SLEEPJ, 2021, 1–3

doi: 10.1093/sleep/zsab071
Advance Access Publication Date: 30 March 2021
Editorial

EDITORIAL

Sleep-tracking technology in scientific research: looking to the future

Michael A. Grandner, Matthew R. Lujan and Sadia B. Ghani*

Sleep and Health Research Program, Department of Psychiatry, University of Arizona College of Medicine, Tucson, AZ, USA

*Corresponding author. Michael A. Grandner, Sleep and Health Research Program, Department of Psychiatry, University of Arizona College of Medicine, 1501 N. Campbell Avenue, Suite 7326, PO Box 245002, Tucson, AZ 85724, USA. Email: grandner@email.arizona.edu.

Since the early 1970s, wearable technology has been used to assess sleep/wake behavior patterns in free-living conditions [1]. Over the past several decades, this technology has improved and has leveraged advancements, including digital accelerometry, microelectromechanical systems, and improved software [2]. As such, wearable sleep assessment technology has become not just an accepted methodology, but also an invaluable tool for characterizing real-world sleep, which has profound implications for health but is difficult to measure under laboratory conditions. Over the past decade, devices designed to objectively assess sleep in free-living situations have proliferated, leading to an increasing number of devices specifically designed for scientific use and/or commercial use by the general public. Also, these next-generation devices increasingly leverage emerging technologies, such as optical plethysmography, proximity-based detection, and artificial intelligence. Importantly, standards for the development of this technology have been codified [3, 4]. As the line between scientifically validated and commercially available devices has increasingly blurred, with devices originally developed for commercial use being deployed in research, it has become especially important to document the validation of these devices under gold-standard, rigorous conditions [5]. This would provide a better understanding of their strengths and limitations relative to more widely accepted devices.

The study in this issue by Chinoy et al. [6] included 34 healthy young adults who underwent three consecutive nights in a sleep laboratory to undergo polysomnographic measurement with concurrent use of actigraphy (Phillips Respironics Actiwatch 2), and assessment with various wearable (Fatigue Science Readiband, Fitbit Alta HR, Garmin Fenix 5S, and Garmin Vivosmart 3) and "nearable" (non-wearables, including EarlySense Live, ResMed S+, and SleepScore Max) consumer sleep-tracking devices. During the three-night study, a sleep disruption protocol was performed on one of the final two nights to assess the effects of fragmented sleep on device performance. The measures for the study were sleep/wake statistics and epoch-by-epoch agreement compared to the gold-standard PSG for both sleep versus wake and individual sleep stages.

For sleep/wake measures, the consumer devices generally performed as well as the Actiwatch, even on nights of fragmented sleep. The Actiwatch performed about as well as it had previously [7] (sensitivity = 0.97, specificity = 0.39, and accuracy = 0.89). In comparison, the consumer devices that performed best (Fatigue Science, Fitbit, EarlySense, Resmed, and SleepScore) also achieved similar levels of sensitivity (0.93–0.96) and nominally better specificity (0.45–0.54), with comparable accuracy (0.88–0.90). Sleep staging was compared to PSG. The Fitbit device performed nominally best for light sleep (sensitivity = 0.76, specificity = 0.67, and accuracy = 0.72) and rapid eye movement sleep (sensitivity = 0.69, specificity = 0.94, and accuracy = 0.89), and the Garmin devices were nominally best for deep sleep (sensitivity = 0.56, specificity = 0.92, and accuracy = 0.87), though in general many of the devices performed similarly. Taken together, there are a few key issues that are relevant to these findings.

The "commercial-grade" devices performed about as well (and sometimes, better) when compared to the commonly used, "scientific-grade" actigraphy device. For this reason, the delineation between devices approved for research, and the devices used by the general public should not be assumed to reflect the accuracy and validity of the devices. Rather, any device that purports to measure sleep should do so in accordance with published technical standards [3], should demonstrate validity according to published guidelines [4], and should be

implemented in the context of published recommendations [8]. Any device that meets these criteria, irrespective of how the device is marketed, should be considered appropriate for scientific use. Once a device and protocol have met these minimum standards, the choice to use that device should be made based on the quality of the data, the demonstrated utility in the population of interest, and other factors that are adjacent to the device accuracy such as access to raw data, ability to modify assessment windows, compatibility with statistical software packages, presence of other useful sensors, and Bluetooth capabilities.

It should be further noted that validation is not an event, it is a process. Devices change, as does their use and both the hardware and software that support them. And validation in one group does not transitively convey validation in all populations and contexts. The details regarding the ability of a specific device to demonstrate validity in the population of interest, for the outcome of interest, should be the most relevant factor in choosing a device. Perhaps "performance" in context is a better way to refer to this process than validation [9].

This paper also addresses the important issue of sleep staging using wearables. Automated scoring of even gold-standard polysomnography (PSG) is controversial, and even two well-trained humans frequently disagree when tasked with scoring the same record. This is likely because often, PSG epochs can contain elements of more than one sleep stage, rendering classification inherently unreliable (as long as probabilistic staging is disallowed). This creates a validation problem—it is difficult to validate against an unreliable target. PSG is not the "gold standard" for validation because it is gold, it is because it is the standard. Furthermore, since the signals obtained using wearables only indirectly correlate with brain-derived waveforms at the outer cortex, it should be assumed that any sleep staging classifications are rough estimates at best. As such, any scientific use of these values should be done with appropriate qualification, recognizing that the sleep staging by these wearables is not a replacement for PSG, but rather a rough estimate. This rough estimate, despite its imprecision, may yet still be quite valuable. Obtaining nightly polysomnographic data for days, weeks, or months in samples of hundreds, thousands, or even millions of people is otherwise impossible under current circumstances. Therefore, these rough estimates remain the best available data for large-scale, real-world, longitudinal sleep stage data at the population level. Additionally, these estimates may be the only form of objective sleep assessment available to some disenfranchised populations and demographics.

Once a device has shown that it performs well relative to PSG at detecting sleep versus wake, other signals may also be useful to derive. Perhaps, there are signals in movement patterns (such as rest-activity rhythms or elements of the raw movement signal) [10] and other peripheral signals [11] (such as heart rate or temperature) assessed by these devices that may deepen our understanding of real-world sleep. Future research might be able to further develop this line of inquiry, identifying metrics and factors present in these real-world signals that predict important outcomes [12]. The paper by Chinoy et al. [6] shows that the sleep–wake agreement is quite good, and the sleep staging agreement is moderate for a number of these consumer devices. Future research may show additional metrics derived from wearables.

A final issue that is relevant to this discussion is one of measurement versus intervention. The study by Chinoy et al. [6] focused on the accuracy of the measurement itself. It is important to note, though, that measurement is not intervention. Available studies show that the feedback provided by sleep trackers may itself be enough of an active intervention to improve sleep [13]. But in general, these effects may be weak and limited to those without much need for intervention. Just as a bathroom scale is not a weight loss program, a sleep tracker is not a sleep improvement program. As the ability of these devices to measure sleep is increasingly recognized, their usefulness will still be limited by the lack of insight about what to actually do with the data we collect. Simply providing feedback is the first step. Recommendations would be a second step, though it is important to note that current sleep health recommendations [14, 15] are based on self-report and not objective data. Even with recommendations, though, much work in behavioral science is needed to develop the educational, motivational, and interventional programs that will optimally make use of these data collected by sleep-tracking devices. Few such studies exist, and more are needed as the advantages of these devices to record continuously for 24 h a day for days, weeks, and even longer and their cost-effectiveness could provide pertinent data at an individual and population level that can be used for sleep improvement interventions.

The paper by Chinoy et al. [6] convincingly demonstrates that a number of commercially available sleep-tracking devices measure sleep and wake about as well as (and sometimes better than) standard actigraphy. Given the lower costs, technological improvements, and ease of use of these devices, this demonstration of relative accuracy should empower researchers to be comfortable using these devices in scientific research. Other factors, such as access to raw data, ability to modify scoring windows, privacy, and Bluetooth capabilities, may serve as important differentiators when choosing devices for a study. As the accuracy of these devices is further established, future work could move beyond just validation against PSG and instead develop novel uses for these real-world, longitudinal sleep data with the goal of assessing and improving population sleep health.

Funding

M.A.G. is supported by R01DA051321 and R01MD011600. S.B.G. is supported by T32HL007249.

Disclosure Statement

Financial disclosure: In the past 12 months, M.A.G. has received grant funding from Jazz Pharmaceuticals and Kemin Foods and has engaged in paid consulting relationships with Merck, Idorsia, Casper Sleep, Athleta, Fitbit, Natrol, and Smartypants Vitamins.

Nonfinancial disclosure: None.

References

1. Foster FG, *et al.* Mobility recording and cycle research in neuropsychiatry. *J Interdiscipl Cycle Res.* 1972;**3**(1):61–72.
2. Grandner MA, *et al.* Actigraphic sleep tracking and wearables: historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep

devices. In: Grandner MA, ed. *Sleep and Health*. Cambridge, UK: Academic Press; 2019: 147–157.
3. Consumer Technology Association Health Fitness and Wellness Technology Committee. *Performance Criteria and Testing Protocols for Features in Sleep Tracking Consumer Technology Devices and Applications*. 2019. ANSI/CTA/NSF-2052.3.
4. Depner CM, *et al*. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;**43**(2). doi:10.1093/sleep/zsz254
5. Menghini L, *et al*. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2021;**44**(2). doi:10.1093/sleep/zsaa170
6. Chinoy ED, *et al*. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*. 2021;**44**(5). doi:10.1093/sleep/zsaa291
7. Marino M, *et al*. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. 2013;**36**(11):1747–1755.
8. Ancoli-Israel S, *et al*. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav Sleep Med*. 2015;**13**(suppl 1):S4–S38.
9. Goldstein CA, *et al*. Miles to go before we sleep…a step toward transparent evaluation of consumer sleep tracking devices. *Sleep*. 2021;**44**(2). doi:10.1093/sleep/zsab020
10. Winnebeck EC, *et al*. Dynamics and ultradian structure of human sleep in real life. *Curr Biol*. 2018;**28**(1):49–59.e5.
11. Stahl SE, *et al*. How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport Exerc Med*. 2016;**2**(1):e000106.
12. Perez-Pozuelo I, *et al*. The future of sleep health: a data-driven revolution in sleep science and medicine. *NPJ Digit Med*. 2020;**3**:42.
13. Berryhill S, *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.
14. Watson NF, *et al*. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep*. 2015;**38**(6):843–844.
15. Hirshkowitz M, *et al*. National Sleep Foundation's updated sleep duration recommendations: final report. *Sleep Health*. 2015;**1**(4):233–243.