SLEEPJ, 2022, 1–10

https://doi.org/10.1093/sleep/zsac002
Advance Access Publication Date: 14 January 2022
Perspective

PERSPECTIVE

Sleep and circadian informatics data harmonization: a workshop report from the Sleep Research Society and Sleep Research Network

Diego R. Mazzotti[1,2,†,*], Melissa A. Haendel[3,†], Julie A. McMurry[3], Connor J. Smith[4], Daniel J. Buysse[5], Till Roenneberg[6], Thomas Penzel[7], Shaun Purcell[8], Susan Redline[9], Ying Zhang[9], Kathleen R. Merikangas[10], Joseph P. Menetski[11], Janet Mullington[12] and Eilis Boudreau[4,†]; on behalf of the Sleep Research Network Task Force

[1]Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, USA, [2]Division of Pulmonary Critical Care and Sleep Medicine, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, USA, [3]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, CO, USA, [4]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA, [5]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA, [6]Institute and Polyclinic for Occupational-, Social- and Environmental Medicine, LMU Munich, Germany, [7]Interdisciplinary Center of Sleep Medicine, Charité University Hospital, Berlin, Germany, [8]Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [9]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [10]Genetic Epidemiology Research Branch, Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA, [11]Foundation for the National Institutes of Health, Bethesda, MD, USA and [12]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

[†]These authors contributed equally to this work.

*Corresponding author. Diego R. Mazzotti, 3901 Rainbow Boulevard, Mail Stop 3065, Kansas City, KS 66160, USA. Email: droblesmazzotti@kumc.edu.

Abstract

The increasing availability and complexity of sleep and circadian data are equally exciting and challenging. The field is in constant technological development, generating better high-resolution physiological and molecular data than ever before. Yet, the promise of large-scale studies leveraging millions of patients is limited by suboptimal approaches for data sharing and interoperability. As a result, integration of valuable clinical and basic resources is problematic, preventing knowledge discovery and rapid translation of findings into clinical care. To understand the current data landscape in the sleep and circadian domains, the Sleep Research Society (SRS) and the Sleep Research Network (now a task force of the SRS) organized a workshop on informatics and data harmonization, presented at the World Sleep Congress 2019, in Vancouver, Canada. Experts in translational informatics gathered with sleep research experts to discuss opportunities and challenges

Submitted: 12 November, 2021; **Revised:** 21 December, 2021

in defining strategies for data harmonization. The goal of this workshop was to fuel discussion and foster innovative approaches for data integration and development of informatics infrastructure supporting multi-site collaboration. Key recommendations included collecting and storing findable, accessible, interoperable, and reusable data; identifying existing international cohorts and resources supporting research in sleep and circadian biology; and defining the most relevant sleep data elements and associated metadata that could be supported by early integration initiatives. This report introduces foundational concepts with the goal of facilitating engagement between the sleep/circadian and informatics communities and is a call to action for the implementation and adoption of data harmonization strategies in this domain.

Key words: sleep; circadian rhythm; informatics; harmonization; ontology

Introduction

Sleep and circadian factors influence health outcomes across the lifespan [1–8]. Obstructive sleep apnea (OSA), insomnia, and circadian rhythm disruptions are very common conditions believed to arise from complex genetic and environmental interactions [9–11], resulting in substantial biological and clinical heterogeneity. Therefore, it is important to identify subpopulations at highest risk if untreated, as well as to tailor treatments according to risk stratification. Fortunately, computer algorithms capable of self-learning and self-refinement (machine learning) are becoming widely available for this task. However, they require learning from "big data"—tens of thousands to millions of observations—to effectively characterize these heterogeneous patterns and impact clinical practice.

Such tasks present an urgent need for access to higher volume and higher quality data. These data are increasingly available due to the widespread adoption of the electronic health records (EHR) and the development of large research repositories designed for promoting data sharing from completed studies. Examples include field agnostic repositories such as the database of Genotypes and Phenotypes (dbGAP) [12], and field specific resources such as the National Sleep Research Resource (NSRR) [13,14], which focus on curating already collected research data. Other initiatives are focused on the aggregation of data from a large number of individuals such as the Million Veteran Program [15] and the National Institutes of Health All of Us Research Program [16]. Despite the increasing availability of such large datasets, substantial barriers still exist, particularly for streamlining access and use of additional data sources, such as EHRs or data curated from research studies not currently available in research repositories. This is particularly relevant in the sleep medicine field, as highlighted in a recent systematic review [17]. Barriers to integrating heterogeneous sources of clinical data for research include: (1) storage in data silos with lack of infrastructure for data sharing within and across institutions (e.g. lack of data standards, data models and limited governance); (2) lack of information about the methods and conditions surrounding collected data (e.g. absence of standardized metadata and data processing workflows); (3) lack of standardized terminologies and structured vocabularies (i.e. ontologies) discouraging pooling data from different studies; (4) time and expenses involved in data curation and harmonization; and lack of attribution or resources for engaging in such activities.

These issues not only impact how previously collected data are handled but also determine how easily data acquired in the future could be shared. The latter is especially important because most current data infrastructure is designed to satisfy the goals of the original research study for which it was collected. Given the time and expense associated with the collection of high-quality data, which is often obtained using public funds, this approach is no longer sustainable. In fact, the National Institute of Health now recommends routine language in consent forms for research studies, as well as new policies about submission of data management and sharing plans [18]. Researchers should be prepared to comply with these policies when they are in effect. Ultimately, due to the diversity of data types routinely used in sleep and circadian biology, such as high-resolution physiological signals, accelerometer-derived activity counts, wearable technology, multi-omics (e.g. genomics, metabolomics) and qualitative data about symptoms and quality of life, the interdisciplinary nature of the field serves as a prototype for other disciplines and could provide an important role model for multidimensional data integration in all areas of medicine.

As part of international efforts to address important data harmonization bottlenecks in our field, a one-day sleep and circadian data harmonization workshop organized by the Sleep Research Society and the Sleep Research Network Task Force was held on September 22, 2019 in conjunction with the World Sleep Meeting in Vancouver, BC. This white paper summarizes key concepts discussed and identifies actionable next steps for the sleep and circadian community.

Key Definitions and Framework for Optimal Data Sharing Data

Best practices for data sharing, barriers to implementing these practices, and potential approaches were outlined by our keynote speaker Dr Melissa Haendel. These included key definitions, infrastructure, and best practices for using findable, accessible, interoperable, and reusable (FAIR) data standards [19] and licensing issues. Dr Haendel outlined existing challenges in the field, including the best way to capture complex, clinically useful representations so they are understandable by both humans and computers. This provided the foundation for the work of the four breakout panel groups in questionnaires, actigraphy, polysomnography, and informatics infrastructure models that followed thereafter.

Key Definitions

Data standards

Developing consensus regarding standards is fundamental to data sharing. Data standards can be technical (e.g. how data should be represented and exchanged) and conceptual (e.g. how knowledge is represented from data). Standards might comprise sets of uniform rules for collecting and exchanging information which should include a descriptive name, abbreviation, common format, and ideally should be usable by both humans and computers [20]. Without strong technical data standards, information exchange becomes difficult or impossible, limiting data reusability. For example, polysomnography data collected using inappropriate sampling

rates and filters will not be amenable to many types of signal processing algorithms useful for characterizing features such as airflow limitation, heart rate variability, and sleep microarchitecture. Similarly, without consensus on methodological standards, integration of data from different studies might be limited or introduce bias. Importantly, data standards are strongly influenced by commercial entities and technology developers.

Metadata

Often described as "data about data," metadata can be considered a type of data standard specifying the minimum amount and type of information that needs to be reported to make a piece of data interoperable and reusable. For example, a study reporting an apnea-hypopnea index (AHI) that failed to specify the percentage of oxygen desaturation required for hypopneas and the version of sleep scoring rules used in the index calculation could not be combined with AHI data from other studies, resulting in interpretation problems when reporting association with clinical outcomes. Careful thought should be given to the long-term durability of metadata as we increasingly need to support not just immediate data sharing but the potential for reuse years in the future. Therefore, both data standards and metadata requirements should be reviewed and updated regularly to ensure that advances in technology are accurately reflected in the standards.

Clinical terminology

This refers to the words (terms) we use in daily language to describe complex concepts in medicine. Adhering to a controlled and standardized terminology is relatively easy and human friendly. However, it may be insufficiently precise for use in machine-readable calculations. Establishing consensus on rules for building terms that are meaningful and self-contained is a challenge, particularly in sleep medicine where data representations have more permutations and variability than classic epidemiological or clinical data. For example, the term "apnea-hypopnea index" is human-readable and interpretable, but its accurate use depends on specific criteria such as how hypopneas were scored (3% desaturation and arousal, 4% desaturation, or any other combination of levels of desaturation and arousals). In order to improve the adoption of terms and retain their consistent meaning across studies, sleep researchers must develop scalable and extendable rules and harmonized terms. Such rules and terms should incorporate specific criteria to improve their accuracy. Approaches that might facilitate long-term maintenance of standardized clinical terminologies also include version control, where updates are tracked, and the history of changes can be accurately traced, when necessary. This helps avoid loss of information as the terminology evolves. An example of a consensus-based approach designed to develop international clinical data standards for common cardiovascular conditions has been recently presented by the European Unified Registries for Heart Care Evaluation and Randomized Trials [21]. This framework could be used as a model to support the development of sustainable clinical terminologies in the sleep and circadian domains.

Ontology

Biomedical ontologies are controlled vocabularies that describe the meaning of biomedical data (i.e. semantics) in human and machine-readable ways. Ontologies are highly structured, often include mappings between different clinical terminology standards, and are resource intensive to develop and maintain. Ontologies describe relationships between terms in a logical way and are a fundamental component of biomedical knowledge representations. Preliminary efforts in the creation of a sleep domain ontology have been reported [22] which also has been used to inform the data structure within the NSRR [14].

Data harmonization

Data harmonization is the process of curating datasets for secondary use, including combining all or parts of a dataset with another. Data harmonization increases the likelihood that research questions requiring a large number of subjects, often larger than is practical for an individual study, can be investigated. Steps involved in data harmonization include (1) identifying the variables for harmonization, (2) assessing the completeness of the study level metadata (e.g. data collection methods, study population, etc.) and variable level metadata (e.g. label, descriptions, unit, version history), (3) developing the "target" harmonized terms and associated metadata for both concept/ontology and data format (e.g. unit, permissible values, etc.), (4) mapping the build variables needed to derive the "target" harmonized term in each dataset, (5) documenting all decisions made in establishing equivalency and deriving the harmonized variables. A priori adoption of technical and methodological standards, as well as the use of well-defined clinical terminologies facilitates data harmonization. Whenever possible, existing terminologies and ontologies should be used to avoid the proliferation of duplicative and overlapping efforts. The process of how the data is being harmonized should be documented and published, facilitating adoption by other users in their own datasets. One example is provided by the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program, that has reported their system for harmonization of a limited set of phenotype data across multiple studies that are part of the program, and published the whole reproduceable harmonization pipeline on a GitHub repository (https://github.com/UW-GAC/topmed-dcc-harmonized-phenotypes) [23].

Best Practices for Data Reusability and Sharing

To improve their value, data resources need to have a unique identifier which is stable over time (**F**indable); be understandable by both humans and computers using accepted conventions (**A**ccessible); use predefined terms, be stored or convertible into an exchangeable file format (**I**nteroperable); and contain enough information that an investigator can sufficiently trace the history of the data (provenance), be assured of the accuracy of the data, and be confident that they have the permission to use the data (**R**eusable). These four components make up FAIR data principles, the accepted gold standard for data sharing [19]. More recently, FAIR has been extended to highlight the importance of having sufficient provenance and a way to correctly cite or attribute the source of the data and any tools used to improve the availability of the data (**T**raceability); address the extent to which data is available for reuse and redistribution (**L**icensure); and identify how well data is interrelated, often seen as a surrogate for how easy data sets can be combined (**C**onnectedness)

[24]. A comprehensive introduction to generating and reusing persistent biomedical data identifiers has been previously reported [25].

One successful example of the application of FAIR in the field of sleep medicine is the NSRR (https://sleepdata.org; [13,14]), a central data repository focused on providing facilitated access to research studies in the sleep domain. NSRR provides researchers with access to the data by completing a simple data use agreement, as well as basic tools for understanding the data collected in each study. Data is organized using standardized terms along with information about how the data was collected by detailed protocols and data dictionaries (metadata). Additional tools designed to facilitate comparison of similar data elements between studies supporting cohort generation (cross-dataset mapping) are also provided (https://x-search.net/; [26]), as well as a matrix indicating data availability across cohorts (https://matrix.sleepdata.org/). The success of the NSRR has resulted in over 270 publications to date (for list of publications, see (https://sleepdata.org/pages/publications).

Another resource available from NHLBI is the BioData Catalyst (BDC) cloud environment (https://biodatacatalyst.nhlbi.nih.gov/) [27]. This platform was designed to facilitate adoption and application of bioinformatics pipelines to the analysis of data funded by the NHLBI. It leverages a flexible cloud-based infrastructure that allows approved users to access datasets and deploy computationally intensive workflows. Originally oriented to facilitate data reuse in the genomics domain, the NHLBI BioData Catalyst scope is expanding to other areas of interest and data types, including images and physiological data. The implementation of standardized workflows for the analysis of high-resolution physiological and activity-related data in sleep and circadian biology in platforms such as BioData Catalyst will democratize data reuse and improve return on research investment. Resources such as the NSRR and the NHLBI BioData Catalyst make it easier for current and future investigators to locate, access, combine, and reuse data from previous studies and therefore support data sharing best practices.

Common Sleep and Circadian Data Domains: Benefits, Barriers, and Next Steps

During the workshop, breakout groups were set up, focused on the three most common data domains in the field (questionnaires, actigraphy and polysomnography) with a fourth group focusing on infrastructure to support data harmonization and sharing. Discussions focused on benefits and barriers to data harmonization, common data sources and data types, as well as next steps to strengthen data sharing were discussed. The following section summarizes these discussions.

Questionnaires

Questionnaires are important components of sleep assessments. Discussion in this section covered self- and observer-reported data. The primary benefits of questionnaires include ease of use, relatively low cost to obtain information, availability of validated scales, and ability to identify issues important to patients; collecting patient-centered outcomes can be used to formulate patient-centered policy changes to interventions and policies. Both unstructured (e.g. single free text question, patient-reported information recorded in clinical notes) and structured questionnaires (e.g. Epworth Sleepiness Scale, Insomnia Severity Index, Pittsburgh Sleep Quality Index, Munich Chronotype Questionnaire) are amenable to the development of terminologies and ontologies, and can be widely incorporated into studies focused on sleep and circadian rhythms, aging and development, and a variety of health conditions.

The working group also identified barriers to harmonizing questionnaire information. There are multiple context-dependent ways to ask about an individual's "sleep quality" or "sleep duration." For example, "good sleep quality" may be perceived differently for young, working adults compared to older, retired adults. More specifically, the lack of context about whether the measure applies to either work or free days (or both) is limited. In addition, even when these questions are standardized, they are strongly influenced by social, ethnic, racial, and national differences. Current normative data, when available, often fail to adequately capture the influence of these factors. Furthermore, use of standardized and validated measures may be limited by access, licensing, fees, and copyright. Lastly, even when high quality data are available, substantial barriers to sharing still exist, and include some technological barriers outlined in this report, as well as lack of incentives for doing the often-time-consuming work associated with data sharing, even when infrastructure and desire to do so are present.

As part of the discussion, common data sources and data types were identified and are listed in **Table 1**. The working group focused on commonly used and available information sources, as well as the most typical data types collected from these sources, rather than a comprehensive list of all sleep and circadian self-reported instruments. Notably, the comparability of findings across studies has been complicated by widespread variability in the phenotypic assessment of sleep variables. For instance, assessments range from single items to comprehensive diagnostic interviews such as the Diagnostic Interview for Sleep Patterns and Disorders, a computerized interview for both sleep patterns and screening for a wide range of sleep disorders in the NIH Toolkit, as well as the more recent Structured Clinical Interview for DSM-5 Sleep Disorders Module [28].

Actigraphy

Actigraphy is an established and non-invasive method to record objective rest activity patterns over several days. Several algorithms have been established to detect sleep and wakefulness using actigraphy data, which has then been compared to polysomnography and self-report measurements under different conditions and patient populations. Actigraphy measures changes in acceleration, which are then converted into activity counts used to distinguish rest (presumed sleep) from active (presumed wake) states. The low subject burden and ability to use over multiple nights have contributed to its widespread use for estimating sleep duration, timing, fragmentation, and several circadian traits of interest.

The major factors driving interest in harmonizing actigraphy data in sleep and circadian research is the increasing availability of easily sharable open-source devices and software, as well as the popularity of consumer-oriented wearable devices used to estimate sleep and physical activity. The landscape of wearable technologies to characterize sleep biomarkers have been

Table 1. Common data sources and terms associated with questionnaire-based sleep and circadian assessments

	Types	Examples
Common data sources	Validated questionnaires	PSQI, ESS, MCTQ
	Specific questions	Do you snore?
	Sleep diaries	Diaries for sleep–wake pattern assessments or for informing actigraphy studies
	Interviews	Structured and unstructured
	Observer reports	Parent reports
	Clinical notes	Notes in electronic health records
	Ecological momentary assessments	Social network data
	Passive data sensing	Smart Speakers
Common data terms	Sleep quality and restfulness	Good sleep quality, feel rested
	Daytime symptoms and consequences	Excessively sleepy
	Sleep timing and chronotype	Morning or evening types
	Self-reported quantitative data	Sleep latency, sleep duration, time spent awake
	Sleep disorder related symptoms	Snoring, witnessed apneas, restless legs, sleep-walking
	Medication or substance use	Sleep aids, melatonin
	Contextual variables	Sleep environment, social and physical factors, use of electronics, consumption of caffeine
	Attitudes and beliefs about sleep	Cultural factors (e.g. “siesta”)

PSQI, Pittsburgh Sleep Quality Index; ESS, Epworth sleepiness scale; MCTQ, Munich ChronoType Questionnaire.

extensively discussed in a previous publication [29]. In distinction to conventional actigraphy, where proprietary algorithms are the standard, open-source frameworks generate activity data from raw accelerometer recordings and support the development of device-agnostic tools. Consequently, data generation is more transparent, prompting the need for harmonization frameworks that allow data from different studies to be integrated in a meaningful way.

Examples of current open-source frameworks include the use of the Axivity device in the UK Biobank that allowed the generation of detailed sleep and circadian information for ~100,000 participants [30], and large cohort studies in Brazil [31] and Spain [32] using Actigraph devices, which were analyzed using the software package GGIR [33]. An exemplary effort for data harmonization of actigraphy data on mood disorders is being developed by the Motor Activity Research Consortium for Health, a collaborative research network from the NIMH and Johns Hopkins Bloomberg School of Public Health that has established common procedures, analyses, and data sharing among multiple international research groups collecting actigraphy and ancillary data on mood disorders (ZIA MH002954-04 Motor Activity Research Consortium for Health—mMARCH). The consortium has identified a common set of measures of clinical state and context for the collection of actigraphy, dissemination of methods to address analytic challenges procedures for defining valid observations, handling of missing data, and analytic tools that address limitations of parametric methods using a family of functional data analytic methods [34]. Through the mMARCH network, Guo et al have established an open-source post GGIR pipeline for processing actigraphy data (https://github.com/dora201888/postGGIR) that facilitates visualization of summary output from GGIR, streamlines the data processing pipeline, and extracts features from sleep, physical activity, and circadian rhythmicity, the three key domains of behavior measurable by wearable accelerometers.

One of the most important lessons from cross study integration efforts is the importance of collecting ancillary diary information, as recommended in guidelines from the American Academy of Sleep Medicine, to provide more valid estimates of actigraphy-derived sleep parameters [35]. Studies that combine actigraphy with real time direct measures of other clinical and contextual states have yielded important insights into the directional links between the core domains extracted from actigraphy including sleep, physical activity and circadian parameters [36]. These efforts will be increasingly relevant to allow integration of large data collections from epidemiological studies, requiring common processing derived from different devices and algorithms. This integration becomes a challenge when proprietary algorithms are combined with open-source algorithms without an established and validated harmonization framework. For example, technological improvement of accelerometer sensors, algorithms and methods over time might limit the integration of data between older and newer devices. In addition, consumer-based wearable technologies are becoming one of the primary sources of sleep data in the general population, but access to raw accelerometer data and implementation of harmonized data processing workflows are not available, due to proprietary algorithms to generate activity counts. Nevertheless, efforts to integrate wearable data into EHR are underway, and if performed under FAIR principles, they could revolutionize how sleep and rest-activity pattern traits are assessed as part of the regular clinical care.

Several common data sources and types were highlighted during the discussion about actigraphy as presented in **Table 2**.

Polysomnography

Polysomnography (PSG) is the gold-standard method for characterizing electrophysiological aspects of sleep. This method relies on high-resolution recording of several physiological signals such as electroencephalogram (EEG), electromyogram (EMG) and electrooculogram (EOG), combined with other sensors to measure respiration, oxygen saturation, thoracic and abdominal effort, heart rate and limb movements. Rich physiological information during sleep is available from the PSG recording, provides the most comprehensive data relevant for the differential diagnosis of several sleep disorders. After digital recording

Table 2. Common data sources and terms associated with actigraphy-based sleep and circadian assessments

	Types	Examples
Common data sources	Actigraphy and tri-axial accelerometer devices	Activity counts, timestamps, event logger, light exposure, skin temperature
	Wearable devices	Steps, estimates of sleep duration, pulse rate
	Smartphones	Activity counts, GPS coordinates
Common data terms	Sleep-related	Sleep duration, onset and offset times, diurnal sleep, stages
	Circadian rhythm-related	Sleep and activity timing, sleep midpoint, amplitude, intradaily variability, interdaily stability
	Physical activity levels	Sedentary, moderate and vigorous physical activity
	Integrated measures	Social jet-leg

became available, specific data representation, technical standards and recording protocols were established, placing PSG ahead of the curve in terms of data harmonization.

Several benefits regarding harmonization of polysomnographic data have been identified during panel breakout sessions. A common data exchange format (i.e. a technical standard) for biological signals generated during polysomnography, the European Data Format (EDF), already exists. It consists of a standard representation of biological signals and a header containing technical specifications of the study (including sampling rates and filters), subject identifier, and the number and time duration of the data records that follow the header. This format tends to be vendor- and platform-agnostic, which facilitates access to raw data and can support the development of new algorithms. Moreover, while the level of physiological information available in a sleep study is enormous, analytical pipelines leveraging the full potential of these biological time series have been only partially explored, making it an area of future opportunities. Advances in signal processing methods and widespread availability of machine learning algorithms are expected to contribute to the development of this field.

However, EDF files exported by different systems may variably display data, and data recorded with slightly different specifications (e.g. sampling rate, analogic band-pass filters) present challenges in combining data as well as consistently using signal processing algorithms. Channel name conventions are not standard, and data processing is often complicated, particularly for large cohort studies and clinical samples. This technical heterogeneity is further complicated by variability in annotation formats and terminology to represent events (e.g. arousals, apneas, arrhythmias). Further difficulties relate to reaching community consensus on defining certain respiratory events such as hypopneas (e.g. 4% vs. 3% associated with arousal) and their relationship with medical reimbursement patterns, particularly in the United States. As a result, integration of data recorded under different settings becomes burdensome and requires extended processing time. Furthermore, standardized processing pipelines do not yet exist for generating well-known biological markers of sleep (e.g. delta power) or for automated sleep scoring. Finally, there are challenges related to storing and transferring high resolution physiological signal data, particularly for datasets containing more than a few thousand subjects (representing terabytes of data). In sum, all these issues limit innovation in the field and substantially lengthening the time between novel algorithm validation and clinical implementation.

Fortunately, there is a growing interest in the community on filling the gap between physiological biomarker identification and clinical utilization. As noted for actigraphy data, the development of open-source tools and availability of repositories hosting polysomnographic data such as the NSRR haves become more common. Two relevant use cases were recently demonstrated by Purcell et al. [37] and Djonlagic et al. [38]. Robust data processing pipelines were applied to polysomnographic data from several different cohort studies hosted at NSRR to characterize sleep spindles [37], which were then related to cognition in older adults [38]. Open-source tools, such as the *luna* software package (http://zzz.bwh.harvard.edu/luna/), have been fundamental to allow large scale processing and integration of signal data across heterogeneous studies. For example, *luna* includes a specific function that harmonizes EDF channel names and creates a canonical set of signals facilitating downstream signal processing. Future development of end-to-end analytical workflows and standardized terminologies to represent sleep-related events are ongoing. Finally, continuing development of wearable sensors that capture the same source of signals as polysomnography but in more ecologically valid settings (e.g. at home) could present opportunities to allow longer recording periods, instead of only being limited to one single night.

Common terms reported in studies using polysomnography are reported in **Table 3**. A comprehensive review of conventional and novel metrics relevant to obstructive sleep apnea have been published elsewhere [10].

Informatics Infrastructure Models

Sleep and circadian biology have substantial data harmonization needs, in part due to the diversity of data types and the volume of data generated. Fortunately, numerous efforts are underway to meet this challenge (for review, see [17]). Many efforts to date have focused on harmonizing research data already collected as described above (e.g. NSRR). However, this represents only one component of the entire data life ecosystem. An important yet underexplored source of sleep and circadian data is the EHR. Cohorts of patients with insomnia have been identified by extracting physician-reported insomnia from clinical notes available in the EHR [39]. In addition, validation of EHR-based algorithms to identify patients with obstructive sleep apnea have been conducted [40]. Efforts like these can support large outcome-based research or genetic association studies using large clinical cohorts with available EHR data. An example was reported in a recent phenome-wide association study (PheWAS), where candidate genetic variants previously associated with obstructive sleep apnea were assessed regarding their associations with other comorbidities identified using the EHR [41].

Table 3. Common data sources and terms associated with polysomnography sleep assessments

	Types	Examples
Common data sources	Electroencephalogram	In-lab/ home sleep studies
	Electromyogram	In-lab/ home sleep studies
	Electrocardiogram/heart rate	In-lab/ home sleep studies, wearables
	Pulse oximetry/pulse rate	In-lab/ home sleep studies, wearables
	Abdominal and thoracic plethysmography	In-lab/ home sleep studies
	Respiratory signals	In-lab/ home sleep studies
Common data terms	Sleep macro-architecture	Total sleep time, sleep stages, arousals
	Sleep micro-architecture	Power spectral analysis, Spindle characteristics
	Respiratory parameters	Sleep-disordered breathing, oxygen desaturations
	Cardiac parameters	Heart rate variability, cardiopulmonary coupling
	Limb movements	Periodic limb movements

While these initial efforts explored the capabilities of leveraging EHR data to support sleep research, they do not address the accuracy of sleep data in the EHR. It is unlikely that adequately structured fields representing data from validated questionnaires, actigraphy and polysomnography are available in EHR systems across the United States and the world. Some early adopters are incorporating electronic data capture as part of clinical workflows with the deployment of structured clinical documentation support toolkits in sleep medicine departments [42]. Other relevant sources of data include wearable devices and continuous positive airway pressure compliance systems. Technical and legal challenges exist when attempting to incorporate such data into EHRs, likely due to lack of consistent standardization. Finally, a more recent challenge identified in the discussion groups relates to integrating data from several independent health systems, including the adoption of common data models such as those governed by the National Patient-Centered Clinical Research Network (PCORnet) [43] and the Observational Health Data Sciences and Informatics (OHDSI) [44]. The lack of standard terminologies in sleep and circadian domains requires additional efforts to adequately represent EHR data for these domains. Early work attempting to map polysomnography results into controlled terminologies such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), which would allow data to be presented in the OHDSI Observational Medical Outcomes Partnership data model, has been recently reported [45].

Looking forward, extending infrastructure to support future data collection within both the EHR (e.g. point of care) and within research studies should be a priority. Examples of work in these areas include ongoing efforts to extra system-wide data with the Veterans Health Administration EHR [46,47]. Ultimately, it will be critical for the sleep and circadian biology communities to work together with health informatics-oriented researchers to ensure that structured language that is both human and machine readable can be applied as part of a learning health system.

Currently Available Resources

One major goal of the workshop was to outline the current landscape of resources available to the sleep and circadian communities, setting the ground for future research. We present a non-exhaustive list of current knowledge and data resources related to human sleep and circadian biology on **Table 4** and **Table 5**, respectively.

Challenges, Opportunities, and Future Directions

The sleep and circadian biology community can take several key steps to facilitate harmonization and adoption of standardized practices for both research and clinical data. We need to establish a process to facilitate the acquisition of standardized data and agree, as a community, on a core set of data for clinical use, including technical and methodological specifications and detailed metadata. In addition, we should encourage researchers to use well documented and open data dictionaries, ideally mapped to controlled clinical terminologies. Moreover, such clinical terminologies should be improved and maintained, so that high-quality sleep and circadian data can be well represented both in clinical and research contexts. These might improve representation of data in unstructured formats such as clinical notes or consult transcripts. Long intervals between the collection of the diagnostic/clinical and actigraphy/polysomnography data have also limited the interpretation of large-scale registry data—for instance, interpreting associations between sleep and circadian measures with clinical phenomena. For example, actigraphy data from the UK Biobank were collected several years after the main clinical assessment of the sample [48]. Clear definitions of the timing, missing data, and selection of subsamples for sleep or actigraphy assessments should also be provided. We encourage vendors of sleep technologies to establish transparent protocols for data representation, processing and sharing, including access to raw signal data to allow effective validation against other methods. We also encourage tool developers to provide open source "research use only" versions of their algorithms, which could then be assessed and validated in larger datasets. The sleep research and clinical communities need to create frameworks to facilitate incorporation of new types of sleep data into clinical practice and EHR, such as continuous positive airway pressure adherence and wearables data. We encourage national and international societies to provide guidance and education to the community regarding data sharing and associated protocols, aligning with expectations from Federal agencies. We also encourage funding agencies to support technological development of standardized process to data sharing and harmonization, including incentives and objective evaluation of data sharing quality. By developing, validating, and adopting standards that are easy to implement (i.e. with reduced technical barriers), we expect that more high-quality shared data could be used, therefore expanding the scientific applicability beyond an individual study and making

Table 4. Non-exhaustive list of knowledge integration resources with potential or direct application do sleep medicine and circadian biology

Name	Access
BioLINCC	https://biolincc.nhlbi.nih.gov/home/
Biomedical Data – Translator	https://ncats.nih.gov/translator
Center for Data 2 Health	https://cd2h.org/
NIH CDE Catalogue	https://www.nlm.nih.gov/cde/index.html
NIH Data Commons	https://commonfund.nih.gov/commons
BioData Catalyst	https://biodatacatalyst.nhlbi.nih.gov/
PhenX	https://www.phenx.org/
PROMIS	https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis
TOPMed	https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program
Sleep Disorder Knowledge Portal	https://sleep.hugeamp.org/

Table 5. Non-exhaustive list of data resources and projects that could support comprehensive collection or integration of sleep related data

Name	More Information
National Sleep Research Resource	https://sleepdata.org/
dbGaP	https://www.ncbi.nlm.nih.gov/gap/
Physionet	https://physionet.org/
National COVID Cohort Collaborative	https://ncats.nih.gov/n3c
All of Us	https://allofus.nih.gov/
Million Veteran Program	https://www.research.va.gov/mvp/
Canadian Sleep & Circadian Network	https://www.cscnweb.ca/
Sleep Apnea Global Interdisciplinary Consortium	https://www.med.upenn.edu/sleepctr/sagic.html
HypnoLaus	https://www.colaus-psycolaus.ch/professionals/hypnolaus/
RAINES	https://rainestudy.org.au/
UK Biobank	https://www.ukbiobank.ac.uk/
ESADA	https://esrs.eu/research-networks/sleep-apnea-network-european-sleep-apnea-database-esada/
NHANES	https://www.cdc.gov/nchs/nhanes/index.htm

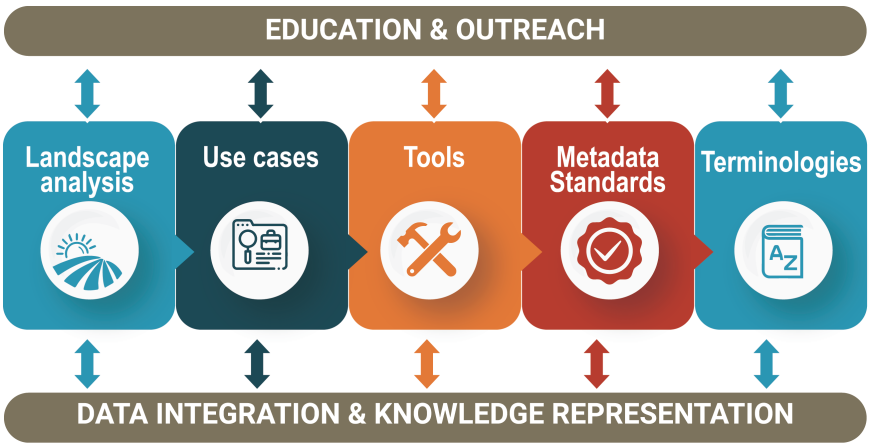

Figure 1. The sleep-circadian data ecosystem. Physician scientists, researchers, and health informaticists must join their knowledge skills to achieve our goals of facilitating harmonization and adoption of standardized practices for improvement of large-scale clinical research in sleep medicine and circadian biology. This is represented by the need to understand the current status of the field (landscape analysis) and develop tools, standards and terminologies to address key questions in the field by leveraging specific use cases. Finally, the sleep research community requires additional training on the importance of clinical research informatics methods, with the ultimate goal of adopting them to support large-scale clinical research.

data more equitable and accessible across the board. A roadmap summarizing key next steps is outlined in Figure 1. Finally, a more specific set of suggested recommendations identified by the Sleep Research Society and Sleep Research Network is also outlined below.

- Identify basic and machine-readable data elements, predefined terms and standards, that can be used by researchers;
- Systematic implementation of standardized devices that measure minimum set of essential sleep and circadian rhythm data in interchangeable electronic format in research studies or in Electronic Health Records;
- NIH and other stakeholders such as the Sleep Research Society (SRS), the Society for Biological Rhythms (SRBR) and other professional medical societies such as the American Academy of Sleep Medicine (AASM), and American Thoracic

Society (ATS), should work together to develop a set of standards for research and clinical data (terminology, data format, and quality metrics);
- Recommend the development and use of standardized common data elements (CDEs) of high quality for inclusion in research and medical data repositories;
- Improve metadata by including sufficient information about each collected CDE in a data library, including a level of detail that will enable the traceability and verifiability of the data. Templates can be developed for researchers to identify the most appropriate set of metadata standards for their data collection and sharing plan;
- Enhance infrastructure for confident data governance and stewardship for use by investigators;
- Collaborate with large public and private healthcare systems and industry to develop best practices within these systems for prospective data collection and incentivize data sharing;
- Require time stamping of behavioral, environmental and biological data;
- Leveraging existing large health systems and their infrastructures and collaborations with industry to share their clinical trials data can help sustain discovery in sleep and circadian medicine and science.

Funding

This work was partially supported by a grant from the American Heart Association (20CDA35310360) and National Institute of Health NIH R35 HL135818 and NIH Contract 75N92019C00011.

Disclosure Statement

None declared.

Acknowledgments

The authors would like to acknowledge all the participants of the Sleep-Circadian Informatics Data Harmonization Workshop, presented by the Sleep Research Society and Sleep Research Network on September 22nd 2019 during the World Sleep Congress 2019 in Vancouver Canada. The Sleep Research Network Task Force is composed by the following members: Nalaka Gooneratne (Chair), Eilis Boudreau, Daniel Buysse, Clete Kushida, Rachel Manber, Diego Mazzotti, Reena Mehra, Lisa Meltzer, Janet Mullington, Sairam Parthasarathy, Shaun Purcell, Susan Redline and John Noel.

References

1. Yong LC, *et al*. Sleep-related problems in the US working population: prevalence and association with shiftwork status. *Occup Environ Med*. 2017;**74**(2):93–104.
2. Kuehn BM. Sleep duration linked to cardiovascular disease. *Circulation*. 2019;**139**(21):2483–2484.
3. Hombali A, *et al*. Prevalence and correlates of sleep disorder symptoms in psychiatric disorders. *Psychiatry Res*. 2019;**279**:116–122.
4. Uehli K, *et al*. Sleep problems and work injuries: a systematic review and meta-analysis. *Sleep Med Rev*. 2014;**18**(1):61–73.
5. Gottlieb DJ, *et al*. Sleep deficiency and motor vehicle crash risk in the general population: a prospective cohort study. *BMC Med*. 2018;**16**(1):44.
6. Youngstedt SD, *et al*. Has adult sleep duration declined over the last 50+ years? *Sleep Med Rev*. 2016;**28**:69–85.
7. Stranges S, *et al*. Sleep problems: an emerging global epidemic? Findings from the INDEPTH WHO-SAGE study among more than 40,000 older adults from 8 countries across Africa and Asia. *Sleep*. 2012;**35**(8):1173–1181
8. Shen X, *et al*. Nighttime sleep duration, 24-hour sleep duration and risk of all-cause mortality among adults: a meta-analysis of prospective cohort studies. *Sci Rep*. 2016;**6**(1):21480.
9. Veatch OJ, *et al*. Pleiotropic genetic effects influencing sleep and neurological disorders. *Lancet Neurol*. 2017;**16**(2):158–170.
10. Mazzotti DR, *et al*. Opportunities for utilizing polysomnography signals to characterize obstructive sleep apnea subtypes and severity. *Physiol Meas*. 2018;**39**(9):09TR–0901.
11. Jansen PR, *et al*. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet*. 2019;**51**(3):394–403.
12. Mailman MD, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;**39**(10):1181–1186.
13. Dean DA, *et al*. Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource. *Sleep*. 2016;**39**(5):1151–1164.
14. Zhang GQ, *et al*. The National sleep research Resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;**25**(10):1351–1358.
15. Gaziano JM, *et al*. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;**70**:214–223.
16. Denny JC, *et al*.; All of Us Research Program I. The "All of Us" research program. *N Engl J Med*. 2019;**381**(7):668–676.
17. Mazzotti DR. Landscape of biomedical informatics standards and terminologies for clinical sleep medicine research: a systematic review. *Sleep Med Rev*. 2021;**60**:101529.
18. NIH Office of Science Policy. Final NIH Policy for Data Management and Sharing. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html. Accessed October 29, 2021.
19. Wilkinson MD, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;**3**:160018.
20. United States Geological Survey. Data Standards. https://www.usgs.gov/products/data-and-tools/data-management/data-standards#examples. Accessed August 9, 2021, 2020.
21. Batra G, *et al*. Methodology for the development of international clinical data standards for common cardiovascular conditions: European unified registries for heart care evaluation and randomised trials (EuroHeart). *Eur Heart J*. 2021:qcab052.
22. Arabandi S. Sleep Domain Ontology. https://bioportal.bioontology.org/ontologies/SDO. Accessed December 14, 2020.
23. Stilp AM, *et al*. A System for phenotype harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. *Am J Epidemiol*. 2021;**190**(10):1977–1992.
24. Melissa H, Andrew S, Julie M, *et al*. FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. 10.5281/zenodo.203295. Accessed December 15, 2020.
25. McMurry JA, *et al*. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to

maximize utility and impact of life science data. *PLoS Biol*. 2017;**15**(6):e2001414.
26. Cui L, *et al*. X-search: an open access interface for cross-cohort exploration of the National Sleep Research Resource. *BMC Med Inform Decis Mak*. 2018;**18**(1):99.
27. BioData Catalyst Consortium. The NHLBI BioData catalyst. http://doi.org/10.5281/zenodo.3822858. Accessed August 31, 2020.
28. Taylor DJ, *et al*. Reliability of the structured clinical interview for DSM-5 sleep disorders module. *J Clin Sleep Med*. 2018;**14**(03):459–464.
29. Depner CM, *et al*. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep* 2019;**43**(2):zsz254.
30. Jones SE, *et al*. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. *Nat Commun*. 2019;**10**(1):1585.
31. Wendt A, *et al*. Sleep parameters measured by accelerometry: descriptive analyses from the 22-year follow-up of the Pelotas 1993 birth cohort. *Sleep Med*. 2020;**67**:83–90.
32. Cabanas-Sanchez V, *et al*. Twenty four-hour activity cycle in older adults using wrist-worn accelerometers: the seniors-ENRICA-2 study. *Scand J Med Sci Sports*. 2020;**30**(4):700–708.
33. Migueles JH, *et al*. GGIR: a research community–driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *J Meas Phys Behav* 2019;**2**(3):188–196.
34. Murray G, *et al*. Measuring circadian function in bipolar disorders: empirical and conceptual review of physiological, actigraphic, and self-report approaches. *Bipolar Disord*. 2020;**22**(7):693–710.
35. 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 Clinical Practice Guideline. *J Clin Sleep Med*. 2018;**14**(07):1231–1237.
36. Merikangas KR, *et al*. Real-time mobile monitoring of the dynamic associations among motor activity, energy, mood, and sleep in adults with bipolar disorder. *JAMA Psychiatry*. 2019;**76**(2):190–198.
37. Purcell SM, *et al*. Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. *Nat Commun*. 2017;**8**(1):15930.
38. Djonlagic I, *et al*. Macro and micro sleep architecture and cognitive performance in older adults. *Nat Hum Behav*. 2020:123–145.
39. Kartoun U, *et al*. Development of an algorithm to identify patients with physician-documented insomnia. *Sci Rep*. 2018;**8**(1):7862.
40. Keenan BT, *et al*. Multisite validation of a simple electronic health record algorithm for identifying diagnosed obstructive sleep apnea. *J Clin Sleep Med*. 2020;**16**(2):175–183.
41. Veatch OJ, *et al*. Characterization of genetic and phenotypic heterogeneity of obstructive sleep apnea using electronic health records. *BMC Med Genomics*. 2020;**13**(1):105.
42. Maraganore DM, *et al*. Quality improvement and practice-based research in sleep medicine using structured clinical documentation in the electronic medical record. *Sleep Sci Practice* 2020;**4**(1).
43. Fleurence RL, *et al*. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;**21**(4):578–582.
44. Hripcsak G, *et al*. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;**216**:574–578.
45. Kim JW, *et al*. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's Common Data Model: a pilot feasibility study. *Sci Rep*. 2021;**11**(1):7013.
46. Sarmiento KF, *et al*. Effects of computer-based documentation procedures on health care workload assessment and resource allocation: an example from VA sleep medicine programs. *Fed Pract*. 2020;**37**(8):368–374.
47. Sarmiento KF, *et al*. National expansion of sleep telemedicine for veterans: the TeleSleep program. *J Clin Sleep Med*. 2019;**15**(9):1355–1364.
48. Lyall LM, *et al*. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. *Lancet Psychiatry*. 2018;**5**(6):507–514.