SLEEPJ, 2021, 1–3

doi: 10.1093/sleep/zsab020
Editorial

EDITORIAL

Miles to go before we sleep…a step toward transparent evaluation of consumer sleep tracking devices

Cathy A. Goldstein[1,*] and Christopher Depner[2]

[1]University of Michigan Sleep Disorders Center, 1500 E. Medical Center Drive, Ann Arbor, MI 48109 and
[2]Department of Health and Kinesiology, University of Utah, Salt Lake City, UT 84112

*Corresponding author. Cathy A. Goldstein, University of Michigan Sleep Disorders Center, 1500 E. Medical Center Drive, Ann Arbor, MI 48109. Email: cathygo@med.umich.edu

A paradox exists in sleep medicine and research—disordered sleep is detrimental through nightly, chronic exposure to sleep disturbances over months to years, but our gold-standard approach to assess sleep, the polysomnogram (PSG) [1], is generally used in the laboratory and is not practical for more than one to two nights of consecutive sleep assessments. Thus, current PSG technology lacks the capacity to assess sleep in an ecologically valid environment outside the laboratory where night-to-night variation in sleep disturbances likely occur over months to years.

Use of US Food and Drug Administration (FDA) cleared actigraphs has long been accepted by the research and medical communities as a sound method to track sleep over days to weeks in the ambulatory environment [2–4]. However, widespread utilization of actigraphy is limited by several factors. For example, actigraphs are expensive, typically monitor only motion without other physiological signals, and usually do not transmit information to clinicians or researchers in real time. Additionally, the information derived from actigraphs requires laborious data cleaning [3, 5]. Despite the resources required, use and interpretation of actigraphy is not typically reimbursed by insurers, which limits clinical utilization. From a research standpoint, the significant time, expertise, and cost of actigraphy data collection and interpretation often restrict study size and duration.

Consumer sleep tracking devices provide a relatively easy, inexpensive way to assess sleep over months to years, and their near ubiquitous use in modern society presents a potential solution to the problem of assessing sleep in the ambulatory environment. However, unclear performance and reliability of consumer sleep technologies (CSTs) has delayed clinical and research implementation [6–8]. Guidelines to assess the performance of CSTs have been disseminated, but strategies to systematically implement these guidelines are lacking and are a stated need to advance sleep medicine and research [7, 8].

In this issue of *SLEEP*, Menghini and colleagues address this need by providing a step-by-step analytic framework, accompanied by open-source R functions, to evaluate the performance of sleep trackers [9]. The procedures detailed by the authors are aimed at standardizing data collection processes, analytical techniques, and terminology used to describe results of studies designed to assess the sleep estimation capabilities of CSTs. Although PSG is the recommended gold-standard comparator, the framework presented here is flexible and adaptable to other reference sleep measures and can be applied to standard actigraphy. Additionally, the presented methods and calculations are applicable to systems that differentiate sleep from wake or classify sleep stages.

Importantly, throughout the manuscript, the authors urge that the term "validity" is replaced by "performance." Validity is defined as "the quality or state of being valid: such as the quality of being well-grounded, sound, or correct." Therefore, use of the term validation to describe the act of comparing two sleep assessment methods is faulty and such use of the term in this context should be abandoned by our field. For example, performance thresholds to "validate" CST devices against PSG are not established and are likely to vary between populations and study outcomes. Findings reporting

performance of a device against PSG, or another device, allow the end-user to determine if the device performance meets their stated needs. Critically, such performance measures must be conducted prior to implementing any CST for clinical or research purposes.

Supporting prior specifications [7, 8], the authors recommend that epoch-by-epoch (EBE) data is sought from the CST manufacturer to allow for comparison of simultaneous time increments between the CST system (device and analytic platform) and reference measure. For systems that provide EBE data, the first step of their pipeline [9] provides instructions to appropriately structure the data. For example, in settings where the epoch duration of the CST system does not match the 30-second epoch used to score PSG, reconciliation is required. Additionally, the start and end of the time in bed period monitored by the CST must be synchronized with lights on and lights off. Finally, annotated data from the CST system and PSG must share the same codes for each state (wake, sleep, etc.) to prepare the data for analysis.

Building upon prior recommendations [7, 8], Menghini and colleagues describe appropriate EBE analysis methods starting with the construction of error matrices that identify correct and misclassified sleep categories (i.e. wake, light, deep, and stage R sleep) by the CST system. Error matrices are then used to compute sleep-stage sensitivity and specificity. To account for individual-level differences, the authors recommend constructing error matrices for each study participant to calculate individual sleep-stage specificity and sensitivity prior to averaging group level values. To facilitate these calculations, both individual and absolute (error matrices calculated with aggregate data) computation methods are included in the corresponding R code.

In addition to sensitivity and specificity, positive predictive value, negative predictive value, prevalence index, and bias index are identified as useful metrics that should be calculated from EBE data to further characterize CST system performance. Other additional EBE computations included in the authors' framework consist of McNemar's test, Kappa coefficient and prevalence-adjusted bias-adjusted kappa (given imbalance between sleep and wake epochs during the usual sleep period), and receiver operating characteristic curves.

The manuscript also sets forth analysis procedures that are applicable even when CST system output is only available in summary form over the course of the night. A discrepancy analysis is necessary in the comparison of any new testing method to gold-standard. Techniques to systematically conduct a discrepancy analysis are provided and the computations that underlie these metrics are described in detail. The measurement differences between CST system and PSG, for each sleep parameter, should first be evaluated on an individual level. Subsequently, systematic and random error are quantified by bias and 95% level of agreement, calculations that are contingent on assumptions of constant bias (bias independent from measurement size), homoscedasticity (error consistent over measurement sizes), and normal distribution of differences. Therefore, tests for these assumptions are also included in the open-source R functions associated with this publication.

After the procedures and rationale for each analysis step are provided, Meghini and colleagues apply their open-source tool to data from an investigation that compared a widely used CST to PSG in 14 healthy adults. The pipeline was successfully deployed for data structuring, and discrepancy and EBE analysis. The discrepancy analysis demonstrated violations of constant bias, homoscedasticity, and normal distribution assumptions and exemplified the role of linear regression to express bias and bootstrapping and logarithmic transformation for heteroscedasticity and non-normal distributions. Bland Altman plots of CST and PSG differences for each sleep parameter (total sleep time, sleep efficiency, wake after sleep onset, light sleep duration, deep sleep duration, REM sleep duration) allow the readers to visualize the scenarios of (1) all assumptions satisfied, (2) proportional bias with homoscedasticity, (3) constant bias with heteroscedasticity, and (4) both proportional bias and heteroscedasticity. Next, error matrices were tabulated from the example EBE data and revealed the sleep state misclassifications underlying the discrepancy analysis findings, highlighting the benefits of this comprehensive approach to assess CST system performance.

The work and companion R functions (https://github.com/SRI-human-sleep/sleep-trackers-performance) by Menghini and colleagues operationalize the recommended techniques for analyzing the performance of CSTs- against PSG (or other reference) for comparison studies. The authors satisfied the need for tools to improve efficiency and reproducibility and reduce heterogeneity of investigations into CST performance.

Numerous limitations translating CST performance from in-laboratory investigations to the free-living environment exist and include passive time in bed detection, home sleep environment factors (e.g. spouse or pets), reliability over time, assessing sleep periods outside of the main sleep bout, output scores that do not correspond to clinical or scientific metrics (e.g. nightly sleep score), and concerns related to use in certain populations with co-morbid disorders. Further, CST system performance is specific to the device and associated firmware and software versions, and the ability to extrapolate these results after sensor, algorithm, and other updates remains unclear. Nonetheless, Meghini and colleagues brought us one step closer to standardized and transparent use of CSTs and we recommend adopting their guidelines.

Objective sleep parameters, recorded over time, provide distinct and valuable information over self-report for the clinical evaluation and management of various sleep disorders [4]. Additionally, a growing body of research has revealed the importance of sleep parameters beyond sleep duration [10–16], which are only feasibly obtained through passive, objective sleep recording over time. Therefore, the utility of CSTs transcends the current use cases for actigraphy as the duration of sleep tracking with CSTs far exceeds the usual one to 2 weeks of sleep recorded with actigraphs. Rational, transparent, and scientifically sound use of CSTs may capture previously unidentified changes in sleep over months, seasons, and years relevant to health and disease. Therefore, a pipeline to improve the efficiency of CST performance assessment is crucial and is well-positioned to advance sleep medicine and research.

Funding

None declared.

Conflict of interest statement

C.A.G. receives royalties from UpToDate and is part-inventor of sleep and circadian tracking app that is licensed to an outside

entity (Arcascope, LLC). C.M.D. reports funding from the NIH and the Colorado Clinical and Translational Sciences Institute that is unrelated to this work.

References

1. Berry RB, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Ver. 2.6*. Darien, IL: American Academy of Sleep Medicine; 2020.
2. Ancoli-Israel S, *et al*. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;**26**(3):342–392.
3. Ancoli-Israel S, *et al*. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav Sleep Med*. 2015;**13**(Suppl 1):S4–s3810.
4. Smith MT, *et al*. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment. *J Clin Sleep Med*. 2018;**14**(7):1209–1230.
5. Patel SR, *et al*. Reproducibility of a standardized actigraphy scoring algorithm for sleep in a US Hispanic/Latino population. *Sleep*. 2015;**38**(9):1497–150310.
6. Khosla S, *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.
7. de Zambotti M, *et al*. Wearable Sleep Technology in Clinical and Research Settings. *Med Sci Sports Exerc*. 2019;**51**(7):1538–1557.
8. 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.
9. Menghini L, *et al*. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2020;**44**(2). doi:10.1093/sleep/zsaa170.
10. Buysse DJ. Sleep health: can we define it? Does it matter? *Sleep*. 2014;**37**(1):9–17.
11. Bei B, *et al*. Beyond the mean: a systematic review on the correlates of daily intraindividual variability of sleep/wake patterns. *Sleep Med Rev*. 2016;**28**:108–124.
12. Huang T, *et al*. Cross-sectional and prospective associations of actigraphy-assessed sleep regularity with metabolic abnormalities: the multi-ethnic study of atherosclerosis. *Diabetes Care*. 2019;**42**(8):1422–1429.
13. Huang T, *et al*. Sleep irregularity and risk of cardiovascular events. *J Am Coll Cardiol*. 2020;**75**(9):991–999.
14. Faust L, *et al*. Deviations from normal bedtimes are associated with short-term increases in resting heart rate. *NPJ Digit Med*. 2020;**3**:39.
15. Wallace ML, *et al*. Which sleep health characteristics predict all-cause mortality in older men? An application of flexible multivariable approaches. *Sleep*. 2018;**41**(1). doi:10.1093/sleep/zsx189
16. DeSantis AS, *et al*. A preliminary study of a composite sleep health score: associations with psychological distress, body mass index, and physical functioning in a low-income African American community. *Sleep Health*. 2019;**5**(5):514–520.