Exposome, 2024, **4(1)**, osae004
https://doi.org/10.1093/exposome/osae004
Advance Access Publication Date: 13 April 2024
Research Article

2022 NIEHS Catalytic Workshop Series on the Exposome

A long and winding road: culture change on data sharing in exposomics

Robert O. Wright [1,*], MD, MPH, Konstantinos C. Makris [2], PhD, Pantelis Natsiavas [3], PhD, Timothy Fennell[4], PhD, Blake R. Rushing[5], PhD, Ander Wilson [6], PhD, and Members of the Exposomics Consortium[†]

[1]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, Institute for Exposomic Research, New York, NY, USA
[2]Cyprus International Institute for Environmental and Public Health, School of Health Sciences, Cyprus University of Technology, Limassol, Cyprus
[3]Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece
[4]Analytical Chemistry and Pharmaceutics, RTI International, Research Triangle Park, NC, USA
[5]Department of Nutrition, Gillings School of Global Public Health, Nutrition Research Institute, University of NC at Chapel Hill, Chapel Hill, NC, USA
[6]Department of Statistics, CO State University, Fort Collins, CO, USA

*To whom correspondence should be addressed: Email: robert.wright@mssm.edu

[†]For full consortium author list, please see: https://www.exposomicsconsortium.org/view/EXPOSOME-2023-010

Abstract

Data sharing requires cooperation from data generators (eg, epidemiologists, lab investigators) and data users (eg, epidemiologists, biostatisticians, computer scientists). Data generation and data use in human exposome studies require significant but different skill sets and are separated temporally in many cases. Sharing will require maintaining a history of data generation and a system to address the concerns of data generators around credit for conducting rigorous work (eg, authorship). Sharing also requires addressing the needs of data users to facilitate harmonization, searchability and QA/QC of data. We present these issues from the perspectives of data generators and data users and include the special case of real-world data (eg, electronic health records). We conclude with recommendations to address how to better promote data sharing in exposomics through authorship, cost recovery and addressing ethical issues.

Keywords: exposome; data sharing; real world data; epidemiology; mixtures; data science.

Introduction

In early 2023, NIH released a set of guidelines for data sharing[1] representing a beginning for the research community rather than an established set of principles. While a long-standing practice in genomics, data sharing is a "Brave New World" in exposomics and requires a change in culture to succeed. Data sharing is highly complex, expensive, and requires significant investment. Findable, Accessible, Interoperable, and Reusable (FAIR) data principles have been previously described in detail and we assume that readers have some familiarity with these issues.[2-6] In this paper, we outline the benefits and challenges of data sharing from the perspectives of exposomic data users and data generators. Currently, data sharing comes with uneven incentives for generators and little infrastructure investment for data users. In addition, some data were not collected for research purposes, and have unclear oversight. Annotating and curating a data set requires considerable time and effort. Analyzing shared data that has not been well annotated, poorly curated or is not searchable, limits the benefits of sharing. Given that all parties want rigorous high-quality data, how can these issues be addressed? We start with a brief summary of data sharing benefits, followed by its costs. We then present the perspectives of different stakeholders—data generators, users of shared data, and the role of real-world data (RWD) (eg, datasets of health care facilities), in which data generation was not organized for research purposes.

Data sharing benefits

Data sharing promotes interdisciplinary research at much-reduced costs, as new data generation is the most expensive component of research. Shared data can enable replication and reproducibility of important findings that can rapidly impact clinical care or public health interventions, as additional studies may not have to be conducted de novo. Data sharing accelerates the timeline for scientific progress by enabling researchers to build upon previously collected data rather than conducting longitudinal research from scratch, saving time and resources. This cannot be understated, as a longitudinal 20-year study not only costs millions of dollars, it takes 20 years to implement- thus, an idea that can be studied with extant data can be published and implemented in 3–5 years, instead of 25 years (ie, in the year 2049 as of this writing). Finally, by depositing data in supported repositories, the research community will build a resource to preserve

Received: August 1, 2023. **Revised:** January 11, 2024. **Accepted:** February 19, 2024

the collective knowledge of their work and ensure that valuable data is available for future generations of scientists.

Data sharing barriers

Many legal and ethical issues arise when medical data are shared, especially given the sensitivity and personal nature of health data. Privacy issues tend to increase when data are to be shared in an international fashion, as international laws may be incompatible. Sharing data can create ethical dilemmas as policies are new and by definition extant data preceded them. Few participants from past studies were told that data sharing was coming. While funding agencies may require data sharing plans, its ethical implications need additional support and research funding agencies will need to provide guidance to local IRBs to address the ethics of data sharing.

Funding is another barrier. NIH funded data are gathered under a grant with a limited life cycle- commonly 5 years. Data generation can occur very close to near the end of the grant, giving researchers limited time to share it before funding ends. Further, data repositories need significant investment to maintain data after study's end particularly if they are to be harmonized with similar studies.

Data generators and data users

Case 1: Researcher who creates a population-based study

Researchers who generate data in a cohort, case control or clinical study represent a large portion of "data generators", with laboratory-based researchers representing a second group (see Case 2). Researchers who collect data have traditionally acted as gatekeepers to data use-which has led to criticism. Those who collect data spend years developing ideas, collecting pilot data, writing grants, hiring and managing personnel, writing IRB proposals, responding to critiques, creating a database, cleaning and creating a data dictionary before analyzing the data themselves.

High quality data generation is an area requiring considerable expertise. Longitudinal data are the gold standard in human subject research and can require decades of effort. This can lead to an imbalance in the amount of work hours needed to generate data versus work hours needed to analyze shared data. Traditional academic systems prioritize publishing research papers rather than sharing data which compounds this problem. Without safeguards for authorship consideration, the present scenario creates negative incentives in which those generating data are reluctant to share, fearing lack of recognition. Further, researchers using extant data analysis for manuscript writing may be years removed from the data collection process. They may not understand or appreciate the strengths and weaknesses of the dataset or may make incorrect assumptions about how the data were generated. Authorship criteria require substantial contributions to the conception, design, execution, or interpretation of the research. Generating data by itself often meets these criteria.

Case 2: Data sharing from a laboratory perspective

A laboratory may generate data from analysis of biosamples without conducting sample collection. Such researchers often have unique expertise and insights on the assay and its interpretation. Assays require processing and researchers may sometimes submit raw data to meet data sharing requirements. This can restrict its use to only researchers who can re-process results, while excluding researchers without such expertise. In many instances, considerable effort has been expended by laboratory faculty and staff in developing methods specific to the sample analysis, as well as ensuring that methods are executed with proper quality control. Without proper annotation and safeguards, data users may not recognize the laboratory contribution to the deposited data. If team science is to be encouraged, appropriate credit should be provided to laboratory investigators. Their contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work merit authorship[2] and often meets the standard for authorship.

Case 3: Data sharing from a data user perspective

Data generators may not have advanced skills in data science. New structures of exposome data and larger datasets will require the development of relevant data science methods. We have seen this in the recent interest in environmental mixtures and the surge of statistical methods developed in response. Exposome data may be assessed from heterogeneous biomarkers among many different scenarios. Different methods are needed for mixtures data that are observed cross-sectionally versus data assessed longitudinally as repeated measures of exposure or grouped by life stage or chemical class. In addition, different methods are needed to address different research questions.[3,4] The vast majority of statistical and data science methods are developed by data scientists to match the data structure and questions that can naturally be answered from the study to which they have access. In part, this is because demonstrating methods using real data is often required for publication. *Sharing data is a key process for overcoming this roadblock to develop methods for emerging exposome data structures and to answer new questions within existing data sources*. As new data structures become available, it is important to share these data so that appropriate methods can be developed to properly analyze data.

Data sharing is essential for benchmarking and comparing exposome analysis methods which informs researchers about the relative abilities and operating characteristics of different methods, including which methods are most appropriate for their analyses. A recent example is the 2021 Exposome Data Challenge Event.[5] This event gathered a diverse scientific audience, including epidemiologists, biostatisticians and computational scientists applying a variety of data science tools to address different research questions and inferential goals related to exposomics, allowing for comparisons of methods. As a result, teams illustrated applications of methods on a common dataset and shared code. Other examples from environmental mixtures research include a National Institutes of Health (NIH) workshop that compared methods on real and simulated datasets.[6] The workshop include simulation studies that quantitatively compared performance in simulation studies[7] and papers that qualitatively compared methods on select datasets.[8] Sharing data allowed for benchmarking for new structures, addressing new research questions, and avoiding biases caused by overusing a limited number of data sources to validate model performance. Innovative means of data sharing include federated data analysis, which is performed on multiple separated datasets without data exchange, maintaining the safeguards of an institution's firewall. Only parameters of the analysis method are exchanged so that sensitive information remains anonymous. Such methods reduce the need for transferring data, reducing risks of compromising personal information.

Case 4: Data sharing internationally

International collaborations are a special case of data sharing. While the possibilities for collaboration are limitless, we present Europe as a case study. The European Health Data Space (EHDS) is a recent European Union (EU) regulatory initiative to support the use of health data. Updated guidelines regulate data sharing beyond EU countries and apply to data collected and shared with a NIH funded study. The Trans-NIH BioMedical Informatics Coordinating Committee maintains lists of domain-specific data repositories in the US, while similar data repositories exist in the EU (eg, Zenodo, etc,).

Beyond EHDS, the European General Data Protection Rule (GDPR), requires protection be ensured when data are transferred to third countries (eg, US). Regardless of the data sharing tool used, transferred data must receive an equivalent level of protection as it would in the European Union. The EU Commission and the USA recently agreed on a Trans-Atlantic Data Privacy Framework (TADPF) to ensure that US surveillance activities are necessary and proportionate in pursuit of national security objectives, by creating an independent complaint mechanism for European citizens.[9]

Case 5: Sharing real-world data

We conclude with the concept of reuse of administrative data i.e., not collected for research- another special case of data sharing. With the proliferation of Information and Computing Technology systems in health (eg, Electronic Health Records), the volume of health data has dramatically increased. Such datasets will be used for purposes different than those for which they were collected, such as epidemiological studies, policy making, etc, These data are called "Real World Data" (RWD) and the conclusions inferred upon them are called "Real World Evidence" (RWE).[10-12] Their importance is widely recognized despite methodological challenges regarding their quality and verifiability. However, using RWD is not trivial as they are often "sensitive" personal data, thus accompanied with legal, ethical, administrative issues. In technical terms, "interoperability" is also a crucial issue, as data exchange or federated analysis would require the use of a Common Data Model (CDM).[13,14]

There are several initiatives supporting the development of observational studies with RWD, including developing a "research network" of Data Partners (DPs) with data available in a specific CDM compatible format. Instead of collecting the data from DPs in a central analysis node, only the non-sensitive and non-personal aggregated results are collected. This prevents moving sensitive data out of the original host/DP reducing legal challenges. Only the query and its aggregated results are transferred enabling observational studies in a federated fashion without sharing raw patient data.

Along these lines, Observational Health Data Sciences and Informatics (OHDSI) should be commended as a global initiative aiming to facilitate the analysis of RWD via observational studies.[15] OHDSI developed a set of open-source tools based on the so-called "OMOP Common Data Model" (OMOP-CDM). The European Health Data & Evidence Network (EHDEN) acts as the European face of OHDSI[16] setting up a community of data partners across Europe upon OMOP-CDM, currently including more than 180 organizations. As a whole, the OHDSI ecosystem has proven that technical means can go a long way in terms of overcoming legal and regulatory barriers in an international scale.[20-22]

Even though technical standards may facilitate the *syntactic interoperability* between data silos (eg, hospitals), the issue of *semantic interoperability still* remains. To this end, widely accepted reference terminologies are used (eg, the Unified Medical Language System—UMLS). Developing and maintaining such terminologies/ontologies and aligning them is a costly and error prone process. *Data quality* is also a significant issue as research and observational studies are secondary use cases for RWD and not the priority of the original data creators. Contextual factors should be considered when RWD are interpreted.

Recommendation 1

Authorship should be a consideration for data generators when sharing is mandated.[17-19,23] This does not mean automatic authorship but a system that facilitates and requires communication between data generators and data users. This system should distinguish between expert (longitudinal study, novel or complex lab assay) and RWD data generation. Citing data that has been deposited is also recommended by the Joint Declaration of Data Citation Principals.[20]

Recommendation 2

The costs of data sharing should be shared. Implementation of fees to use data would help cover some costs of data sharing. Even with such fees, significant infrastructure investment is needed to support data harmonization, ontologies and search engines that allow researchers to find datasets. Data sharing fees may not be sufficient. Training programs in these methods are needed if we are to reach a critical mass of data generators and data users.

Recommendation 3

Ethical issues in sharing data need guidance from the Federal government to ensure privacy. At present, there are few guidelines to assist IRBs in determining whether data can or should be shared and ethical issues may be highly contextual requiring development of an expert review board to evaluate cases as they arise. Adding "data donation" routinely to consent forms as an opt out might accelerate data sharing.

Recommendation 4

Infrastructure is needed to facilitate federated data analysis as some data cannot be shared in a centralized database. Finally, data repositories should facilitate advanced computational approaches that enable full processing of data while still keeping data protected. Methods beyond deidentification need to be developed in scale. For example, homomorphic encryption allows data processing while keeping data encrypted, but its use is computationally intensive.

Summary

Data sharing will have its greatest impact through the implementation of incentives that benefit both generators and users in order to promote sharing and recover costs. Significant infrastructure investment is needed to accomplish this. This is a critical moment for exposomic data sharing and not addressing these issues will likely impede progress. Data sharing to date has been conducted with a bottom-up approach, and some top-down direction on incentives, fee structures and investment is needed. A more nuanced approach, bridging cultural and scientific

disciplinary differences to promote data sharing will be challenging, but the benefits of large-scale data sharing will clearly justify the costs.[25]

Funding

Research reported in this publication was supported by the National Institute Of Environmental Health Sciences of the National Institutes of Health under Award Number P30ES023515, U2CES026561, U2CES030857, U2CES030859 and R01 ES028811. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Robert Wright (Conceptualization [equal], Methodology [equal], Project administration [equal], Writing—original draft [lead], Writing—review & editing [equal]) Konstantinos Makris (Conceptualization [equal], Writing—review & editing [equal]) Pantelis Natsiavas (Conceptualization [equal], Formal analysis [equal], Methodology [equal]) Timothy Fennell (Conceptualization [equal], Methodology [equal], Writing—review & editing [equal]) Blake Rushing (Conceptualization [equal], Writing—review & editing [equal]) Ander Wilson (Conceptualization [equal], Formal analysis [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal])

Data availability

This is an opinion paper and no data were used to generate the manuscript.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

References

1. National Institutes of Health. 2023. 2023 NIH Data Management and Sharing Policy. Accessed March 25, 2024. https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy
2. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet. 2018;26(7):931–936. https://doi.org:10.1038/s41431-8–0160-0
3. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of the findable, accessible, interoperable, and reusable data principles in health data stewardship: scoping review. J Med Internet Res. 2023;25:e45013.https://doi.org:10.2196/45013
4. Rootes-Murdy K, Gazula H, Verner E, *et al.* Federated analysis of neuroimaging data: a review of the field. Neuroinformatics 2022;20(2):377–390. https://doi.org:10.1007/s12021-1–09550-7
5. Vesteghem C, Brøndum RF, Sønderkær M, *et al.* Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief Bioinform. 2020;21(3):936–945. https://doi.org:10.1093/bib/bbz044
6. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.https://doi.org:10.1038/sdata.2016.18
7. International Committee of Medical Journal Editors. 2023. Defining the Role of Authors and Contributors. https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html
8. Braun JM, Gennings C, Hauser R, Webster TF. What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environ Health Perspect. 2016;124(1):A6–9. https://doi.org:10.1289/ehp.1510569
9. Hamra GB, Buckley JP. Environmental exposure mixtures: questions and methods to address them. Curr Epidemiol Rep. 2018;5(2):160–165. https://doi.org:10.1007/s40471-8–0145-0
10. Maitre L, Guimbaud J-B, Warembourg C, *et al.* State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. Environ Int. 2022;168:107422. https://doi.org:10.1016/j.envint.2022.107422
11. Taylor KW, Joubert BR, Braun JM, *et al.* Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environ Health Perspect. 2016;124(12):A227–A229. https://doi.org:10.1289/EHP547
12. Hoskovec L, Benka-Coker W, Severson R, Magzamen S, Wilson A. Model choice for estimating the association between exposure to chemical mixtures and health outcomes: a simulation study. PLOS One. 2021;16(3):e0249236.https://doi.org:10.1371/journal.pone.0249236
13. Gibson EA, Nunez Y, Abuawad A, *et al.* An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. Environ Health. 2019;18(1):76.https://doi.org:10.1186/s12940-9–0515-1
14. European Data Protection Board. 2020. Frequently Asked Questions on the judgment of the Court of Justice of the European Union in Case C-311/18 - Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems. Accessed March 25, 2024. https://edpb.europa.eu/our-work-tools/our-documents/other/frequently-asked-questions-judgment-court-justice-european-union_en
15. Kondylakis H, Kalokyri V, Sfakianakis S, *et al.* Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp. 2023;7(1):20.https://doi.org:10.1186/s41747-3–00336-x
16. Stenzinger A, Moltzen EK, Winkler E, *et al.* Implementation of precision medicine in healthcare—a European perspective. J Intern Med. 2023;294(4):437–454. https://doi.org:10.1111/joim.13698
17. You SC, Lee S, Choi B, Park RW. Establishment of an international evidence sharing network through common data model for cardiovascular research. Korean Circ J. 2022;52(12):853–864. https://doi.org:10.4070/kcj.2022.0294
18. Fernandez-Luque L, Imran M. Humanitarian health computing using artificial intelligence and social media: a narrative literature review. Int J Med Inform. 2018;114:136–142. https://doi.org:10.1016/j.ijmedinf.2018.01.015
19. Long C, Tcheng JE, Marinac-Dabic D, Iorga A, Krucoff M, Fisher D. Developing minimum core data structure for the obesity devices Coordinated Registry Network (CRN). BMJ Surg Interv Health Technol. 2022;4(Suppl 1):e000118.https://doi.org:10.1136/bmjsit-2021-000118

20. Observational Health Data Sciences and Informatics. 2023. Welcome to OHDSI!. Accessed March 25, 2024. https://www.ohdsi.org/
21. European Health Data & Evidence Network. 2022. Becoming the trusted open science community built with standardised health data via a European federated network, https://www.ehden.eu/
22. Publications Office of the European, U. & Jessop, P. *Data Citation: A Guide to Best Practice*. Publications Office of the European Union; 2022.
23. Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. Nature 2019;570(7759):30–32. https://doi.org:10.1038/d41586-9–01715-4
24. Saito H, Kobayashi H, Takeno S, Sakai T, Ishii H. In vivo and in vitro studies on fetal toxicity of benzodiazepines in rats. Res Commun Chem Pathol Pharmacol. 1986;52(3):295–304.
25. *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles*. 2014. Accessed March 25, 2024. https://doi.org/10.25490/a97f-egyk

Exposome, 2024, 4, 1–5
https://doi.org/10.1093/exposome/osae004
Research Article