NIH STRATEGIC PLAN FOR DATA SCIENCE 2023-2028

Introduction

Modern biomedical and behavioral science benefits from the fundamental transformation of basic biological and biomedical experiments and data science-enabled clinical studies that drive new discoveries. Data enable new opportunities for scientific inquiry; Hence, this updated National Institutes of Health (NIH) Strategic Plan for Data Science sets a bold vision for the future, one in which data generated in the course of care of individuals and data generated from biomedical and basic research become powerful inputs that enhance our understanding of fundamental biology and enable the development of new clinical treatments and diagnostic technologies. Data science includes genomics, transcriptomics, proteomics, metabolomics, imaging, and other data that underly basic biological experimentation. Data science also consists of clinical trial data; real-world data including electronic health data, wearable data and geospatial data, and health derived data; survey data; data from social and observational studies; and data on social and environmental determinants of health. The vision, articulated in this strategic plan, supports the NIH Policy for Data Management and Sharing[1] and embraces data-driven discovery as a powerful tool to elucidate biological processes, better characterize the health and health consequences of all people and fosters ethical use of new methodologies arising from artificial intelligence (AI) and machine learning (ML). Progress towards the promise of data-driven discovery requires a unified effort across the NIH Institutes, Centers, and Offices (ICOs) that is coordinated by and stimulated with the resources of the Office of Data Science Strategy (ODSS). To accomplish the goals set forth with this vision, NIH will address key challenges and outline opportunities relevant to:

- Generate and disseminate FAIR[2] Data in a manner that will foster greater sharing and add value to NIH research investments.
- Enact cost-effective strategies for sustainable, secure, and accessible biomedical data repositories and knowledgebases.
- Acquire and protect data obtained from electronic health records and other real-world data, including data captured outside of traditional health care settings, that preserves privacy and promotes participant consent.
- Promote emergence of innovations in trustable AI approaches that reduce bias and risks, and are FAIR, validated, and explainable.
- Create opportunities for exploration of new technologies and computing paradigms for biomedical research.
- Decrease disparities across institutions, regions, and global partners in data science.

In support of the NIH mission and the goals of the Department of Health and Human Services (HHS) to increase data sharing, modernize data infrastructure, and develop AI capacity, the 2023-2028 NIH Strategic Plan for Data Science articulates the NIH's strategic views, goals, and objectives to advance data science in the next five years. By addressing these challenges, NIH will pioneer robust data

[1] sharing.nih.gov/data-management-and-sharing-policy.

[2] FAIR data denotes Findable, Accessible, Interoperable, and Reusable datasets.

governance frameworks, ensuring data integrity, security, and accessibility, while promoting cross-disciplinary collaborations that accelerate scientific discovery.

The 2023-2028 NIH Strategic Plan for Data Science builds on accomplishments from significant collaborations of NIH ICOs under the initial NIH Strategic Plan for Data Science[3]. Experiences with public-private partnerships and alignment with activities across the Federal sector demonstrate that the NIH need not solve all data science challenges alone, but rather that it must ensure that solutions advanced in the private sector are sufficiently robust to be applied within the biomedical research enterprise. Advancing data science requires new partnerships including, but not limited to, health care delivery systems, private sector industries in technology and pharmaceuticals, non-profit patient representative groups and community partners, and other government agencies. This Strategic Plan will prepare NIH to face the acceleration of sophisticated new technologies and address the rapid rise in the quantity and diversity of data by accomplishing five overarching goals:

Goal 1: *Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing*

Goal 2: *Develop Programs to Enhance Human Derived Data for Research*

Goal 3*: Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence*

Goal 4: *Support for a Federated Biomedical Research Data Infrastructure*

Goal 5*: Strengthen a Broad Community in Data Science*

This document includes a summary of emerging opportunities and challenges facing NIH and delineates strategic objectives for each of the five goals. Associated with each strategic objective are suggested implementation tactics and evaluation schemes. Achievement of the vision set forth in this plan will position the NIH for accelerating discovery in biomedicine and health, mitigating health disparities, improving health equity through more relevant, comprehensive scientific findings, and achieving a workforce of researchers and clinicians sophisticated in the use of data science methods for discovery and care.

Emerging Opportunities for Biomedical and Behavioral Data Science

Significant advances in data science have been made since the initial NIH Strategic Plan for Data Science (**Appendix I**). For example, NIH has maintained data sharing policies for several decades and has taken a bold step forward with the final NIH Policy for Data Management and Sharing (DMS), which articulates the need to prospectively plan for how scientific data and accompanying metadata will be managed and shared. NIH defines metadata as information intended to make scientific data citable, interpretable, and reusable. NIH will continue to support data management and sharing capabilities that enable researchers to appropriately share data in ways that reduce barriers and overall cost. New capabilities and resources are needed that enable researchers to improve the automated collection of valuable metadata during the research process. These capabilities and resources should be consistent with community expectations and standards and should enable easier sharing of these data in appropriate

[3] datascience.nih.gov/strategicplan

repositories. Moreover, new opportunities and guidelines are needed to enhance trustworthy data repositories in a manner that aligns with global community expectations and contains open metrics that illustrate the impact of data sharing. Finally, developing new methodologies to allow for computational interoperability across data repositories and knowledgebases enables greater research from the underlying data.

Today there is potential to create federated networks that connect the billions of data points stored in electronic health records (EHRs), other real-world data such as wearable data, and clinical trial data obtained from medical systems and medical research institutions across the country. However, to maximize the potential of these data to discover new treatments and cures, there needs to be broad adoption of standardized data exchanges and integration. Through the Health Level Seven International (HL7®)[4] Fast Healthcare Interoperability Resources (FHIR®)[5] specification, certified health IT products will have standardized API capabilities to facilitate health data sharing. Leveraging and building on the FHIR® standard to exchange and share not only EHR data, but also phenotypic data obtained from clinical and genomics studies, clinical records and related social determinates of health data, and eventually other data from medical devices and wearable sensors, provides promising new avenues for clinical research. The ability to gather individual health data over time offers tremendous opportunities to accelerate research and medical breakthroughs and enable individualized preventions and treatments and is the vision of the NIH's *All of Us*[6] program. Recently the National COVID Cohort Collaborative (N3C)[7] illustrated the power of a collective data initiative. The N3C pulls data in 4 common models from 77 health systems, which represents more than 230 organizations. Data are harmonized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model on a weekly basis. N3C represents the largest de-identified limited datasets for COVID-19 research and uses privacy-preserving linkages to other Real-World Data (RWD), such as Centers for Medicare & Medicaid Services (CMS) and mortality data. Advances such as those seen in the *All of Us* and N3C programs require standardized vocabularies and ontologies that include communities from different areas of biomedical science and medicine. Experiences during COVID-19 also emphasized the importance of using and promoting Common Data Elements (CDEs), as was illustrated in the Rapid Acceleration of Diagnostics (RADx)[8] initiative which developed a core set of CDEs used across all RADx-funded projects. In addition, RADx's Mobile At-Home Reporting through Standards (MARS)[9] program established a core set of CDEs and a common HL7® specification to facilitate standardized public health reporting of at-home COVID-19 test results. CDEs continue to lack standardized semantics and ontologies. As a goal, this updated Strategic Plan for Data Science advocates for the creation of minimal sets or core CDEs, enabled by creating standardized concepts with allowable responses and data representations, that would enable and broaden the use of clinical and health data.

Another challenge for the data science research community is leveraging the massive datasets derived from the same individual across multiple data repositories and resources in a way that preserves participant identity and their intent for sharing. The situation is further complicated by inclusion of time-

[4] www.hl7.org/

[5] datascience.nih.gov/clinical-informatics

[6] allofus.nih.gov/

[7] ncats.nih.gov/n3c

[8] www.nih.gov/research-training/medical-research-initiatives/radx

[9] www.nibib.nih.gov/covid-19/radx-tech-program/mars

dependent participant data collection methods and requires data linkages. This can only be accomplished if these data have compatible standards and data models. Addressing these challenges requires new governance policies for data linkage and approaches to ensure participants' autonomy is respected. Technical capabilities are needed as well to support data harmonization and aggregation across different sources, including new methodology for collecting, integrating, and sharing social and environmental determinants of health data. These challenges open the door for new algorithms that incorporate secure data governance and participant consent, including privacy preserving computing, generative AI, foundation models, and blockchain methods.

Machine learning, deep learning, and AI technologies hold significant opportunities to advance basic and clinical research and to improve health and health care at individual and community levels. Recognizing these opportunities, NIH launched the Bridge to Artificial Intelligence[10] (Bridge2AI) program in 2022 to produce new flagship biomedical and behavioral datasets that adhere to FAIR principles and integrate ethical considerations (**see textbox for Bridge2AI**). Creating AI-ready data requires the necessary tools to collect FAIR-data at the beginning of the research process (FAIR by design-intentional integration of FAIR principles from the beginning of the data lifecycle) and methods that are complementary to the Collective benefit, Authority to control, Responsibility, and Ethics (CARE) principles[11] for Indigenous Data sovereignty and data governance. New capabilities to facilitate tribal data sovereignty that respects cultural needs and expectations are needed.

Bridge to Artificial Intelligence (Bridge2AI) will propel biomedical research forward by supporting widespread adoption of AI that tackles complex biomedical challenges beyond human intuition.

1. Generate new flagship biomedical and behavioral data sets
2. Develop software and standards to unify data attributes
3. Create automated tools to accelerate the creation of FAIR and ethically sourced data sets
4. Provide resources to disseminate data, ethical principles, tools, and best practices
5. Create training that bridges the AI, biomedical, and behavioral research communities

Efforts are also ongoing to utilize cloud service providers for data storage and management and to create interoperable data systems, efforts to enhance biomedical-artificial intelligence for ethical and unbiased data and algorithms[12] [13], and efforts to create tools that enable researchers to collect, find, and utilize FAIR data and software. Through the STRIDES[14] initiative, the NIH partnership with cloud service providers Amazon Web Services (AWS), Google Cloud Services (GCP), and Microsoft Azure has resulted in significant increases in data storage, data access, and the use of computational data platforms. Many NIH ICOs have leveraged STRIDES and have created cloud-based data repositories. However, much of these data remain siloed and are not utilized to their fullest potential. Addressing this challenge will require NIH to leverage a modern, federated data architecture approach. This approach will enable NIH to create cost-effective and sustainable practices that are tailored to the needs of

[10] commonfund.nih.gov/bridge2ai
[11] www.gida-global.org/care
[12] www.nimhd.nih.gov/resources/schare/
[13] ncats.nih.gov/funding/challenges/bias-detection-tools-in-health-care
[14] datascience.nih.gov/strides

individual ICOs and will allow researchers to take full advantage of biomedical data in the cloud with shared scientific analysis capabilities at unprecedented scales.

Today, technological innovations in AI and new capabilities to optimize large language models have generated considerable interest in the possibility of AI to recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets. Yet, challenges remain in creating transparent and explainable datasets and models. The need for ethical principles and frameworks, and connecting those principles/frameworks to practice, for developing and using AI in biomedical research remains an important priority and an unmet need. New paradigms in data discovery and knowledge generation[15] that utilize the integrative power of foundational models would enable researchers and citizen scientists to explore and use data to address complex questions involving diverse and heterogeneous datasets.

Beyond AI, other technological breakthroughs are emerging including advances in quantum computing, quantum sensing, privacy enhancing computing such as federated learning, and privacy preserving data sharing such as blockchain. Other applications are emerging in scientific fields such as physics, engineering, computer science, and in other industries from manufacturing to finance. However, at the intersection of these emerging technologies and biomedical research remains a relatively less-explored area. Expanding NIH investments in these emerging technologies will better position the biomedical research community and the agency to take full advantage of these and other new capabilities.

To make progress in the next five years, NIH must leverage the exponential growth in the amount and variety of data by developing and using sophisticated approaches to data management and embracing new technologies, including new methods in data reduction and compression for downstream analysis. Cloud computing will continue to play a significant role in the ability to share and utilize scientific workflows at an unprecedented scale. For example, the National Library of Medicine's (NLM) entire compendium of the Sequence Read Archive database is available on Amazon Web Services (AWS) through the Open Data Sponsorship Program. As a result, researchers can search and align over 19 million diverse samples (16.8 petabytes) to answer questions about evolutionary and comparative biology. By taking advantage of current and emerging data and computing technologies from the commercial and public sector, NIH could realize exabyte scale data science in the next decade.

Plan Content and Implementation

This updated Strategic Plan for Data Science is organized into five overarching goals, corresponding to strategic objectives and implementation tactics. These goals will ensure that NIH's strategic approach will address concerns to reduce unintended biases, protect participant privacy, and increase transparency in AI while collectively improving data and tools for research. This strategic plan also supports NIH's Policy for DMS by developing new capabilities to streamline data access with renewed emphases on metadata consistency and accuracy, use of CDEs, and support for utilizing community driven schemas and ontologies to enable data discovery across collections and repositories. This strategic plan also addresses data science gaps in the intramural research program and encourages increased and targeted collaborations to realize sharable opportunities for data and software. New to this strategic plan are objectives to enhance NIH's ability to leverage AI/ML technologies for biomedical

[15] datascience.nih.gov/sites/default/files/NIH%20Search%20Workshop%20Summary%20Final.pdf

and behavioral research, enhance clinical and health care data for research, and integrate policy, ethics, and health equity into its visions and objectives.

The implementation tactics are a roadmap for how the overarching goals and strategic objectives will be achieved. Details of these implementation tactics will be determined by the NIH Associate Director for Data Science in collaboration with working groups established by the NIH Scientific Data Council (SDC) and NIH Data Science Policy Council (DSPC), in consultation with the NIH ICOs, other federal and international agencies, the research community, the private sector, and other key stakeholder groups. NIH will continually assess and adjust these priorities based on the needs of NIH and its stakeholders and new opportunities in response to new technologies and capabilities.

Through implementation of this strategic plan during the next five years, NIH will:

- Develop new programs to support innovative approaches to data curation, harmonization, and validation and increase support for communities to develop and implement new CDEs and standards in priority disease areas.
- Increase support for research on clinical and health care data science, including new methods for privacy protection, participant informed consent, and data governance.
- Increase support for developing tools to collect and analyze data from wearable devices and other new RWD technologies.
- Develop new research, training programs, and collaborations in AI and Bioethics.
- Provide new ways for researchers to search, discover, access, and analyze data across resources and enhance the accuracy, validity, transparency, and reproducibility of these capabilities.
- Engage researchers and communities in data science training across biomedical, social, environmental, and behavioral disciplines.

This strategic plan aligns with the recently released document ***Digital NIH**: Innovation, Technology, and Computation for the Future*[16]. Digital NIH proposes new approaches to manage and govern NIH technology investments; describes a framework to guide implementation of high-priority, high value capabilities; and identifies cross-cutting capabilities that will support data science within NIH (see textbox **Digital NIH**).

This strategic plan will provide direction to encourage greater integration of data science to improve access to and use of biomedical and behavioral data that supports strong ethical, transparent, and anti-bias frameworks. These efforts will increase the scientific community's ability to address new challenges in accessing, managing, analyzing, integrating, and making reusable the huge amounts of data being generated by the biomedical research enterprise.

Digital NIH identifies new governance and funding approaches as well as capabilities organized by four functional areas: Extramural Research Management, Intramural Basic Research, Intramural Clinical Research, and Administration and Management. **Digital NIH is** supported by efforts in the five cross-cutting themes:

- A common architecture with well-defined standards to enable integration
- Innovative, cutting-edge storage, analytics, and computational infrastructure
- Increased technical competency of the workforce at all levels
- Technology to support an anywhere, anytime workplace of the future
- Risk-based, embedded cybersecurity protections

[16] ocio.nih.gov/Documents/Digital%20NIH%20Strategy_2023.02.06_Final_508C.pdf

Overarching Goals, Strategic Objectives, and Implementation Tactics

GOAL 1

Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing

NIH has a long-standing commitment to data management and sharing across two decades of policies with the goal to create and support a data sharing culture. For example, the final NIH Data Management and Sharing Policy emphasizes the importance of good data management practices and encourages data management and data sharing that reflect practices within research communities. Data management and sharing should reflect practices consistent with FAIR principles to be most beneficial. NIH-supported and NIH-managed repositories are the building blocks of the NIH data ecosystem and one of the primary mechanisms by which NIH makes the results of federally funded data available to the research community and the public. Federally funded data repositories should adopt the Office of Science, Technology, and Policy (OSTP) Desirable Characteristics of Data Repositories[17] and should align with community standards such as the Transparency, Responsibility, User focus, Sustainability, and Technology (TRUST) principles.[18] The TRUST principles provide a framework for formalizing the capabilities of a repository to efficiently serve its intended scientific community. Together the FAIR, CARE, and TRUST principles, the National Science and Technology Council (NSTC) Guidance provides a framework for formalizing the capabilities of a repository (see textbox **NSTC Desired Characteristics of Data Repositories**). NSTC ensures that federal investment in research that results in scientific data is accessible to accelerate biomedical discoveries, advance human health, and maximize America's return in dollars invested in scientific research. NIH will continue to promote and support researchers' ability to comply with the NIH Policy for Data Management and Sharing expectations by providing resources and guidance to researchers.

NSTC Desired Characteristics of Data Repositories are designed to be relevant to all repositories that manage and share data resulting from Federally funded research. The characteristics are organized across three themes:

- Organizational Infrastructure
- Digital Object Management
- Technology

In addition, additional characteristics for repositories storing human data must be able to address privacy protections, confidentiality, and security.

NIH will also develop new frameworks to handle the needs of modern data science challenges. For example, NCATS is developing the Maintainable, Observing, Securing, and Timing (MOST) framework to augment the FAIR and CARE principles. The MOST framework is a new paradigm for management of data "in use" that emphasizes the importance of maintainable data infrastructure and policies, observing and understanding data as it is generated. This ensures data security compliance during data ingestion, validation, and utilization. This framework has been used to guide several of the largest initiatives such as the Rare Diseases Clinical Research Network (RDCRN)[19], National Covid Cohort

[17] www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf
[18] Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* **7**, 144 (2020).
[19] www.rarediseasesnetwork.org

Collaborative (N3C), and A Specialized Platform for Innovative Research Exploration (ASPIRE)[20]. In addition, NIH will promote data repository interoperability. NIH seeks to create a FAIR-enabled data ecosystem that will break down data silos and promote greater findability and accessibility of data, thereby preventing unnecessary duplication of efforts and maximizing NIH investments.

Objective 1-1: Support the Biomedical Community to Manage, Share, and Sustain Data

The NIH Policy for Data Management and Sharing established requirements that emphasize the importance of good data management practices. It also established the expectations for maximizing the appropriate sharing of scientific data generated from NIH-funded or -conducted research, with justified limitations or exceptions. This policy applies to research funded or conducted by NIH-supported researchers that results in the generation of scientific data. By requiring researchers to anticipate their needs for managing and sharing scientific data, NIH will ensure that researchers develop data management and sharing plans that include where and how their scientific data will be shared and any anticipated limitations. This forward-thinking policy integrates data management and sharing into the routine conduct of a scientific project, and in the process, NIH aims to shift the biomedical research culture into one in which data sharing and data reuse are the rule rather than the exception.[21] NIH is developing resources to support the Data Management and Sharing Policy compliance activities of their funded investigators, including the NIH Office of Extramural Research (OER) sharing website,[22] the NIH Biomedical Informatics Coordinating Committee's portal to key data repositories,[23] and the National Institute of Child Health and Human Development's (NICHD) Data Repository Finder to support the development of data management and sharing plans. Supporting the NIH Policy for Data Management and Sharing requires a coordinated effort between the OSP, OER, Office of Intramural Research, ODSS, and other NIH ICOs. As such, NIH will establish and enhance guidelines, processes, data sharing tools and training in data management and sharing and will explore funding and governance models to ensure a sustainable NIH data sharing infrastructure, for researchers, NIH staff, and for data stewards and librarians, including those at low resourced institutions.

Implementation Tactics

- Strengthen the core data management competencies of researchers, data stewards, data librarians, and NIH program and grants management officers with tools and training:
  - For **researchers**: core data management and sharing competencies.
  - For **data stewards[24] and data librarians**: promote and enhance FAIR data sharing at their institutions.
  - For **NIH staff**: evaluate and improve data management and sharing practices and plans.
- Enhance programs that provide for credit and incentives for sharing data, including working with publishers, academic institutions, and other funding organizations and agencies.
  Develop metrics to measure data sharing, reuse, and impact.

[20] ncats.nih.gov/aspire

[21] Jorgenson LA, Wolinetz CD, Collins FS. Incentivizing a New Culture of Data Stewardship: The NIH Policy for Data Management and Sharing. *JAMA.* 2021;326(22):2259–2260

[22] sharing.nih.gov

[23] www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

[24] For this document, data stewards provide guidance on data quality, inclusion of appropriate metadata, appropriate data standard usage, data governance and limitations of data use.

- Establish a **data steward program** to guide data sharing and leverage existing activities at the NNLM National Center for Data Services and support additional partnerships, including with societies and associations, for training.
- Support tools that will assist researchers with the process of preparing, annotating, and sharing their data.

Objective 1-2: Enhance FAIR Data and Greater Data Harmonization

Through enhanced data sharing efforts and the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative, NIH-supported investigators have generated and made available over 200 Petabytes of data in the cloud. This represents a significant amount of biological data including genomics data, clinical study data, phenotypic and 'omics data, fitness measures and survey data, and data derived from electronic health care systems and social determinants of health. These data are most valuable when researchers can combine datasets to answer challenging questions such as "What are the physiological changes evident in long COVID?" or "What is the relationship between obesity and diabetes in populations with inadequate health care and how does this differ across geographic areas?" To address questions such as these requires aggregation and comparison of datasets with supported efforts in the standardization of collections, formats, and data models/data dictionaries. A key to success in data interoperability is the development and use of agreed upon data standards, standardized terminologies, and CDEs. The NIH Common Data Elements Task Force includes a governance committee that reviews and provides NIH endorsements to submitted CDEs, which are then deposited into the NIH CDE Repository,[25] hosted by the NLM, for use by researchers to help share and combine datasets. Other resources for CDEs include the National Cancer Institute's (NCI) Enterprise Vocabulary Services (EVS)[26] and the cancer Data Standards Registry and Repository (caDSR).[27]

HEAL Common Data Element (CDE) Program

The NIH HEAL Initiative research portfolio spans a broad array of data types that are a rich resource for future studies. The NIH HEAL Initiative's CDE Program supports the initiative's Public Access and Data Sharing policy, which requires researchers to develop plans to share their project's underlying primary data through a repository that is appropriate for the data type and research discipline, and will connect and expose data via the HEAL Platform.

To facilitate cross-study comparisons and improve the interpretability of findings, clinical pain research grantees collaborate and agree to use common data elements across nine core pain domains for patient-reported outcomes (PROs).

During the past five years a number of large NIH programs have undertaken an effort to enhance data harmonization, including Helping End Addiction Long-Term[28] (**see text box HEAL**) that is standardizing metadata; NCI's Cancer Data Aggregator[29] that is mapping harmonized data elements to Fast Healthcare Interoperability Resources (FHIR), Observational Medical Outcomes Partnership (OMOP), and other data models; National Institute of Dental and Craniofacial Research's (NIDCR) FaceBase[30] that is providing guides to scientists to produce

[25] cde.nlm.nih.gov/home
[26] evs.nci.nih.gov/
[27] datascience.cancer.gov/resources/metadata
[28] heal.nih.gov/
[29] datacommons.cancer.gov/cancer-data-aggregator
[30] www.facebase.org/

harmonizable metadata/data; and other large programs such as *All of Us*, which is harmonizing clinical data to the OMOP data model. In addition, community efforts such as the Minimal Common Oncology Data Elements (mCODE)[31] provides an agreed upon data standard that can be widely adopted and can increase high-quality data for all cancer types. International funders, such as the International Alliance for Mental Health Research Funders, have established a community of funders, medical journals and data measurement experts that are committed to adopting an agreed upon set of common measures for mental health science.[32] NIH applauds these efforts to support the development of standardized outcome measures in basic and clinical research. Creating standardized outcome measures, when appropriate, will allow for unbiased analysis, interpretation, and reporting of results. These standardized measures of behavioral and health outcomes facilitate cross-study comparisons and improve the interpretability and reporting of research findings and translation into evidence-based clinical practice but will need to be balanced with the flexibility that enables innovative clinical research.

To further enhance the data ecosystem, NIH will encourage the use of community agreed upon standard schemas and metadata, enhance automated ontologies and automated curation processes, and create capabilities for greater data discovery and interoperability across data multiple data repositories and knowledgebases. This objective will inform future activities under the NIH Plan for Public Access to Research Results.

Implementation Tactics

- Enhance abilities to improve data and metadata quality, including Data QA/QC.
- Encourage usage of open and standardized schemas, ontologies, and data formats.
- Enhance automated processes for ontology use and for enhanced data curation including enrichment of metadata.
- Create a minimal set of consistent and computable common data elements, or concepts, with consistent data models.

Objective 1-3: Strengthen NIH's Data Repository and Knowledgebase Ecosystem

Data repositories and knowledgebases are essential to increasing the information value of the scientific research enterprise and serve as important components of the implementation of the Policy for NIH Data Management and Sharing for preserving, archiving, and sharing scientific data. As the size and diversity of data collected and stored from biomedical research continues to increase and we transition to a modernized data ecosystem, the need for making scientific research data and information FAIR and the important role of repositories and knowledgebase in bringing this to fruition is even more evident. As articulated in the first Strategic Plan for Data Science, NIH makes a distinction between data repositories and knowledgebases as follows:

> **Biomedical data repositories** accept the submission of relevant data from the research community and store, organize, validate, archive, preserve, and distribute data in compliance with the FAIR data principles. Curation focuses on quality assurance and quality control.

[31] health.mitre.org/mcode
[32] iamhrf.org/projects/driving-adoption-common-measures

Biomedical knowledgebases extract, accumulate, organize, annotate, and link the growing body of information that is related to and relies on core datasets. Curation of information is often required in knowledgebases.

NIH has separately supported data repositories and knowledgebases as valuable assets and recognizes that these unique resources require funding mechanisms and review panels tuned to the needs of data science resources. Data resources and good data management practices are the key to data and knowledge discovery, data integration, and data reuse. To sustain a healthy and productive data resource ecosystem, it is critical to ensure that data repositories and knowledgebases:

- Deliver scientific impact to the communities that they serve.
- Employ and promote good data management practices and efficient operation for quality and services.
- Engage with the user community and continuously address their needs.
- Implement, adopt, or contribute to openly shared metric.
- Provide sufficient metadata and semantic annotation.
- Supports a process for data life-cycle analysis, long-term preservation, and trustworthy governance.

NIH supports both unrestricted access and controlled access data repositories. Controlled access data repositories manage and share research participant data, at the individual or aggregate/cohort-level, in order to respect research participants' privacy and autonomy. Controlled access datasets often have data use limitations requiring NIH authorization for data access, which, although necessary, can pose challenges to accessibility by the research community and access to controlled data in a timely fashion. Controlled access processes are currently labor- and resource-intensive, which limits their scalability. To accelerate research, and to maintain participants' data protections, NIH seeks to streamline and semi-automate controlled access processes by developing, testing, and deploying the use of emerging technological advancements where feasible and appropriate. In addition, NIH seeks to develop common approaches and infrastructure for addressing data management incidents across NIH supported repositories.

Generalist Repository Ecosystem Initiative is a collaborative effort with Dataverse, Dryad, Figshare, OSF, Mendeley Data, Vivli, and Zenodo to:

- Establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure
- Raise general awareness and help researchers to adopt FAIR principles to better share and reuse data

The aim of the GREI initiative is to establish consistent metadata, develop use cases for data sharing, train and educate researchers on FAIR data and the importance of sharing.

In recent years, NIH has developed a number of capabilities to promote a data ecosystem, including a collaborative approach for data management and sharing with seven generalist repositories (see textbox **Generalist Repository Ecosystem Initiative**) and support for the use of persistent unique data identifiers through a consortium membership with DataCite.[33] By partnering with DataCite, NIH data

[33] datascience.nih.gov/news/nih-joins-datacite-consortium

resources will be able to enhance data sharing and enable researchers to cite and reuse research outputs. These efforts aim to strengthen data management and sharing by enhancing data visibility, data citation in scholarly publications, data preservation, future data reuse, and data access.

As the size and diversity of data collected and stored from biomedical and behavioral research continues to grow and NIH enhances its modernized data ecosystem, the need for making these research data and information FAIR underscores the important role of data repositories and knowledgebases. Moreover, to maintain the scientific value of data, repositories and knowledgebases are increasingly required to embrace trustworthy principles. The recently formulated TRUST principles provide a framework for formalizing the capabilities of a repository to efficiently serve the intended scientific community. Developing sustainable data resources requires an understanding and use of metrics for evaluating the usage, utility, and impact of a given repository. Moreover, promoting equitable access to research products with appropriate security controls, privacy protections, including human subjects' protections, as outlined in the "Desirable Characteristics of Data Repositories for Federally Funded Research" will continue to be central to the NIH goals.

As important in adopting FAIR principles are the principles for the governance of data generated by or specific to American Indians and Alaska Natives (Indigenous data). Indigenous data are intrinsic to Indigenous Peoples' capacity and capability to realize their human rights and reflect the crucial role of data in advancing Indigenous innovation and self-determination. The CARE Principles (Collective benefit, Authority to control, Responsibility, and Ethics) outline goals for Indigenous Data Governance that reaffirm the principles of Indigenous self-governance and self-determination. As a first step, NIH developed supplemental information to the DMS Policy on "Responsible Management and Sharing of American Indian/Alaska Native Participant Data"[34] as a result of Tribal consultation. Similarly, international data sharing, especially involving data generated in low- and middle-income countries, should be respectful of regional and population-specific data governance considerations.

The ubiquitous use of data resources in biomedical research, coupled with a greater emphasis on data management and sharing, has greatly amplified the need for NIH to ensure the stability and robustness of widely used data resources. Over the last five years, NIH has made advances in understanding its portfolio of supported data resources (i.e., data repositories and knowledgebases), but this has also revealed their vulnerabilities with respect to long-term support – especially in light of their growing size, complexity, and demands from the research community. With growing concerns about the sustainability of data resources, NIH aims to articulate a coherent framework for their long-term support.

The challenges that NIH faces with respect to the support of widely used data resources are mirrored at the federal and international level. NIH provides the largest amount of support for the most widely used biomedical data resources, with resources managed by NLM serving as an important node in the international biomedical data ecosystem. For this reason, NIH has been involved in a number of efforts including CoreTrustSeal,[35] Research Data Alliance,[36] DataOne,[37] Open Science Framework,[38] DataCite,[39] the Wellcome Trust,[40] and the Global Biodata Coalition.[41] In addition, for more than 30 years, NLM has

[34] grants.nih.gov/grants/guide/notice-files/not-od-22-214.html
[35] www.coretrustseal.org
[36] www.rd-alliance.org
[37] www.dataone.org
[38] osf.io
[39] datacite.org
[40] wellcome.org

worked globally to preserve data and enable broad data sharing by coordinating with critical resources such as those comprising the International Nucleotide Sequence Database Collaboration,[42] and continuing to develop relationships with important global actors, such as the World Health Organization. These organizations provide a platform for the international community to work together to better coordinate the management and sharing of scientific data. Over the next five years, NIH will work closely with these and new efforts to help ensure the long-term sustainability of the global biodata ecosystem that is relied upon by NIH-funded and all other biomedical researchers worldwide.

Implementation Tactics:

- Enhance data repositories and knowledgebases that promote equitable access to all in alignment with the OSTP memo about Desirable Characteristics of Data Repositories for Federally Funded Research.
- Enhance FAIR, CARE, and TRUST capabilities that ensure secure and effective data management and promote data governance and data sovereignty.
- Support methods and programs with tribal communities to develop tribal data governance and sharing that recognize tribal rights in data.
- Promote shared data management practices, utilize open metrics for impact including enhancing data citation practices, and provide guidance on data preservation and long-term data archiving.
- Develop a comprehensive, coherent, and acceptable sustainability framework for identifying and supporting the portfolio of the most widely used and impactful NIH data resources.
- Develop a single policy framework that governs controlled data access repositories and standardized language for institutions and researchers.
- Streamline controlled data access processes across NIH repositories, including greater use of automation.
- Develop a common approach and infrastructure for addressing data management incidents across controlled access data repositories.
- Develop a single approach to help investigators find and appraise the relevance of controlled access data in NIH repositories, which enable meta data sharing.
- Enhance the visibility and use of NIH intramural research datasets and data resources.
- Develop methods to promote computational interoperability across data repositories and knowledgebases.

Goal 1: Partnerships and Measuring Progress

Potential measures of progress for this goal include data-resource key performance indicators for both data resources and for individual datasets, quantity and interoperability of databases and knowledgebases, quantity and citations of datasets deposited (over baseline), ability to find datasets across multiple resources, and data lifecycle FAQs. NIH will support and engage in partnership and collaboration across multiple stakeholders including Research Data Alliance, GO-FAIR, biomedical societies, and international partnerships such as with GA4GH, ELIXIR, and Global Biodata Coalition.

[41] globalbiodata.org/
[42] www.insdc.org/

GOAL 2

Develop Programs to Enhance Human Derived Data for Research

Data discoveries that aim to improve human health and underpin new treatments require a wide range of participant data including clinical data, gathered for the broad purpose of clinical research: health care data including medical history, records, and information that is necessary for care and treatment of patients, enhanced through linkages to social determinants of health (SDoH) and environmental determinants of health (EDoH) data. During the last decade, the United States has seen an increase in the generation and usage of these data in research including through efforts such as *All of Us*. These efforts are enhanced by large scale data collection and curation that utilize agreed upon common standards and data models. While progress has been made, integration of multiple types of real-world data with other data sources remains a challenge because the interpretation of health care-derived data for research purposes is highly dependent on the context of the interactions between patients, their health care providers, and their health environment.

To enable the biomedical and behavioral research community to take full advantage of the multitude of health-derived data requires the adoption and integration of health care data standards with research data standards. NIH will work with federal agencies, medical institutions, and health IT developers and vendors, where appropriate, to bridge the technology or data gaps between health care settings and clinical research. To enable researchers to gather and integrate data of interest to address health related questions, NIH will improve access to data repositories that hold participant-derived data and will enhance abilities to link real-world data from multiple sources, with appropriate informed consents from the participants. NIH will support approaches to leverage or build on existing programs, bring new partnerships together to enhance clinical data science, and support cross-training between clinician researchers, data scientists, and other technical experts/stakeholders. A major goal is to increase the use and utility of health care-derived data for research, with proper security and privacy safeguards. To achieve this goal, activities that integrate clinical data and real-world data including data from wearables and data originating from health care settings such as mental health, dental, pathology, and ophthalmology settings should be developed.

Objective 2-1: Improve access to and use of clinical and Real-World Data

The health care enterprise is a rich source of data for biomedical and behavioral researchers. However, methods of and policies for sharing these data with the wider research community differ in complexity from more traditional research settings and from data sharing expectations and approaches. Unique challenges in data quality, privacy and confidentiality, policy, regulatory, and ethical issues associated with health care and administrative data will require considerations for data sharing and its uses. Informed consent for collecting, using, and sharing these data is essential for respecting participant rights and maintaining public trust. NIH will increase capabilities for informed consent processes and transparency in how participant data is used in research. This is particularly pertinent to the specific challenges for data science in clinical use cases to build trust, explainability, and transparency into the systems and processes leveraging participant data. These activities are consistent with and building on recent NIH guidance and templated informed consent language[43] for secondary research use of data and specimens.

In addition, wearable device data require substantial efforts to extract, transform, and structure the data in order to reduce the risk of exposing personal information and improve its suitability for research. There are several existing models for sharing health data with researchers: independent hospitals forming networks or consortia to exchange data with each other and with select external researchers; professional societies engaging with their member institutions and membership to establish data sharing agreements and new channels for data sharing (e.g., NIBIB Medical Imaging and Data Resource Center or MIDRC);[44] data enclaves or secure networks that support federated learning where computational tools can be sent and data can be stored or disseminated without the need for data exchange; and NIH supported enclaves (e.g., NCATS N3C and *All of Us*). Although each approach comes with benefits and challenges, all are important components of the NIH data ecosystem. NIH is committed to improving data FAIRness, transparency of data governance and stewardship expectations, and requirements for accessing and using data derived from care.

In understanding the relationship between health, environment, and lifestyle, researchers are finding that linking and combining individual-level health data with other real-world data and digital sources improves our understanding. However, challenges remain in developing multi-modal data from richly characterized research participants. In addition, linked data provides greater opportunities for researchers to study epidemiological factors. For example, the National Eye Institute recently articulated the need to include vision-specific data missing from large-scale research efforts, such as the NIH *All of Us* Research Program and the Genotype-Tissue Expression Project (**NEI Strategic Plan**).[45] Similarly, new and improved sources of environmental data continue to emerge from industry, federal agencies, and the research community, as well as through new AI/ML methods for estimating individual-level exposures. However, there remains a need to better understand the ethical, legal, and social implications of data linkage. Additionally, researchers need to know how to define and apply rules to adequately protect study participants, and to navigate relationships with the public and private sectors that will ensure that linked data is appropriately governed, shared, and used in an ethical and sustainable manner. NICHD recently commissioned a report on the technology and governance considerations for pediatric COVID-19 record linkages[46] that articulate a need to define collaborative governance approaches, technical requirements, and the data elements required to ensure high-quality linkage. NIH will collaborate with other federal, academic, and private partners to explore avenues for researchers to appropriately use and combine health care data sources where allowable.

Implementation Tactics

- Enhance methods for informed consent in cases where data are combined from multiple sources and/or combined over longitudinal studies, with additional considerations for populations with health disparities.

[43] sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/considerations-for-obtaining-informed-consent

[44] www.midrc.org

[45] www.nei.nih.gov/sites/default/files/2021-12/NEI-StrategicPlan-VisionForTheFuture_508_edit.pdf

[46] www.nichd.nih.gov/about/org/od/odss#projects

- Create, test, validate, and adopt methods to enable researchers to use multi-modal and digital data combined from multiple sources including through partnerships with other agencies, where appropriate.
- Establish and promote standards for new types of health data, such as data captured from home health care devices.
- Enable federated frameworks that will allow sensitive data to be utilized in clinical research, including fostering data linkages and interoperability across existing NIH supported real-world data platforms.
- Develop ethical, governance, and policy frameworks to guide data linkages in different use case scenarios.
- Leverage existing agreements and infrastructure to create avenues for researchers to use and access health care and administrative datasets, enhancing participant awareness and consent of data use, especially for vulnerable populations.

Objective 2-2: Adopt Health IT Standards for Research

Data sharing is essential to expedite the translation of research into knowledge, products, and procedures that will improve human health and accelerate the development and improvement of treatments for diseases. While there may be benefit to biomedical and behavioral research in connecting and sharing the billions of data points stored in EHRs and clinical trial records across thousands of medical systems, there are significant challenges in making use of these data for research. For example, these data lack consistency in standardization. NLM maintains the Unified Medical Language System (**UMLS**)[47] to distribute key terminology, classification, and coding standards, supports to key terminologies that are now required for use in certified EHRs (e.g., **SNOMED, LOINC, RxNoRM**, among others), and associated resources to promote more effective and interoperable biomedical information systems and services.

Fast Healthcare Interoperability Resources (FHIR®) standard enables electronic healthcare data exchange through an application programming interface (API). An API is a specified set of protocols and data standards that establish the ground rules by which one information system directly communicates with another. Software developers can seamlessly connect their system to another through a FHIR API to transmit electronic health data.

Data sharing has made significant progress in the health care community, in part due to the development and adoption of terminology and exchange standards. In 2020, NIH held a virtual workshop entitled *Advancing the Use of Fast Healthcare Interoperability Resources (FHIR®) in Research*. This workshop brought together leaders in data science and research from across federal agencies to develop a framework for increasing the use of FHIR for research (see textbox **Fast Healthcare Interoperability Resources (FHIR®)**). The workshop discussed the interplay needed between policy and technical advances, the opportunities for FHIR to expand the sources of data that can be integrated into the larger 'system of care' to support both clinical care and clinical research, the opportunity that FHIR® presents to

[47] www.nlm.nih.gov/research/umls/index.html

increase data reuse across both clinical care and research settings and is enabling patients to access their own clinical data. In addition, FHIR or other such systems should facilitate population science and social determinants of health standards to foster the integration of applied research.

To further advance NIH's goal to bridge the gap between health care settings and applied and clinical research, NIH will strengthen the use of ontologies with vocabularies and terminologies (e.g., SNOMED, LOINC) and exchange standards such as FHIR®. NIH will partner with health data standards bodies and organizations and other federal agencies that work with health data standards. A successful example of this is RADx's *Mobile At-home Reporting through Standards (MARS)* program, which coordinated with federal agencies (ONC, FDA, and CDC) and test manufacturers to establish HL7® v2 and FHIR® standards for capturing data from at-home COVID-19 tests.

Implementation Tactics

- Implement agile programs that convene researchers and developers to develop, test, validate and adopt health IT technologies and standards based on scientific use cases and provide feedback based on lessons learned.
- Promote development, training, and adoption of FHIR® to enable further tools for clinical research and for data exchange in research infrastructure, cohort discovery, and applied real-world research.
- Partner with other agencies such as the Office of the National Coordinator for Health Information Technology (ONC), large health care systems, health technology groups, and researchers to develop use cases outlining how health data standards can benefit and enhance scientific data analysis.

Objective 2-3: Enhance the Adoption of Social and Environmental Determinants of Health for Health Equity

Technological advances have made a significant impact on positive health outcomes; however, advances have not benefited all Americans equally. Health disparities persist, disproportionately affecting racial and ethnic minority populations, individuals of less privileged socioeconomic status (SES), underserved rural residents, sexual and gender minorities (SGMs), individuals with disabilities, and any subpopulations that can be characterized by two or more of these descriptions. It has long been recognized that health and treatment outcomes are not solely determined by clinical procedures but that environmental, behavioral, and social factors also play a crucial role. Environmental risk factors such as exposure to pollutants, air and water contamination, and health impacts from climate change are entangled with SES and may increase health disparities. These factors affect health at both the individual and community level as evidenced by the health disparity in many communities when compared against the national health indices for a spectrum of diseases and conditions. Advances in data science can help researchers better understand the social and environmental factors associated with racial or ethnic minority group health outcomes and can lead to more effective interventions. This can be accomplished by adopting standardization, collection, reporting, and leveraging of measures of health determinants in both existing and emerging data sources and fostering appropriate data linkages between clinical research data, SDoH data[48] [49]and environmental determinants of health (EDoH).[50] In addition, the National Academies of Sciences, Engineering, and Medicine is developing a vision for data infrastructure for federal statistics and social and economic research in the 21st century.[51] In this new strategic plan for data science, there is an opportunity to broaden and enhance consensus-driven SDoH[52] and EDoH standards for data capture and integration across a variety of systems. NIH will engage communities and stakeholders to develop demonstration projects, real world pilots and use cases to identify and implement SDoH and EDoH data and common data elements for specific diseases/conditions. This objective is aligned with the NIH UNITE[53] initiative to facilitate new research in health disparities and minority health research (HD/MH).

Implementation Tactics

- Identify SDoH/EDoH of health data and their associated value set.
- Support demonstration projects to test how best to capture SDoH/EDoH of health data for interoperable electronic data exchange.
- Develop infrastructure and tools for extracting structured and unstructured SDoH/EDoH from multiple sources and enable iterative models to include SDoH /EDoH in training.
- Enable linkage of SDoH/EDoH with other data such as clinical, Real-World Data (RWD), wearable sensor, ‘omics data, and administrative data and develop demonstration projects to show technical feasibility of such linkages when appropriate and when there are no increase risks of reidentification for small communities.
- Support real-world pilots to integrate social and environmental determinants with clinical common data elements.
- Support training programs and activities for under-represented groups to expand use of SDOH/Behavioral/EDoH data models and data collections.

Objective 2-4: Cross-disciplinary Training to Empower Clinical Data Science

NIH recognizes that to maintain and enhance clinical research informatics as a career path requires not only clinical training but also training in informatics, analytics, ethics, and standards. This training will focus on appropriate use of data generated from clinical, health care, and real-world settings to better understand the regulatory and policy standards in the generation and use of these data. Equally important is the need to provide health science training to individuals with strong backgrounds in data science. Cross-training between data scientists and clinical researchers would pave the way for interdisciplinary research and could help to reach across new research areas (**NIDDK Strategic Plan[54] NINDS Strategic Plan[55])** Data management and data linking requires partnerships across ethics, social, technological and data science fields. Research involving linking multiple data types, and exposure to new opportunities in technologies, will require a diverse cadre of colleagues for future collaborations. In

[48] Science Collaborative for Health disparities and Artificial intelligence bias Reduction (ScHARe): https://www.nimhd.nih.gov/resources/schare/
[49] www.nimhd.nih.gov/about/strategic-plan
[50] www.niehs.nih.gov/about/assets/files/niehs_strategic_plan_20182023_508.pdf
[51] www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century
[52] www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
[53] www.nih.gov/ending-structural-racism/unite
[54] www.niddk.nih.gov/about-niddk/strategic-plans-reports/niddk-strategic-plan-for-research
[55] www.ninds.nih.gov/modules/custom/ninds/assets/files/NINDS_Strategic_Plan_2021-2026_Final_508C.pdf

addition, other health related research fields, including dental and ophthalmology, can benefit from enhanced data science training, with a goal to integrate clinical data, imaging data, and -omics data with diverse data types from other health-related fields including the SDoH.

Implementation Tactics

- Support cross-training between data scientists, clinical researchers, and nurses engaged in research at various stages of the academic tracks.
- Develop training on consent practices and ethical use of data that go beyond legal and regulatory requirements with special considerations for linked/merged data and data from underrepresented communities.
- Develop trainings on data sharing, management, transparency, provenance, and data quality for clinical research.
- Create networking opportunities for clinical and data science researchers to develop collaborations, build teams, and learn from experts on these topics.

Goal 2: Partnerships and Measuring Progress

NIH understands the strength of partnerships and collaborations for innovation in biomedical and behavioral research. NIH will seek partnership and collaboration across multiple stakeholders including the ONC in implementing relevant data standards, large health care systems and heath technology groups, relevant standards development organizations, and to develop collaborations with bioethics organizations. Importantly, NIH will increase and improve opportunities for community engagement and partnership with underserved communities to collectively build tools and frameworks for biomedical and clinical data science uses.

For this goal, potential measures of progress need to address how these activities are advancing biomedical research and include greater use of RWD and SDoH/EDoH data, increased utilization of FHIR for data exchange including creating or fine-tuning implementations to address research needs, new examples of discovery and harmonization, new or enhanced common data elements for interoperability, and increased use of existing and new standards in clinical and research applications and increasing the number of and reducing processing time for data access requests.

GOAL 3

Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence

Immense amounts of data are generated throughout the biomedical research enterprise from fundamental experiments using cells and research organisms to clinical studies and community-level epidemiological research. These data have value not only for the original research question, but also for secondary data analyses for study replication or for other researchers asking different questions. Harnessing research data for data-driven discovery remains a major challenge that requires attention to data quality, quantity, computability, and standards as well as new methods in computational and AI modeling.

AI/ML are a collection of data-driven technologies with the potential to significantly advance biomedical research. Advances in the field of AI have led to exciting opportunities, including improvements in protein structure prediction and protein design, computer-aided diagnosis on medical images, better understanding of Long-COVID phenotypes, and large language models to interpret clinical, electronic health care records and reports to aid in clinical decision support. AI algorithms can analyze va***st amounts of data, identify complex patterns, and gain deeper insights into fundamental scientific phenomena. This approach allows researchers to unlock new avenues for exploration, drive scientific discoveries, and further our understanding of the underlying principles in various fields of study. With the ability to process billions of parameters, AI could significantly improve future health research in recommender systems, rapid annotations including medical and tissue image processing, and the daunting task of organizing large bodies of medical information. However, utilizing AI for biomedical research and health care practices is still hampered by inconsistent, incomplete, biased, and low-quality data. The task of making data FAIR and AI/ML-ready is not only algorithmic. It requires multi-disciplinary expertise, experimentation and, often, iterative feedback from AI/ML applications and experts. In particular, ground-truth, standardization, and validation of training datasets is particularly important in biomedical applications where bias and inaccuracies could have misleading and inaccurate results.

Across the federal landscape, AI is seen as a priority and as such the National AI Initiative Act of 2020[56] called on the National Science Foundation (NSF), in coordination with OSTP, to form a National AI Research Resource (NAIRR) Task Force. This Task Force laid out a plan to establish the NAIRR with four measurable goals in mind, namely to (1) spur innovation, (2) increase diversity of talent, (3) improve capacity, and (4) advance trustworthy AI. The roadmap to implementing the NAIRR[57] calls for an all-of-government approach to leveraging resources, such as massive compute infrastructures, large data, and a growing talent pool of researchers, to realize this vision.

To enhance the robustness and utility of data analysis and processing methods, NIH will take advantage of new innovations in open and FAIR software and algorithms. NIH will support partnerships to co-design emerging capabilities including new methods in AI including generative AI, computational image analysis, and machine vision; new infrastructures such as quantum information sciences; automated workflows new tools for researchers to leverage data in a transparent, explainable, fair, and ethical

[56] www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1218
[57] www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

manner; and new ways of enabling communities to develop software through collaborative projects. In a world that could unintentionally create a technological divide, NIH must strengthen and diversify its data and software expertise and the technological workforce and make these resources accessible to underrepresented populations in data science.

Objective 3-1: New Opportunities to enhance Artificial Intelligence, including ethical AI for biomedicine

AI (which includes knowledge representation, ML, natural language processing, computer vision and perception, deep learning, and language models) has made progress in medical diagnoses and in better understanding of underlying biological processes. AI methods require attention to transparency; data and algorithm biases; and ethical, legal, and social implications. Making data FAIR and AI/ML-ready also requires interdisciplinary skills. Particularly for biomedical and behavioral research, AI/ML-readiness should be guided by attention to individual and societal impacts of datasets used to train the AI/ML models. While different classes of AI may have unique data requirements, in general AI requires machine readable data that are well described with ontologies and schema so that data can be parsed by the algorithm. Including data quality, such as accuracy, completeness, consistency, and reliability, in AI metadata standards will help address the trustworthiness of the information provided as a result of the AI algorithms. Biases in datasets, algorithms, and applications raise risks and increase potential harms related to privacy, confidentiality, and adverse cultural and social impacts with consequences to people, organizations, and communities, particularly for disadvantaged, disempowered, or marginalized groups. The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)[58] identifies risks and/or potential harms and discusses how managing these risks will lead to more trustworthy AI systems and enable AI developers and users to better understand and take responsibility for the potential limitations and uncertainties in their models and systems.

[58] doi.org/10.6028/NIST.AI.100-1

There are many challenges that hinder the widespread use and deployment of AI/ML capabilities. AI/ML algorithms need big and diverse datasets, yet many underrepresented communities have a long history of being absent or misrepresented in existing biomedical and behavioral datasets including clinical, observational, and data generated in the course of care. Additionally, there is a lack of diversity among researchers, which may unintentionally lead to bias in the design, use, and deployment of AI/ML models in health care and research. Further, non-traditional measures, including SDoH, are important for disease outcomes and health care delivery and when missing may adversely affects predictions. SDoH, including poverty, education level, stress level, access to healthy foods and health care, and exposure to hazards, may play an important role in the diagnoses and treatments of patients from underrepresented communities. Their omission from consideration may lead to fatal outcomes, misdiagnosis, and lack of generalization. Under-represented communities, which are often disproportionately affected by diseases and health conditions, have the potential to contribute expertise, data, diverse recruitment strategies, and cutting-edge science; and to inform the field on the most urgent research questions; but may lack financial, infrastructure, and data science training capacity to apply AI/ML approaches to research questions of interest to them. Recognizing these challenges, NIH launched the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) in 2021 to address health disparities and health inequities using AI/ML (**see AIM-AHEAD textbox**) and **Science Collaborative for Health disparities and Artificial intelligence bias Reduction (**ScHARe). Ethical and unbiased data and algorithms are necessary to create safe, secure, and trustworthy AI[59] for people and for a civil society. According to NIST, trustworthy and responsible AI includes essential building blocks of accuracy, explainability and interpretability, privacy, reliability, robustness, safety, security, and mitigation of harmful bias. This requires the development of assessment frameworks for measuring bias attributes in existing datasets and algorithms across the continuum of AI development and use. Connecting these principles to frameworks that can be used in practice by researchers is essential for continuing to build trust and transparency.

Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) fosters and supports mutually beneficial partnerships to increase the participation of underrepresented researchers and communities, and build capacity and capabilities of AI/ML in these communities through:

- Access to high-quality AI/ML-ready data from diverse populations
- Coordinate federated data approaches and computing infrastructure
- Train diverse data science workforce
- Support research questions that connect EHRs, SDoH and other related datasets to detect and mitigate biases, develop predictive models, and incorporate community-engaged research

Important goals for NIH are to enhance AI methodology and technologies that expand on the unique opportunities for biomedical and health research; and to ensure that AI/ML capabilities are equitably beneficial across populations in the United States and globally. NIH activities align with and support the Blueprint for an AI Bill of Rights[60] to ensure that AI algorithms and systems are used and designed in an equitable way. In partnership with the NIH ICOs, the agency will support emerging technologies and AI to integrate multiple streams of data including genomic, nutritional, sensor-based, social and behavioral, exposure, and community-level data to develop explanatory theoretical models, to inform prevention

[59] Executive Order 13960, Promoting Use of Trustworthy AI in Federal Government: www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

[60] www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

efforts, and to address health disparities (***Eunice Kennedy Shriver* National Institute of Child Health and Human Development Strategic Plan,**[61] **National Institute of Mental Health Strategic Plan,**[62] **National Heart, Lung, and Blood Institute Strategic Vision**[63]). Support to increase opportunities for research communities to include concepts of diversity, equity, and inclusion in the development of trustworthy AI-enabled infrastructures and training is a priority.

Implementation Tactics:

- Develop socio- technical- solutions, including guidelines and principles, for ethical AI and to redress biases in training sets (containing ground truth), and algorithms, and support their effective assessment, validation, and adoption.
- Establish and operationalize community engagement for diverse, equitable and inclusive data, methods, and sources for AI.
- Support research in the development, validation, and use of synthetic clinical datasets for AI training and applications, when appropriate.
- Develop tools and training opportunities to help researchers create and prepare data that are FAIR and AI-Ready, including ontologies, schema, and data quality measures.
- Support the development of AI models, with appropriate metadata (model cards) that are explainable, transparent, and FAIR.
- Leverage new technologies and methods for foundational models to accelerate biomedical and behavioral research.
- Support opportunities to develop new AI technologies that will enable the translation of data to knowledge, including AI tools to enable data cleaning, harmonization, integration, and metadata collection.
- Enhance NIH capabilities in AI through partnerships across federal agencies and communities to develop new methods in AI.

Objective 3-2: Develop cutting edge software technologies

NIH is poised to take advantage of the integration of real-world devices, the increased scale of computational resources and significant automation in software and algorithms to advance biomedical discoveries and innovation. For example, new methods that can integrate multidimensional data from a variety of sources including molecular, wearable sensors, environmental, and survey data are needed to develop predictive and actionable models of weight gain, weight loss, and weight loss maintenance and to clarify the role of obesity in the risk, prevention, and treatment of cardiopulmonary and sleep disorders (**National Heart, Lung, and Blood Institute Strategic Vision**[64]). Multi-dimensional data integration remains a significant challenge for biomedical and behavioral research.

Additionally, low code no code technologies provide a growing opportunity for trainees and citizen scientists to develop functional applications via ‘drag-and-drop’ software platforms or on the web, with appropriate training. New opportunities to enhance biomedical and behavioral research through the

[61] www.nichd.nih.gov/sites/default/files/2019-09/NICHD_Strategic_Plan.pdf

[62] www.nimh.nih.gov/sites/default/files/documents/about/strategic-planning-reports/NIMH%20Strategic%20Plan%20for%20Research_2022_0.pdf

[63] www.nhlbi.nih.gov/about/strategic-vision

[64] www.nhlbi.nih.gov/sites/default/files/2017-11/NHLBI-Strategic-Vision-2016_FF.pdf

support of digital twinning approaches to model organs, systems, individuals, and populations; new capabilities for privacy preserving computing and privacy preserving technologies; and quantum computing should be explored. Finally, ethical considerations for transparency in software and algorithms should be supported.

Implementation Tactics:

- Adopt and adapt emerging and specialized methods, algorithms, tools, software, and workflows for biomedical and behavioral scientific discovery.
- Enhance tools and workflows for greater automation, while maintaining robust ethical standards
- Leverage new passive and mobile devices and technologies for data collection and analysis with improved practices for informed consent.
- Facilitate FAIR software, with sufficient documentation and metadata, and enhance ethical frameworks.
- Leverage advances in computational methodology and studies to create new opportunities for ethical and social science research.
- Investigate the potential of digital twinning approaches to organs, systems, individuals, and populations.
- Explore opportunities to combine theory-based modeling and simulations with data-driven capabilities.
- Promote opportunities to engage new communities in software development and make these resources accessible to under-represented communities interested in data science.

Objective 3-3: Supporting FAIR Software Sustainability

Software is an integral component of biomedical behavioral research due in part to the speed and growth of new technology innovations in the software and computing fields including AI, computer transistors, and microchips. NIH collaborates across 19 ICOs to support the development and enhancement of software tools for open science[65] by fostering new collaborations between biomedical and clinical scientists and software engineers. For example, significant progress has been made in developing computing models for client-server architectures for data acquisition and progress in developing cloud-based data management and data analytics. Through partnerships with Cloud Service Providers Google, AWS and Microsoft Azure, NIH has realized over 275 million compute hours for data analysis in the cloud. Yet challenges remain in creating FAIR software.[66] The FAIR software principles, similar to the FAIR Data principles, ensure that software will be usable beyond a single laboratory or investigator. FAIR software principles foster practices to ensure that research software is sustained by larger biomedical research communities over time. NIH recently issued best practices for software sharing[67] that align with the FAIR software principles.

[65] datascience.nih.gov/tools-and-analytics/administrative-supplements-to-support-enhancement-of-software-tools-for-open-science

[66] Barker, M., Chue Hong, N.P., Katz, D.S. *et al.* Introducing the FAIR Principles for research software. *Sci Data* **9**, 622 (2022).

[67] datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq

To develop FAIR and sustainable software at a scale beyond single academic laboratories requires multi-disciplinary collaborations from biomedical, computer science, and related fields. Today NIH and other federal agencies and nonprofits are tackling software sustainability head on, including the NSF program on Cyberinfrastructure for Sustained Scientific Innovation,[68] the recent NIH supplements to support enhancement of software tools for open science, the Schmidt Futures Virtual Institutes for Scientific Software,[69] and the Chan Zuckerberg Initiative's program for Essential Open-Source Software for Science.[70] A long-standing program at NIH is NCI's Information Technology for Cancer Research (ITCR)[71] program. The ITCR program serves the informatics needs of cancer research continuum and provides support for informatics resources across the development lifecycle (**see ITCR textbox**).

ITCR supports investigator-initiated, research-driven informatics technology development spanning all aspects of cancer research. The ITCR lifecycle approach includes separate funding in the following areas:

- Algorithm Development
- Protype and Hardening of Software
- Enhancement and Dissemination of Software
- Software Sustainability

These programs have a common theme: to enable investigators to adapt and enhance software and tools to take advantage of new technologies and computing paradigms and to optimize research software for robustness and ultimately to increase software sustainability.

Implementation Tactics:

- Enhance community-focused software development and dissemination.
- Improve visualization tools to support the scale and variety of modern biomedical data.
- Establish metrics and best practices for software sustainability and integrate these into software development lifecycle.
- Facilitate research activities for software engineers and biomedical and computational researchers to collaborate.
- Develop mentorship programs that pair experienced software engineers with early-career researchers and software developers.
- Explore innovative models for public-private partnerships to support software and data innovation and sustainability.

Goal 3:

Partnerships and Measuring Progress

Potential measures of progress for this goal include an increase in the number of software tools that align with the FAIR principles and have a measurably enhanced user experience, increased citation of NIH software across broader communities, software tools that support an increase of use cases across

[68] beta.nsf.gov/funding/opportunities/cyberinfrastructure-sustained-scientific
[69] www.schmidtfutures.com/our-work/virtual-institute-for-scientific-software
[70] chanzuckerberg.com/eoss
[71] itcr.cancer.gov/about-itcr

various scientific domains, and integration of tools from other domains into biomedical research. Additional measures of progress include increased representation of health disparity populations in AI development and auditing AI models, development of ethical frameworks and tools for software and algorithms, and greater transparency in the development and processing of data and models. NIH will seek partnerships and collaborations with other federal agencies such as NSF and DOE, non-profit organizations, and societies and communities such as the Research Software Alliance.[72]

[72] https://www.researchsoft.org

GOAL 4

Support for a Federated Biomedical Research Data Infrastructure

Throughout the last five years, NIH has seen a remarkable growth in the support for and use of biomedical data repositories and platforms for biomedical research. Today, more and more of these data infrastructures are now moving entirely to the cloud. By moving data infrastructures to the cloud, NIH utilizes advanced cybersecurity controls and scales data management and computation that can take advantage of new technologies while simultaneously creating cost efficiencies and enhancing a positive user experience. The challenge now is to provide greater connections across NIH cloud-based data platforms for easier access to multiple datasets, streamline of similar functions, and enabling more facile analytics. Current cloud-based data platforms include the NHLBI's BioData Catalyst®,[73] an ecosystem that offers data, analytic tools, applications, and workflows in secure workspaces to accelerate reproducible biomedical research to drive scientific advances that can help prevent, diagnose, and treat heart, lung, blood, and sleep disorders on over 400,000 individuals; NCI's Cancer Research Data Commons (CRDC)[74] which provides secure access to a large, comprehensive, and expanding collection of cancer research data; the Kids First Data Resource[75] which houses data on 44 childhood cancer and structural birth defects cohorts; the National Human Genome Research Institute's (NHGRI) Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)[76] genomic data sharing and analysis platform; *All of Us*[77] which has collected data from over 500,000 participants; the NIH database of Genotypes and Phenotypes (dbGaP)[78] which has controlled access data including sequence, genotype, and/or phenotype data from over 3.2 million research participants; and other NIH-supported data platforms. These resources are cloud-based data infrastructures that provide the research community with data and analytical tools, applications, and workflows in secure environments.

[73] biodatacatalyst.nhlbi.nih.gov
[74] datascience.cancer.gov/data-commons
[75] kidsfirstdrc.org
[76] anvilproject.org
[77] allofus.nih.gov
[78] www.ncbi.nlm.nih.gov/gap/

The NIAID Data Ecosystem enables simultaneous search across 15 infectious- and immune-mediated disease and general data repositories based on metadata. The NIAID Data landscape is highly distributed and requires an ecosystem approach that allows for freedom to operate regarding system, syntactic, and semantic interoperability while requiring a minimal set of FAIR-compliant metadata about existing data access protocols used by the repositories. The NIAID approach to the ecosystem is to leverage FAIR metadata to describe data as well as API's and other data access protocols to create a FAIR compliant, interoperability layer on top of a diverse landscape of data, software, and services.

NIH ICOs are also developing data ecosystems including the **NIH Cloud Platform Interoperability** initiative;[79] the National Institute of Biomedical Imaging and Bioengineering (NIBIB) **Medical Imaging and Data Resource Center** (MIDRC)[80], which provides open access to 300k+ curated, AI-ready COVID-19 imaging studies and has demonstrated interoperability with BioData Catalyst and the N3C**; National Institute of Allergy and Infectious Diseases** (**see textbox**); the National Institute on Minority Health and Health Disparities and the National Institute of Nursing Research **ScHARe,** which hosts population science, SDoH, behavioral and environmental data sets to advance health disparity, health care delivery and health outcomes research and foster strategies to mitigate AI biases; the **Common Fund Data Ecosystem;**[81]and NCBI's **Comparative Genomic Resource,**[82] which aims to integrate genomic data across all eukaryotic species.

With recent and significant migrations of data resources to the cloud, and the ability to enable petabyte scale data analytics, NIH has the responsibility to integrate these resources into a federated data infrastructure that leverages ideas from industry and cutting-edge research. The benefit of federating NIH data resources includes: 1) easier access to and use of data across multiple Institutes supported data platforms, 2) economies of scale for NIH to support and maintain shared tools and capabilities, 3) opportunities for communities to collaboratively develop and share new methods and workflows, and 4) oversight by the community for greater transparency and autonomy of data use. In collaboration with the NIH ICOs, the agency will support development of innovative data sharing platforms, data analytics, and their integration. This is integral to the missions of each NIH ICO (specific examples found in the **National Institute of Environmental Health Sciences Strategic Plan)**[83] and to the overall mission of NIH. The broad use of big data frameworks and FAIR principles, with continued emphasis on partnerships within and outside NIH, will result in new discoveries.

Objective 4-1: Develop, test, validate, and implement ways to federate NIH data and infrastructure

As articulated in the **National Institute on Aging Strategic plan**,[84] NIH needs to develop comparable databases on health outcomes, risk factors, and determinants of health disparities. An emerging paradigm in data science is data-oriented infrastructure that moves away from a traditionally centralized data infrastructure. A data-oriented infrastructure is distributed between data repositories, or nodes, and has shared capabilities and services to allow for maximum interoperability and economies of scale.

[79] datascience.nih.gov/nih-cloud-platform-interoperability-effort
[80] www.midrc.org
[81] commonfund.nih.gov/dataecosystem
[82] www.ncbi.nlm.nih.gov/comparative-genomics-resource
[83] www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf
[84] www.nia.nih.gov/sites/default/files/2020-05/nia-strategic-directions-2020-2025.pdf

This requires a common and coordinated data access process with shared policy and governance. The challenge of a distributed data infrastructure is to create a fabric of harmonized services (e.g., identity and data access management (Authentication [AuthN]/Authorization [AuthZ]), data catalogues, search capabilities, and application programming interfaces (APIs) that are commonly shared, or federated, across the data repositories. In a federated paradigm, NIH ICOs, and organizations supporting biomedical and behavioral research data infrastructures, will control and manage their own data, adopt common processes and interfaces, analysis tools, and services that can be used broadly for biomedicine and behavioral research. In line with this vision, the goal of this objective is to improve efficiencies and maximize researchers' ability to find, access, and use data that are generated from federally funded research and stored in cloud-based data repositories. NIH's vision is to build a connected and federated data ecosystem to ensure that data repositories can be used together rather than in isolation. Several NIH ICOs have collaborated and developed early capabilities including a common approach for researcher's access to control access data across a set of data repositories (**see textbox on RAS**), and an approach to implement guidelines and technical standards to empower end-user analyses across participating cloud platforms. In 2020, these interoperability standards were piloted and as a result researchers were able to demonstrate data access across multiple cloud-based NIH data repositories and perform combined analysis with meaningful results.[85] These initial piloted efforts are the genesis of a NIH-wide federated data ecosystem and are articulated as priorities in the Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap [86] priorities.

The **Researcher Auth Service (RAS)** Initiative is advancing NIH's data infrastructure and ecosystem. RAS is an identity and data access and management service provided by NIH's Center for Information Technology to facilitate consistent and user-friendly researcher access to NIH's controlled-access data. RAS has adopted the Global Alliance for Genomics and Health (GA4GH) standards for integrating researcher-focused applications and data repositories over the OpenID Connect (OIDC) platform. RAS supports the FY 2023 Federal Cybersecurity R&D Strategic Plan Implementation Roadmap to protect systems and ensure confidentially, integrity, availability and privacy of data and Executive Order on Improving the Nation's Cybersecurity. The RAS initiative is advancing data infrastructure and ecosystem goals as defined in the 2018 NIH Strategic Plan for Data Science

To capitalize on these early successes, NIH will support and enhance a federated biomedical data research infrastructure that will create, test, validate, and implement a set of sharable services (e.g., common search capabilities, application programming interfaces (APIs), identify and access management (IAM) services, and workspaces/sandboxes). By doing so NIH will improve efficiencies, reduce duplicative funding, and maximize researchers' ability to find, access, and use data that is generated from federally funded research and stored in cloud-based data repositories.

Implementation Tactics:

- Enhance utilization of cloud and hybrid computing architectures and provide opportunities for low-resource institutions to access and utilize NIH supported cloud capabilities.

[85] Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) (biorxiv.org)

[86] www.nitrd.gov/pubs/FACE-SP-FY22-Implementation-Roadmap.pdf

- Support efficiencies and sharable technologies across in NIH data platforms, including increasing new and existing technology industry partnerships.
- Expand RAS to increase researchers' ability to access data and ensure accountabilities for privacy protection and cybersecurity of systems.
- Ensure a robust and connected data resource ecosystem that includes supporting linkages and interoperability across NIH supported cloud platforms for curation, analysis, and sharing of data and metadata.
- Develop new capabilities for data search and discovery by enhancing metadata standards, indexing techniques, and improving data interoperability and harmonization.
- Explore new paradigms in computing for biomedical and behavioral applications.

Goal 4: Partnerships and Measuring Progress

For this goal, the potential measures of progress should advance biomedical research through the use of an integrated infrastructure and include ease of findability of datasets and the ability of researchers to integrate NIH data, increase ability to use data across the NIH data ecosystem as measured by publications. Additional metrics include guidance and best practices on interoperability so that data, analysis tools and models of biological or population systems can be shared more easily. NIH will require partnership and collaboration across multiple NIH ICOs and scientific organizations such as GA4GH, Research Data Alliance, and Research Software Alliance.

GOAL 5

Strengthen a Broad Community in Data Science

As data science is a necessity in most biomedical and behavioral research, there is a need to develop and nurture data science talent from a diverse array of scientific interests. NIH is committed to growing a stronger and broader community of data scientists including:

- Data science literate researchers who feel comfortable reading and understanding reported outcomes resulting from data science approaches.
- Data science savvy researchers who are data science literate and can actively use data science approaches in designing research projects and initiating and/or participating in collaborations with data scientists.
- Data scientists who are skilled in areas that include bioinformatics, AI/ML, clinical informatics, cloud computing, statistics, computational science, software design and programming, bioinformatics, foundational models, visualization, predictive analytics, modeling and simulation, and data management and sharing.

Following the first strategic plan for data science, NIH, through ODSS, has worked collaboratively with ICOs on programs that train and educate researchers in data science. These collective efforts will continue to grow, with a particular focus in enhancing the diversity of the data science community so that it better reflects the diversity of the United States. Diversity of backgrounds and scientific areas expands the range of research questions, facilitates the translation of scientific data and findings to different communities and helps to build trust in all communities. The time has never been better for all biomedical and behavioral researchers to take full advantage of data science including new innovations from cloud computing, utilizing the availability of significant amounts of biomedical and behavioral data, and new advances in AI/ML. In working to ensure that data science advances in biomedical research can benefit all populations, NIH will help to create a vibrant, innovative, and inclusive data science community.

Objective 5-1: Increase training opportunities in Data Science

Drawing on the foundation of the network of existing extramural training programs, NIH will coordinate across the ICOs to promote data science training and education. In alignment with the Notice of NIH's Interest in Diversity (NOT-OD-20-031), NIH will boost investment in programs that increase the number of underrepresented individuals in data science, including but not limited to racial/ethnic minorities (Blacks or African Americans, Hispanics or Latinos, American Indians or Alaska Natives, Native Hawaiians and other Pacific Islanders), individuals with disabilities, individuals from disadvantaged backgrounds, and women. The aim is to strengthen the support for students and scientists from pre-college through early investigator levels and provide them with a continuum of competitive funding opportunities in data science. Studies show that bright minds may be lost to science long before reaching the college years. Early intervention strategies in research education are therefore necessary to provide a foundation on which essential data skills and visions can develop. Early intervention strategies in research education, such as those supported by the Science Education Partnership Award (**SEPA**[87]) at NIGMS and the Youth Enjoy Science program [88]at NCI, are therefore necessary to provide a foundation on which essential data skills and visions can develop. Promoting focused data science training for graduate students and postdoctoral fellows will help these early researchers develop into independent investigators with data science acumen. In addition, professional and career development support such as access to mentors, soft skills training, and resilience and wellness support are critical to retain the data science trainees in biomedical and behavioral research.

Implementation Tactics:

- Support data science training for students and scientists at all academic and career levels from pre-college through early investigators.
- Enhance diversity among data science trainees by promoting diversity-focused training and education initiatives.
- Increase pairing of technical data science training with domain-specific knowledge training in NIH training programs.
- Increase the use of hands-on training in new areas such as AI/ML.
- Develop requirements of foundational elements in data science training such as data ethics and cybersecurity.

Objective 5-2: Develop and Advance Initiatives to Expand the Data Science Workforce

Since the first publication of the strategic plan for data science, significant progress has been made within NIH to enhance its administrative and programmatic data science workforce. For example, in 2020 NIH launched the Data and Technology Advancement National Service Scholars (DATA Scholars) Program[89]. These scholars spend one to two years transforming NIH programs by applying cutting-edge methods to health-related challenges. NIH has also implemented the Civic Digital Fellowship [90]program in collaboration with the non-profit organization Coding it Forward to bring to the NIH early-career technologists to spend a summer in data-related projects in NIH program offices. This program successfully supported 80 fellows over four years and provides a solid foundation for NIH to expand to a longer-term program. In addition to these programs, some NIH ICOs have initiated new Offices of Data Science to oversee data management and sharing, the responsible use of data, data science training to staff, and new funding programs in data science. These efforts strengthen the data science workforce within NIH and provide a strong foundation for continued growth. In the extramural community, NIH will focus on promoting the use data science approaches for established investigators, enhancing the diversity of the data scientists, and supporting the growth of data science skills among clinician scientists.

[87] nigms.nih.gov/capacity-building/division-for-research-capacity-building/science-education-partnership-awards-(sepa)
[88] www.cancer.gov/about-nci/organization/crchd/about-health-disparities/resources/yes-r25-fact-sheet.pdf
[89] datascience.nih.gov/data-scholars-2022-closed#overview
[90] www.codingitforward.com/fellowships

Implementation Tactics:

- Enhance the diversity of data science investigators and broaden the reach of data science in the biomedical and behavioral research community.
- Facilitate cross-disciplinary trainee programs in data and biomedical sciences.
- Enhance the data science knowledge and skill building for biomedical and clinician scientists including cross-disciplinary skillsets.
- Facilitate recruitment and retention of diverse data science talents at the NIH.
- Develop a pathway for early-career data scientists to join the NIH.

Objective 5-3: Enhancing Data Science Collaboration within the NIH Intramural Research Program

In addition to promoting data science training in the extramural community, NIH will also work to enhance the recruitment of data science trainees from diverse backgrounds in the Intramural Research Program (IRP).[91] With approximately 1,150 Principal Investigators, more than 2,600 Non-Principal Investigators and more than 5,000 trainees conducting basic, translational, and clinical research, NIH IRP is the largest biomedical research institution and conducts long-term and high-impact science that would otherwise be difficult to undertake. Moreover, NIH supports Biowulf, [92] a high-performance computing system specifically for use by the intramural NIH community. Biowulf is consistently ranked in the top 100-200 of the Top 500 computing infrastructures worldwide and provides access to a wide range of computational applications for genomics, molecular and structural biology, mathematical and graphical analysis, image analysis, and other scientific fields. NIH will build a strong and diverse cohort of intramural data science students and researchers, develop a supportive network for the data science trainees in the IRP and enhance the intramural computational capabilities to realize new opportunities and partnerships not only across NIH, but also with industries and other organizations.

Implementation Tactics:

- Coordinate with the NIH Office of Intramural Training and Education to develop a data science-focused intramural cross-disciplinary training program that supports mentored research experiences for postbaccalaureate, post-master's and postdoctoral fellows from diverse backgrounds.
- Support cross-Institute intramural data science projects and enhance interconnectivity of data scientists of all levels.
- Enable federated capabilities, for data and software, within the NIH IRP.
- Facilitate opportunities for intramural researchers to partner with the private sector.
- Enhance NIH's intramural computing environment to utilize new opportunities in cloud computing, AI/ML, and other data science and computing initiatives.

[91] irp.nih.gov/
[92] hpc.nih.gov/

Objective 5-4: Broaden and Champion Capacity Building and Community Engagement Efforts

Developing and sustaining a biomedical and behavioral research workforce that is reflective of the communities being served and supported in an environment that nurtures their success is essential to truly advancing health equity. However, for some investigators and institutions, including Minority Serving Institutions (MSIs) and low-resourced institutions, data science challenges remain, including easy access to cloud computing environments, sufficient training and mentoring in data science, and opportunities to apply unique expertise to conduct data science focused health disparities research. NIH is committed to broaden the participation of MSIs and low-resource institutions in the data science community and support efforts to increase human capacity, build partnerships and strengthen research infrastructure.

Following the first strategic plan for data science, new programs have resulted from data science partnerships with NIGMS, including enhancement to the INBRE program (**see textbox on INBRE**) to support new data science cores and the development of cloud-based learning modules for the NIH CloudLab.[93] Support for NHGRI's **Educational Hub for Enhancing Diversity in Computational Genomics and Data Science**, partnership with NIMHD to **Enhance Data Science Capacity Research Centers** in Minority Institutions (RCMIs), and new data science training efforts in the Ruth L. Kirschstein National Research Service Award (**NRSA**) Institutional Research Training Grant. Partnership with the NIH Common Fund and Fogarty International Center has enhanced data science capacity in low- and middle-income countries through the Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa[94]) program. These efforts offer a platform for research and collaboration as well as a way to inspire interest in aspiring new data scientists. NIH will continue to develop and expand activities and events to attract a wider community.

IDeA Networks of Biomedical Research Excellence (INBRE) fosters the development, coordination and sharing of research resources, and expertise that will expand research opportunities and increase the number of competitive investigators in IDeA-eligible states.

Recently the INBRE program has required a Data Science Core for Biomedical Research. The Data Science Core will provide resources for research, education, and training to expose undergraduate students to data science research and engage a broader community with expertise in biomedical data sciences and related disciplines such as machine learning, artificial intelligence, and virtual reality technologies.

Implementation Tactics:

- Collaborate with existing NIH programs, such as the Institutional Development Award (IDeA) at NIGMS, the Research Centers in Minority Institutions (RCMI) program at NIMHD and the Partnerships to Advance Cancer Health Equity (PACHE) program at NCI, to develop and expand programs to enhance data science capacity, particularly in MSIs and low-resource institutions.
- Leverage datasets in NIH supported data repositories and data platforms as training resources.
- Build synergies across government, academic, nonprofit, international, and industry stakeholders focused on data science workforce development and training.

[93] cloud.nih.gov/resources/cloudlab/
[94] commonfund.nih.gov/africadata

Goal 5: Partnerships and Measuring Progress

For this goal, potential measures of progress include an increase in the number and diversity of data science trainees and the number of trainees leveraging NIH-supported data platforms, increased number of trainees who matriculate to data science careers, increased number of data scientists recruited to the NIH and increased numbers of intramural scientists developing and utilizing NIH supported software. Additional measures may include the products of the trainees and scientists, including publications, patents, models and technologies. NIH will seek partnerships and collaborations with other federal agencies, non-profit organizations, and private sector industries.

Appendices

I. Accomplishments from the First NIH Strategic Plan for Data Science

Appendix I Accomplishments from the First NIH Strategic Plan for Data Science

Goal 1: Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

- NIH Partnership with Google Cloud Services, Amazon Web Services, and Microsoft Azure through the STRIDES program has resulted in over 200PB of biomedical data on the cloud, 320 compute hours, 5,000 researchers trained, 1,300 program working in the cloud and the development of the NIH CloudLab
- Development of the Research Auth Services for single-sign on and efficient data access across NIH data platforms includes integrating over 30 data programs into RAS, and partnership with Internet2.
- NIH has made further efforts to connect NIH data platforms through the NIH Cloud Platform Interoperability program with a partnership between NLM, NHGRI, NHLBI, NCI, and the Common Fund, with a result of single sign-on and cross platform analysis of data.

Goal 2: Promote Modernization of the Data-Resources Ecosystem

- NIH has supported funding opportunities for data resources (databases and knowledgebases) that has resulted in 17 new awards across 7 NIH Institutes and Centers and NIH has supported supplemental funding to existing databases to align with the OSTP characteristics for FAIR data repositories, resulting in 21 awards across 12 NIH Institutes and Centers.
- NIH has also launched the Generalists Repository Ecosystem Initiative (GREI), partnering with 7 generalists repositories to establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure across these repositories. This initiative conducted a number of webinars with over 1,100 attendees, enabled open metrics in the MakeDataCount Project and created “Search by Funder and Grant ID” metadata fields in participating repositories.
- NIH has also partnered with DataCite to support the ability to find and cite NIH funded data, via the use of persistent unique identifiers.
- NIH has partnered with NLM and the Data Curation Network to provide on-going training in data management and sharing for researchers, data resource staff, and NIH program staff
- NIH has partnered with FASEB to offer the first ever Data Sharing and Data Reuse prize, resulting in over 100 applicants and 12 finalists, with two grand prize winners.
- NIH has also partnered with HL7 to support training in Fast Healthcare Interoperable Resources and supported NIH Institutes to leverage FHIR for clinical data platforms. NIH partnered with the Research Data Alliance (RDA) and Health Level 7 (HL7) to develop and publish a Fast Healthcare Interoperability Resources (FHIR®) implementation guide with 6 real-world use cases by assessing the impact of FHIR® implementation using FAIR data metrics.

Goal 3: Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools

- NIH has supported supplemental funding for NIH funded software, tools, and workflows to develop robust, sustainable, ‘cloud-ready’, capabilities, resulting in 94 awards across 19 NIH Institutes and Centers.
- NIH partnered with NSF on the Smart and Connected Health for AI and data science resulting in 17 awards across 11 NIH Institutes and Centers.

- NIH developed a Software Best Practices document for sharing research software and source code, developed under research grants in any stage of development, in a free and open format.

Goal 4: Enhance Workforce Development for Biomedical Data Science

- NIH launched the Data and Technology Advancement (DATA) National Service Scholar Program to bring experts in data and computer scientists and engineers to tackle challenging biomedical data problems with the potential for substantial public health impact, resulting in 17 DATA Scholars across 13 NIH Institutes and Centers.
- NIH partnered with Civic Digital Fellows program to bring 80 Coding-it-Forward fellows to NIH for four consecutive summers.
- NIH has supported code-a-thons to engage underrepresented communities and increase their participation in data science, including coding partnerships with the African society for Bioinformatics and Computational Biology.
- NIH has expanded the SEPA, NARCH, INBRE programs to include new initiatives in data science resulting in 9 new awards.
- NIH has also expanded diversity supplements in data science to existing grants, resulting in 15 new awards.

Goal 5: Enact Appropriate Policies to Promote Stewardship and Sustainability

- NIH published the 2023 NIH Data Management and Sharing Policy, with new training and infrastructure support for the implementation of this policy.