

NIH AD/ADRD Platforms Workshop: FAIRness Within and Across Infrastructures

National Institute on Aging
June 20–21, 2023

TABLE OF CONTENTS

Executive Summary	i
Meeting Summary	1
Session 1: Overview and Keynote Address.....	1
Welcoming Remarks	1
Overview of NIH Data Strategy	1
Keynote Address: Principles of a Data Biosphere	2
Session 1 Q&A.....	4
Session 2: Current NIH Data Ecosystems	5
Introduction to Current NIH Data Ecosystems	5
NCI Cancer Research Data Commons	5
NIH Common Fund Data Ecosystem/Kids First	6
National COVID Cohort Collaborative	7
Genomics at Scale with NHGRI AnVIL	8
BioData Catalyst.....	9
NIH Data Ecosystem Panel Discussion	9
Session 3: Ethical, Legal, and Social Implications (ELSI) of Data Sharing.....	11
Introduction to the ELSI of Data Sharing.....	11
Establishing an Ecosystem that Fosters DEIA: Connecting Pieces of the Puzzle	12
“Nothing About Us Without Us:” Meaningful Engagement of Tribal Communities in Research	13
Technical Solutions for De-identification	15
ELSI Panel Discussion	16
Session 4: AD/ADRD Knowledge Environments.....	18
Current State of AD/ADRD Knowledge Environments	18
NIH-sponsored AD/ADRD Data Platforms	18
AD Knowledge Portal: Enabling FAIR Data Infrastructure for Research on AD/ADRD	19
NIA Genetics of Alzheimer’s Disease Data Storage Site	20
National Alzheimer’s Coordinating Center	22
Laboratory of Neuro Imaging.....	23
NIH Center for Alzheimer’s and Related Dementias: Cloud Forward Data Sharing—Update and "Limit Testing" with Long Reads.....	24
AD/ADRD Knowledge Environments Panel Discussion	25
Session 5: Breakout Groups	26
Use Case #1.....	26
Use Case #2.....	27
Breakout Group Reports	27
Closing Remarks	29
Post-Workshop Discussion.....	30
Acronyms	31

EXECUTIVE SUMMARY

The National Institute on Aging (NIA) supports valuable infrastructures that provide data resources to enable and accelerate research on Alzheimer's disease and Alzheimer's disease-related dementias (AD/ADRD). The NIA Division of Neuroscience hosted NIH AD/ADRD Platforms Workshop: FAIRness Within and Across Infrastructures to convene major NIA-supported AD/ADRD data infrastructures along with data ecosystems from across the National Institutes of Health (NIH). The goal of the workshop was to explore opportunities and challenges for nurturing data ecosystems that (1) promote FAIR (i.e., findability, accessibility, interoperability, and reusability) data principles within and across constituent platforms and (2) support open and inclusive data access and compute resources for researchers and other stakeholders. The summary provides an overview of presentations and discussions occurring as part of the workshop and is not intended to provide comprehensive coverage of this topic.

The two-day workshop explored key themes, including (1) how biomedical data ecosystems ensure effective discovery and use of data and analytical resources; (2) technical, social, and policy solutions that enable such practices; and (3) ongoing challenges and opportunities to increase the impact of valuable data resources. Several observations were made in a post-workshop discussion, including the fact that existing data platforms—while impressive—are currently underutilized in proportion to their value. Establishing linkages between these platforms would increase their utility and thus make them more attractive to researchers. Creation of an integrated data ecosystem will require collaboration around technical and policy issues, adoption of standards and shared norms, and exploration of new approaches (e.g., large language models) to address challenges.

Session 1: Overview and Keynote Address

Drs. Jennie Larkin and Richard Hodes welcomed participants on behalf of NIA. Dr. Susan Gregurick then began the workshop by providing an overview of the current NIH data strategy. The number of controlled-access biomedical data repositories and platforms that provide the research community with data and analytical tools, applications, and workflows in secure environments has grown substantially in recent years. NIH aims to create an agile integrated cloud-based ecosystem while maintaining the autonomy of individual platforms, in part through the ongoing work of the NIH Cloud-Based Platforms Interoperability (NCPI), an effort established in 2019 to facilitate interoperability among genomic analysis platforms. In addition, NIH is committed to ensuring its data systems adhere to [FAIR](#) and [CARE](#) (i.e., collective benefit, authority to control, responsibility, ethics) principles. Dr. Benedict Paten then delivered the keynote address, in which he reviewed the four governing principles of a [Data Biosphere](#). Specifically, a Data Biosphere should be *modular*, comprising functional components with discrete capabilities and clear rules of interaction; *community focused*, with projects free to determine which modules best serve the needs of their communities; *open* so that resources can be reused, which enables extensibility; and *standards based* to allow components to interface with one another.

Session 2: Current NIH Data Ecosystems

Dr. Jack DiGiovanna moderated presentations and discussions related to current NIH data ecosystems. The cloud is an empowering technology, allowing for a paradigm shift from bringing data to researchers to bringing researchers to the data. Challenges for data ecosystems are wide ranging and include optimizing storage strategies and achieving interoperability. To begin, Dr. Tanja Davidsen described the Cancer Research Data Commons ([CRDC](#)), the primary data science platform of the National Cancer

Executive Summary

Institute (NCI). CRDC includes six data commons and components focused on interoperability, search, and analysis. Dr. George Papanicolaou then presented on the NIH Common Fund Data Ecosystem ([CFDE](#)), a gateway platform that enables users to query federated datasets generated by numerous NIH Common Fund programs; CFDE is working to harmonize and integrate these datasets. This was followed by an overview of the National COVID Cohort Collaborative ([N3C](#)) by Dr. Kenneth Gersing. N3C compiles and harmonizes data from clinical and research sites across the country. Dr. Gersing's presentation included an example of a privacy preserving record linkage, in which records for the same research participant across multiple datasets are linked via an independent honest broker. Dr. Michael Schatz then presented on the National Human Genome Research Institute (NHGRI) Analysis, Visualization, and Informatics Lab-space ([AnVIL](#)), a federated data storage, analysis, and sharing platform that includes genomic data generated by multiple NHGRI-funded initiatives (e.g., Centers for Common Disease Genomics [CCDG]). A final presentation, led by Dr. Regina Bures, focused on BioData Catalyst, which hosts several National Heart, Lung, and Blood Institute (NHLBI) datasets and provides several analysis and collaboration tools for researchers. Common themes emerged across these diverse data ecosystems, including user-friendly interfaces, flexibility to allow work in different cloud environments, developer-friendly tools, support for collaboration, ability to incorporate resources from external repositories, scalability, and sustainability. Key challenges identified during the session included data harmonization, education and training, the economics of cloud-based computing, and tools for identification of relevant data and analysis tools.

Session 3: ELSI of Data Sharing

Dr. Patricia Jones moderated presentations and discussions regarding ethical, legal, and social implications (ELSI) of data sharing and AD/ADRD data ecosystems. ELSI considerations are aligned with the need for DEIA (diversity, equity, inclusion, and accessibility) in research and the scientific workforce. Dr. Michael Griswold began the session by noting the interplay among DEIA, FAIR, TRUST (i.e., transparency, responsibility, user focus, sustainability, technology), and [OPEN](#) (i.e., open, permissioned, entrusted, nonshared) principles. Diversity among research participants and transparency are essential to ensure that research results can be applied to individual patients. To achieve diversity in research, intentionality is required. Furthermore, technological advances integrated with this intentionality would result in data platforms fostering progress toward DEIA across and within data ecosystems. Dr. Spero Manson followed with a presentation that described examples of research misconduct that undermined the trust of the American Indian/Alaska Native (AI/AN) communities. Dr. Manson noted that researchers can build collaborative relationships with AI/AN Tribes through respective engagements and acknowledgement of organizational and cultural differences between the research and Tribal communities. Federal affirmations and new activities are working to guide discussions and ensure research collaborations with AI/AN Tribes benefit all parties involved. Dr. Lucila Ohno-Machado rounded out the session with a presentation on ethical issues and considerations for de-identification of clinical data. With advancements in data science, prior methods for ensuring data privacy and security (such as maintaining anonymity of research participant personal identifiable information) have become insufficient. For example, the ability to link de-identified data sets from different research studies, while valuable for scientific inquiry, increases the risk of re-identification of participants. Strategies to mitigate re-identification risks are available but must be coupled with strong data privacy and security policies.

Session 4: AD/ADRD Knowledge Environments

Dr. Maryann Martone moderated presentations and discussions related to current AD/ADRD knowledge environments. The goals of the session were to assess available resources, consider use cases that will

Executive Summary

drive further progress on FAIR principles, and discuss ways to improve user experiences across data ecosystems in meaningful ways. In the opening presentation, Dr. Eliezer Masliah provided an overview of NIH-sponsored AD/ADRD platforms and the research they support, ranging from basic science to clinical trials. Dr. Anna Greenwood then described the [AD Knowledge Portal](#), which supports an array of NIA-funded AD data translational centers and systems biology consortia that generate multi-omic data from humans and model systems. The AD Knowledge Portal is the entry point into a larger infrastructure system underpinned by the Synapse platform, an NIH-designated cloud-based generalist repository that performs all data cataloging, storage, and access control for consortia contributing data. Dr. Li-San Wang followed with a presentation on the NIA Genetics of Alzheimer's Disease Data Storage Site ([NIAGADS](#)), the designated repository for AD genetic data and coordinating center for the Alzheimer's Disease Sequencing Project (ADSP). In addition to sharing data, NIAGADS has developed tools to view analyses, derived data, and annotations. The next speaker, Dr. Sarah Biber, shared information on the National Alzheimer's Coordinating Center ([NACC](#)). As the data, collaboration, and communication hub for the NIA Alzheimer's Disease Research Centers ([ADRC](#)) Program, NACC houses the most comprehensive AD/ADRD longitudinal, standardized, clinical, and neuropathology datasets in the world. Through its Data Front Door platform, NACC is working to achieve interoperability with other AD/ADRD research ecosystems—including NIAGADS, the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD), and the AD Knowledge Portal. Dr. Arthur Toga then presented on the Laboratory of Neuroimaging ([LONI](#)) Image and Data Archive ([IDA](#)), a repository of imaging and other data that facilitates data searching, mapping, and harmonization. IDA includes data from the Alzheimer's Disease Neuroimaging Network ([ADNI](#)); LONI collaborates with the ADSP Phenotype Harmonization Consortium to facilitate integration of ADSP and ADNI data. Dr. Mike Nalls ended the session with an overview of data platforms used by the Center for Alzheimer's and Related Dementias ([CARD](#)) to accelerate global collaborations for AD/ADRD research. CARD is facilitating access to multi-omic cell line, genetic, cell model, and long-read sequencing data through several existing data ecosystems. The resources presented aspire to adhere to FAIR principles and acknowledge opportunities to improve interoperability and integration. Workshop organizers suggested that a viable first step toward creation of an integrated data ecosystem could be linkage of a core set of AD/ADRD data resources and enabling user access to the array of data in these resources via a single entry point.

Session 5: Breakout Groups

Workshop participants were divided into groups to consider two research use cases through the lens of one of five topics: (1) data security and access, (2) data linkage and harmonization, (3) data ecosystem architectures, (4) promoting DEIA in data ecosystems, and (5) data analysis within data ecosystems. The first use case focused on the impact of sex differences on brain aging and AD/ADRD, and the second focused on understanding the molecular processes that underlie healthy versus pathological aging. Participants emphasized the importance of data standards; common data elements (CDEs) and ontologies for repository architecture, data, and metadata; tools and policies to facilitate data linkage; opportunities for artificial intelligence and machine learning to accelerate data linkage and harmonization (and the need to ensure data and models are amenable to these approaches); transparency and meaningful engagement with underserved communities; meaningful study design that facilitates thoughtful data collection and data provenance; and user-friendly training materials.

MEETING SUMMARY

Session 1: Overview and Keynote Address

Welcoming Remarks

Jennie Larkin, PhD, National Institute on Aging (NIA)

Richard Hodes, MD, NIA

Dr. Larkin welcomed participants from across the National Institutes of Health (NIH) and the broader research community to the workshop. The goal of the workshop was to explore opportunities and challenges for nurturing data ecosystems that promote FAIR (i.e., findability, accessibility, interoperability, and reusability) data principles both within and across platforms for Alzheimer's disease and Alzheimer's disease related dementias (AD/ADRD) data while supporting open and inclusive data access. Dr. Hodes noted that the workshop will inform NIA efforts to optimize the value and interoperability of its data infrastructure for AD/ADRD. Dr. Hodes highlighted the opportunity for workshop participants to learn from each other and expressed enthusiasm for the upcoming sessions focused on current NIH data ecosystems; ethical, legal, and social implications (ELSI) associated with data ecosystems; and AD/ADRD knowledge environments.

Overview of NIH Data Strategy

Susan Gregurick, PhD, NIH Office of Data Science Strategy (ODSS)

During the past five years, NIH support for and use of controlled-access biomedical data repositories and platforms that provide the research community with data and analytical tools, applications, and workflows in secure environments has grown substantially. For example, between Fiscal Year 2020 (FY20) and FY22, the [Data Repository and Knowledgebases Program](#) supported the creation of 19 investigator-initiated sustainable data resources to meet critical research needs. In total, NIH currently has about 206 petabytes of data across three cloud service providers (Amazon Web Services [AWS], Google Cloud, and Microsoft Azure).

This wealth of data presents both scientific opportunities and challenges related to data management, discovery, and analytics. A [2023 study](#) found that rates of data repository utilization to access published datasets have steadily increased in recent years, and utilization growth rates will likely accelerate given recent changes to the [Final NIH Policy for Data Management and Sharing](#).

Recent examples of large-scale data activities comprise the completion of the human pangenome by the Human Pangenome Reference Consortium, improved large language model outcomes (with pretraining using Medical Information Mart for Intensive Care data), and creation of the Federated European Genome-phenome Archive (which facilitates access to data across countries while maintaining national protections and regulations).

Many research questions would best be addressed by bringing together diverse data types and/or data from diverse populations housed across multiple platforms. NIH is actively exploring ways to create an agile integrated cloud-based ecosystem that meets the lofty goals of a federated data ecosystem—such as facilitating search, collaborations, analysis, visualization of results, outreach, and training—while maintaining the autonomy of individual platforms. For example, the NIH Cloud Platforms Interoperability ([NCPI](#)) effort seeks to enable a federated data ecosystem for researcher-driven analyses of datasets

Meeting Summary

across multiple cloud-based platforms and repositories through testing and implementing standards and approaches for systems interoperability and universal authentication and authorization. NCPI has implemented the Researcher Auth Service ([RAS](#)), which facilitates single sign-on across platforms, and is working on Global Alliance for Genomics and Health (GA4GH) [passport](#) implementation through RAS, which will enable verification of data access authorization. More than 10 NIH-supported resources are already using RAS for authentication or authorization, and others are in various stages of adopting it. Upon finding and accessing data, researchers can analyze it on cloud compute platforms such as Terra or Seven Bridges. NIH is also exploring privacy preserving record linkage (PPRL) methods to enable data from an individual subject to be linked over time and/or across datasets. NIH held a [workshop](#) on the policy and ethics of record linkage and released a [report](#) on PPRL for pediatric COVID-19 studies.

NIH is investing in improvement of its datasets and data systems; this includes efforts to address challenges related to data quality, management, sharing, and analysis. For example, an imputation super learning method has been applied to address missing data and class imbalances to ensure that National Institute of Environmental Health Science Puerto Rico Test site for Exploring Contamination Threats (PROTECT) datasets are artificial intelligence (AI)/machine learning (ML) ready. NIH joined DataCite as a consortium member; through this collaboration, NIH-funded repositories can mint digital object identifiers (DOI) and support data citation and tracking of data provenance.

Data are most meaningful when they are representative of diverse populations and embody FAIR and CARE (i.e., collective benefit, authority to control, responsibility, ethics) principles. CARE principles acknowledge that data should benefit and be controlled by contributing communities and that researchers maintain the responsibility of fostering trusting, positive relationships with communities and avoid causing harm. NIH is committed to ensuring its data systems adhere to FAIR and CARE principles. NIH will be putting CARE principles into practice as it creates a Tribally managed and governed data repository. In addition, NIH invested more than \$4 million from FY21 through FY22 to help existing data repositories and knowledgebases adopt FAIR and TRUST (i.e., transparency, responsibility, user focus, sustainability, technology) principles.

NIH is committed to training and career development of the scientific workforce, with an emphasis on increasing diversity. The Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program was created to enhance the participation and representation of researchers and communities currently underrepresented in the development of AI, address health disparities using AI/ML, and improve the capabilities of AI/ML. In 2022, AIM-AHEAD supported 22 research fellowships, 25 leadership fellowships, and 22 pilot research projects, along with launching a virtual hub for research at the intersection of AI/ML and health equity.

Keynote Address: Principles of a Data Biosphere

Benedict Paten, PhD, University of California, Santa Cruz

Biomedical research data have grown too big for any single institution to store and manage, and researchers need access to datasets stored in different places to answer interesting scientific questions. Dr. Paten was part of a group of data engineers and scientists who conceived the idea of creating a vibrant ecosystem of interoperable platforms, called the Data Biosphere, to address this problem. The approach calls for an inverted model of data sharing that involves bringing researchers to the data rather than bringing the data to each individual researcher. This cloud-based strategy reduces data duplication,

Meeting Summary

enables active security monitoring, and democratizes data access. It also facilitates collaboration across institutions and allows for flexible shared computing resources. The following four governing principles were proposed to guide development of the Data Biosphere: it should be modular, community focused, open, and standards based.

Modular. The Data Biosphere should comprise functional components, each with discrete capabilities and clear rules of interaction. Components include data assets, gatekeepers that manage data access, platforms that allow users to perform various functions (e.g., search for data, create data subsets, share data, create and store workflows, execute analyses), and portals that provide a user-friendly interface for interacting with and visualizing data. The National Human Genome Research Institute (NHGRI) [AnVIL](#) (Analysis, Visualization, and Informatics Lab-space) is an example of a modular ecosystem. The AnVIL Dataset Explorer indexes and enables data search. Its workspace module, Terra, provides a secure, compliant environment to analyze and share data. Analytical modules include Galaxy, Jupyter, and RStudio. Its tool repository module, Dockstore, integrates with Terra and enables thousands of analysis workflows. All AnVIL modules operate and interact within a Federal Information Security Management Act (FISMA)-compliant cloud environment.

Community focused. Every data repository has distinct community needs. Projects should be free to determine which modules best serve the needs of their respective research communities. The National Heart, Lung, and Blood Institute (NHLBI) BioData Catalyst, for example, uses many of the same modules as AnVIL. However, BioData Catalyst includes an additional commercial workspace environment and clinical exploration tools to better serve its specific research community.

Open. Data Biosphere resources should be open source to enable extensibility. More than 200 registered Data Biosphere engineers currently share software and code via GitHub, which allows other groups to adapt and reuse the software being developed. The principle of openness also refers to the need for open science. In one example of open science, Trans-omics for Precision Medicine (TOPMed) researchers made their containerized genome-wide association study (GWAS) workflows available on Dockstore through AnVIL. Users can pull this workflow into Terra and apply it to any dataset from AnVIL or other compatible data resources. Tutorials and interactive notebooks lower barriers to use. In a second example, a preprint publication of a neurodegeneration study provides links to new datasets and workflows in Dockstore for complex genomic and epigenetic analytical methods that can be applied to long-read sequencing data.

Standards based. Use of standards is essential to enable Data Biosphere components to interface with one another. Many groups are working to develop standards, including [GA4GH](#), which brings key stakeholders together to develop technical standards for effective data management, sharing, and analysis with an interest in standardizing policies that expand the use of genomic data. GA4GH provides many core interoperability standards for the Data Biosphere, including standards for accessing data, executing individual tasks, executing workflows, and sharing tools and workflows. GA4GH uses Driver Projects to test and demonstrate these standards; for example, one project demonstrated application of the same workflow across different analysis platforms. NCPI is also establishing and implementing guidelines and technical standards based on GA4GH protocols to improve interoperability across NIH data systems.

Session 1 Q&A

Nonhuman Data

- The data ecosystems introduced in this session focus on human data, but interest in resources for nonhuman data was acknowledged.
- The NIH Generalist Repository Ecosystem Initiative is developing common metadata and use cases for generalist repositories, many of which store nonhuman model system data. The goal is to promote FAIR data and practices so researchers can search and work across repositories.
- The AD Knowledge Portal includes large amounts of data from mouse model representative of late onset AD.
- The Data Biosphere work is foundational; there may be species-specific considerations, but the governing principles apply to all types of data.

Resource Sustainability

- NIH recognizes the importance of long-term sustenance and sustainability of data generated by NIH-funded projects. Generalist repositories and cloud-based data storage help address this issue. In addition, funded researchers are required to outline timeframes for sharing and long-term plans for the data in their data management and sharing plans.
- NIH ODSS is exploring cloud-based deep archiving of data in partnership with the National Center for Biotechnology Information (NCBI).
- NIH welcomes input from the research community on these issues.

Cloud-Based Computing Concerns

- Many researchers are concerned about the financial costs of using cloud-based data and computational resources. Currently, sufficient guardrails do not exist to prevent researchers from spending more than planned for their cloud instances.
- Financial incentives often are not aligned with cloud use. Many institutions make it cheaper for researchers to use on-premises versus cloud-based computing resources as institutional indirect costs, though this does not necessarily reflect true costs. In addition, data egress charges from cloud-based platforms, which can be substantial, are often covered by data platform providers.
- Researchers need an easier way to understand and budget for their cloud use; complicated economics make researchers hesitant to move to the cloud. The NIH Cloud Lab will provide a sandbox for researchers to explore and benchmark different cost models for analyzing data in the cloud. A [June 23, 2023, webinar](#) will introduce the Cloud Lab. NIH also is exploring hybrid computing through its Biowulf cluster.

Training and Education

- Data resources should be made broadly available beyond the small group of researchers who have deep technical expertise. The work that data scientists currently do to organize data and data systems will make it easier for researchers to contribute data and make it findable and usable. Education is key to helping users understand how to find and access data and apply analytical tools. NCPI may facilitate training opportunities and fellowships to help. Mentorship of

researchers from communities underrepresented in AI/ML research is already being provided through the AIM-AHEAD program.

Session 2: Current NIH Data Ecosystems

Introduction to Current NIH Data Ecosystems

Session Chair: Jack DiGiovanna, PhD, Seven Bridges

Data ecosystems help researchers gain new insights through an iterative process of data analysis, interpretation of results, and refinement of research questions based on results. Common themes emerged across diverse ecosystems; these included collaboration, user-friendly interfaces, flexibility to support work in diverse cloud environments, developer-friendly tools, ability to incorporate resources from external repositories, scalability, and sustainability.

Challenges for data ecosystems include optimizing storage strategies and achieving interoperability. Storage may be centralized (i.e., all data are stored and managed by a single party within the ecosystem) or federated (i.e., data exist in multiple locations within or outside the ecosystem) or be a mix of the two. The best approach depends on the ecosystem. Creating an “ecosystem of ecosystems” through interoperability could create incredible opportunities for the research community, but there are significant challenges, especially in the policy space. NCPI, which was created to facilitate interoperability among the genomic analysis platforms established across NIH, focuses both on demonstrating scientific progress and building generalizable technology standards to achieve that progress.

NIH data ecosystems support multiple pillars of the NIH Strategic Plan for Data Science. Workforce development is one of five pillars; data ecosystems can be used to train the next generation of data scientists, including high school, undergraduate, and graduate students. They also present opportunities to engage with underrepresented communities and encourage people in those communities to use available data to advance their own research questions.

NCI Cancer Research Data Commons

Tanja Davidsen, PhD, National Cancer Institute (NCI)

The [Cancer Research Data Commons](#) (CRDC) is the primary NCI data science platform for cancer research. The CRDC mission is to empower researchers by providing access to a cancer data ecosystem with state-of-the-art visualization, analysis, and interoperability tools in a flexible, cloud-based computational environment. The goals of CRDC include preserving the long-term value of NCI-funded data; improving data submission, access, and interoperability; and accelerating cancer research through integrative analysis of multi-modal data. Moreover, the vision for CRDC includes a focus on all stakeholders who generate and use data, including scientists, clinicians, intramural NIH research staff, and NIH program staff managing extramural research. CRDC aims to lower barriers with respect to data submission; data access, search, and retrieval; data integration across platforms; and access to platforms, tools, and workflows. CRDC also focuses on infrastructure and sustainability to ensure security and appropriate access for sensitive data; creation of sustainable, reusable, and uniform architecture; and development of a comprehensive plan for long-term data storage and accessibility to tools. CRDC data come from a variety of sources including surveillance, patient participation (e.g., clinical trials, cohort studies), and discovery (e.g., genomic research).

Meeting Summary

CRDC is a federated system that currently includes six data commons, including five data type-specific commons for genomic, proteomic, imaging, canine, and clinical trial data, as well as the Cancer Data Service, which houses data types not currently accepted by the other commons. Each of the data commons uses an application programming interface (API) and graphical user interface (GUI) for data exploration. CRDC service components focus on data interoperability and analysis. For example, the Data Commons Framework includes reusable components (e.g., authorization and authentication). The Data Standards Services helps make data more findable, accessible, and reusable by standardizing associated metadata. The Cancer Data Aggregator provides an API that conducts searches across the CRDC data commons, which currently spans genomic, proteomic, and imaging data. CRDC has three cloud-based resources that provide analysis workspaces (i.e., Seven Bridges Cancer Genomic Cloud, Broad Institute FireCloud, Institute for Systems Biology Cancer Gateway in the Cloud); researchers can bring their own data or tools to these workspaces and share data with collaborators.

CRDC data have been FAIR since the launch of the platform in 2016. Determining the best way to measure FAIRness is challenging, but several CRDC metrics highlight the resource's success. The NCI Genomic Data Commons currently houses about 3 petabytes of data and serves about 70,000 users per month. There are about 12,000 registered users across the three CRDC cloud-based resources. These users have collectively conducted more than 2,300 years of computational analyses.

The long-term vision of the CRDC is to move toward a learning health care system for cancer. This will include increasing data sharing across the cancer data ecosystem and with ecosystems in other disease areas; addressing inequities in cancer care and research; fostering a collaborative research community; expanding data commons to capture the most valuable multimodal cancer data; and improving the usability of analytical and visualization tools. Usability will be enhanced through the availability of centralized, intuitive tools that can be used for all data and by all users, as well as one-on-one concierge services. CRDC also aims to create a single data submission site and a single data discovery dashboard to allow users to query the entire CRDC ecosystem.

NIH Common Fund Data Ecosystem/Kids First

George Papanicolaou, PhD, NIH Office of the Director

Common Fund programs generate a vast array of data types from different organ systems and species. The goal of NIH [Common Fund Data Ecosystem](#) (CFDE) is to make Common Fund data sets more impactful by enabling easier data access and use. To this end, CFDE enables users to query across and use multiple Common Fund data sets; conducts training and outreach to increase awareness of available Common Fund Data; and coordinates and integrates infrastructure and activities into a cohesive ecosystem. CFDE, now in its fourth year of funding, currently consists of 12 Data Coordinating Centers (DCCs) that have Common Fund programs and a CFDE Coordinating Center that is responsible for technical and logistical coordination as well as training and outreach. The DCCs partner with each other and the CFDE Coordinating Center to harmonize and integrate datasets and develop new ways to make the datasets and other digital resources interoperable. CFDE also integrates with other NCPI datasets (e.g., GTEx data on AnVIL) through [Kids First](#), one of the Common Fund DCCs.

Three key challenges facing data ecosystems include finding and accessing the data, harmonizing files into a single data set, and creating and analyzing cross-cohort tables. To address findability and accessibility, the CFDE Coordinating Center created the [CFDE Search Portal](#). This searchable catalog of

Meeting Summary

metadata for all Common Fund program data sets was generated using the Crosscut Metadata Model. In addition, Kids First data, tools, and appyters (a no-code solution to execute bioinformatics workflows and visualizations) are available on [CAVATICA](#). For harmonization, CFDE is establishing common pipelines and developing appyters for data analysis.

CFDE begins phase two in October 2023. As part of this transition, CFDE Coordinating Center functions will be spread across five interconnected centers, each with a defined area of focus. The Data Resource Center will establish and maintain the CFDE landing page and resources for integrating and searching data sets, tools, and digital resources. The Knowledge Center will provide a portal for knowledge graphs and enable users to import assertions (data quality test queries) into CFDE workspaces. The Integration and Coordination Center will support integration among the CFDE centers and DCCs and oversee sustainability and evaluation activities. The Cloud Workspace Implementation Center will enable users to import their own data to co-analyze with Common Fund data and resources. The Center for Training will coordinate training activities and provide on-demand training; training will focus on underrepresented groups and underfunded institutions. The goal is for these centers to use portals to interconnect in a single cohesive, interoperable interface.

Common Fund programs have a 10-year funding limit, and therefore require a federated data storage approach. CFDE is not a repository but a gateway platform to help users find, use, and work across federated data. While data are federated, metadata and assertions are centralized, which should help retain long-term access to annotations. DCCs will use persistent identifiers and the GA4GH Data Repository Service.

CFDE has enabled research discovery across and within the datasets of its ecosystem by empowering the creativity of both data generators and data users. A program that funded partnerships of at least three Common Fund DCCs engendered creation of tools and annotations including an RNA-seq pipeline; a knowledge graph database; and a workflow playbook of reusable data, computation, and visualization APIs. These tools and annotations have been adopted by groups beyond those who created them. Furthermore, R03 grants that require incorporation of at least two Common Fund program data sets spurred the use of the data and led to the creation of products that are returned to DCCs and shared via repositories and GitHub. These grants also provide opportunities to solicit feedback on the program from external users.

National COVID Cohort Collaborative

Kenneth Gersing, MD, National Center for Advancing Translational Science

The National COVID Cohort Collaborative ([N3C](#)) includes 26 billion rows of data representing nearly 20 million patients from all 50 states. Contributors to N3C include Clinical and Translational Science Award sites, Clinical and Translational Research sites, and Federally Qualified Health Centers. N3C harmonizes data submitted by these sites and stores them in a central repository. Development of harmonization methods began many years ago, and this investment allowed rapid implementation of N3C when the COVID-19 pandemic began in 2020.

To gain access to N3C data, investigators submit a Data Use Request, which is reviewed by a federally staffed Data Access Committee. Upon approval of the Data Use Request, investigators are permitted to access N3C data and tools through an N3C workspace. To date, about 3,500 investigators from more than 360 institutions have used N3C data for more than 480 research studies.

Meeting Summary

The question of whether enclaves should be centralized or federated is a challenging one, and features of both architectures may be needed for data platforms to function most effectively. N3C was created with both centralized and federated features. All N3C data are harmonized and stored centrally. The harmonization process is critical due to differences in data collection across sites (e.g., units of measure). Records are also linked centrally through an independent honest broker, which would not be possible in a purely federated system. As part of the Tenant pilot project, N3C data are returned to contributing sites, allowing data to be used and analyzed locally. Sites can also collaborate with one another through closed N3C workspaces.

It is not feasible or sustainable to store all data in a single repository as different organizations will need to continue to maintain enclaves that meet their unique needs. Thus, data sharing efforts should focus on enabling data linkage and sharing across enclaves. The Freeport model put forth by Alistair Thompson (NHLBI) includes a jointly owned and managed ephemeral workspace that can be used for harmonization and analysis of datasets from multiple sources. The biomedical research community should carefully consider the governance of depositing data in these ephemeral spaces. N3C is exploring this issue in partnership with Medical Imaging Data Resource Center (MIDRC), with funding from the Patient-Centered Outcomes Research Institute (PCORI) Trust Fund and Advanced Research Projects Agency for Health (ARPA-H).

Genomics at Scale with NHGRI AnVIL

Michael Schatz, PhD, Johns Hopkins University

[AnVIL](#) is a FedRAMP-certified secure data storage, analysis, and sharing platform that offers a rich compute environment for the research community to collaborate in ways that were not possible with previous tools and infrastructure. AnVIL is a federated system featuring linked applications (e.g., through common APIs or data management tools). Different applications provide unique functions, such as workspaces, data catalogs, batch workflows, interactive tools, harmonization tools, workflow storage and sharing, analysis, and visualization. There are several entry points for AnVIL (e.g., APIs, Docker containers, Kubernetes, Helm Charts). AnVIL currently has almost 5 petabytes of data from more than 600,000 genomes contributed by the Centers for Common Disease Genomics (CCDG), the Centers for Mendelian Genomics, and many other consortia; an additional 5 petabytes of data are available through NCPI. NHGRI recognizes the need to link data from different ecosystems and is actively working toward standards and systems to make these linkages as seamless as possible.

Many research studies have been powered by AnVIL. For example, AnVIL enabled joint analysis of nearly 140,000 whole genomes sequenced at four CCDG institutions. The platform also was used by the Telomere-to-Telomere Consortium and the Human Pangenome Consortium in their efforts to sequence the full human genome and the pangenome, respectively. AnVIL enabled these consortia to merge their data with other extant datasets (e.g., 1000 Genomes) to enrich their own data. AnVIL is partnering with the [Genomic Data Science Community Network](#) to facilitate use of the data by groups that have historically been left out of genomic research, such as community colleges, historically Black colleges and universities, and Tribal colleges and universities. As part of this effort, students from across the United States will work with researchers to build a unified map of the diversity of the microbiome.

More information on AnVIL is available on the [AnVIL portal](#) and in a [2022 Cell Genomics paper](#) by Schatz, et al.

BioData Catalyst

Regina Bures, PhD, NHLBI

[BioData Catalyst](#) was developed to bring the abundance of available clinical and research data into the cloud for storage and analysis. The platform specifically seeks to support heart, lung, blood, and sleep researchers (including through the BioData Catalyst Fellows Program) by developing and integrating advanced cyber infrastructure, leading edge tools, and FAIR data. It aims to support personalized user experiences and data access while emphasizing data security and maintaining a focus on ethical considerations. NHLBI is dedicated to building new functionality to support scientific innovation and implementing data science solutions that will democratize access to data and computational tools.

BioData Catalyst is a secure cyber infrastructure of modular interoperable platforms. Users can search and browse data, access hundreds of workflows, and use the University of Michigan Trans-Omics for Precision Medicine ([TOPMed](#)) Imputation Server. Users also have access to cloud-based secure workspaces where they can implement their own tools or leverage those already available within the ecosystem. BioData Catalyst platforms related to data searches and cohort creation include [Gen3](#) and Patient-centered Information Commons: Standard Unification of Research Elements ([PIC-SURE](#)). Workspaces and analysis tools are provided through Seven Bridges and Terra, with Seven Bridges providing the primary platform. Analysis tools include JupyterLab, RStudio, Annotation Explorer, and Hail, as well as tools for image visualization and analysis. NHLBI is working with Terra to make BioData Catalyst interoperable with the National Institute on Minority Health and Health Disparities (NIMHD) Science Collaborative for Health disparities and Artificial intelligence bias REduction ([SchARE](#)) platform.

The BioData Catalyst ecosystem currently hosts several controlled and open datasets. Initial development focused on the needs of TOPMed as one of NHLBI's largest data investments; however, BioData Catalyst has since expanded to include other NHLBI programs and partnerships. For example, BioData Catalyst is partnering with the Biologic Specimen and Data Repository Information Coordinating Center ([BioLINCC](#)), which houses biological specimens and clinical trial data. BioData Catalyst also allows researchers to upload their own data and workflows into the ecosystem, which can then be shared with authorized collaborators within their established workspaces. The BioData Catalyst Data Management Core, funded in 2022, provides support and concierge services for data generators and research consortia.

BioData Catalyst is a federated data ecosystem comprising the Renaissance Computing Institute (RENCI), Dug, Gen3, PIC-SURE, Seven Bridges, and Terra, as well as its partnership with BioLINCC. BioData Catalyst leverages GA4GH standards (e.g., Passport, Data Repository Service) and is expanding use of Fast Healthcare Interoperability Resources (FHIR) standards. This ecosystem is also working to implement the NIH Center for Information Technology RAS Initiative via NCPI. The architecture of BioData Catalyst has many advantages, including the ability to work with data in one cloud-based location, increased data availability and diversity, improved scalability, and implementation of governance and standards consistent with federated partners.

NIH Data Ecosystem Panel Discussion

Data Ecosystem Knowledge and Usability

- The breadth of NIH data infrastructure can be overwhelming, particularly for people who are not data scientists. Panelists noted that many similarities exist among the various resources currently

Meeting Summary

available. It may be helpful for people to focus on learning about a single data ecosystem to start, as many components will translate to other ecosystems. Many ecosystems have training resources to help users get started: for example, AnVIL and CAVATICA have produced short videos that describe their tools and resources, and CFDE offers training sessions on how to use CFDE data and available tools.

- Ideally, data ecosystems would be easily usable by people with limited data science expertise, but NIH ecosystems have not yet achieved this goal. Establishment of best practices and business rules will reduce the risk that people who are not expert users of the ecosystems will unintentionally misuse or misinterpret data.

Data Harmonization

- N3C has focused many resources on data harmonization. Data cleaning and harmonization are common needs across all research, particularly for real-world data. Some panelists advocated investing more resources in this area. Other panelists noted that while harmonization is important, it may not be a good use of resources to fully harmonize all datasets, particularly older datasets. There must be a balance between harmonization of existing data and generation of new data. There should be a focus on making it easier to harmonize and integrate new data through thoughtful study design, increased use of standards, strong documentation, and searchable metadata. (The availability of searchable metadata is a notably helpful feature of BioData Catalyst.)
- CRDC Data Standards Services is working to set metadata standards throughout CRDC. Once standards are in place, some aspects of harmonization can be automated, but much of the work will continue to be manual. Large language models like ChatGPT may facilitate automated harmonization, although human checks on the results of these models will be needed. Panelists acknowledged that while automated approaches may achieve some milestones in the near term, meaningful progress is a long-term goal. One goal that should be achievable in the near future is generation of local ancestry data for all samples to allow comparison of genetic data from different studies.
- Data scientists and researchers should learn from and build on previous scientific efforts. CFDE requires users to return their results—including results of harmonization—to data generators and repositories. Importantly, these results should be thoroughly annotated.

Education and Training

- Education for current and future generations of scientists can be transformational. Providing learning opportunities and teaching people how to interact with data and resources will ensure that these ecosystems are well utilized. Once people learn to use these ecosystems effectively, they are more likely to promote these resources to their colleagues.
- Ecosystem utilization should be easy and effective. To enable this, users need education in the use of cloud-based tools, including an understanding of the value of using these tools as well as the goals that may otherwise be unachievable if not for the availability of these tools.
- Education in cloud-based data resources should go beyond technical usage considerations; these trainings should also cover ethical and regulatory issues.

Meeting Summary

Cloud Concerns

- Research teams often are hesitant to use the cloud; a key concern is the potential to inadvertently incur extremely large costs. These concerns need to be addressed to convince people to use cloud-based resources.¹

Identifying Data and Tools

- Helping researchers find the best data and computational tools to answer their research questions is a challenge that ecosystems are working to address. The NCPI Dataset Catalog lists and provides key information (e.g., disease relevance, number of subjects) for major cohorts available through NCPI. AnVIL is piloting new technology to streamline dataset identification and access. One way to address this challenge would be to develop a concierge system, which some universities use to help potential external partners find the best faculty members for a given project.

Value of Data Ecosystems

- Large datasets are critical for detecting the often small signals that may elucidate the pathophysiology of diseases. The ability to aggregate data and apply powerful computational tools will enable incredible scientific insights that were not previously possible.
- To date, the panel was unaware of any data ecosystem or coordinating center that has achieved the level of FAIR compliance needed to enable autonomous data integration. This is a goal for the future.

Session 3: Ethical, Legal, and Social Implications (ELSI) of Data Sharing

Introduction to the ELSI of Data Sharing

Session Chair: Patricia Jones, DrPH, NIA

The concept of ELSI with respect to human subjects research gained popularity in the United States in the 1980s. In 1990, the Human Genome Project illustrated the impact that research findings could have on participants as well as their family members, communities, and society as a whole. ELSI considerations are not unique to the United States. In 1994, the European Union established an ELSA framework (in which A stood for “aspects”). ELSI is now included under the concept of Responsible Research Innovation, which emphasizes socioeconomic impacts of research. *The Journal of Responsible Innovation* welcomes submissions of work in this area.

ELSI considerations are aligned with the need for diversity, equity, inclusion, and accessibility (DEIA) in the scientific workforce. DEIA increases research team productivity and research quality, as researchers from diverse backgrounds have unique and complementary skills and areas of expertise. A lack of diverse representation of human subjects and human subjects research compromises the ability to generalize clinical research findings. It also leads to greater financial costs, compounds existing health disparities, and can undermine public trust in research enterprises and individual research studies.

¹ Following the workshop, Dr. DiGiovanna indicated that Terra and Velsera have cost prediction tools available to alleviate this concern. Other cost controls (e.g., Billing Group limits) are also available or under development.

Meeting Summary

When conducting human subjects research of any kind, ELSI considerations comprise:

- Privacy, security, verifiability, confidentiality, and data safety.
- Appropriate and complete informed consent and assent.
- Unfair bias; alignment of values between researchers, study participants or patients; and ensuring mutual benefit.
- Data ownership, accessibility, and rights to sharing and future use.
- Transparency, traceability, and “explainability” for AI/ML algorithms.

Neglecting ELSI leads to medical mistrust, inaccurate research findings, and further mischaracterization of historically marginalized groups. Harm has been done when research was carried out without thoughtful, intelligent engagement with the individuals and populations directly affected by the work. In addition, multiple examples exist in which failure to consider ELSI resulted in ML algorithms with inconsistent results across racial/ethnic groups.

NIA currently has several open Notices of Funding Opportunities relevant to ELSI ([PAR-20-257](#), [PAR-20-255](#), [PAR-20-254](#)).

Establishing an Ecosystem that Fosters DEIA: Connecting Pieces of the Puzzle

Michael Griswold, PhD, University of Mississippi

Data ecosystems should balance considerations for the individual and the collective (or aggregate). From a research and statistical perspective, an individual is an unsolvable, unpredictable puzzle; however, in the aggregate, one can predict what “the average” person will do. From a health and medicine perspective, a great physician treats the individual patient, who will likely have both similarities with and differences from the average person.

DEIA principles help connect FAIR, TRUST, and OPEN (i.e., open, permissioned, entrusted, nonshared) data ecosystem principles. Fostering diversity requires transparency, yet research studies often report an overall result. A *transparent* overall result enables understanding of the data behind that result. Transparency, in kind, requires accessibility. For example, a *New York Times* report, “The Voting Habits of Americans Like You,” used data from the 2012 presidential election to compare voting patterns with other voter characteristics. The article emphasized differences in voting between racial/ethnic groups. However, a closer look at the data revealed that differences in voting patterns were similarly large *within* racial/ethnic groups when analyses were done by state of residence. This additional insight was possible because the data were available to readers through an interactive tool.

Intentionality also appears necessary to foster DEIA. A comparison of two recent clinical trials illustrates the importance of being intentional about adhering to DEIA principles in research. The Systolic Blood Pressure Intervention Trial (SPRINT) examined the impact of several biologically plausible effect modifiers on cardiovascular disease. Diversity was prespecified as a study objective. Among the more than 9,000 participants, 59% were White Non-Hispanic, 31% were Black Non-Hispanic, and 11% were Hispanic. Because of the large number of intentionally included diverse individuals, tools such as interactive visualizations of the SPRINT dataset enable doctors and patients to look at outcomes among participants who are similar to the individual patient seeking treatment (e.g., an 80-year-old Black woman who does

not smoke and has no history of cardiovascular disease). Conversely, a trial of lecanemab for treatment of early Alzheimer's disease was described by the drug's sponsor as having inclusive eligibility criteria but was not intentionally designed to enroll diverse populations. Consequently, Black/African American people made up 2.8% of the study population, with only 22 African American participants receiving lecanemab. Lecanemab was recently approved by the U.S. Food and Drug Administration based on the observed 27% reduction in global cognition decline. The study also reported an adverse event rate for amyloid-related imaging abnormalities-edema (i.e., brain swelling) for the lecanemab group that was over 8 times the rate among placebo groups (33% compared with 4%) among homozygous *APOE4* carriers. African Americans are more likely than other racial/ethnic populations to be homozygous *APOE4* carriers but, unfortunately, the small number of Black participants in the study precludes any additional useful study of their adverse event rates. The comparison between these two trials demonstrates the importance of intentionality with respect to diversity and inclusivity. Diversity was prespecified as an important component in SPRINT; however, the lecanemab sponsors relied on "inclusive eligibility criteria," which does not necessarily result in a diverse study population.

In the same way, the data ecosystem community must be intentional about DEIA. Many of the current data ecosystems highlight the number of available file formats or volume of data (e.g., number of terabytes) or number of downloads, but very few provide even basic characteristics about the people whose data are housed in these repositories (e.g., age, sex, race, geography). While there are challenges to allowing transparency and accessibility for data ecosystems, there are governance models that can help address public sharing issues such as the Baltimore Longitudinal Study of Aging (BLSA) FAIR and OPEN platform.

In summary, transparency and accessibility are essential for promoting diversity and inclusivity to advance knowledge from the collective to the individual. Similar to individual studies, ecosystems should be intentional about diversity. Providing open access, pass-through logins, user-friendly visuals, and clear information about study populations and phenotypes should be the norm for data platforms. Intentionality coupled with technology can result in platforms that foster progress toward DEIA in data ecosystems.

"Nothing About Us Without Us:" Meaningful Engagement of Tribal Communities in Research *Spero Manson, PhD, University of Colorado Anschutz Medical Campus*

The slogan "Nothing About Us Without Us" is at the heart of meaningful research engagement with Tribal communities. Two signal events illustrate the impact of research misconduct in American Indian and Alaska Native (AI/AN) communities. In one event, findings from the Barrow Alcohol Study at Tulane University were released simultaneously to the press and the community of Barrow, Alaska. As a result, newspapers published articles that inaccurately and problematically characterized the local population, leading to drastic economic loss for the city of Barrow and incalculable harm to the Native community. Furthermore, all attempts to correct the public record failed. The second signal event stemmed from the Havasupai Tribe's agreement to provide biospecimens and personal health information to Arizona State University researchers. The study was presented to the Tribe as a study of diabetes, with promises of insights that could inform future health care for participants. The investigators abused the trust of their participants and continued to use the provided samples for additional research into stigmatizing topics for which they did not have consent. The emotional consequences of this violation of trust have

Meeting Summary

reverberated throughout the Havasupai Tribe and Indian country. Today, the Navajo Nation is only just beginning to consider the conduct of genetic research among its Tribal members.

To clarify federal and Tribal relationships, particularly regarding research, the Presidential Executive Memorandum of April 29, 1994, reaffirmed the government-to-government relationship and necessary procedural steps federally sponsored efforts should take in consultation with federally recognized Tribes. This Memorandum led to the creation of new Tribal and Indian Health Service Institutional Review Boards that collaborate with their respective communities to ensure ethical conduct and dissemination of research. A 2004 Memorandum from President George H. W. Bush expanded and reaffirmed the nature of the federal obligation to Tribal communities, as did a 2009 Memorandum from President Barack Obama on Tribal Consultation. President Obama was among the first in his office to dedicate time to visit Tribal communities, an act that bolstered trust. This visit also led to a concentrated effort by the U.S. Department of Health and Human Services to develop specific policies and procedures for Tribal consultation and engagement in research, along with establishment of a new Tribal Health Research Office at NIH.

When searching for common ground, universities, the federal government, and Tribal communities should acknowledge organizational, political, and cultural differences. In universities and the federal government, decision-making is hierarchical and rule-driven; processes emphasize efficiency, replicability, and written documentation; leadership is formal and centralized; analytic thinking is prominent; instrumental competence is highly valued, and standard for assessment. In Tribal communities, decision-making is horizontal, precedent-oriented, and consensual; processes are fluid, iterative, documented orally, and benchmarked by key events; leadership is shared, diffused, and ascribed rather than attributed; the emphasis is on distributive cognition, with the assumption that one question may have multiple answers; and authority is rooted in collective decision-making. Political and cultural differences are equally stark, with a focus on problem-solving and the individual in universities and the federal government and an orientation toward collective wellbeing and mutuality in Tribal communities. Three specific points of tension arise from these differences: respect for persons, beneficence, and justice. These principles are outlined in The Belmont Report of 1979. Practical areas of friction include the review and approval of research, the distribution of resources, the manner of informed consent, the nature and extent of accountability and control, data ownership and sharing, and requirements for continued collaboration.

These organizational, political, and cultural differences do not prevent the creation of Data Use Agreements between institutions and federal bodies and Tribal communities. Dr. Manson serves on the Tribal Collaboration Working Group of the *All of Us* Research Program and is an investigator on the AIM-AHEAD Initiative, helping guide discussion and ensure collaborations that benefit all parties involved. Workshop attendees and researchers should familiarize themselves with important documents regarding Tribal engagement with research. The section of the [Final NIH Policy for Data Management and Sharing](#) on research engagement with American Indian and Alaska Native communities is a meaningful step forward.²

² Supplementary information to the policy for AI/AN populations is available in [NOT-OD-22-214](#).

Technical Solutions for De-identification

Lucila Ohno-Machado, MD, PhD, Yale University

Under the Health Insurance Portability and Accountability Act (HIPAA), de-identified electronic health record data are those from which 18 different identifier categories, including dates, biometrics, and names, have been removed (i.e., the Safe Harbor Approach), or those that have been certified by an expert as having a low risk of re-identification. Limited data sets include de-identified data plus dates and can be accessed through Data Use Agreements.

Ethical questions concerning the use of clinical data for research include:

- Is it right to share these data? The individuals to whom the records belong may not be aware of how their records are being used.
- Is it right not to share, which may impede scientific progress and advancements in health?
- Can data be accessed without being downloaded?
- Can we ask people what they want? Is this practical and fair?

Data de-identification is fraught with technical challenges. Combinations of values such as age, income, or diagnosis can be used to identify people even when biometric data or subject identifiers have been removed. Unique combinations of values do not guarantee privacy or prevent an individual from being re-identified. With advancements in data science, prior methods for ensuring anonymity are now insufficient, and linking sets of de-identified data also increases the risk of re-identification.

Strategies to mitigate re-identification risks are available, though none protect privacy completely and they must be coupled with strong privacy policies. Institutional and people-centric strategies to reduce the risk of re-identification include the institution acting as a data broker and use of patient-defined sharing permissions and consent management. An example of an institution acting as a data broker is the *All of Us* Research Program's data management center, which implemented three tiers of data access: public, registered, and controlled. All data, even those in the controlled tier, have been stripped of certain identifiers, and thorough privacy measures are applied to the entire dataset. Implementation of patient-defined data sharing permissions and consent management can be facilitated by tools such as immutable blockchain ledgers and automated smart contract infrastructure. A system implemented at Dr. Ohno-Machado's previous institution allowed research participants to consent to specific requests for their data and see how their data were being used. Approximately 50% of participants reported being satisfied with how permissions were handled and wanted to share their data and biosamples for research. Some categories of research, including family history, met with greater refusal to consent.

Data-centric strategies to reduce the risk of re-identification include federated computation; adding noise to data using techniques such as differential privacy; and operating on encrypted data (e.g., homomorphic encryption). Use of federated computation bypasses the need for a central repository; central repositories increase the risk that a large amount of data will be accessed via a single system breach. In one successful example of federated computation from early in the COVID-19 pandemic, several institutions contributed their data to a distributed analysis to determine whether acetylcholinesterase (ACE) inhibitors could be safely used for hypertensive patients hospitalized with COVID-19. Because data did not have to be centralized, researchers had access to records of 45 million

patients. Another example of distributed computing is the collaborative secure, federated GWAS analysis being conducted by the Veterans Affairs Million Veteran Program and *All of Us* Research Program.

ELSI Panel Discussion

Engagement with Tribal Communities

- The trend is progressing from individual data system architecture toward separate entities addressing common questions in a parallel yet independent fashion. The data ecosystem architectures identified in AIM-AHEAD included a federated approach that is conducive to protecting data from AI/AN populations, for whom protections and governance of data rest primarily with Tribal entities. The federated approach enables collective governance of multiple independent datasets while maintaining data sovereignty of each individual jurisdictional authority.
- Recent guidance on sharing AI/AN data is provided in the context of the Rapid Acceleration of Diagnostics-Underserved Populations ([RADx-UP](#)) initiative that invests substantial resources to enhance its coordinating center, outreach, engagement, and continued support of 11 Tribal Nation grantees. NIH funds a separate but related activity that will establish a Tribal data repository governed by participating Tribal nations. In addition, the *All of Us* Research Program is deeply invested in developing relationships with Tribal Nations with specific data sharing guidelines. In addition, the NIH Tribal Health Research Office has established guidelines, use case examples, and other resources for investigators who are interested in further collaboration with AI/AN communities.
- Codification of regulatory requirements is progressing rapidly in federally recognized Tribal communities such that contracts, Data Use Agreements, and oversight requirements address permission to undertake research and enable flow of research-related resources to communities; who has access to those data and for what purposes; and resulting products. These codes require an authentic attempt on the part of scientists to ensure that findings make their way back to the communities that contributed the data and to remain engaged with advocates from those communities to translate the meaning and utility of those findings to their health and welfare.

De-identification/Re-identification

- De-identification relies on data infrastructure, policy, and researcher (i.e., data submitters and data users) conduct. Risk of re-identification can be managed but not guaranteed.
- Linking ecosystems increases the risk of re-identification. The actual risk of re-identification is a complex function of what data are available and how it might be re-connected to individuals. This risk was currently viewed as low by some; however, there was pushback on this view.

Data Sharing and Informed Consent

- Participants who are unwilling to share their data with for-profit organizations may not realize that this includes insurance and social media companies in addition to pharmaceutical companies. They also may not realize the proportion of treatments that come from for-profit organizations. Increased awareness of these factors may alter participant preferences.

Meeting Summary

- AI/AN groups would be more willing to share with for-profit groups if they received a share of profits resulting from commercialization of their contributions. Dr. Ohno-Machado noted that some professional societies receive royalties from drugs developed with their cohort data; there is some debate about whether this practice represents a conflict of interest because the society benefits from higher prices for the drugs.
- Whether participants should be specifically consented for linkage is an open question. IRBs want consent forms to address three key types of data sharing: sharing of genetic data, sharing of data with for-profit organizations, and sharing of data with other researchers. Consent for linkage is inherent in agreeing to share data with other researchers.
- The risks and benefits of blanket permissions versus on-demand or dynamic consent for data sharing must be weighed. A dynamic consent process involves obtaining consent from individuals as research progresses and changes. This reaffirms the trust relationship between the registry as data stewards, keeps scientists honest, and keeps local stakeholder communities informed. Dynamic consent is resource intensive; however, blanket consent for research has been historically abused with respect to marginalized communities. Given the lack of trust in research among members of some underrepresented communities, a blanket permission approach may skew research data further toward those already well represented.³

Suggestions for Data Ecosystem

- Data ecosystems offer a channel to increase communication not only with potential data users but also with research participants. Many data repositories provide information on who uses the data and for what purpose, such that interactive tools could be developed for participants to see who has published using their data.
- Diverse cohorts in AD/ADRD data ecosystems do exist, although demographic information is not always transparent. The AD/ADRD research community should encourage intentional inclusion and transparent visualization of diversity metrics on AD/ADRD data platforms.

Social Determinants of Health

- The Area Deprivation Index (ADI) provides census-block level information for a set of socioeconomic indicators, is readily obtainable, and has been used as one measure of social determinants of health. It allows researchers to evaluate and use neighborhood-level data rather than race alone and provides an entry for research on social determinants of health.⁴

³ Following the workshop, Dr. Griswold noted that dynamic consent processes may further reduce representation from individuals from marginalized populations that are already under included in research studies.

⁴ Following the workshop, Dr. Griswold shared that social and structural determinants of health (SSDoH) have been shown to impact health as much as or more than factors like biology, behavior, and health care. He suggested that SSDoH should be more widely incorporated into research studies and made more accessible by data ecosystems.

Session 4: AD/ADRD Knowledge Environments

Current State of AD/ADRD Knowledge Environments

Session Chair: Maryann Martone, PhD, University of California at San Diego

In addition to the top-level principles of findable, accessible, interoperable, and reusable, the FAIR framework includes 15 nested subprinciples that provide a deeper understanding of what it means to be FAIR. For example, findability is linked to unique identifiers and machine-readable metadata, and interoperability requires resolvable vocabularies and ontologies. The FAIR framework also provides specific recommendations for how to achieve its goals.

Presenters in this session were asked to assess their respective AD/ADRD knowledge environments using a rubric based on FAIR principles. Many of these platforms predate the [release of the FAIR principles in 2016](#). Nonetheless, these knowledge environments have made clear progress in achieving many of the goals of FAIR, particularly with respect to cross-platform interoperability. AD-relevant common data elements and data dictionaries have helped build a foundation for this crosstalk.

There is an impressive array of data resources available to researchers from these knowledge environments; however, both human and technical challenges remain when working within and across these resources. Furthermore, trying to navigate this large pool of resources can lead to frustration among data users. Steps can and should be taken to make it easier for users to take advantage of the available data.

The overall goals of this session were to:

- Assess available resources with respect to FAIR.
- Consider use cases that will drive further progress on FAIR.
- Discuss ways to smooth user experiences in meaningful ways across ecosystems.

NIH-sponsored AD/ADRD Data Platforms

Eliezer Masliah, MD, NIA

Progress toward understanding the pathogenesis of AD/ADRD requires integrating knowledge of mechanisms of aging and genetics as well as the interactions of these mechanisms with the environment. This research is critical to achieve the goal of preventing and effectively treating AD by 2025, as set forth by the National Plan to Address Alzheimer's Disease. New drugs targeting amyloid- β demonstrate progress toward this goal, but more targets, drugs, and clinical trials are needed to prevent and treat the heterogeneous pathology of AD. NIA programs span the research continuum, from basic science through clinical trials. An array of enabling infrastructure platforms support these efforts.

The Alzheimer's Disease Sequencing Project (ADSP) is on track to generate 100,000 whole genomes by 2025, including data on diverse populations. To date, ADSP has identified more than 75 risk loci and 20 genes linked to AD through processes such as immune signaling, lysosomal function, and lipid metabolism. ADSP is supported by several infrastructure components, such as the NIA Genetics of Alzheimer's Disease Data Storage Site (NIAGADS), and includes several other research consortia such as the AI/ML Consortium, the Phenotype Harmonization Consortium, and the Functional Genomics

Meeting Summary

Consortium. All data generated by ADSP consortia and projects are deposited in NIA data repositories for use by the research community.

The national network of Alzheimer's Disease Research Centers (ADRCs) gathers fluid biomarker, neuroimaging, neuropathology, and clinical data. All ADRC data are stored in the National Alzheimer's Coordinating Center (NACC) while biospecimens are housed in the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD). NACC has coordinated extensively with NCRAD and NIAGADS, as well as the Standardized and Centralized Alzheimer's and Related Dementias Neuroimaging (SCAN) and Global Alzheimer's Association Interactive Network (GAAIN).

Recent research has identified biological markers for AD/ADRD in cerebrospinal fluid, imaging, and plasma. This work has been facilitated by publicly available data from NIA-supported longitudinal studies. One such study is the Alzheimer's Disease Neuroimaging Network (ADNI), which is supported by the Laboratory of Neuro Imaging (LONI). Other longitudinal studies include the Alzheimer's Prevention Initiative, the Dominantly Inherited Alzheimer Network (DIAN), and the Alzheimer Biomarkers Consortium-Down Syndrome (ABC-DS) focused on AD; the Advancing Research and Treatment for Frontotemporal Lobar Degeneration (ARTFL)-Longitudinal Evaluation of Familial Frontotemporal Dementia Subjects (LEFFTDS) Longitudinal Frontotemporal Lobar Degeneration (ALLFTD) study focused on frontotemporal dementia (FTD); the North American Prodromal Synucleinopathy (NAPS) Consortium focused on Lewy body disease; and Atherosclerosis Risk in Communities (ARIC) and the Vascular Contributions to Cognitive Impairment and Stroke in a Bi-racial National Cohort (REGARDS) focused on vascular dementia.

NIA has several programs focused on drug development, from target identification and validation through phase III clinical trials. NIH currently funds more than 400 clinical trials for AD/ADRD, including pharmacological, nonpharmacological, and neuropsychiatric trials. NIA also supports several systems/network biology resources, including Accelerating Medicines Partnership®-Alzheimer's Disease (AMP-AD), the Translational Center for Model Development and Evaluation for Late Onset Alzheimer's Disease (MODEL-AD), and others. Multi-omics data collected through these programs are stored in the AD Knowledge Portal.

NIA recognizes that it is critical that study participants are representative of the U.S. population and is working to prioritize diversity and inclusion in research. Recent funding for ADNI and the ADRCs is focused on obtaining data from more diverse populations. NIA also funded a large consortium in Texas to measure biomarkers in a Latino population.

NIA is working to link its four major data hubs—NIAGADS, NACC, LONI/ADNI, and the AD Knowledge Portal—to support research that will lead to precision medicine for AD/ADRD. This will be done in part through data harmonization. Application of FAIR principles will help achieve this.

AD Knowledge Portal: Enabling FAIR Data Infrastructure for Research on AD/ADRD

Anna Greenwood, PhD, Sage Bionetworks

The [AD Knowledge Portal](#) supports an array of NIA-funded AD data translational centers and systems biology consortia that generate large amounts of multi-omics data from human and model system studies. As the data coordination center for these consortia, Sage Bionetworks conducts two primary activities: dissemination of resources to the broader AD research community and support of consortium-

Meeting Summary

wide efforts through application of team science approaches. Data coordination comprises infrastructure and people—scientists, data engineers, governance analysts, software developers, and usability designers who perform the sociotechnical work required to coordinate the science of these large consortia.

The AD Knowledge Portal is the entry point into a larger infrastructure system underpinned by the Synapse platform, an NIH-designated cloud-based generalist repository that performs all data cataloging, storage, and access control for the consortia contributing data to the system. These tools support data discoverability, results exploration (Agora, Model AD explorer), and data processing and analysis (CAVATICA, cloud workspaces, and direct downloads). Contributors include 13 NIA programs, 105 projects, and 124 studies, with AMP-AD contributing the most data files to date. The portal contains specimens from a variety of species (e.g., human, mouse, and *Drosophila*) and multi-omics data types (e.g., genomics, transcriptomics, proteomics, metabolomics, metagenomics) and diverse levels of process (raw, processed, analytical outputs). In addition to data, the resource shares other information about these projects, including grant abstracts, methodological details, experimental models, and computational tools.

A streamlined user interface offers two steps to identify resources of interest: (1) using the Explore menu to select studies, data, or experimental models, and then (2) browsing and filtering. A controlled metadata vocabulary and annotations are provided to facilitate searching. Annotations of individual demographics and phenotype are being enhanced, which requires navigating governance concerns.

Between November 2018 and November 2022, more than 1,800 research teams downloaded data from the portal. More than 650 publications have cited the portal since March 2023. Data contributors authored 44% of these publications, while the other 56% of the publications represent secondary data use. In addition to a help site, user support includes regularly updated documentation, data release notes, a newsletter, a discussion forum, and hands-on workshops. Proposed portal improvements include a cohort builder, improved high-level navigation, full site-level search capability, linkage to compute platforms, and interoperability with AD repositories. Plans to enhance interoperability include connecting directly with other repositories, including NIAGADS and NACC proposed connections with CAVATICA and Terra.

An assessment of the AD Knowledge Portal system demonstrated that it meets criteria set forth for FAIR principles. However, maintaining FAIRness is an ongoing process as the resource continues to expand.

NIA Genetics of Alzheimer's Disease Data Storage Site

Li-San Wang, PhD, University of Pennsylvania

The [NIAGADS](#) mission is to serve as a continuing data resource for AD genetics/genomics research. As a designated data repository for sharing AD genetic data with the research community, NIAGADS harmonizes and curates new sample data into a comprehensive genomic resource for AD genetics discovery (with associated phenotypic data). In addition to individual-level raw data, NIAGADS has developed platforms for sharing genomic knowledge so investigators can view analysis findings, derived data, and annotations. In addition, NIAGADS is the coordinating center providing data production, data sharing, and user support for the ADSP.

Meeting Summary

NIAGADS houses and manages sharing of nearly 100 datasets (e.g., whole genome, whole exome, GWAS, analysis results, omics, genome-wide summary statistics) for over 172,000 samples from multiple cohorts with more than 200 Institutional Certifications defining the various research use limitations based on consent levels. Many of these data are only available through the collaborations with NACC and NCRAD.

ADSP is the flagship project instituted by NIA in 2012 to understand the genetic architecture of AD/ADRD. This project involves a large sample of participants from ethnically diverse populations and will capture and analyze common and rare genetic variation. ADSP components include the Follow-Up Study (FUS) 2.0: Diversity Initiative, gene discovery U01 Cooperative Agreements, the Phenotype Harmonization Consortium, Functional Genomics Consortium, and AI/ML Consortium. NIAGADS supports 14 work groups, 5 consortia, and 309 investigators at 49 institutions. The October 2022 ADSP data release includes 36,000 genomes (19,456 new in this release) from 40 cohorts and harmonized clinical data (cognition, fluid biomarker, neuropathology) for eight cohorts. The next release will include 23,128 new genomes from 12 new cohorts, bringing the total to nearly 60,000 genomes from 52 cohorts. All genomes are harmonized by joint calling.

A key goal of the ADSP FUS 2.0 initiative ([PAR-21-212](#)) is to increase the diversity of ADSP data. Achieving statistical significance for rare variants requires at least 18,500 cases and 18,500 controls. FUS 2.0 aims to reach these numbers for each of three populations: African, Hispanic, and Asian ancestry. That will bring the total of ADSP genomes to about 15,000 within the next 5 years, an ambitious goal involving cohorts from six continents. The ADSP [summary of sequencing progress](#) is updated regularly.

Researchers can apply for data access through the NIAGADS Data Sharing Service, which is aligned with the NIH Genome Data Sharing policy framework. Data sharing is done in compliance with initial informed consent as reflected in research use limitations. Some ADSP subjects would be difficult to reconsent. The platform is housed on AWS and has achieved FISMA moderate compliance appropriate for human subjects data.

The NIAGADS AD Genomics Knowledgebase comprises the GenomicsDB, a full-featured genomics database for AD genetics with genome browser visualization that enables users to see all the omics, analysis summaries, and p values for the region of interest; Functional gEnomics Repository (FILER), a comprehensive harmonized, searchable human functional genomics data collection from more than 20 data sources ideal for ML uses; and the AD Variants Portal, a harmonized collection of high-quality, suggestive genetic association findings curated from the literature. A NIAGADS API is in development for programmatic access to all NIAGADS AD Genetics and Genomics Knowledge bases.

A FAIR assessment of NIAGADS resources found that:

- Data are findable. Detailed dataset summaries and comprehensive metadata can be browsed and filtered online, a manifest of all files with MD5 checksum and file site ensures integrity, and users can link data via subject/sample IDs.
- Data are accessible. All data are on AWS for on-cloud analysis or secure download. Data access requests are reviewed by the NIA-chartered ADSP Data Access Committee made up of NIH program officers who are familiar with human subjects protection policy.

Meeting Summary

- Data are interoperable. Standard file formats are fully documented with references and contact. Data processing and quality control procedures and workflows are available.
- Data are reusable. Findings are returned to NIAGADS for re-sharing as derived data.

National Alzheimer's Coordinating Center

Sarah Biber, PhD, University of Washington

[NACC](#) is the data, collaboration, and communication hub for the 37 independent research and exploratory centers in the NIA ADRC Program. Over the past 23 years, NACC has collected and integrated a broad range of data from these centers and shared it with AD researchers around the world, compiling one of the largest, most comprehensive AD/ADRD longitudinal, standardized, clinical, and neuropathology datasets. Rich multidomain neurocognitive data are available for more than 47,000 participants in over 7,000 neuropathology datasets representing diverse and mixed etiologies. NACC data are known for robust, gold standard criteria-based diagnoses. Data utilization increased dramatically over the past year as new tools and initiatives were introduced. NACC data have enabled many impactful and innovative research discoveries, including in the areas of disease heterogeneity, early detection and differential diagnosis, and genomic analysis.

NACC is a FAIR resource. The consortium sets community-driven clinical research standards for AD/ADRD data. Sociodemographic, neuropathology, and neurocognitive test data are standardized prior to submission to NACC. Every few years, the Clinical Task Force, a group of leading experts, convenes to define the latest clinical research standards for the field, including co-developing tools to collect and ensure quality control for these data. The Uniform Data Set (UDS) is used by ADRCs and has been translated into dozens of other languages for use in international research studies.

A dramatic modernization of NACC is under way. The consortium is building new infrastructure that will place NACC on the leading edge of FAIR innovation, including an electronic data capture system powered by REDCap, a modern data platform that is easily scalable to new data streams and capable of linkages with other data repositories, and an advanced data search and access portal. Tools are being developed iteratively in collaboration with key stakeholders across the ADRD field with the goal of meeting the growing needs of the research community.

NACC is working to achieve a more seamless AD/ADRD research ecosystem. Recent accomplishments include moving NACC data to a new cloud-based data platform powered by Flywheel and building pipelines to integrate new data streams for electronic health record, digital biomarker, and digital neuropathology data. NACC is collaborating with other NIA repositories to expand interoperability between platforms and is a key partner on the SCAN initiative that collects standard magnetic resonance imaging (MRI) and positron emission tomography (PET) data from the ADRCs. NACC is linking SCAN metadata and analysis data with existing data for thousands of participants using the NACC ID and collaborating with SCAN and LONI to make these newly integrated data available to researchers. NACC has a long-standing partnership with NCRAD and NIAGADS, enabling NACC to share more genomic, genetic, and blood-based biomarker data. NACC is piloting work to build interoperability with Sage.

Linking these data streams together will enable researchers to ask and answer new questions. NACC is collaborating with other NIA repositories and AD/ADRD researchers to develop a one-stop shop for all ADRC data modalities called Data Front Door. This tool allows researchers to search and access a wider

Meeting Summary

range of standard data and metadata modalities for ADRC participants. The Data Front Door enables data sharing through several self-service components, including quick access files, a real-time dashboard, and an innovative cohort selection for building multimodal datasets.

A FAIR assessment of NACC found that:

- NACC data are findable. Standard metadata for all new data streams are being integrated, and NACC is exploring options for assigning persistent identifiers to ADRC participants to facilitate linkages with other NIH studies.
- NACC data are easily accessible via a web-based, self-service interface with a single sign-on. It currently takes about 15 minutes to request a quick access file and sign a Data Use Agreement; once completed, all data are provided within 24 to 48 hours at no cost. Other self-service data access options are being built to make it easier for researchers to locate and access data for ADRC participants across platforms. Ultimately, the Data Front Door will serve as the single data discovery dashboard for the ADRC program. Down the line, researchers can access, analyze, and build algorithms with this data using secure enclaves and tools within the NACC cloud.
- NACC data are interoperable in several important ways. NACC and the Clinical Task Force have set standards for data collection and quality across all ADRCs. NACC is transitioning to electronic data capture; this will streamline data submission and quality control processes and make it easier to use UDS forms in other studies. NACC aims to increase interoperability with the broader NIH ecosystem; for example, by adopting more NIH common data elements and open biomedical ontologies such as the National Library of Medicine's RxNorm, which will be used in the January 2024 UDS release.
- NACC data are reusable. NACC Data Use Agreements and robust data dictionaries will support increased multimodality. NACC handbooks and resident experts help researchers understand how to use NACC data for different types of analyses and hypotheses. New transparent quality control processes will be available via GitHub. The cohort selection tool is designed to enable reproducible cohort selection queries. NACC is exploring options for standardizing consent across the ADRCs, which will simplify the addition of new modalities in the future.

NACC is collaborating with the AD Knowledge Portal and others to conduct an interactive AD/ADRD Multi-Modal Data Workshop focused on identifying the best way to leverage multimodal data. They are seeking examples of compelling research in the AD/ADRC space and new hypotheses researchers want to test. Webinar information is available on the [NACC website](#).

Laboratory of Neuro Imaging

Arthur Toga, PhD University of Southern California

The [LONI](#) Image and Data Archive (IDA) currently houses 16.3 petabytes of data and over 9 billion files from about 160 studies comprising nearly 100,000 subjects from 164 countries. Roughly 45 million files are added per month. The data are regularly backed up to the cloud and multiple sites. There currently are about 150,000 active IDA users. In addition to images, IDA includes genetics, clinical, demographic, and other data. The repository is searchable and enables data mapping and harmonization. Rules of engagement for the data—including for data access and data use—are determined by each project for their data.

Meeting Summary

ADNI was one of the earliest studies to deposit data in LONI. From the beginning, ADNI data—including clinical genetic, and imaging data—were made available to researchers almost immediately. Project-specific dashboards provide information on the composition of the dataset; for example, the ADNI dashboard lists numbers of subjects with AD, mild cognitive impairment (MCI), or one of several categories of cognitive impairment. Applications for data and data downloads have been increasing for ADNI and other projects; to date, there have been more than 300 million downloads of ADNI data. More than 4,000 manuscripts using ADNI data have been submitted to IDA and are searchable via keywords.

LONI is collaborating with the ADSP Phenotype Harmonization Consortium and NACC to facilitate integration of ADNI and ADSP data. The consortium is working to harmonize data dictionaries and the images themselves. A set of common data elements across ADNI and NACC helps with interoperability. Use of common identifiers and enabling permission token systems to communicate makes integration of the platforms possible. Importantly, individual projects continue to control access to their data.

Another LONI partner is GAAIN, which includes 61 data partners and more than 560,000 subjects. Unlike many other LONI projects, GAAIN operates with a federated model, with most data housed outside of LONI. This allows the data partners to retain control over their data, giving them the ability disconnect from and reconnect to LONI as desired. A graphical interface for exploration, analysis, and visualization of GAAIN data is available through LONI. Similar interfaces are available for some other projects. For example, the Fox Data Exploration Network (Fox DEN) provides investigators with a tool to explore, download, and apply statistical models on aggregated Parkinson's disease (PD) data. The graphical interface with simple search fields and drag-and-drop tools makes it easy to create cohorts and perform statistical analyses. These graphical interfaces improve accessibility and utilization of the data.

Use of a universal identifier, such as a globally unique identifier (GUID), would help ensure that subjects are not duplicated when combining data across platforms.

NIH Center for Alzheimer's and Related Dementias—Cloud Forward Data Sharing: Update and "Limit Testing" with Long Reads

Mike Nalls, PhD, Data Tecnica International

Biomedical research has grown more complex over the years, progressing from studies of kilobyte-scale data on individual single nucleotide polymorphisms (SNPs) in the 1990s to current analyses of terabyte and petabyte sized imputed and whole genome sequencing datasets. Early studies of smaller datasets could be done on individual computers, but eventually cluster computing was needed. The emerging model of centralized cloud-based computing is making it easier to work with very large datasets.

Center for Alzheimer's and Related Dementias ([CARD](#)) uses cloud platforms to accelerate global collaborations and decrease community activation energy for research. Rather than build its own cloud-based platform, CARD elected to use existing platforms. Different platforms are used for different datasets. Selected platforms were required to meet criteria for safety, inclusivity, fairness, clarity, interoperability, flexibility, and accessibility. The goal was to use platforms that can provide compute resources at no cost or allow sponsorship to make it more accessible to researchers from around the world. Security and transparency help projects maintain control of their data and create accountability. A current focus is allowing tokens to move between platforms to allow for single sign-in.

Meeting Summary

As of June 2023, CARD is sharing multi-omics cell line data from the Induced Pluripotent Stem Cell Neurodegenerative Disease Initiative (iNDI) projects, imputed genomes and clinical data from several countries, proteomic biomarker data, and hundreds of terabytes of long-read sequencing data. Data in process include data from diverse longitudinal cohort studies of aging and neurodegeneration, cell imaging data, and single-cell and single-nucleus brain samples for AD/ADRD cases and controls. CARD also is preparing to share common data elements developed for neurodegenerative diseases and a schema for sample IDs.

One flagship CARD project is the Global Parkinson's Genetics Program (GP2), an Aligning Science Across Parkinson's (ASAP) project. ASAP brings together initiatives for PD, Lewy body dementia, FTD, and other neurodegenerative diseases to perform population-scale genotyping with harmonized clinical data across 150 sites; the goal is to have data on up to 0.5 million samples in the next 5 years. GP2 data are available through Terra. Users can take advantage of Terra workspaces and notebooks to analyze the data.

iNDI is engineering more than 130 genetic variants associated with Alzheimer's and related dementias into isogenic cell models of disease. Multi-omic data from these cell lines are being shared in near real time through the Alzheimer's Disease Workbench. Data can be analyzed for free thanks to subsidies from the Alzheimer's Disease Data Initiative.

CARD is generating long-read genetic sequencing data to study structural variation and other genetic architecture in AD/ADRD. Long-read sequencing data files are large (~1.5 terabytes/sample), and the processing pipeline for these data is complex. CARD is using a hybrid strategy to manage these data. Data are processed locally on Biowulf. Data are then moved to the Alzheimer's Disease Workbench and Terra through collaborations with the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative and AnVIL, respectively. Raw data are kept in cold storage to limit costs, while derived data are more easily accessible. CARD has developed a benchmarking app for long-read sequencing datasets to help manage resources for cloud-forward efforts.

AD/ADRD Knowledge Environments Panel Discussion

Nonhuman Data Resources

- The current workshop focused primarily on human data, but NIA recognizes the importance of model organisms and systems.
- The AD Knowledge Portal includes a significant and growing amount of data on nonhuman model systems, including mouse models.

Subject IDs

- Use of GUIDs has increased over the last few years, and many research groups have aligned with the NIA GUID system. Uniform adoption of the NIA system by the research community would improve the quality of the science.
- Data archivists cannot assign GUIDs. This must be done by the studies.
- Cross-linking subjects between cohorts can be done without GUIDs, although this is more labor intensive.

Meeting Summary

- NACC is working to become more FAIR in collaboration with partners at NIA and across the community. There are sociocultural, technical, and governance challenges associated with making changes, including adopting a new subject ID system. For NACC, any changes would need to be made in collaboration with all the ADRCs.

FAIR Standards for AD Research

- At present, there are no AD-specific FAIR standards across the research community. FAIR leaves a lot of room for interpretation, but when trying to integrate across resources, a shared understanding of common FAIR practices will help improve the operation of the overall ecosystem, particularly for those who want to reuse the data. Thus, identifying a set of standards or core elements would be useful to promote interoperability among the AD/ADRD resources, especially for any new resources that are developed. A community-led FAIR standard would not necessarily need to be mandatory to support more uniformity across studies; resources should be motivated to adhere to these standards because it will make the resource better.
- Standards are emerging from Europe and GA4GH, as well as across other platforms. Rather than starting from scratch, Sage Bionetworks is adopting existing standards as much as possible.

Global Searches

- NACC is working to create a home for ADRC data that is accessible and searchable across platforms. This will involve pulling in and incorporating metadata and other fields from external repositories to create a user-friendly one-stop shop for the data.
- CARD is developing the DIVER tool to enhance data discovery by functioning as a lightweight form-based FAIR browser and inventory. This streamlined application facilitates access by both intramural CARD researchers and visiting scientists to curated internal data. The tool also permits natural language queries of connected external datasets, which is enabled by large language models.

Session 5: Breakout Groups

Workshop participants split into breakout groups to discuss practical considerations for FAIR data ecosystems through consideration of two use cases. Each of the five breakout groups discussed the use cases from a different perspective: Data Security and Access, Data Linkage and Harmonization, Data Ecosystem Architectures, Promoting DEIA in Data Ecosystems, and Data Analysis Within Data Ecosystems.

Use Case #1

Identify the impact of sex differences on brain aging and AD/ADRD.

Sex differences have a major impact on the molecular trajectories of brain aging and AD/ADRD. All NIA data platforms contain data that are relevant to understanding these differences. However, in many cases, data from the same individuals are represented in a disparate manner across these platforms. What are the major barriers to investigating this topic within the NIA data ecosystem, and what tools would enable researchers to address these questions using data pooled from these resources?

Use Case #2

Understand the molecular processes that underlie healthy versus pathological aging and identify determinants of resilience to AD/ADRD.

To understand how to promote healthy aging, we need to understand the trajectory of healthy aging as well as determinants that promote transitions to pathological aging. These questions would be best addressed by combining data from NIA platforms with data residing in other places throughout the NIH data ecosystem. What are the major barriers to working with data from multiple sources within the NIH data ecosystem, and what tools would enable interoperability?

Breakout Group Reports

A representative of each breakout group provided a brief summary of the group's discussion.

Data Security and Access

The group identified several major challenges to addressing the proposed use cases with respect to data security and access. These include the complexity of sharing -omic data among repositories due to differences in consent groups and related restrictions; challenges linking patients across repositories; privacy concerns for facial images; and the need to harmonize data across repositories.

Potential solutions to address these challenges include the creation of a process to combine expertise across Data Access Committees to simplify approvals. Processes that allow approval of groups of researchers to access data from two or more repositories would promote collaborative research. Techniques to reface images are needed to reduce the risk of identification. Linkage of data across repositories would be bolstered by open-source technology for PPRL. The NIH Office of Science Policy (OSP) and ODSS could provide support for this. API-based services could be used to link data across repositories, and AI/ML or federated ML could be used to improve data harmonization across platforms.

Data Linkage and Harmonization

The group considered tools needed to assist in automated harmonization of data. Basic standards for metadata definitions are needed, as are common metadata models informed by domain experts. Graph models should be explored for data harmonization, and the research community should think about how additional data types will be integrated in the future (e.g., electronic health records, social determinants of health). There was discussion about the potential for generative AI to support harmonization and/or data linkage. Data readiness for generative AI is an important consideration; protocol-based harmonization prior to AI will be important, and AI-ready models should be created. Imaging data may provide the best opportunity for exploring AI for harmonization. Harmonized data and AI models should be returned to the data source for validation.

With respect to use case #1, the group discussed challenges linking data for an individual across platforms. Various factors—such as metadata/data linkage, harmonization, and data coding—should be considered through the lens of sex. It also would be important to determine how data were harmonized with respect to sex. Gender differences in disease may be confounded by social determinants of health or bias in human clinical studies; for example, there could be sex bias in cohort selection tools or sex differences in the scoring of AD pathology or cognitive measures.

With respect to use case #2, the group discussed the potential for collaborations with longitudinal cohorts that collect cognitive data or relevant brain imaging data. Potential collaborations with the Framingham Heart Study, *All of Us* Research Program, and Centers for Medicare & Medicaid Services

Meeting Summary

were suggested. Standards and structured vocabularies would facilitate interoperability between the data platforms. There also was discussion about what model systems may be useful for studying healthy aging and AD/ADRD; nonhuman primate models were suggested.

Data Ecosystem Architectures

The group reflected on the multidimensional nature of data ecosystem architecture and considered what idealized architecture would help address the use cases. Substantial overlap was noted in the architectural challenges associated with the use cases. For example, policies regarding data use and linkage are relevant in both scenarios. The need for standard ontologies to describe data ecosystem architectures and capabilities was discussed; this would make it easier to interface with platforms within and outside biomedical research. A major theme of the discussion was the realization that, regardless of the integral architecture of any given system, integration of data across platforms necessitates a federated approach. Thus, algorithms and other tools must be portable. Better standards are also needed to facilitate movement of cohort descriptions and other information between systems. Consideration must be given to the sustainability of data ecosystems; NIH STRIDES will be helpful in this regard.

The group also discussed emerging opportunities for transformational advancement. They concluded that interoperability should be driven by research needs and researcher-defined use cases; prioritizing interoperability efforts in this way will ensure the greatest possible gain with limited resources. There likely will be opportunities to harness the power of generative models to do some of the tedious work of linking datasets across federated systems.

Promoting DEIA in Data Ecosystems

The group identified several challenges regarding the use cases with respect to DEIA. These include the lack of transparency of data ecosystems as they relate to diversity and representation; lack of representation among research participants and research staff across racial/ethnic and sex/gender groups; inconsistencies and limitations in data reporting by race/ethnicity (e.g., broad categorization of race/ethnicity, and lack of social determinants of health relevant to underserved communities); lack of harmonization and standardization of sexual and gender minority variables, and other pertinent variables; and insufficient sample sizes for minority groups (particularly understudied populations such as intersex groups) to perform meaningful analyses and create generalizable results.

Potential solutions to these challenges include fostering trust and community engagement (through community-based participatory research efforts), which could help the collection of more diverse, equitable, and inclusive research data. The development of an ontology for sex as a biological variable that acknowledges that sex and gender are not synonymous would support studies such as the one outlined in use case #1. Researchers should use NIH common data elements (CDEs) for sex, gender, race, ethnicity, and other variables. Data ecosystems should be considered through the lens of DEIA when they are established and as they evolve to meet changing research needs. Metrics are needed to validate and promote DEIA within NIA data ecosystems (i.e., “DEIA metrics” to certify the data ecosystem with respect to DEIA). The collection of more lived experience data (e.g., social determinants of health) and use of PPRL to link data across ecosystems would also be helpful. Further, democratizing access and use of NIA data ecosystems by underrepresented researchers and students from low-resourced institutions or communities (e.g., minority-serving institutions) through training and concierge services should be prioritized.

Meeting Summary

The group discussed emerging opportunities for transformational advancement. Interactive visualizations of datasets and results will increase transparency of data ecosystems and engagement, including of underserved communities, and encourage participation in research. Underserved communities often prefer federated data models to centralized models because it allows them to maintain privacy and ownership. Researchers should explore how large language models can promote DEIA and combat inaccuracy and bias against underserved communities; one application may be to extract unstructured data on social determinants of health.

Data Analysis Within Data Ecosystems

Many nontechnical issues create challenges to integrated multimodal research. Many of these challenges start with study design and data collection (e.g., sample bias). Data quality is critical. It is important to understand the provenance of all data when doing analyses. Sampling must be well defined with consideration given to overcoming bias. Issues of governance, policy, and consent are relevant across both use cases and all research that draws on data stored in different places. With respect to use case #1, there are specific considerations about whether sex is listed as a variable across all repositories. For use case #2, researchers must be cognizant of heterogeneity in definitions and clinical presentation of AD/ADRD. Integration across resources would be particularly important for comparisons of healthy and pathological aging. It is difficult to obtain high-throughput, high-quality, harmonized molecular data, particularly for neurodegenerative diseases.

Potential solutions to the challenges discussed include containerized approaches that can be executed anywhere and use of searchable metadata across platforms. User-friendly training materials and opportunities are also critical to help users understand and feel comfortable with platforms and tools. Better harmonization of how data are governed across modalities would be helpful. Release of synthetic data sets across datasets may help improve data dictionaries while avoiding governance issues. Potential solutions specifically relevant to use case #2 include collection of more longitudinal data from healthy controls and creation of harmonized collections of longitudinal samples.

Emerging opportunities for transformational advancement include collaborations across NIH-funded cloud platforms, creation of tools to allow cross-repository searches (e.g., NACC Front Door), use of PPRL and GUIDs, large biobanking, data-driven approaches to inclusive recruitment, and linkage of research data with real-world data. There are opportunities to learn from the open data science community with respect to standardized tools and metadata.

Stimulating researchers to undertake cross- or multi-platform research requires funding, supportive NIH data science policies, accessibility of tools, training, and a better articulation of scientific priorities. Highlighting investigators and initiatives performing this type of research well could be helpful. The research community should not limit itself to research questions that are amenable to analysis within current data ecosystems. Some of the most important questions cannot easily be addressed with current systems. Data systems should be improved so they can support the most important research questions.

Closing Remarks

Workshop co-chairs Drs. Stanley Ahalt and Lara Mangravite shared their reflections on the proceedings. The NIH data ecosystem—including the AD/ADRD data ecosystem—supports data distribution and analysis across multiple research communities and dozens of platforms. Data resources, platforms, and sharing practices must be built to be inclusive, transparent, diverse, and culturally appropriate to ensure

Meeting Summary

an appropriate foundation for future scientific findings and medical recommendations. Data platforms are continuously being built to meet the evolving needs of the communities they serve; these resources reuse established components and are actively innovating new ones. Adherence to FAIR principles, modular architecture, and common standards for both content and platform architecture is necessary but not sufficient for interoperability. Engagement between platforms, driven by scientific need, must be actively promoted; this will require granular discussions to establish strategies to drive useful solutions.

Dr. Larkin thanked the workshop co-chairs, session chairs, and participants for contributing to the workshop. NIA appreciates the thoughtful discussion and looks forward to continued collaborative work on data ecosystems for AD/ADRD research.

Post-Workshop Discussion

Following the workshop, the workshop co-chairs, session chairs, and organizers met to reflect on the productive discussions and consider next steps. The large and growing number of data platforms represents a rich resource, but these platforms are currently underutilized in proportion to their value. Bringing platforms together to function as a collective would broaden their accessibility and utility. Achieving this goal will require a more thorough assessment of architecture gaps among the platforms as well as collaborative development of standards and policies. The continuing evolution of the technologies and architectures underlying the various data platforms poses a challenge, but there are opportunities to drive progress toward the goal of an integrated data ecosystem. For example, API alignment could enhance interoperability while adoption of GUIDs and utilization of large language models could facilitate cross-platform searches. Integration strategies should consider how to leverage the diversity of architectural models across current platforms and what lessons can be gleaned from other fields beyond AD/ADRD research. Importantly, data ecosystems should be accessible to researchers from different sectors, including academia as well as biotechnology and pharmaceutical companies. As a next step, linkages among a core set of AD/ADRD data resources could be enhanced to establish an ecosystem with a single entry point for users. A possible use case for this integrated system is the creation of a cross-platform cohort comprising subjects from populations underrepresented in research.

ACRONYMS

Acronym	Definition
ABC-DS	Alzheimer Biomarkers Consortium-Down Syndrome
ACE	Acetylcholinesterase
AD	Alzheimer's disease
ADI	Area Deprivation Index
ADRC	Alzheimer's Disease Research Center
ADRD	Alzheimer's disease-related dementia
ADSP	Alzheimer's Disease Sequencing Project
AI	Artificial intelligence
AI/AN	American Indian/Alaska Native
AIM-AHEAD	Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity program
ALLFTD	ARTFL-LEFFTDS Longitudinal Frontotemporal Lobar Degeneration
AMP-AD	Accelerating Medicines Partnership-Alzheimer's Disease
AnVIL	Analysis, Visualization, and Informatics Lab-space
API	Application programming interface
ARIC	Atherosclerosis Risk in Communities
ARPA-H	Advanced Research Projects Agency for Health
ARTFL	Advancing Research and Treatment for Frontotemporal Lobar Degeneration
AWS	Amazon Web Services
CARD	Center for Alzheimer's and Related Dementias
CARE	Collective benefit, authority to control, responsibility, ethics
CCDG	Centers for Common Disease Genomics
CDE	Common data element
CFDE	Common Fund Data Ecosystem
CRDC	Cancer Research Data Commons
DCC	CFDE Data Coordinating Center
DEIA	Diversity, equity, inclusion, accessibility
DIAN	Dominantly Inherited Alzheimer Network
DOI	Digital object identifier
ELSA	Ethical, legal, and social aspects
ELSI	Ethical, legal, and social implications
FAIR	Findability, accessibility, interoperability, and reusability
FHIR	Fast Healthcare Interoperability Resource
FILER	Functional gEnomics Repository
FISMA	Federal Information Security Management Act
FTD	Frontotemporal dementia
FUS 2.0	ADSP Follow-Up Study 2.0
GA4GH	Global Alliance for Genomics and Health
GUID	Globally unique identifier
GWAS	Genome-wide association study
HIPAA	Health Insurance Portability and Accountability Act of 1996

Acronyms

Acronym	Definition
IDA	LONI Image and Data Archive
iNDI	Induced Pluripotent Stem Cell Neurodegenerative Disease Initiative
LEFFTDS	Longitudinal Evaluation of Familial Frontotemporal Dementia Subjects
LONI	Laboratory of Neuro Imaging
MIDRC	Medical Imaging Data Resource Center
ML	Machine learning
MODEL-AD	Translational Center for Model Development and Evaluation for Late Onset Alzheimer's Disease
MRI	Magnetic resonance imaging
N3C	National COVID Cohort Collaborative
NACC	National Alzheimer's Coordinating Center
NAPS	North American Prodromal Synucleinopathy
NCI	National Cancer Institute
NCBI	National Center for Biotechnology Information
NCPI	NIH Cloud-Based Platforms Interoperability
NCRAD	National Centralized Repository for Alzheimer's Disease and Related Dementias
NHGRI	National Human Genome Research Institute
NHLBI	National Heart, Lung, and Blood Institute
NIA	National Institute on Aging
NIAGADS	NIA Genetics of Alzheimer's Disease Data Storage Site
NIH	National Institutes of Health
NIMHD	National Institute on Minority Health and Health Disparities
ODSS	NIH Office of Data Science Strategy
OPEN	Open, permissioned, entrusted, nonshared
OSP	NIH Office of Science Policy
PCORI	Patient-Centered Outcomes Research Institute
PET	Positron emission tomography
PIC-SURE	Patient-centered Information Commons: Standard Unification of Research Elements
PPRL	Privacy preserving record linkage
PROTECT	Puerto Rico Testsite for Exploring Contamination Threats study
RADx-UP	Rapid Acceleration of Diagnostics-Underserved Populations
RAS	Researcher Auth Service
REGARDS	Vascular Contributions to Cognitive Impairment and Stroke in a Bi-racial National Cohort
RENCI	Renaissance Computing Institute
SCAN	Standardized and Centralized Alzheimer's and Related Dementias Neuroimaging
SPRINT	Systolic Blood Pressure Intervention Trial
SSDoH	Social and structural determinants of health
STRIDES	NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability
TRUST	Transparency, responsibility, user focus, sustainability, technology
UDS	Uniform Data Set