi:10.1186/s12888-022-03800-x
144. Nahum-Shani I, Smith SN, Spring BJ, *et al.* Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2018;**52**(6):446–462. doi:10.1007/s12160-016-9830-8
145. Colgate J, Maisel J. The right not to share: weighing personal privacy threat vs. promises of connected health devices. *Women Securing the Future with TIPPSS for Connected Healthcare: Trust, Identity, Privacy, Protection, Safety, Security*. Springer; 2022: 135-157
146. Maaß L, Freye M, Pan C-C, Dassow H-H, Niess J, Jahnel T. The definitions of health apps and medical apps from the perspective of public health and law: qualitative analysis of an interdisciplinary literature overview. *JMIR Mhealth and Uhealth*. 2022;**10**(10):e37980. doi:10.2196/37980
147. Hofmann B. Ethical challenges with welfare technology: a review of the literature. *Sci Eng Ethics*. 2013;**19**(2):389–406. doi:10.1007/s11948-011-9348-1
148. Camara C, Peris-Lopez P, Tapiador JE. Security and privacy issues in implantable medical devices: a comprehensive survey. *J Biomed Inform*. 2015;**55**:272–289. doi:10.1016/j.jbi.2015.04.007
149. Liu J, Sun W. Smart attacks against intelligent wearables in people-centric internet of things. *IEEE Commun Mag*. 2016;**54**(12):44–49. doi:10.1109/mcom.2016.1600553cm
150. Sadhu PK, Yanambaka VP, Abdelgawad A. Internet of things: security and solutions survey. *Sensors (Basel)*. 2022;**22**(19):7433. doi:10.3390/s22197433
151. Ray PP, Dash D, Kumar N. Sensors for internet of medical things: state-of-the-art, security and privacy issues, challenges and future directions. *Comput Commun*. 2020;**160**:111–131. doi:10.1016/j.comcom.2020.05.029
152. Müller R, Kuhn E, Ranisch R, Hunger J, Primc N. Ethics of sleep tracking: techno-ethical particularities of consumer-led sleep-tracking with a focus on medicalization, vulnerability, and relationality. *Ethics Inf Technol*. 2023;**25**(1). doi:10.1007/s10676-023-09677-y
153. Lanzing M. "Strongly recommended" revisiting decisional privacy to judge hypernudging in self-tracking technologies. *Philos Technol*. 2019;**32**:549–568. doi:10.1007/s13347-018-0316-4
154. Ravichandran R, Sien S-W, Patel SN, Kientz JA, Pina LR. *Making Sense of Sleep Sensors: How Sleep Sensing Technologies Support and Undermine Sleep Health*. New York, NY: Association for Computing Machinery:2017: 6864–6875. doi:10.1145/3025453.3025557
155. 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):351–354. doi:10.5664/jcsm.6472.
156. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;**28**(1):31–38. doi:10.1038/s41591-021-01614-0
157. Liu W, Ploderer B, Hoang T. *In Bed with Technology: Challenges and Opportunities for Sleep Tracking*. New York, NY: Association for Computing Machinery:2015: 142–151. doi:10.1145/2838739.2838742
158. 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):841–848. doi:10.5664/jcsm.7120
159. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;**21**(11):e16273. doi:10.2196/16273
160. Plews DJ, Scott B, Altini M, Wood M, Kilding AE, Laursen PB. Comparison of heart-rate-variability recording with smartphone photoplethysmography, polar H7 chest strap, and electrocardiography. *Int J Sports Physiol Perform*. 2017;**12**(10):1324–1328. doi:10.1123/ijspp.2016-0668
161. Støve MP, Hansen ECK. Accuracy of the Apple watch series 6 and the whoop band 30 for assessing heart rate during resistance exercises. *J Sports Sci*. 2023;**40**(23):2639–2644. doi:10.1080/02640414.2023.2180160
162. Pasadyn SR, Soudan M, Gillinov M, *et al.* Accuracy of commercially available heart rate monitors in athletes: A prospective

study. *Cardiovasc Diagn Ther*. 2019;**9**(4):379–385. doi:10.21037/cdt.2019.06.05

163. Altini M, Plews D. What is behind changes in resting heart rate and heart rate variability? A large-scale analysis of longitudinal measurements acquired in free-living. *Sensors (Basel)*. 2021;**21**(23):7932. doi:10.3390/s21237932
164. Lu G, Yang F, Taylor J, Stein JF. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. *J Med Engineering Technol*. 2009;**33**(8):634–641. doi:10.3109/03091900903150998
165. Constant I, Laude D, Murat I, Elghozi J-L. Pulse rate variability is not a surrogate for heart rate variability. *Clin Sci (Lond)*. 1999;**97**(4):391–397
166. Bellenger CR, Miller DJ, Halson SL, Roach GD, Sargent C. Wrist-based photoplethysmography assessment of heart rate and heart rate variability: validation of WHOOP. *Sensors (Basel)*. 2021;**21**(10):3571. doi:10.3390/s21103571
167. Buchheit M. Monitoring training status with HR measures: do all roads lead to Rome? *Front Physiol*. 2014;**5**:73. doi:10.3389/fphys.2014.00073