Building Community Through Data: The value of a Researcher Driven Open Science Ecosystem

Building Community Through Data: The Value of a Researcher Driven Open Science Ecosystem

Meredith C.B. Adams, MD, MS[1*], Carla M. Bann, PhD[2], Emine Ozgur Bayman, PhD[3], Maria Chao, DrPH, MPA[4], Georgene W. Hergenroeder, PhD[5], Charles Knott, MPA[2], Martin A. Lindquist, PhD[6], Z. David Luo, MD, PhD[7], Rosemarie Martin, PhD[8], Maryann E. Martone, PhD[9], John McCarthy, MS[2], Micah McCumber, PhD[10], Sharon B. Meropol, MD, PhD[11], Ty A. Ridenour, PhD[2,12,13], Lissette M. Saavedra, PhD[2], Abeed Sarker, PhD[14], Kevin J. Anstrom, PhD[10], Wesley K. Thompson, PhD[15]

1. Wake Forest University School of Medicine, Winston-Salem, NC, USA

2. RTI International, Research Triangle Park, NC, USA, Bann: ORCID: 0000-0001-9305-8305, Knott: ORCID: 0000-0002-3340-1850, Ridenour: ORCID: 0000-002-9709-6808, Saavedra: ORCID: 0000-0001-8880-0624, McCarthy: ORCID: 0000-0003-0765-7884

3. University of Iowa, Iowa City, IA, USA, ORCID: 0000-0003-2615-0915

4. University of California, San Francisco, San Francisco, CA, USA, ORCID: 0000-0001-9846-7044

5. McGovern Medical School at the University of Texas Health Science Center Houston ORCID: 0000-0002-6170-2191

6. Johns Hopkins University, Baltimore, MD, USA, ORCID: 0000-0003-2289-0828

7. Anesthesiology & Perioperative Care, School of Medicine, University of California, Irvine, CA, USA, ORCID: 0000-0001-6811-9380

8. UMass Chan Medical School, Worcester, Massachusetts, USA, ORCID: 0000-0002-0412-8074

9. University of California, San Diego, San Diego, CA, USA, ORCID: 0000-0002-8406-3871

10. Gillings School of Public Health, University of North Carolina, Chapel Hill, NC, USA, McCumber: ORCID: 0000-0003-2570-1180, Anstrom: ORCID: 0000-0001-6452-2172

11. New York University Grossman School of Medicine, New York, NY, USA ORCID: 0000-0003-2025-1766

12. University of North Carolina, Chapel Hill, NC, USA

13. University of Pittsburgh, Pittsburgh, PA, USA

14. Emory University School of Medicine, Atlanta GA, USA, ORCID: 0000-0001-7358-544X

15. Laureate Institute for Brain Research, Tulsa, OK, USA, ORCID: 0000-0002-1148-1976

***Corresponding Author**:

Meredith C. B. Adams, MD, MS

Department of Anesthesiology, Artificial Intelligence, Translational Neuroscience, and Public Health Sciences

Pain Outcomes Lab

Wake Forest University School of Medicine

Medical Center Boulevard, Winston-Salem, NC 27157

meradams@wakehealth.edu

Tel: +1- 336-716-2255

Running Head: Building Community Through Data

Key words: data ecosystem, NIH HEAL Initiative, opioid, chronic pain, datasets, open science

Words: 1481

Figure: 1

Exponential scientific data growth presents challenges and opportunities for addressing complex public health issues like the opioid epidemic and chronic pain management. Despite the vast amount of research conducted globally, many datasets remain inaccessible or underutilized due to publication access policies and stringent data use agreements. The amount of data generated through research activities is enormous. Limited access to scientific datasets stifles discovery and delays the translation of proven scientific advances into real-world applications.(1) To address these challenges, we argue for the critical importance of open science ecosystems, using the National Institutes of Health Helping to End Addiction Long-term® Initiative (NIH HEAL Initiative®) as a case study. We discuss how building community around data can accelerate scientific discovery by enabling dataset integration, increasing statistical power, and fostering interdisciplinary collaboration.

Open science ecosystems represent a fundamental shift in how research is conducted, shared, and utilized. An open science ecosystem is a comprehensive research environment that combines technological infrastructure, standardized protocols, and collaborative networks to enable transparent sharing and integration of scientific data, methods, and findings. These interconnected networks integrate data collection, storage, processing, analysis, and use across organizations while adhering to FAIR (Findability, Accessibility, Interoperability, and Reusability) principles. The ecosystem encompasses not just the technical components for data sharing, but also the human elements: researchers, institutions, funding bodies, and community partners who work together under shared governance frameworks and data standards to accelerate scientific discovery. The NIH HEAL Initiative® demonstrates how such ecosystems can accelerate discovery in critical public health areas through transparent research methods, open access to data and findings, collaborative approaches to complex problems, and standardized data elements enabling cross-study analyses.

Shifting Culture and Confronting Barriers

Data sharing is crucial to HEAL's mission of addressing the interconnected public health crises of chronic pain and opioid use disorder (OUD). While appropriate medical use of opioids remains important for pain management, the rise in OUD presents distinct challenges requiring comprehensive research approaches. By distinguishing between therapeutic opioid use and OUD, researchers can better target interventions and support both pain management and addiction treatment needs.

HEAL researcher and community partner data dashboards actively identify and monitor emerging threats in the drug supply, such as highly potent fentanyl and xylazine.(2) This capability enables rapid resource deployment where needed, potentially saving lives. More broadly, combining datasets can increase sample size, provide more statistical power to make conclusions, help testing hypotheses that a single dataset alone would be insufficient to generate, and increase result generalizability. Data can be analyzed differently by multiple groups to answer numerous research questions and consolidate trust in the validity of reported results.(3) Secondary analysis of shared data can also uncover patterns in disparate datasets, such as signatures of different chronic pain types,(4) genetic factors that may underlie similar pain phenotypes (or symptoms), or qualities likely to make an individual respond to treatment.(5)

The impact of data inaccessibility on research progress is particularly evident in pain management studies. For example, when clinical trial data remains siloed, researchers cannot identify subtle patterns in treatment response across different patient populations. This limitation has historically hampered the development of personalized pain management approaches and delayed the identification of risk factors for OUD development. Through the HEAL Data Ecosystem, researchers have begun breaking down these barriers. In one instance, the combination of datasets from multiple pain management clinics revealed previously unrecognized patterns in treatment outcomes across diverse demographic groups, leading to more targeted intervention strategies.(6)

The competitive nature of scientific research, characterized by a culture of individual studies and siloed networks, has hindered widespread acceptance and adoption of open science ecosystems. Additionally, legitimate privacy concerns present complex challenges to carefully navigate. Many scientists and research teams remain skeptical or reluctant to participate in open science ecosystems, due in part to the current passive data sharing culture. Active data sharing, in contrast, requires adjusting to externally imposed standards. Creating an inclusive scientific community around data requires both interdisciplinary blending of people and unique approaches. Technical aspects of a functional open science ecosystem include tools for data ingestion, integration, and visualization; real-time data pipelines; cloud access; metadata management; and open-source code-sharing tools. People are essential drivers of such ecosystems by interacting and collaborating through investigator meetings, webinars, and other interpersonal strategies. For HEAL, these offerings have enabled highly productive synergies, joining researchers working in a range of diverse settings, such as neonatal intensive care units and emergency departments, jails and prisons, Indigenous communities, and basic research laboratories.

Identifying Strategies to Maximize the Value of the HEAL Data Ecosystem

HEAL research will achieve its greatest scientific impact when researchers leverage the Data Ecosystem to analyze large-scale datasets across studies of chronic pain, opioid use disorder, and co-occurring conditions. Building and sustaining a functional data ecosystem presents numerous challenges beyond cultural barriers. These challenges span from intellectual property concerns to the complexities of data curation and secondary data analyses. Researchers must grapple with the need for a priori harmonization of data collection across distinct studies, as well as operational issues related to sharing, managing, and analyzing large-scale heterogeneous datasets. Ethical considerations regarding data use and community involvement add additional complexity. Pragmatic limitations tied to NIH grant funding cycles and budgeting for data management further complicate the landscape. Additionally, researchers often lack understanding and acceptance for NIH processes to obtain metadata, case report forms, and protocol-level data, which can hinder effective participation in open science initiatives.

There is no one-size-fits-all solution to support full participation in an open science ecosystem. Unsurprisingly, perhaps the most urgent task for the scientific community is to fully embrace the use of data standards,(7-9) such as common data elements, data dictionaries, and data repositories compliant with FAIR (findability, accessibility, interoperability, and reusability) principles.(10) Embracing data standards will broadly advance the value of data across the research spectrum, including basic research with model organisms, genomic and imaging analyses, wearables, social media monitoring, clinical studies, and multimodal research projects (**Figure 1**). Similarly, data harmonization efforts will advance a diverse array of projects ranging from imaging studies to analyzing registry data. Expected data use agreement norms are needed to break the logjam of access to public health data, corporate data, and electronic health record (EHR)/medical claims data.

The era of data overload has evolved, with datasets now serving as crucial inputs for AI and advanced data science methods. Machine learning, natural language processing, and other algorithms show promise in analyzing diverse data sources, revealing clinical trends absent in traditional research trials and enabling continuous health monitoring for early interventions. HEAL investigators can advance the mission of open science by prioritizing transparency in response to the opioid and overdose crisis. Building community around large-scale data analysis is essential, involving creative activities like hackathons and question-focused sessions that encourage multimodal data collaboration. HEAL has developed resources for metadata registration and tools to guide investigators toward appropriate FAIR repositories, facilitating effective data sharing while adhering to usage agreements. Available tools also help investigators navigate appropriate FAIR repositories where data will most likely be found and used effectively and collaboratively. These repositories typically ensure data use agreement adherence.

This paper's authors, as members of the HEAL Data Ecosystem Collective Board, contend that the benefits of an expansive and open data ecosystem outweigh perceived challenges. By embracing open science principles, the research community can more effectively leverage diverse datasets to generate insights, inform policy, and ultimately improve outcomes for individuals affected by chronic pain and addiction. (**Figure 1**). Growing and sustaining a researcher driven HEAL Data Ecosystem will not only advance health equity but also provide other broad benefits for data contributors and users. Individuals living with these complex and often stigmatized conditions deserve to understand their effective health options based on what rigorous research shows. Policy makers need scientific proof to support policies aiming to help individuals and communities gain access to new treatments and prevention interventions.

A researcher-driven HEAL Data Ecosystem serves as a catalyst for scientific innovation directly impacting millions affected by chronic pain, addiction, and overdose (Table 1). Research institutions can support these efforts through resources for data standardization, recognition of data sharing contributions, and clear policies protecting scientific integrity and patient privacy. By embracing open science principles, the research community can more effectively leverage diverse datasets to generate insights, inform policy, and ultimately improve outcomes for individuals affected by chronic pain and addiction.

The HEAL Data Ecosystem represents a powerful tool for advancing our understanding and treatment of chronic pain and addiction. By fostering a culture of open science and data sharing, we can accelerate discovery, improve patient outcomes, and address the urgent public health crises of opioid addiction and chronic pain. As members of the scientific community, it is our responsibility to embrace these principles and actively contribute to building a more collaborative, transparent, and impactful research environment.

Disclosure Statement: The authors include Dr. Adams (Chair) and members of the NIH HEAL Data Ecosystem Collective Board. The views expressed in this manuscript represent those of the authors and do not necessarily reflect the official views of the National Institutes of Health or the NIH HEAL Initiative.

Acknowledgments: The authors appreciate the editorial contributions of Alison F. Davis, PhD. The authors are current and past members of the NIH HEAL Initiative Data Ecosystem Collective Board, which guides the overall strategy and direction of HEAL's data management and sharing efforts. The authors also recognize the diligence and dedication of Jessica N. Mazerik, PhD, Anthony Juehne, and the HEAL Stewards.

Funding Disclosures: Research reported in this publication was supported by National Institutes of Health HEAL Initiative and National Institute on Drug Abuse under grant numbers U24DA058606, U24DA057612, R24DA055306, R25DA061740 (Meredith Adams). Additional funding was provided by the National Institutes of Health through the following grants: U24AR076730 (Kevin Anstrom, Micah McCumber); PL1HD101059 (Carla Bann); UG3AR076387 and UH3AR076387 (Emine Bayman); R61MD018333 and R33MD018333 (Maria Chao); R61NS113329 and R33NS113329 (Georgene Hergenroeder); UM1DA049394 (Charlie Knott, John McCarthy); U54DA049110 (Martin Lindquist); R01DE029202 and R01DE029202-01S2 (Z. David Luo); U01DA050442 (Rosemarie Martin); U24NS135547 and OT2OD030541 (Maryann Martone); U24NS113844-05S2 (Sharon Meropol); U24DA050182 (Ty Ridenour); R24DA057611 (Lissette Saavedra); R01DA057599 (Abeed Sarker); and U24DA055330 (Wes Thompson).

References

1. Khan S, Chambers D, Neta G. Revisiting time to translation: implementation of evidence-based practices (EBPs) in cancer control. *Cancer Causes Control* 2021;**32**(3)**:** 221-30.
2. Wu E, Villani J, Davis A, Fareed N, Harris DR, Huerta TR, LaRochelle MR, Miller CC, Oga EA. Community dashboards to support data-informed decision-making in the HEALing communities study. *Drug Alcohol Depend* 2020;**217:** 108331.
3. Gao Y, Staginnus M. Cortical structure and subcortical volumes in conduct disorder: a coordinated analysis of 15 international cohorts from the ENIGMA-Antisocial Behavior Working Group. *Lancet Psychiatry* 2024;**11**(8)**:** 620-32.
4. Davis KD, Aghaeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, Burczynski ME, Crean C, Edwards R, Gaudilliere B, Hergenroeder GW, Iadarola MJ, Iyengar S, Jiang Y, Kong JT, Mackey S, Saab CY, Sang CN, Scholz J, Segerdahl M, Tracey I, Veasley C, Wang J, Wager TD, Wasan AD, Pelleymounter MA. Discovery and validation of

biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nat Rev Neurol* 2020;**16**(7)**:** 381-400.

5. Simons L, Moayedi M, Coghill RC, Stinson J, Angst MS, Aghaeepour N, Gaudilliere B, King CD, López-Solà M, Hoeppli ME, Biggs E, Ganio E, Williams SE, Goldschneider KR, Campbell F, Ruskin D, Krane EJ, Walker S, Rush G, Heirich M. Signature for Pain Recovery IN Teens (SPRINT): protocol for a multisite prospective signature study in chronic musculoskeletal pain. *BMJ Open* 2022;**12**(6)**:** e061548.
6. Alter BJ, Anderson NP, Gillman AG, Yin Q, Jeong J-H, Wasan AD. Hierarchical clustering by patient-reported pain distribution alone identifies distinct chronic pain subgroups differing by pain intensity, quality, and clinical outcomes. *PLOS ONE* 2021;**16**(8)**:** e0254862.
7. Adams MC, Brummett CM, Wandner LD, Topaloglu U, Hurley RW. Michigan body map: connecting the NIH HEAL IMPOWR network to the HEAL ecosystem. *Pain Medicine* 2023;**24**(7)**:** 907-09.
8. Adams MC, Hassett AL, Clauw DJ, Hurley RW. The NIH Pain Common Data Elements: A Great Start but a Long Way to the Finish Line. *Pain Medicine* 2024**:** pnae110.
9. Adams MC, Hurley RW, Siddons A, Topaloglu U, Wandner LD. NIH HEAL Common Data Elements (CDE) implementation: NIH HEAL Initiative IDEA-CC. *Pain Medicine* 2023;**24**(7)**:** 743-49.
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3:** 160018.

Figure 1. NIH HEAL Initiative Data Management Challenges and Opportunities A comprehensive overview of data management challenges and strategies across 13 key data types in the HEAL Initiative ecosystem. The figure illustrates specific challenges for each data type (clinical, corporate, EHR, genomic, imaging, medical claims, public health, qualitative, registry, social media, software, wearables, and animal models) and corresponding strategies to address these challenges. For each data type, challenges

related to standardization, privacy, integration, and technical implementation are matched with specific solutions focused on improving data sharing, harmonization, and utilization within the research community.

Table 1. Accessing and Contributing to the Ecosystem Overview of key entry points and resources available through the HEAL Data Platform (healdata.org) for researchers interested in data sharing and collaboration. The table outlines essential tools and services including step-by-step data preparation instructions, standardized templates, metadata creation tools, repository connections, and support services. These resources are designed to facilitate efficient data contribution while maintaining high standards of quality and compatibility within the HEAL Data Ecosystem.

NIH HEAL Initiative Data Management Challenges and Opportunities

Data Type	Challenges	Strategies
Clinical	- Data quality across different health centers affects standardization - Standard of care might differ across study sites	- Use HEAL common data elements to allow researchers to combine datasets - Reuse data elements from other studies, to promote usability between studies
Corporate	- Data access, confidentiality, intellectual property concerns	- Develop community standards for data use agreements and outcomes with industry to support balanced innovation and commercial interests
Electronic Health Records (EHRs)	- Variable, nonstandardized across health systems - Need to protect personal health information (PHI) - Legal issues because hospital owns medical record	- Develop/generalize tools to map EHR data between systems (e.g. OMOP common data models and FHIR) - Material transfer agreements
Genomic	- Genetic variations among species (human vs other organisms) - Genetic variants of specific genes	- Leverage data standards and common data models (e.g. GA4GH)
Imaging	- Lack of harmonization across scanners, sites, and studies can be difficult - Large file sizes	- Use standards for organizing and describing imaging datasets. - Develop and share code to translate data into common structured data elements between studies
Medical Claims	- Protecting PHI and business-confidential data has severely limited public sharing of medical claims data - Incomplete representation because of multiple billing sources	- Share codebook and data use lan guage to streamline access to claims data that can provide pat terns and trends for population level health
Public Health	- Lack of data harmonization - Inconsistencies from time delays in data release, access, and reporting	- Develop data use agreements, deidentification approaches, and recognize data source limitations
Qualitative	- Heterogeneous clinical domains - Data collection challenges - Privacy protection - Data storage	- Facilitate development of community standards for which portion of qualitative work can be shared
Registry	- Missing data, long-term follow-up, versioning issues	- Harmonize registry data collection to enable long-term follow-up and facilitate connections with other health conditions
Social Media	- Privacy, user platform shifts - Data reliability - Free-text data are unstructured and lack predefined formats	- Share experiences and lessons learned - Develop data standards for real time analysis of trends and patterns - Use natural language processing
Software	- Data interoperability, technological versioning	- Develop and share foundational algorithmic principles to support consistency across institutions and projects
Wearables	- Data accuracy and user compliance - Privacy and safety concerns - Proprietary control of software and hardware	- Develop research standards that can be shared with industry and other wearables developers
Animal Models	- Many different types and formats - No commonly used data standardi zation criteria - Outcome data can be animal-, sex-, and species-specific - Use of specialized data repositories that are not interconnected	- Remove siloed data repositories - Develop data standards - Enhance rigor by reporting all bio logical variables that can affect outcomes

Table 1. Accessing and Contributing to the Ecosystem
The HEAL Data Ecosystem provides multiple entry points for researchers interested in data sharing and collaboration. New contributors can access comprehensive guidance through the HEAL Platform (healdata.org), which offers: • Step-by-step instructions for data preparation and submission • Templates for common data elements and standardized formats • Tools for metadata creation and dataset documentation • Direct connections to FAIR-compliant repositories • Support services for data harmonization and integration. These resources ensure that data contributors can efficiently prepare and share their data while maintaining high standards of quality and compatibility.