Published in final edited form as:

Sleep Health. 2022 June ; 8(3): 263–269. doi:10.1016/j.sleh.2022.02.006.

Rigorous performance evaluation (previously, “validation”) for informed use of new technologies for sleep health measurement

Massimiliano de Zambotti, PhD[a], **Luca Menghini, PhD**[b], **Michael A. Grandner, PhD**[c], **Susan Redline, MD, MPH**[d,e], **Ying Zhang, PhD**[d], **Meredith L. Wallace, PhD**[f], **Orfeu M. Buxton, PhD**[g,*]

[a]Center for Health Sciences, SRI International, CA, USA

[b]Department of Psychology, University of Bologna, Italy

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

[d]Department of Medicine, Brigham and Women’s Hospital, Boston, MA, USA

[e]Harvard Medical School, Boston, MA, USA

[f]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

[g]Department of Biobehavioral Health, The Pennsylvania State University, University Park, PA, USA

Abstract

New sleep technologies have become pervasive in the consumer space, and are becoming highly common in research and clinical sleep settings. The rapid, widespread use of largely unregulated and unstandardized technology has enabled the quantification of many different facets of sleep health, driving scientific discovery. As sleep scientists, it is our responsibility to inform principles and practices for proper evaluation of any new technology used in the clinical and research settings, and by consumers. A current lack of standardized methods for evaluating technology performance challenges the rigor of our scientific methods for accurate representation of the sleep health facets of interest. This special article describes the rationale and priorities of an interdisciplinary effort for rigorous, standardized, and rapid performance evaluation (previously, “validation”) of new sleep and sleep disorders related technologies of all kinds (eg, devices or algorithms), including an associated article template for a new initiative for publication in Sleep Health of empirical studies systematically evaluating the performance of new sleep technologies. A structured article type should streamline manuscript development and enable more rapid writing, review, and publication. The goal is to promote rapid and rigorous evaluation and dissemination of new sleep technology, to enhance sleep research integrity, and to standardize terminology used in Rigorous Performance Evaluation papers to prevent misinterpretation while facilitating comparisons across technologies.

[*]Corresponding author: Orfeu M. Buxton, PhD, The Pennsylvania State University, 221 Biobehavioral Health Building, University Park 16802 PA, USA. orfeu@psu.edu (O.M. Buxton).

Conflicts of interest
The remaining authors have no conflicts to disclose.

Keywords

Performance evaluations; Actigraphy; Wearables; Sensors; Algorithms; Machine learning

Evaluation is not validation

"The very word validation implies an affirmative result, that the process of validation will somehow validate the model…," whereas "Evaluation implies an assessment in which both positive and negative results are possible, and where the grounds on which a model is declared good enough are clearly articulated."[1] Such a distinction can be meaningfully transferred from the field of modeling to that of technology evaluation, and specifically the evaluation of new sleep technology, whose fast-paced development is increasingly accompanied by studies claiming to "validate" new devices, algorithms and measurement systems against gold standard technologies (eg,[2–4]).

Rationale

New sleep technologies have become common in research and clinical sleep settings.[5] These technologies include, for example, sleep wearable and nearable tracking devices, wireless measurement systems, novel analytic methods based on artificial intelligence (AI) designed to measure sleep, sleep-related physiology, and/or other physiological events occurring during and around sleep. They have become increasingly visible in the consumer space and within the internet of things framework. The rapid, widespread use of largely unregulated and unstandardized technology (eg, clinical or wellness tools) has enabled the quantification of many different facets of sleep health.

Promising technologies can drive innovation and new discovery. In sleep research, new technology is evolving and elaborating the very concept of the definition of sleep health, accounting for its multiple subjective, behavioral, and physiological determinants,[6] and providing enriched information on psychosocial and other contextual factors affecting sleep (eg,[7–8]). Sleep is now considered a global, multidimensional, integrated aspect of health, and the "sleeping" period is being seen as a window of opportunity to measure a broad range of typical and atypical psychophysiological processes and events.[9] As sleep scientists, it is our responsibility to inform principles and practices regarding proper evaluation of any new technology used in the clinical and research fields, and by consumers.

However, the limited coordinated response from the scientific community (eg, lack of standardized methods for evaluating technology performance) challenges the rigor of our scientific methods for accurate representation of the sleep health facet of interest potentially degrading the integrity of scientific discoveries. Particularly, limitations of sleep devices and algorithms that employ undisclosed, proprietary, and otherwise obscured methods challenge scientific evaluation of these technologies despite ongoing use in research and clinical sleep settings. Additional emerging challenges include addressing privacy issues such as exchanging and storing personal health information, which may have potential long-term unintended consequences.[10]

Goals

Sleep Health recognizes the need to systematically understand and promote the appropriate use of sleep technology that shows promise in advancing sleep science.[5,10–12] As the first step in this journey, *Sleep Health* is coordinating an interdisciplinary effort for rigorous, standardized, and rapid performance evaluation (previously, “validation”) of new sleep technology that proposes to measure sleep and associated psychophysiological processes (eg, respiratory events such as oxygen desaturations, seizures, arrhythmias, and heart attacks) and health outcomes (eg, cardiovascular risk assessment). Technology includes, but is not limited to, consumer-focused and clinical/research devices and machine/deep learning methods, sleep/wake/stages/events classifiers, motion-based, wearables and “nearables” (noncontact), electrocardiographic-, photoplethysmographic-, electroencephalographic-based, and hybrid approaches. Reference methods include, but are not limited to, the accepted standards in the respective fields (eg, polysomnography sleep assessment for sleep/wake and sleep staging, electroencephalography for seizure detection).

This special article describes the rationale, the priorities, and the article template for the new *Sleep Health* initiative concerning empirical studies systematically evaluating the performance of new sleep technologies. Examples of such investigations could involve but are not limited to the performance evaluation of wearable devices for sleep assessment against polysomnographic sleep staging, apnea detection, device-agnostic multisensor algorithms for sleep staging, sleep technology performance across different populations and conditions, etc. Studies evaluating conditions and events (eg, menstrual cycle tracking, seizure detection) that use sleep as a window of opportunity to better track them, are also within scope. To this end, we have created a streamlined, structured article type that will undergo rigorous but expedited review and publication. This new article type is designed to serve the following purposes:

- Promote rapid but rigorous evaluation and dissemination of new sleep technology to bridge the gap between technology development and evidence-based protocol for their use and adoption by the research community.
- Enhance research integrity while reducing researcher burden by explicitly stating both minimally acceptable and best practices to evaluate the performance of sleep technologies against reference technologies.
- Standardize the terminology used in performance evaluation to prevent misinterpretation of study outcomes (eg, abandoning the term “validation,” which implies that a new technology is generally appropriate for research or clinical use simply because a “validation” study exists, regardless of performance criteria). Readers will be better able to determine the “most valid” device for their particular context and use case.
- Prioritize information requiring reporting in scientific publications, and thus facilitating comparisons of study outcomes.
- Provide credible references for scientists and practitioners considering the use of new technology (eg, methodological abstract, outline of critical study outcomes).

- Reduce variability in language and assessments that occurs as authors and reviewers discuss study technology and performance metrics (eg, debatable statements for accuracy of specific performance metrics, or devices overall). This will minimize subjective interpretation while maximizing comparability across publications and systematically facilitate dissemination.

Priority areas

The *Sleep Health* expert task force has identified priorities in sleep technology performance assessment. Here we outline some top priorities, as a result of a consensus reached among the task force working group of opinion leaders and a literature review (see Table 1).

Key aspects of the new *Sleep Health* structured article type for empirical studies evaluating the performance of new sleep technologies

In the following section, we outline the main *required* and *recommended* information for any new Rigorous Performance Evaluation paper submission. In addition, a template will be provided for reference (in Guidelines for Authors).

In this new article type, *Sleep Health* prioritizes high-quality research using standardized and replicable methods for analysis, standard outcomes, and standard terminology and definitions (see Table 2 for examples). Empirical objective results are prioritized over interpretation, and methodological aspects are prioritized over background. Specifically, *Sleep Health* recommends the use of metric-specific, data-driven details (eg, “Technology A has X sensitivity for assessing Y”) over qualitative statements on the overall “goodness/ badness” of a new technology (eg, “This device provides good performance in …” or “The new technology is accurate in …”). Clearly refer to specific contexts of application as appropriate to the evaluation performed (eg, “Technology A is useful for detecting differences in sleep metric Y between populations 1 and 2”). Crucially, the authors should describe the appropriate context of technology use, such as particular environments or experimental settings, or types of participants (eg, patient vs. population level, infants vs. adults) for which the sleep technology performance has been evaluated.

Sleep Health promotes transparency and rigor, and would consider industry-sponsored or industry-led studies (Table 3). When meeting ethical standards (eg, Human Subjects Research approval), and when conducted using rigorous procedures, industry-supported or industry-internal studies are valued. Indeed, it could be considered a responsibility of corporations to demonstrate the quality of their devices across all aspects of performance. Authors should disclose sources of support as already required, as well as the role of the funder in study design and manuscript review. Moreover, it is recommended that datasets used for analysis be made available for transparency and to enable reproducibility.

In addition, *Sleep Health* welcomes studies replicating previous evaluations (eg, a similar study from different research/clinical institutions or in different populations/conditions).

Title Page/Abstract

Title—Starting from the title, *Sleep Health* aims to promote clarity and consistency. Thus, the title should be informative. The focus technology should be named and/or spelled out, together with the main reference and main object of the comparison (eg, *Performance of a specific technology, against a specific reference* technology*, to estimate a specific outcome variable, in a specific sample under a specific condition*). If the technology has a commercial brand and model, *Sleep Health* encourages specifying both brand and model in the title.

Summary/abstract—Different from other journal article types in *Sleep Health*, *Rigorous Performance Evaluations of Sleep Technologies* are designed to capture key information for an immediate understanding of outcomes, facilitating dissemination and replicability. Thus, the front matter must convey all critical information to offer the reader a clear objective picture of the study, free from interpretation biases (see Template for details).

Introduction: Rationale, significance, background, and aims of the study

Introduction should be brief. Extensive background literature review is discouraged.

Sleep Health requires a brief structured introduction, which includes the following headers: (a) **Rationale**, (b) **Significance**, (c) **Background**, and (d) **Aims** of the study. The use of subheaders (eg, “Why it matters”, “What still is not known”), also related to background topics (eg, “Challenges in measuring sleep in adolescence”), is recommended (see Template for details).

The introduction should clearly state the measured object/process (eg, sleep quantity, sleep apnea), the current standards of measurement (eg, polysomnography), and the rationale for using the new technology (eg, large-scale population screening). Aims should clearly highlight the core aspects investigated in the study (eg, population, conditions). The introduction should focus on the importance of the new technology vs. gold standard technology(ies) (eg, polysomnography), algorithms (eg, if the purpose is to create new O_2 desaturation classification based on XYZ, the intro should focus on existing classificators), or devices (eg, if the study evaluates a new device model, the intro should refer to previous comparison with previous device versions, or with other devices using similar features).

Methods and results: Standardize methodological study information and core data processing

Sleep Health requires a structured methods and results section, with the following subsections: **Sample; Focus technology; Reference technology; Design, study setting, and procedures; Core analytics and main outcome variables;** and **Additional analytics and exploratory analyses.** The use of sub-headers is encouraged (see Template for details).

Sample—*Sleep Health* requires reporting relevant demographic information that may affect the performance of the technology under observation. Reports should include frequency of participants by sex, participants’ age (by cohort if multiple), race/ethnicity, underlying health characteristics, inclusion/exclusion criteria, are required. Detailed description of training vs testing samples where appropriate (eg, in use of machine learning and deep learning neural networks) is required. Details about the sample size, including design analysis, sampling strategy and/or sample size justification for core analytics and main outcomes, are encouraged. Information should be included about participant consent, ethical board (e.g., IRB) approval, and compliance with the Declaration of Helsinki.

Focus technology—*Sleep Health* requires details about the technology under assessment, such as brand, model, firmware, and functioning of a device, and if available, sampling frequency, automatic and post-processing filters, and artifact processing procedures; method, rationale, features, and/or the validation strategy (eg, cross-validation) of new classification algorithms, etc.

Reference technology—*Sleep Health* requires details and supporting justification for the chosen reference technology (eg, polysomnography, electrocardiography, piezoelectric breathing band), including data processing procedures (eg, filtering, artifact identification and removal).

Design, study setting, and procedures—*Sleep Health* requires details about the study setting (eg, in-lab, free-living), protocol (eg, single vs. multiple nights), timeline for data collection, known updates that may affect the performance of the technology under evaluation, key settings as used in practice, and, if applicable, the procedure used for temporal synchronization between the new technology and the reference technology.

Core analytics and main outcome variables—*Sleep Health* requires a clear statement of the main outcomes under evaluation (eg, 30-s sleep annotations, total sleep time, oxygen desaturation) using the accepted definition for the main variables under observation (eg, American Academy of Sleep Medicine definitions of sleep parameters) prior to describing the core analytics.

Sleep Health requires that the study provides epoch-by-epoch (EBE) analysis (confusion matrix) when applicable, Bland-Altman (BA) plots and coefficients, and receiver operator characteristic (ROC) curves with area under the curve (AUC) indicators for the main outcomes considered by the study (see Marino et al.[3] and Menghini et al.,[13] for examples). For processing these outcomes, an analytical pipeline has been described by Menghini et al.[13] and implemented as open-source R-based code available at https://github.com/SRI-human-sleep/sleep-trackers-performance. The pipeline includes the inspection of individual-level data for data cleaning and filtering (eg, excluding total sleep time values above a certain threshold), data recoding (eg, differentiating pre- and post-sleep onset wake epochs), and data aggregation (eg, summary sleep measures from EBE data), in addition to the automatic computation of group-level performance metrics.

Sleep Health recommends reporting summary results in well-structured tables and figures (see Guidance for Authors). Additional details to be included in supplemental materials and online appendices are strongly encouraged. Core analytics should be implemented based on the following recommendations:

- Clearly specify core analytics used, and consider preregistering the analysis.

- Provide sufficient details for data analysis reproducibility, including publicly available data, meta-data, and analytical code (with short descriptions), in order to make readers able to read the manuscript and reproduce the analyses.
- Report all performance metrics with the corresponding 95% confidence intervals, and by following the statistical guidelines provided by *Sleep Health*, for example, use correction methods for multiple comparisons (eg, Bonferroni) when appropriate.
- When evaluating classification algorithms, use independent training data and evaluation data for computing core performance metrics, or best practices in machine learning methods for independence of training and evaluation.

Additional analytics and exploratory analyses—*Sleep Health* welcomes additional analytics and exploratory analyses, particularly to explore potential confounders of main outcomes, or consider secondary outcomes. Secondary and exploratory analyses should be clearly described as such, and follow the same recommendations listed above.

Discussion

Sleep Health requires a critical discussion of the outcomes in relation to the purpose and the background in the introduction. The structured discussion must include the following subsections: **Main results and implications, Additional results and implications, Limitations and future perspectives**, and **Core conclusion**.

Main results and implications—Discuss the implications of the main performance metrics of the study, in the context of the previous studies evaluating the performance of the same or a similar technology, with data-driven recommendations for the informed use of the evaluated technology.

Additional results and implications—When appropriate, discuss the implications of the results of additional analyses in the context of the previous literature, with data-driven recommendations for future studies.

Limitations and future perspectives—Disclose the main study limitations, their implications, generalizability for future use of the evaluated technology, and recommendations for future studies evaluating the performance of the same or a similar technology.

Core conclusion—Succinctly summarize the outcomes and significance of the study, according to the main aims of the study.

Data sharing policy

Sleep Health encourages providing de-identified individual-level data from both the reference and new technology data, per funder policies. *Sleep Health* encourages depositing de-identified raw and summary data and associated study documentation in publicly accessible data repositories (eg, National Sleep Research Resource; https://sleepdata.org). Code could be made publicly available using GitHub or similar sites.

Sleep Health strongly recommends that for any evaluation of an ad-hoc generated analytical model or classifier, the final algorithm is made publicly available (eg, if a new classifier is built for apnea event detection based on a combination of accelerometry and photoplethysmography [PPG] features, the algorithm should be disclosed). In recognition of the challenges of publicly releasing algorithms from consumer-oriented technology corporations attempting to defend their intellectual property in a broken system for such protections, algorithm releases are strongly encouraged but not always required. A justification for not releasing algorithms must be provided if they are not disclosed.

Disclosure of sponsor

Sleep Health requires discussing any potential financial interests and/or conflicts of interest, particularly when related to an industry sponsor, which could be perceived as influencing the study design, execution, and publication.

Sleep Health expert task force

The *Sleep Health* task force brought together a multidisciplinary group of experts in sleep, sensing technology, AI, and statistics. Here we provide an outline of the core competencies and expertise of the team.

Team members	Expertise	Link to Publications
 Massimiliano de Zambotti, PhD	Dr. de Zambotti is a Principal Scientist at SRI International and an expert in wearable sleep technology. He has been involved in several initiatives and international collaborations to investigate the performance, standardization, informed use, and regulation of sleep technology. Of relevance, he co-authored the position statement following the "*International Biomarkers Workshop on Wearables in Sleep and Circadian Science*" that was held at the 2018 SLEEP Meeting of the Associated Professional Sleep Societies. He introduced and operationalized practical guidelines for evaluating the performance of sleep technology vs. reference and reviewed capability, rationale, and limitation of sleep technology.	https://scholar.google.com/citations?user=PsR6NFYAAAAJ&hl=en
 Luca Menghini, PhD	Dr. Menghini is a postdoctoral research fellow at University of Bologna working on the development and improvement of innovative methods in occupational health and sleep assessment, particularly focusing on Ecological Momentary Assessment designs. Having evaluated the performance of several multi-sensor wearable devices for both diurnal and nocturnal psychophysiological, he has recently proposed an R-based analytical pipeline for evaluating consumer sleep technologies.	https://scholar.google.it/citations?user=ZFRn4ssAAAAJ&hl=en

Team members	Expertise	Link to Publications
 Michael Grandner, PhD	Dr. Grandner is the Director of the Sleep and Health Research Program at the University of Arizona College of Medicine and the Behavioral Sleep Medicine Clinic at the Banner-University Medical Center in Tucson, Arizona. He is Associate Professor of Psychiatry, Medicine, Psychology, Nutritional Sciences, and Clinical Translational Science. His research has focused on real-world implications of sleep health, including the development, evaluation, and implementation of sleep health technology.	https://scholar.google.com/citations?user=ePsVuAgAAAAJ&hl=en
 Susan Redline, PhD	Dr. Susan Redline is the Farrell Professor of Sleep Medicine, Harvard Medical School, and Director of the Program in Sleep Medicine Epidemiology at Harvard Medical School. She has directed multiple large cohorts and clinical trials and co-directs the National Sleep Research Resource, a NIH funded sleep data repository that shares about 2TB of sleep data per week to the greater community.	https://scholar.google.com/citations?user=xWLZ_wgAAAAJ&hl=en
 Ying Zhang, PhD	Dr. Ying Zhang is a data scientist is a Data Scientist for the National Sleep Research Resource (NSRR), a NHLBI-data and tool resource repository offering free access to large collection of polysomnography, actigraphy and other phenotype data at Brigham and Women's Hospital. She is currently leading the data harmonization and metadata standards development at the NSRR. Prior to joining Brigham and Women's Hospital, Dr. Zhang led data collection, harmonization, as well as construction and management of a consortium database of multiple national cohorts at the Harvard/MGH Center on Genomics, Vulnerable Populations, and Health Disparities. Dr. Zhang has previously worked as a research fellow at various non-profit organizations including the National Academy of Medicine, specialized in data modeling and visualization to inform policy decisions.	https://scholar.google.com/citations?hl=en&user=snjAPHMAAAAJ
 Meredith Wallace, PhD	Dr. Wallace is a biostatistician and Associate Professor of Psychiatry, Statistics, and Biostatistics at the University of Pittsburgh. With funding from the National Institute on Aging, her primary research program is focused on developing, adapting, and applying state-of-the-art machine learning approaches to determine which dimensions of multidimensional sleep health predict mental and physical health outcomes in older adults. Through this work, she harmonizes data across cohorts to perform rigorous external evaluations of predictive algorithms. In addition, Dr. Wallace is a statistical co-investigator on several studies evaluating the use of machine learning for improving sleep-related technologies.	https://scholar.google.com/citations?user=VxqPl8IAAAAJ&hl=en&oi=sra
 Orfeu Buxton, PhD	Dr. Buxton, Elizabeth Fenton Susman Professor of Biobehavioral Health at Penn State, directs the Sleep, Health, & Society Collaboratory. His completed and ongoing interdisciplinary studies in free-living humans of all ages address sleep health and wellbeing across the life course, with sleep usually measured by wearable devices. In addition to extensive experience with large-scale, longitudinal studies, he has co-authored device, algorithm, and machine learning algorithm evaluations.	https://scholar.google.com/citations?user=DP_YDXoAAAAJ https://pennstate.pure.elsevier.com/en/persons/orfeu-m-buxton

Funding & declaration

MdZ has received funding from Noctrix Health, Inc., Lisa Health Inc., Verily Life Sciences LLC., the National Science Foundation (NSF) IIP-2111818; he has patents issued: US222921894, US20140316191, and pending:

EP3591659, JP2020014841, US20200013511, US20200222699, AU2018302101, EP3654839, WO2019018400, WO2022006119 WO2022032121. MG received grants from Kemin Foods, Jazz Pharmaceuticals, and CeraZ; he has received consulting fees from Idorsia Pharmaceuticals, Fitbit, Smartypants Vitamins, Natrol, National Sleep Foundation, American Sleep Medicine Foundation; He has received publishing royalties from Elsevier; he has received speaking honoraria and/or travel support from New York University, University of Maryland, Stavis & Cohen, University of Miami, Clinical Education Alliance, and Jazz Pharmaceuticals. SR received consulting fees from Jazz Pharma and Eli Lilly, and has received consulting fees for serving as a member of the Scientific Advisory Board for Apnimed, Inc. Outside of the current work, OMB received subcontract grants to Penn State from Proactive Life LLC (formerly Mobile Sleep Technologies) doing business as SleepScape (NSF/STTR #1622766, NIH/NIA SBIR R43-AG056250, R44-AG056250), received honoraria/travel support for lectures from Boston University, Boston College, Tufts School of Dental Medicine, New York University, University of Miami, and Allstate, consulting fees from SleepNumber, and receives an honorarium for his role as the Editor-in-Chief of Sleep Health (sleephealthjournal.org). MLW received consulting fees from HealthRhythm, Sleep Number Corporation, and Noctem, unrelated to this work. Outside of the current work, MAG received grants from Kemin Foods, Jazz Pharmaceuticals, and CeraZ; he has received consulting fees from Idorsia Pharmaceuticals, Fitbit, Smartypants Vitamins, Natrol, National Sleep Foundation, American Sleep Medicine Foundation; he has received publishing fees from Elsevier; and he has received speaking honoraria and/or travel support from New York University, University of Maryland, Stavis & Cohen, and Jazz Pharmaceuticals. SR and YZ are supported by NIH contract, 75N92019C00011; National Sleep Research Resource.

References

1. Oreskes N. Evaluation (not validation) of quantitative models. Environ Health Perspect 1998;106(Suppl 6):1453–1460. (suppl 6). [PubMed: 9860904]
2. Khademi A, El-Manzalawy Y, Master L, Buxton OM, Honavar VG. Personalized sleep parameters estimation from actigraphy: A machine learning approach. Nat Sci Sleep 2019;11:387–399. [PubMed: 31849551]
3. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013;36(11):1747–1755. [PubMed: 24179309]
4. 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): zsaa045. [PubMed: 32215550]
5. 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. [PubMed: 30789439]
6. Scarlett S, Nolan HN, Kenny RA, O'Connell MDL. Discrepancies in self-reported and actigraphy-based sleep duration are associated with self-reported insomnia symptoms in community-dwelling older adults. Sleep Health 2021;7(1):83–92. [PubMed: 32732155]
7. Etindele Sosso FA, Holmes SD, Weinstein AA. Influence of socioeconomic status on objective sleep measurement: a systematic review and meta-analysis of actigraphy studies. Sleep Heal 2021;7(4):417–428.
8. Rezaei N, Grandner MA. Changes in sleep duration, timing, and variability during the COVID-19 pandemic: large-scale Fitbit data from 6 major US cities. Sleep Heal 2021;7(3):303–313.
9. Chung J, Goodman M, Huang T, Bertisch S, Redline S. Multidimensional sleep health in a diverse, aging adult cohort: concepts, advances, and implications for research and intervention. Sleep Heal 2021;7(6):699–707.
10. 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).
11. Khosla S, Deak MC, Gault D, et al. Consumer sleep technology: an American academy of sleep medicine position statement. J Clin Sleep Med 2018;14(5):877–880. [PubMed: 29734997]
12. 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. [PubMed: 32005346]
13. 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).

Table 1

Priority	Description
Sleep metrics	Improved and standardized approaches to the definition of sleep intervals (ie, when deliberate sleep opportunities exist, such as times in and out of bed). Improved detail regarding performance of technology to determine when sleep opportunities exist (eg, when individuals are in or out of bed and/or attempting to sleep) with and/or without user input.
	Improved and standardized approaches to the classification of sleep continuity variables, including time in and out of bed, sleep latency, awakenings, wake time after sleep onset, etc
	Improved and standardized approaches to the classification of sleep stages.
	Improved characterizations of daytime sleep, unintended sleep, and/or naps, including differentiation between sedentary behavior and sleep and specification for timing and duration criteria in the classification of naps whether intentional or unintentional.
	Oxygen desaturation and sleep-related breathing, particularly in at-home settings.
Study populations	Evaluations in populations where activity patterns and relationships between activity and other physiological measures may vary from “young healthy adults” that are routinely included in studies.
	Example priority populations include individuals with pacemakers, those undergoing medical treatments affecting motion and/or cardiovascular hemodynamics, people with physical (eg, back injury) and mental (eg, depressive disorders) conditions and sleep disorders (eg, restless legs syndrome, narcolepsy, insomnia disorder, sleep disordered breathing), individuals from underrepresented cultural backgrounds, individuals with atypical sleep-wake patterns (eg, nursing home residents, shift workers), individuals with atypical sleeping arrangements (eg, people who don’t sleep in a bed).
Characteristics and/or confounders	Exploration of the role of contextual factors on measurement properties, including sex-specific variables (eg, menstrual cycle effects on sleep, or on primary data for algorithm development), age-related variables (eg, studies conducted during puberty, transition to nursing home), and other demographic variables (eg, race, ethnicity, skin color, body size), environmental variables (eg, temperature and light, seasonal changes).
	Evaluations in conditions of pronounced variation in the targeted variable of interest (eg, accuracy of a sleep technology in measuring wake in condition of low and high sleep fragmentation).
	Evaluations under different conditions known to alter the physiological features used in sleep classification (eg, evaluating the impact of caffeine/alcohol consumption on performance).
	Examples of factors affecting sleep technology performance are outlined in de Zambotti et al.[12]
Data source	Evaluation of accuracy and reliability of raw signals derived from the sensors versus derived and/or summarized signals.
	Unidimensional vs. multidimensional approaches to sleep and sleep event classification.
	Evaluation of outputs integrating multiple physiological measurements (eg, oxygen desaturation, heart rate variability [HRV]) to improve (and extend) activity-based measurements.

Table 2

Promoting standard definitions and operationalization of common performance outcome metrics.

Required
Report level-specific specificity and sensitivity: When classifications with more than 2 levels are involved, *Sleep Health* requires the use of the terms ‘Specificity’ and ‘Sensitivity’ in relation to any level of a classification (e.g., 4-level sleep classification with wake-, N1 + N2, N3, and REM sleep specificity and sensitivity outputs).
Employ evaluation methods including more than accuracy alone, which is affected by disproportionate numbers of cases underlying sensitivity and specificity. To avoid mischaracterizing a new technology as having high accuracy while there is an imbalance in sensitivity/specificity performance, *Sleep Health* requires providing accuracy together with sensitivity and specificity, and/or other methods for balancing.
Sample-based confusion matrix: Sleep Health requires that the confusion matrix be calculated on an individual level, and then summarized by providing mean, SD, and 95% CI[9].
Recommended
Do not use “ *Validation*”: Deprecated terminology to be usually avoided. *Sleep Health* recommends the use of “evaluation” to avoid the improper use of the concept “validation” to state that a new technology is ‘valid’ simply because there is a study evaluating its performance against a reference. Exceptions include and are not limited to machine learning or deep learning algorithms, where the term “validation” has a different meaning.
Use independent samples: When the performance of an empirically-derived algorithm is tested, *Sleep Health* recommends that independent samples be used for training and testing. This can include, but is not limited to, both internal cross-validation and external validation in a separate sample. For machine learning algorithms, using a k-fold process for selecting repeated, balanced training and testing datasets is an accepted standard method. Further evolution of rigorous methods is expected.
Report proportional bias and limits of agreement (LOAs): When reporting Bland-Altman plots and statistics, *Sleep Health* recommends distinguishing between uniform bias (differences are uniformly distributed across the size of measurement) and proportional bias (differences are proportional to the range of measurement) and distinguishing between homoscedastic LOAs (bias ±1.96*SD of the differences) and heteroscedastic LOAs (expected LOAs proportional to the range of measurement).
Report level-specific receiver operating characteristic (ROC): Sleep Health recommends (where appropriate) plotting a ROC curve and reporting the corresponding area under the curve (AUC) for each level of relevant classification (e.g., wake, N1 + N2, N3, and REM sleep).
Report measures of classification agreement not due to chance, such as the coefficient and the prevalence-adjusted bias-adjusted kappa (PABAK) coefficients.

In studies evaluating the performance of sleep technology, we frequently encounter inconsistency in the use of evaluation terminology. We define recommended and required elements in the use of a set of standardized terminology to address commonly “misused” language in the field of technology evaluation. Please refer to Menghini and colleagues[13] for a comprehensive discussion on the topic and a detailed description and operationalization of the key analytics and outcomes. We recognize that the field is fluid. We will update our Guide for Authors and Template as these metrics and methods evolve.

Table 3

Talking to industry. What do we need to know?

Sleep Health still recognizes critical barriers in the evaluation and use of consumer sleep technology, which are highly promising, powerful tools to advance the field of sleep research and clinical sleep medicine. With this in mind, we outline some key barriers to adopting this technology, with the hope of opening a meaningful discussion amongst industry, academia, and other stakeholders:

– Raw data are still largely unavailable.

– The frequent undisclosed/proprietary nature of algorithms has limited their evaluation and use.

– Undisclosed updates and/or no control of this versioning process disrupts research and classifier development, and further limits their dissemination and adoption.

– Unclear claims regarding wellness / diagnostics / therapeutics are reflected in a lack of clear outcomes.

– Peer review should be preferred to company-internal public press.

– Importance of implementation, including access to application programming interfaces (APIs), research-level access to data, availability of real-time data

– Privacy considerations, including ability to de-identify data, avoid deductive re-identification, etc