Nature and Science of Sleep

Dovepress
open access to scientific and medical research

Open Access Full Text Article

EDITORIAL

Publicly Available Health Research Datasets: Opportunities and Responsibilities

Ahmed S BaHammam [1], Michael WL Chee [2]

[1]Department of Medicine, University Sleep Disorders Center and Pulmonary Service, King Saud University, Riyadh, Saudi Arabia; [2]Centre for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Correspondence: Ahmed S BaHammam, University Sleep Disorders Center, Department of Medicine, College of Medicine, King Saud University, Box 225503, Riyadh, 11324, Saudi Arabia, Tel +966-11-467-9495, Fax +966-11-467-9179, Email ashammam2@gmail.com

Nature and Science of Sleep downloaded from https://www.dovepress.com/
For personal use only.

> We are moving slowly into an era where big data is the starting point, not the end.
>
> – Pearl Zhu, author of the "Digital Master".

The drive to share scientific data related to human health began with the Open Science movement in the 1990s and has recently been supported by governments; the move aims to advance science and scientific communication and transform contemporary culture and decision-making. Also, data sharing serves to accelerate discovery, offering accessibility to high-value large, complex datasets that many researchers cannot readily collect. Undoubtedly, this international effort represents one of the most significant developments in evidence-based practices this century.[1] Publicly available data have created a fundamental and significant change in how we conduct research, make decisions, develop policy, and assess our actions. It is proposed that several tasks, including disease surveillance and signal identification, risk prediction, therapeutic intervention targeting, and last but not least, disease comprehension, may be accomplished with big data.[2] Another advantage of publicly available datasets is the ability to utilize, share, and integrate these data with other datasets, which opens up new avenues for scientific interaction and cooperation. The reuse and scrutiny of collected data by parties other than the original researchers enhance transparency and rigor in the documentation and conduct of data collection. Reanalysis of original or aggregated datasets ensures the reproducibility and robustness of inferences made.

In addition, health-related research, including sleep medicine research, is currently moving towards individual-tailored therapeutic interventions, and the big data era.[3,4] Sleep medicine-related research is suitable for large digital data due to the nature of data recording and collection; for example, polysomnography (PSG) provides several physiological data that aid in clinical research and therapeutic decision-making. Moreover, wearable devices and self-quantification systems are other sources of large data. In addition, a growing number of large datasets pertaining to sleep are publicly accessible for analysis to researchers worldwide, such as PhysioNet, which provides large libraries of recorded physiologic signals that are available for free on the web, and the "Montreal Archive of Sleep Studies" (MASS).[3] Besides, the "National Sleep Research Resource" (NSRR) is a new "National Heart, Lung, and Blood Institute" ad hoc site created to give the community of sleep researchers access to colossal data.[5,6] The NSRR is a system for exchanging and reusing large-scale physiological signals developed with a single point of access (NSRR; R24HL114473). De-identified data for more than 35,000 patients (at the time of writing) related to sleep medicine from 22 US-based datasets, including PSGs and connections to risk factors and outcome data for research participants, are available on the NSRR's free and public web platform.[7] The functional architecture was created and implemented to allow ongoing data sharing and integration. It has been effectively adapted and enhanced to allow the collection of prospective data for epidemiological cohort studies supported by the "National Institutes of Health (NIH)", such as the "Sleep Heart Health Study", the "MrOS dataset" with its primary focus being PSG data, "Heart Biomarker Evaluation in Apnea Treatment", and

Received: 18 September 2022
Accepted: 20 September 2022
Published: 27 September 2022

© 2022 BaHammam and Chee. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (http://creativecommons.org/licenses/by-nc/3.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

the "Multi-Ethnic Study of Atherosclerosis (MESA Sleep)", which included a sleep questionnaire, actigraphy, and a full overnight unattended PSG, and to combine data from independent research groups, such as the "Wisconsin Sleep Cohort".[8] The UK Biobank (data on half a million participants) and the "Adolescent Brain Cognitive Development (ABCD)" studies are two prominent examples of government-funded programs developing an interest in sleep.[9,10] In addition, scientist-led data-collecting consortia-like ENIGMA (50 active ENIGMA working groups) have also recently developed sleep sections.[11]

The availability of such colossal health data, once only available to the country or province of high-end well-funded laboratories, has stimulated a surge in work on automated sleep staging systems and opened new vistas for the early detection of sleep-disordered breathing disorders and other sleep disorders.

Although open data provide golden opportunities for scientists and researchers, they also have shortcomings that need to be discussed. For example, despite all the unprecedented advantages of publicly available datasets, there have been reports of low-quality association studies with limited clinical utility that utilized public database data.[12,13] In addition, the machine learning community has recently discovered a concerning number of potential ethical and legal issues with many of the most widely used picture datasets, including representational harms, bias effects, invasions of privacy, and ambiguous or questionable downstream uses.[14,15] Moreover, there are some public concerns about preserving privacy and confidentiality,[16,17] particularly with healthcare and public health databases. However, there is an understanding among the research community that the desire for openness and transparency must be balanced with the requirement to protect confidentiality. Additionally, because few datasets have high visibility, easy access, and usability, there is a risk that researchers may choose a tiny, biased pool of data, which could result in significant biases.

Therefore, publicly available data need precise standards to assure transparency about the source, how the data were developed, the credibility of the collected data, proper analysis and its potential combinability with other datasets, and their weaknesses.

Due to the intrinsic nature of the secondary analysis, the available data are not gathered to answer a specific research question or to test a specific hypothesis. As a result, some significant variables frequently are not available for analysis. Similar to this, not all population groupings or geographic areas of interest may have their data collected.

Moreover, the fact that the researchers evaluating the data are frequently different from those who were part of the data collection process is another significant constraint of data analysis. As a result, they are probably unaware of details or flaws unique to the study that may affect how certain the dataset's variables are interpreted. In addition, users may overlook crucial information if it is not prominently shown in the documents since there is sometimes an overwhelming amount of material (primarily designed for sophisticated, extensive surveys conducted by government bodies).

However, in *Nature and Science of Sleep*, we recognize the importance of publicly available datasets in advancing public health and sleep medicine research. Therefore, we proposed guidelines for authors submitting studies utilizing publicly available databases in the Scope and Aims of the journal to ensure that valuable and high-quality scientific work derived from publicly available databases reaches the readers of *Nature and Science of Sleep* and the sleep medicine community.

Nature and Science of Sleep aims to ensure that the used data should meet the values recognized as pertinent to big data in health and research on a substantive and procedural level, with accepted definitions utilizing an ethical decision-making framework.[18,19] Data analysis must meet certain requirements, which should include a thorough description of the population being studied, a sample plan and strategy, a time period for data collection, assessment tools, response rates, and quality control procedures; additionally, it should be specified if the data analysis follows a "question-driven" or "data-driven" approach and how missing variable were managed. While data imputation techniques can help cover gaps from missing data, variables that are key to testing focused hypotheses may not have been collected.[20] Further, the statistical power in large numbers cannot make up for gaps in representation (for example, of racial groups and social status) that limit the generalizability of the data.

Moreover, dataset's strengths and limitations should be clearly described to the readers. This task is mutual between the curators and the researchers. On the one hand, curators need to be open about the limitations or flaws in the collected data, some of which get uncovered by users or following updates in technology or standards. On the other hand, researchers need to thoroughly study all essential documentation offered to database users to

evaluate the data's internal and external validity and decide whether there are enough cases in the dataset to produce accurate estimates about the topic of interest. Moreover, creators need to oversee the usage of their datasets, make license and documentation adjustments as needed and, if required, restrict access.[21] Additionally, the researchers should legibly define the exposure variables, outcome variables, covariates, and confounding factors that will be considered in the analysis before beginning the investigation.

A significant shift in dataset approach is required going forward[14,21] since the ethical implications of a dataset are challenging to predict and handle at the time. Peng et al underlined the need for damage reduction and stewardship throughout the dataset's life cycle of dataset construction, as well as moral and societal standards that may evolve over time.

Nature and Science of Sleep shall continue to observe the evolving community standards closely and encourage authors to submit high-quality research papers generated using publicly available research datasets. Furthermore, *Nature and Science of Sleep* will closely monitor for any articles that employ retracted datasets and pay close attention to data citation and availability statements.

Disclosure

Ahmed S BaHammam is the Editor-in-Chief for *Nature and Science of Sleep*. The authors have no other conflicts of interest to declare for this work.

References

1. Huston P, Edge VL, Bernier E. Reaping the benefits of Open Data in public health. *Can Commun Dis Rep*. 2019;45(11):252–256. doi:10.14745/ccdr.v45i10a01
2. Dolley S. Big data's role in precision public health. *Front Public Health*. 2018;6:68. doi:10.3389/fpubh.2018.00068
3. Bragazzi NL, Guglielmi O, Garbarino S. SleepOMICS: how big data can revolutionize sleep science. *Int J Environ Res Public Health*. 2019;16(2):291. doi:10.3390/ijerph16020291
4. Holst SC, Valomon A, Landolt HP. Sleep pharmacogenetics: personalized sleep-wake therapy. *Annu Rev Pharmacol Toxicol*. 2016;56:577–603. doi:10.1146/annurev-pharmtox-010715-103801
5. Dean DA, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*. 2016;39(5):1151–1164. doi:10.5665/sleep.5774
6. Imtiaz SA, Rodriguez-Villegas E. An open-source toolbox for standardized use of PhysioNet sleep EDF expanded database. *Annu Int Conf IEEE Eng Med Biol Soc*. 2015;2015:6014–6017. doi:10.1109/EMBC.2015.7319762
7. National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). National sleep research resource; 2022. Available from: https://sleepdata.org/. Accessed September 11, 2022.
8. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;25(10):1351–1358. doi:10.1093/jamia/ocy064
9. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779
10. Adolescent Brain Cognitive Development (ABCD) Study. About the study. ABCD study; 2022. Available from: https://abcdstudy.org/about/. Accessed September 16, 2022.
11. The ENIGMA Consortium. Enhancing neuro imaging genetics through meta-analysis; 2022. Available from: https://enigma.ini.usc.edu/. Accessed September 16, 2022.
12. Ledford H, Van Noorden R. High-profile coronavirus retractions raise concerns about data oversight. *Nature*. 2020;582(7811):160. doi:10.1038/d41586-020-01695-w
13. Salmi J. *Study on Open Science. Impact, Implications and Policy Options*. Brussels: EUROPEAN COMMISSION; 2015.
14. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. *Patterns*. 2021;2(11):100336. doi:10.1016/j.patter.2021.100336
15. Editorial Nat Mach Intell. The rise and fall (and rise) of datasets. *Nat Mach Intell*. 2022;1–2. doi:10.1038/s42256-42022-00442-42252
16. Goben A, Sandusky R. Open data repositories: current risks and opportunities. *Coll Res Libraries News*. 2020;81:62. doi:10.25860/crln.24281.24271.24262
17. Hand D. Open data is a force for good, but not without risks. The Guardian; 2012. Available from: https://www.theguardian.com/society/2012/jul/2010/open-data-force-for-good-risks. Accessed September 20, 2022.
18. Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. *Shanghai Arch Psychiatry*. 2014;26(6):371–375. doi:10.11919/j.issn.1002-0829.214171
19. Xafis V, Schaefer GO, Labude MK, et al. An ethics framework for big data in health and research. *Asian Bioeth Rev*. 2019;11(3):227–254. doi:10.1007/s41649-019-00099-x
20. Zhang Z. Missing data imputation: focusing on single imputation. *Ann Transl Med*. 2016;4(1):9. doi:10.3978/j.issn.2305-5839.2015.12.38
21. Peng K, Mathur A, Narayanan A. Mitigating dataset harms requires stewardship: lessons from 1000 papers. *Adv Neural Inf Process Syst*. 2021. doi:10.48550/arXiv.2108.02922

Nature and Science of Sleep

Dovepress

Publish your work in this journal

Nature and Science of Sleep is an international, peer-reviewed, open access journal covering all aspects of sleep science and sleep medicine, including the neurophysiology and functions of sleep, the genetics of sleep, sleep and society, biological rhythms, dreaming, sleep disorders and therapy, and strategies to optimize healthy sleep. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/nature-and-science-of-sleep-journal