NIH Strategic Plan For Data Science

2025–2030

National Institutes of Health

Page 1

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Letter from NIH Director

Monica Bertagnolli

Biomedical research relies on an interplay between scientific observation, hypothesis development, experimental design, data analysis, and interpretation. The increasingly powerful data resources produced by National Institutes of Health (NIH)-funded research and the rapid proliferation of data science technologies available to modern biomedical and behavioral researchers have advanced our understanding of living systems and led to medical breakthroughs. But we can do so much more. Currently, optimal data use for research remains limited by challenges with managing, sharing, and integrating data and technology in a usable, secure, and equitable way. But data science's opportunity to dramatically accelerate improved health outcomes for all is more inspiring than daunting.

With its mandate to galvanize the energy, talent, and creativity of the broadest possible research community, the NIH must play a critical role in establishing and maintaining a Biomedical Data Ecosystem that moves beyond data collection and sharing to actively fostering practical and ethical data use. The 2025-2030 NIH Strategic Plan for Data Science provides a bold vision for the future of data science at the NIH. The five goals – built on the solid foundation laid by the initial data science strategy and rooted in NIH's mission to seek fundamental knowledge to protect and enhance health – prepare NIH to adapt and advance in a changing data science ecosystem.

This strategic plan supports the vision of NIH to enhance data-driven discovery to advance biological knowledge, better characterize the health and health outcomes of all people, and foster the development of new technologies and care delivery approaches. Moreover, this strategy encourages the scientific community to share data and better integrate data science practices into their research. However, this work must extend beyond the mainstream scientific community. For example, every person has something important to contribute to advancing health by allowing their health data to be used for research. NIH must, therefore, earn and maintain the public's trust through engagement and empowerment of their contributions to data and data science. The plan addresses this imperative by emphasizing cross-disciplinary collaborations, including public-private and researcher-clinician-community partnerships, to find opportunities and solve challenges together.

I am thrilled to introduce this new strategic plan that will carry us forward through the next five years. This plan embraces unprecedented technological changes like generative artificial intelligence (AI) development, availability, and growing interest in quantum information sciences. Growth in the diversity and extent of biomedical and behavioral data is equally rapid, with the research community gaining access to powerful, new, NIH-supported assets like the *All of Us* research program, Human Pangenome, Human Connectome Project, and so much more. The NIH Biomedical Data Ecosystem, with its development advanced by this new strategic plan, will bring increasingly effective data and tools that enable the broadest research community possible to contribute to our mission to bring better health to all people. It is an incredible time to engage in biomedical and behavioral research, and I cannot wait to see what the future brings!

Monica M. Bertagnolli, M.D.
Director, National Institutes of Health

Table of Contents

Introduction 2
Emerging Opportunities for Biomedical and Behavioral Data Science 4
Plan Content and Implementation 8

GOAL 1: **Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing** 10
OBJECTIVE 1-1: Support the Biomedical Community to Manage, Share, and Sustain Data 12
OBJECTIVE 1-2: Enhance FAIR Data and Greater Data Harmonization 14
OBJECTIVE 1-3: Strengthen NIH Data Repository and Knowledgebase Ecosystem 16
Partnerships and Measuring Progress 19

GOAL 2: **Develop Programs to Enhance Human Derived Data for Research** 20
OBJECTIVE 2-1: Improve Access to and Use of Clinical and Real-World Data 22
OBJECTIVE 2-2: Adopt Health IT Standards for Research 24
OBJECTIVE 2-3: Enhance the Adoption of Environmental and Lifestyle Exposome in Health Research 25
OBJECTIVE 2-4: Empower Clinical Data Science through Cross-Disciplinary Training 26
Partnerships and Measuring Progress 27

GOAL 3: **Provide New Opportunities in Software, Computational Methods, and AI** 28
OBJECTIVE 3-1: Enhance Health for All through AI 30
OBJECTIVE 3-2: Develop Cutting Edge Software Technologies 32
OBJECTIVE 3-3: Supporting FAIR Software Sustainability 33
Partnerships and Measuring Progress 35

GOAL 4: **Support a Federated Biomedical Research Data Infrastructure** 36
OBJECTIVE 4-1: Develop, Test, Validate, and Implement Ways to Federate NIH Data and Infrastructure 38
Partnerships and Measuring Progress 41

GOAL 5: **Strengthen a Broad Community in Data Science** 42
OBJECTIVE 5-1: Increase Training Opportunities in Data Science 44
OBJECTIVE 5-2: Develop and Advance Initiatives to Expand the Data Science Workforce 45
OBJECTIVE 5-3: Enhance Data Science Collaboration within the NIH Intramural Research Program 46
OBJECTIVE 5-4: Broaden and Champion Capacity Building and Community Engagement Efforts 47
Partnerships and Measuring Progress 49

Endnotes 50
Appendix 1: Accomplishments from the First NIH Strategic Plan for Data Science 52
Appendix 2: Abbreviations 54

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 discoveries. To put it simply, data enable new opportunities for scientific inquiry. This updated NIH Strategic Plan for Data Science sets a bold vision for the future, in which data generated in individual care and 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 underlying basic biological experimentation. Data science also consists of clinical trial data; real-world data (RWD) from electronic health records (EHRs), sensors, and wearables; geospatial data; public and community health research; surveys; and data from social, environmental, and observational studies. The vision articulated in this strategic plan supports the NIH Policy for Data Management and Sharing (DMS Policy).[1] It embraces data-driven discovery as a powerful tool to elucidate biological processes, better characterize the health needs of the American people, and foster the effective and efficient use of new methodologies such as those arising from artificial intelligence (AI) and machine learning (ML). Progress toward the promise of data-driven discovery requires a unified effort across the NIH Institutes, Centers, and Offices (ICOs) that is coordinated and stimulated by the resources of the NIH Office of Data Science Strategy (ODSS). To accomplish the goals set forth with this vision, NIH will address key challenges and explore opportunities to:

- **Generate and disseminate** Findable, Accessible, Interoperable, and Reusable (FAIR) data to foster greater sharing and add value to NIH research investments.[2]
- **Develop cost-effective strategies** for sustainable, secure, and accessible biomedical data repositories, workspaces, and knowledgebases.
- **Acquire and protect data** obtained from EHRs and other RWD, including data captured outside of traditional health care settings, to preserve privacy and confidentiality and promote participant consent.
- **Promote the emergence** of innovations in AI approaches that are FAIR, validated, and accurate.
- **Create opportunities** for innovative technologies and computing paradigms such as quantum computing and digital twins to advance biomedical research.
- **Foster a data science workforce** across institutions and regions.

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 2025-2030 NIH Strategic Plan for Data Science articulates the NIH's strategic views, goals, and objectives to advance data science over the next five years. By addressing these challenges, NIH will pioneer robust and efficient data governance frameworks, ensuring data integrity, security, and accessibility while promoting cross-disciplinary collaborations that accelerate scientific discovery.

The 2025-2030 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 collaboration is key to developing robust solutions for leveraging data science in the biomedical and health research enterprise. Advancing data science requires NIH to leverage its partnerships with many actors, including 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 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 outlined in this plan will position the NIH to accelerate discovery in biomedicine and health, improving health for all Americans through more relevant, comprehensive scientific findings and developing a workforce of researchers and clinicians skilled in the use of data science methods for discovery and care that contribute to US leadership at the frontier of science and technology.

Goals

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 AI

GOAL 4
Support a Federated Biomedical Research Data Infrastructure

GOAL 5
Strengthen a Broad Community in Data Science

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 is taking a bold step forward with the final DMS Policy – which articulates the need to prospectively plan for how scientific data and accompanying metadata[4] 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 reduce barriers to and the cost of sharing research data, advancing NIH's goals to promote research safety, transparency, and reproducibility.

> "New opportunities and guidelines are needed to enhance trustworthy data repositories that align with community expectations and contain open metrics that demonstrate the impact of data sharing."

New capabilities and resources are needed to 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 enable easier data sharing in appropriate repositories. Moreover, new opportunities and guidelines are needed to enhance trustworthy data repositories that align with community expectations and contain open metrics that demonstrate the impact of data sharing. Finally, developing new methodologies for computational interoperability across data repositories and knowledgebases enables more research on complex, hard-to-collect data, increasing the productivity of NIH's research and development investments.

Today, there is potential to create federated networks that connect the billions of data points stored in EHRs, other RWDs such as wearable data, and clinical trial data obtained from medical systems and medical research institutions nationwide. 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®)[5] Fast Healthcare Interoperability Resources (FHIR®)[6] specification, certified health information technology (IT) products will have standardized application programming interface (API) capabilities to facilitate health data sharing. Leveraging and building on the FHIR® standard allows for the exchange and sharing of EHR data, phenotypic data obtained from clinical and genomics studies, clinical records and community data, and eventually, other data from medical devices such as wearables and sensors. These all represent 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. This is the vision of several NIH initiatives, including the *All of Us*[7] Research Program.

The National Clinical Cohort Collaborative (N3C)[8] illustrated the power of a collective data initiative. N3C pulls data in 4 standard models from 77 health systems representing more than 230 organizations. Data are harmonized with the *Observational Medical Outcomes Partnership* (OMOP) Common Data Model every week. N3C represents the largest de-identified limited datasets for COVID-19 research and uses privacy-preserving linkages to other RWD, such as Centers for Medicare & Medicaid Services (CMS) and mortality data. Advances like those seen in the *All of Us* and N3C programs require standardized vocabularies and ontologies, including communities from different biomedical science and medicine areas. Experiences during the pandemic also emphasized the importance of using and promoting common data elements (CDEs), as was illustrated in the Rapid Acceleration of Diagnostics (RADx)[9] 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)[10] program established a core set of CDEs and a standard 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 creating minimal sets, or core CDEs, which can be achieved by creating standardized concepts with allowable responses and data representations that would enable and broaden clinical and health data use.

Another challenge for the data science research community is leveraging the massive datasets derived from the same individual across multiple data repositories and resources to preserve participant

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

identity and their intent for sharing. The situation is further complicated by including time-dependent participant data collection methods and requires data linkages. This can only be accomplished if these data have compatible standards and models. Addressing these challenges requires new governance policies for data linkage and approaches to ensure participants' autonomy is respected. Technical capabilities are also needed to support data harmonization and aggregation across different sources, including lifestyle data. These challenges open the door for new algorithms that ensure data security, governance, quality management, and participant consent, utilizing privacy-preserving computing, generative AI, foundation models, and blockchain methods.

Machine learning, deep learning, and AI technologies are significant opportunities to advance basic and clinical research and improve health and health care at individual and community levels. Recognizing these opportunities, NIH launched the Bridge to Artificial Intelligence[11] (Bridge2AI) program in 2022 to produce new flagship biomedical and behavioral datasets (see textbox for Bridge2AI). Creating AI-ready data requires the tools to collect FAIR data at the beginning of the research process (FAIR by design-intentional integration of FAIR principles from the start of the data lifecycle). It also requires methods complementary to the Collective benefit, Authority to control, Responsibility, and Ethics (CARE) principles[12] for Indigenous data sovereignty and governance. New capabilities to facilitate Tribal data sovereignty that respects cultural needs and expectations are also needed.

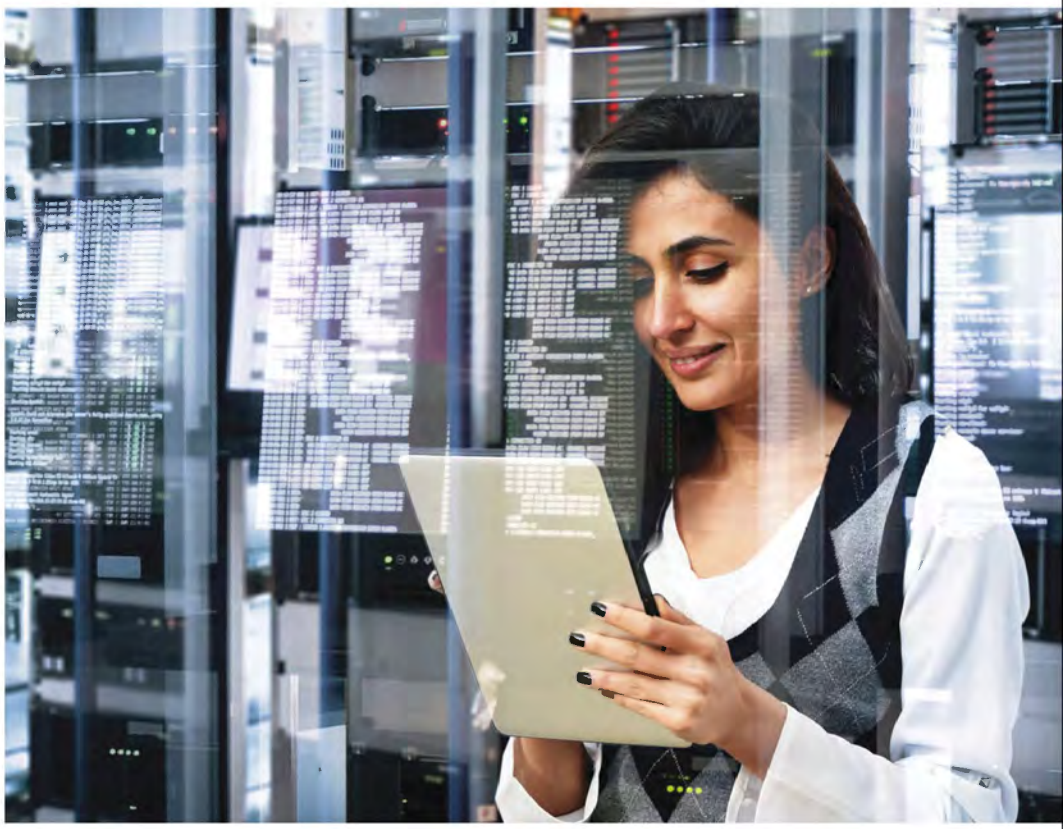
Efforts are also ongoing to utilize cloud service providers for data storage and management and to create interoperable data systems, efforts to enhance biomedical AI for ethical and unbiased data and algorithms,[13,14] and to develop tools that enable researchers to collect, find, and utilize FAIR data and software. Through the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES)[15] initiative, the NIH partnership with cloud service providers Amazon Web Services (AWS), Google Cloud Services, 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

Bridge to Artificial Intelligence (Bridge2AI) will propel biomedical research forward by supporting the 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.

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 tailored to individual ICOs' needs and 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 (LLMs) have generated considerable interest in the possibility of AI recognizing, summarizing, translating, predicting, and generating 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 transparency, validation, and unbiased principles and frameworks, as well as 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[16] that utilize the integrative power of advanced AI methods and 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, digital twins, privacy-enhancing computing such as federated learning, and privacy-preserving data sharing using blockchain. Other applications are emerging in scientific fields such as physics, engineering, computer science, and other industries, from manufacturing to finance. The intersection of these emerging and re-emerging technologies and biomedical research has incredible innovative potential. Expanding NIH investments in these advanced technologies will better position the biomedical and health research community and the agency to take full advantage of these and other new capabilities.

To progress in the next five years, NIH must leverage the exponential growth in the amount and variety of data by developing and deploying sophisticated approaches for data management and embracing advanced technologies, including new methods in data reduction and compression for downstream analysis, while preserving information. Cloud computing will continue to play a significant role in sharing and utilizing scientific workflows at an unprecedented scale. By taking advantage of current and emerging data and computing technologies from the commercial and public sectors, NIH could realize exabyte-scale data science in the next decade.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Plan Content and Implementation

This updated Strategic Plan for Data Science is organized into five goals corresponding to strategic objectives and implementation tactics. These goals will ensure that NIH's strategic approach will address reliability, protect participant privacy and confidentiality, and increase efficiency, trust, and transparency in AI while collectively improving data and tools for research. This strategic plan also supports the DMS Policy by developing new capabilities to streamline data access with a renewed emphasis on metadata consistency and accuracy, the 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. This strategic plan aims to enhance NIH's ability to leverage AI technologies for biomedical and behavioral research, improve clinical and health care data for research, and integrate efficiency, responsibility, and transparency into its visions and objectives.

The implementation tactics are a roadmap for achieving the overarching goals and strategic objectives. 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 to address the health of the nation.

> These efforts will increase the scientific community's ability to address new challenges in accessing, managing, analyzing, integrating, and reusing the enormous amounts of data generated by the biomedical research enterprise.

Through the 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 developing and implementing 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 that can be leveraged for health.
- **Develop** new research, training programs, and collaborations to advance the impact of responsible AI in biomedical/health research. Provide new ways for researchers to search, discover, access, and analyze data across resources and enhance these capabilities' accuracy, validity, transparency, and reproducibility.
- **Engage** researchers and communities in data science training across biomedical, social, environmental, and behavioral disciplines.

This strategic plan aligns with Digital NIH: Innovation, Technology, and Computation for the Future. Digital NIH proposes new approaches to managing and governing NIH technology investments. It also describes a framework to guide the implementation of high-priority, high-value capabilities. Finally, it identifies cross-cutting capabilities that will support data science within NIH (see textbox for Digital NIH).

This strategic plan will encourage greater integration of data science to improve access to and use of biomedical and behavioral data. These efforts will increase the scientific community's ability to address new challenges in accessing, managing, analyzing, integrating, and reusing the enormous amounts of data generated by the biomedical research enterprise.

Digital NIH identifies new governance and funding approaches and capabilities organized by four functional areas: Extramural Research Management, Intramural Basic Research, Intramural Clinical Research, and Administration and Management. Efforts in the five cross-cutting themes support Digital NIH with the following:

- 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Goal 1

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

NIH is committed to data management and sharing across two decades of policies to create and support a data-sharing culture. For example, the final DMS Policy emphasizes the importance of good data management practices and encourages data management and 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] They should also 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 serve its intended scientific community efficiently. Together, the FAIR, CARE, and TRUST principles and the National Science and Technology Council (NSTC) Guidance provide a framework for formalizing the capabilities of a repository (see textbox for 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 DMS Policy expectations by providing resources and guidance to researchers.

NIH will also develop new frameworks to handle the needs of modern data science challenges. For example, the National Center for Advancing Translational Sciences (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 managing 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 guided several significant initiatives, including the Rare Diseases Clinical Research Network (RDCRN), N3C, and A Specialized Platform for Innovative Research Exploration (ASPIRE).[19] 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.

» **NSTC Desired Characteristics of Data Repositories** are designed to be relevant to all repositories that manage and share data 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

OBJECTIVE 1-1

Support the Biomedical Community to Manage, Share, and Sustain Data

The DMS Policy established requirements that emphasize the importance of good data management practices. It also established the expectations for maximizing the appropriate sharing of relevant 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 to promote transparency, reproducibility, and replicability. This forward-thinking policy integrates data management and sharing into the routine conduct of a scientific project. In the process, NIH aims to shift the biomedical research culture into one in which data sharing and reuse are the rule rather than the exception.[20] NIH is developing resources to support the DMS Policy compliance activities of their funded investigators, including the NIH Office of Extramural Research (OER) sharing website,[21] the NIH Biomedical Informatics Coordinating Committee's portal to key data repositories,[22] and the National Institute of Child Health and Human Development's (NICHD)'s Data Repository Finder to support the development of data management and sharing plans. Supporting the DMS Policy requires a coordinated effort between the Office of Science Policy (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 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**[23] **and 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 providing 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.
- **Establish** a data steward program to guide data sharing, leverage existing activities at the Network of the National Library of Medicine (NNLM) National Center for Data Services, and support additional training partnerships, including with societies and associations.
- **Support** tool development and leverage existing tools that will assist researchers with preparing, annotating, and sharing their data.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

> NIH will encourage using community-agreed-upon standard schemas and metadata, enhance automated ontologies and curation processes, and create capabilities for greater data discovery and interoperability across multiple data repositories and knowledgebases.

OBJECTIVE 1-2

Enhance FAIR Data and Greater Data Harmonization

Through enhanced data-sharing efforts and the **NIH STRIDES Initiative**, NIH-supported investigators have generated and made over 200 petabytes of data available in the cloud. This represents a significant amount of biological data, including genomics, clinical studies, phenotypic and omics data, fitness measures and survey data, and data derived from electronic health care systems and lifestyle data. These data are most valuable when researchers combine different data to answer challenging questions such as "What is the relationship between obesity and diabetes in children?" Answering these questions requires wrangling, aggregating, and comparing datasets with supported efforts in standardizing collections, formats, and data models/dictionaries. The key to success in data interoperability is the development and use of agreed-upon data standards, standardized terminologies, and CDEs. The NIH CDE Task Force includes a governance committee that reviews and provides NIH endorsements to submitted CDEs, which are then deposited into the NIH CDE Repository,[24] hosted by the National Library of Medicine (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)[25] and the Cancer Data Standards Registry and Repository (caDSR).[26]

During the past five years, several large NIH programs have undertaken an effort to enhance data harmonization, including Helping End Addiction Long-Term (HEAL)[27] (see textbox for HEAL CDE program) that is standardizing metadata; NCI's Cancer Data Aggregator[28] that is mapping harmonized data elements to FHIR®, OMOP, and other data models; National Institute of Dental and Craniofacial Research (NIDCR)'s FaceBase[29] that is providing guides to scientists to produce 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)[30] provide an agreed-upon data standard that can be widely adopted and 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 committed to adopting an agreed-upon set of mental health science standards.[31] 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 normalized analysis, interpretation, and more accurate 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 using community-agreed-upon standard schemas and metadata, enhance automated ontologies and curation processes, and create capabilities for greater data discovery and interoperability across multiple data repositories and knowledgebases. This objective will inform future activities of the NIH Plan for Public Access to Research Results.

Implementation Tactics

- **Enhance** abilities to improve data and metadata quality, including data quality assurance and quality control.
- **Encourage** usage of open and standardized schemas, ontologies, and data formats.
- **Enhance** automated processes for ontology use and data curation, including metadata enrichment.
- **Create** a minimal set of consistent and computable CDEs or concepts with consistent data models.

HEAL Common Data Element (CDE) Program

The NIH HEAL Initiative research portfolio spans many data types that are rich resources for future studies. The initiative's CDE Program supports its Public Access and Data Sharing policy, which requires researchers to develop plans to share their project's underlying primary data and will connect and expose data via the HEAL Platform.

Clinical pain research grantees collaborate and agree to use CDEs across nine core pain domains for patient-reported outcomes (PROs) to facilitate cross-study comparisons and improve the interpretability of findings.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

OBJECTIVE 1-3

Strengthen NIH Data Repository and Knowledgebase Ecosystem

Data repositories and knowledgebases are essential to maximizing the value of the scientific research enterprise and serve as critical components in the implementation of the DMS Policy, preserving, archiving, and sharing scientific data. As the volume, velocity, and variety of data collected and stored from biomedical research continue to increase, the need for making scientific research data and information FAIR and the vital role of repositories and knowledgebases will continue to play a critical role in accelerating scientific discoveries to improve health. 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.

Biomedical knowledgebases extract, accumulate, organize, annotate, and link the growing body of information related to and reliant on core datasets. Information curation 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 keys to discovering, integrating, and reusing data and knowledge. 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 metrics.
- **Provide** sufficient metadata and semantic annotation.
- **Support** a process for data life-cycle analysis, long-term preservation, and trustworthy, responsible stewardship.

NIH supports both unrestricted access and controlled access to data repositories. NIH controls access to some data to protect participants who have provided information to advance NIH's research efforts. Such data can contain sensitive information or Controlled Unclassified Information that requires safeguarding. To respect research participants' autonomy, privacy, and confidentiality, data are stored in Controlled-Access Data Repositories (CADRs) to manage and share research participant data at the individual or aggregate/cohort level. Controlled access datasets often have data use limitations requiring NIH authorization which, although necessary, can pose challenges to accessibility by the research community. Controlled access processes are labor- and resource-intensive, which limits their scalability and speed to access data for research. To accelerate research while protecting sensitive data, NIH is developing, testing, and deploying advanced technologies and restructuring data management processes to streamline controlled-access procedures. In addition, NIH seeks to develop common approaches and infrastructure and bring all CADRs to NIH-wide system standards for security postures for addressing streamlined data access and more easily detect and mitigate data management incidents across NIH-supported repositories.

In recent years, NIH developed several capabilities to promote a data ecosystem, including a collaborative approach for data management and sharing with seven generalist repositories (see textbox for Generalist Repository Ecosystem Initiative (GREI) and support for the use of persistent unique data identifiers through a consortium membership with DataCite.[32] By partnering with DataCite, NIH data resources will be able to enhance data sharing and enable researchers to cite and reuse research outputs. These efforts strengthen data management and sharing by enhancing visibility, citation in scholarly publications, preservation, future data reuse, and 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 to make these research data and information FAIR underscores the critical role of data repositories and knowledgebases. Developing sustainable data resources requires understanding and using metrics to evaluate a repository's usage,

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

- **Establish** a standard set of cohesive and consistent capabilities, services, metrics, and social infrastructure.
- **Raise** general awareness and help researchers adopt FAIR principles to share better and reuse data.
- **Establish** consistent metadata, develop use cases for data sharing, and train and educate researchers on FAIR data and the importance of sharing.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

utility, and impact. Moreover, promoting easier access to research products with appropriate security controls and 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.

Just as necessary 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 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[33]" due to Tribal consultation. Similarly, international data sharing should respect regional and population-specific data governance considerations, especially involving data generated in low- and middle-income countries.

The ubiquitous use of data resources in biomedical research, coupled with a greater emphasis on data management and sharing, has dramatically 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 concerning 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 NIH faces in supporting widely used data resources are mirrored at the federal and international levels. NIH provides the most significant support for the most commonly used biomedical data resources, with resources managed by NLM serving as an essential node in the international biomedical data ecosystem. For this reason, NIH has been involved in several efforts, including CoreTrustSeal,[34] Research Data Alliance (RDA),[35] DataOne,[36] Open Science Framework,[37] DataCite,[38] the Wellcome Trust,[39] and the Global Biodata Coalition.[40] For more than 30 years, NLM has 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,[41] and continuing to develop relationships with critical global actors. These organizations provide a platform for the international community to work together better to coordinate the management and sharing of scientific data. Over the next five years, NIH will continue efforts to facilitate the long-term sustainability of the global biodata ecosystem that is relied upon by all biomedical researchers, including NIH-funded ones.

Implementation Tactics

- **Enhance** data repositories and knowledgebases that promote impartial 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 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 guide data preservation and long-term data archiving.
- **Develop** a comprehensive, coherent, acceptable sustainability framework for identifying and supporting the most widely used and impactful NIH data resource portfolio.
- **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 enables metadata 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 data resources and 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 various stakeholders, including RDA, GO-FAIR, biomedical societies, and international partnerships such as with the Global Alliance for Genomics and Health (GA4GH), ELIXIR, and Global Biodata Coalition.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 necessary for patient care and treatment. During the last decade, the United States has seen an increase in the generation and usage of this data in research, including through efforts such as the All of Us Research Program. 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, integrating multiple types of RWD 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.

Adopting and integrating health care data standards with research data standards is required to enable the biomedical and behavioral research community to take full advantage of the multitude of health-derived data. Where appropriate, NIH will work with federal agencies, medical institutions, health IT developers, and vendors 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 enhance abilities to link RWD from multiple sources with appropriate informed consent 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 primary goal is to increase the use and utility of health care-derived data for research with proper security and privacy safeguards. Activities that integrate clinical data and RWD should be developed to achieve this goal, including data from wearables and data originating from health care settings such as mental health, dental, pathology, and ophthalmology settings.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

OBJECTIVE 2-1

Improve Access to and Use of Clinical and Real-World Data

The health care enterprise is a rich data source for biomedical and behavioral researchers. However, methods and policies for sharing these data with the broader research community differ in complexity from more traditional research settings and data-sharing expectations and approaches. Unique challenges in data quality, privacy and confidentiality, policy, regulation, and ethics require considerations when sharing and using health care and administrative data. Informed consent for collecting, using, and sharing these data is essential for respecting participant rights and maintaining public trust. NIH will increase the capabilities of 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 build on recent NIH guidance and templated informed consent language[42] for secondary research use of data and specimens.

In addition, wearable device data require substantial efforts to extract, transform, and structure owing to challenges like securing personal information and the suitability of this data 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., National Institute of Biomedical Imaging and Bioengineering [NIBIB]'s Medical Imaging and Data Resource Center [MIDRC]);[43] 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 has benefits and challenges, all are essential for the NIH data ecosystem. NIH is committed to improving data FAIRness, transparency of data governance and stewardship expectations, and streamlining requirements for accessing and using data derived from care.

In understanding the relationship between health, environment, and lifestyle, researchers find that linking and combining individual-level health data with other RWD and digital sources improves our understanding of health and disease. 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 (NEI) 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).[44] Similarly, new and improved environmental data sources continue to emerge from industry, federal agencies, and the research community, as well as through new AI methods for estimating individual-level exposures.

However, a need remains to better understand data linkage's ethical, legal, and social implications. Additionally, researchers must know how to define and apply rules to protect study participants and navigate relationships with the public and private sectors to ensure that linked data is appropriately governed, shared, and used. NICHD recently commissioned a report on the technology and governance considerations for pediatric pandemic record linkages[45] 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 use and appropriately combine health care data sources where allowable.

"In understanding the relationship between health, environment, and lifestyle, researchers find that linking and combining individual-level health data with other RWD and digital sources improves our understanding of health and disease."

Implementation Tactics

- **Enhance** methods for informed consent in cases where data are combined from multiple sources and/or combined over longitudinal studies.
- **Create**, test, validate, and adopt methods to enable researchers to use multi-modal and digital data combined from multiple sources, including 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 RWD platforms.
- **Develop** 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 control of their data and how they are used.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 benefits 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)[46] to distribute key terminology, classification, and coding standards, supports to key terminologies that are now required for use in certified EHRs (e.g., Systematized Nomenclature of Medicine [SNOMED], Logical Observation Identifiers Names and Codes [LOINC], RxNorm, among others), and associated resources to promote more effective and interoperable biomedical information systems and services.

Data sharing has made significant progress in the health care community, partly due to the development and adoption of terminology and exchange standards. In 2020, NIH held the virtual workshop, *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 for 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 increase data reuse across both clinical care and research settings and is enabling patients to access their clinical data. In addition, FHIR® or other such systems should facilitate population science to foster the integration of community-based 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, organizations, and other federal agencies that work with health data standards, including developing research use cases for United States Core Data for Interoperability Plus (USCDI+) and Trusted Exchange Framework and Common Agreement (TEFCA).

The Fast Healthcare Interoperability Resources (FHIR®) standard enables electronic health care 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 systems to one another through FHIR® API to transmit electronic health data.

Implementation Tactics

- **Implement** agile programs that convene researchers and developers to develop, test, validate, and adopt health IT and standards based on scientific use cases and provide feedback based on lessons learned.
- **Promote** development, training, and adoption of FHIR® and related technologies, such as SMART on FHIR®, to enable further clinical research tools and data exchange in research infrastructure, cohort discovery, and applied real-world research.
- **Partner** with other agencies such as the Office of the Assistant Secretary for Technology Policy (ASTP), 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 Environmental and Lifestyle Exposome in Health Research

The environmental and lifestyle exposome encompasses all environmental exposures an individual experiences throughout life, including diet, physical activity, social factors, pollution, and toxins. It provides a fuller picture of health influences beyond genetics and clinical data alone. Several studies have shown that environmental exposures (e.g., pollution and toxins) and lifestyle factors (e.g., diet and substance use) are significant contributors to the development and progression of various chronic diseases such as cardiovascular disease, cancer, and neurodegenerative conditions. By leveraging lifestyle and environmental data alongside traditional biomedical inputs, researchers can unlock more profound insights into the root causes of chronic diseases and health disparities, ultimately improving prevention, diagnosis, and treatment strategies. Incorporating these factors into AI-driven health research can lead to more accurate and personalized models for predicting disease risk, understanding health trajectories, and designing targeted interventions. However, this can be accomplished by adopting standardization, collection, reporting, and leveraging lifestyle and environmental data measures in both existing and emerging data sources and fostering appropriate data linkages with EHRs, clinical research, and genomic data.

Implementation Tactics

- **Identify** lifestyle and environmental determinants of health.
- **Support** demonstration projects to test how best to capture these data for interoperable electronic data exchange.
- **Develop** infrastructure and tools for extracting structured and unstructured data from multiple sources and enable iterative models to include these data in training.
- **Enable** the linkage of these data with other data, such as clinical, RWD wearable sensor, omics data, and administrative data, and develop demonstration projects to show the technical feasibility of such linkages when appropriate and when there are no increased risks of reidentification for small communities.
- **Support** real-world pilots integrating environmental and lifestyle data with clinical CDEs.

> By leveraging lifestyle and environmental data alongside traditional biomedical inputs, researchers can unlock more profound insights into the root causes of chronic diseases.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

› OBJECTIVE 2-4

Empower Clinical Data Science through Cross-Disciplinary Training

NIH recognizes that maintaining and enhancing clinical research informatics as a career path requires clinical training and training in informatics, analytics, data models, and standards. This training will focus on using data from clinical, health care, and real-world settings to understand better the regulatory and policy standards for generating and using this 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 help reach new research areas (National Institute of Diabetes and Digestive and Kidney Diseases [NIDDK] Strategic Plan[47], National Institute of Neurological Disorders and Stroke [NINDS] Strategic Plan[48]). Data management and linking require social, technological, and data science partnerships. Research involving linking multiple data types and exposure to new opportunities in technologies will require a diverse cadre of colleagues for future collaborations. In addition, other health-related research fields, including dental and ophthalmology, can benefit from enhanced data science training to integrate clinical data, imaging data, and omics data with diverse data types from other health-related fields, including social science.

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 responsible data use beyond legal and regulatory requirements, with special considerations for linked/merged data.
- **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 biomedical and behavioral research innovation. NIH will seek partnerships and collaboration across multiple stakeholders, including ASTP, in implementing relevant data standards, large health care systems and health technology groups, and applicable standards development organizations. Significantly, NIH will increase and improve community engagement and partnership opportunities to collectively build trust, 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, 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 CDEs for interoperability and reproducibility, 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Goal 3

Provide New Opportunities in Software, Computational Methods, and AI

The biomedical research enterprise generates Immense amounts of data from fundamental experiments using cells and research organisms to clinical studies and community-level epidemiological research. These data are valuable for the original research question and secondary data analyses for study replication or for other researchers asking different questions. Harnessing research data for data-driven discovery remains a significant challenge that requires attention to data quality, quantity, computability, standards, and new computational and AI methods.

AI[49] consists of a collection of data-driven technologies with the potential to advance biomedical research significantly. Advances in AI have led to exciting opportunities, including improvements in protein structure prediction and protein design, computer-aided diagnosis with medical images, better understanding of cancer phenotypes, and LLMs to interpret clinical, electronic health care records and reports to aid in clinical decision support. AI algorithms can analyze vast amounts of data, identify complex patterns, and gain deeper insights into fundamental scientific phenomena. These approaches allow researchers to unlock new avenues for exploration, drive scientific discoveries, and further our understanding of underlying principles in various fields of health. With the ability to process billions of parameters, AI could significantly improve future health research via decision support systems, rapid data annotation, including medical and tissue image processing, and the daunting task of organizing large bodies of disparate medical information. However, utilizing AI for biomedical research and health care practices is still hampered by inconsistent and incomplete data. Making data FAIR and AI-ready requires multi-disciplinary expertise, experimentation, and, often, iterative feedback from AI applications and experts. In particular, establishing ground truth, standardization, and validation of training datasets is fundamental in biomedical applications where inaccuracies have amplified repercussions for patients, providers, and researchers.

AI is a priority across federal agencies, and as such, the National AI Initiative Act of 2020[50] 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: (1) spur innovation, (2) increase the breadth of talent, (3) improve capacity, and (4) advance trustworthy AI. The roadmap to implementing the NAIRR[51] calls for an all-of-government approach to leveraging resources, such as massive computing infrastructures, extensive data, and a growing talent pool of researchers, to realize this vision. For the initial pilot period, NIH contributes high-quality data assets, privacy-preserving methods, responsible practices, and methods for growing capacity, broadening participation in AI, and co-leading NAIRR Secure with the Department of Energy (DOE).

To enhance the robustness and utility of data analysis and processing methods, NIH will take advantage of innovations in open and sustainable software and algorithms. NIH will support partnerships to co-design emerging capabilities, including new techniques in AI, including generative AI, computational image analysis, and machine vision; new infrastructures such as quantum information sciences; automated workflows, and new tools for researchers to leverage data in a transparent, explainable, responsible, and effective manner; and new ways of enabling communities to develop software through collaborative projects. NIH is removing barriers to innovation and empowering a broader research community to make AI work for all Americans.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

OBJECTIVE 3-1

Enhance Health for All through AI

AI has shown promise in improving medical diagnoses and understanding underlying biological processes. The rapid growth in the volume of data generated through EHRs and other biomedical research provides a rich foundation for AI applications to improve health for all. However, realizing this potential requires overcoming several challenges. Key considerations include ensuring transparency and reproducibility, accounting for ethical, legal, and social implications, and engaging in risk mitigation in AI systems. Making data FAIR and AI-ready also requires interdisciplinary skills. Particularly for biomedical and behavioral research, AI readiness necessitates attention to individual and societal impacts of datasets used to train the AI models. While different classes of AI may have unique data requirements, AI generally requires machine-readable data that are well described with ontologies and schema so that the algorithm can parse data. Including data quality, such as accuracy, completeness, consistency, and reliability, in AI metadata standards will help address the trustworthiness of the information provided by AI algorithms. Misrepresentation in datasets, algorithms, and applications raises risks and increases potential harms related to privacy, confidentiality, and the health of all Americans. The National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF 1.0)[52] identifies risks and/or potential harms. It discusses how managing these risks will lead to more trustworthy AI systems and enable AI developers and users to understand better and take responsibility for their models' and systems' potential limitations and uncertainties.

Many data-related challenges hinder fully leveraging AI capabilities. Data quality is critical because datasets often contain inconsistencies, errors, and missing values, leading to inaccurate model outputs. AI algorithms require large and comprehensive datasets to train models. Achieving reliable and repeatable results from these models across different care settings and environments requires consistent and representative data. Additional challenges arise when using multiple data types with AI (i.e., multimodal AI), including integrating and aligning various data types (including text, images, audio, video, and omics) and managing the heightened computational and storage requirements. In the biomedical research domain, an additional data-related challenge is that, without proper protection, AI models can inadvertently leak sensitive health data through mechanisms such as training data memorization, model inversion attacks, and membership inference attacks.

Additionally, the shortage of skilled AI researchers and clinicians and the persistence of siloed scientific approaches further hinder progress. To fully leverage AI in biomedicine, fostering innovative, cross-disciplinary frameworks, building trusted partnerships that empower researchers and communities across the U.S., and promoting responsible and inclusive AI adoption in health care is essential. Recognizing these challenges, NIH launched the AIM-AHEAD program in July 2021 (see textbox for AIM-AHEAD). It is a nationwide network of institutions and organizations designed to build AI talent and technology and support multidisciplinary research projects that harness AI to improve the health of the public. The program strengthens the AI capabilities and infrastructure of institutions and hospitals that otherwise would not have had the resources or the capacity to investigate advances in AI.

To fully take advantage of AI innovations, creating reproducible, trustworthy, and accurate AI tools, complete data, and robust practices is critical. Living

> To fully leverage AI in biomedicine, fostering innovative, cross-disciplinary frameworks, building trusted partnerships that empower researchers and communities across the U.S., and promoting responsible and inclusive AI adoption in health care is essential.

resources (frameworks, tools, best practices) – based on agile iteration with the research community – across the continuum of AI development can lower barriers to entry and enable researchers to focus on innovation. Connecting principles embraced across the private, non-profit, and academic sectors to these resources will allow practical approaches for researchers by building trust, explainability, and transparency.

Essential 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 capabilities contribute to improving the health of all Americans. In partnership with the NIH ICOs, the agency will support advanced 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 and to inform prevention efforts (NICHD Strategic Plan,[53] National Institute of Mental Health [NIMH] Strategic Plan,[54] National Heart, Lung, and Blood Institute [NHLBI] Strategic Vision[55]).

NIH's AIM-AHEAD is a large nationwide network of institutions and organizations designed to:

- **Build** talent of AI researchers and clinicians and promote AI literacy for non-practitioners.
- **Improve** health by supporting multidisciplinary research projects that harness AI.
- **Enhance** the AI capabilities and infrastructure of institutions and communities across the US.

Implementation Tactics

- **Develop** socio-technical solutions, including resources and guidelines, to improve the quality of AI training sets and algorithms and support their rigorous assessment, validation, and adoption.
- **Build** a robust pipeline of AI researchers and clinicians while promoting literacy of AI among non-practitioners.
- **Establish and operationalize** community engagement for improved data, methods, and sources for AI. When appropriate, support research in developing, validating, and using synthetic clinical datasets for AI training and applications.
- **Develop** tools and training opportunities to help researchers create and prepare FAIR and AI-ready data, including ontologies, schema, and data quality measures.
- **Support** the development and use of responsible 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 enable data translation to knowledge, including AI tools for 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 (NHLBI Strategic Vision[56]). Multi-dimensional data integration remains a significant challenge for biomedical and behavioral research.

Additionally, low-code, no-code technologies allow 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 support of digital twins approaches to model organs, systems, individuals, and populations; new capabilities for privacy-preserving computing and preserving technologies; and quantum computing should also be explored. Finally, best practices for transparent, open, and reusable software and algorithm development 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, responsible, and transparent standards.
- **Leverage** new passive and mobile devices and technologies for data collection and analysis with improved practices for informed consent.
- **Facilitate** FAIR software and algorithms with sufficient documentation and metadata and enhance ethical frameworks.
- **Leverage** advances in computational methodologies and studies to create new opportunities for ethical and social science research.
- **Investigate** the potential of digital twins approaches to organs, systems, individuals, and populations.
- **Explore** opportunities to combine theory-based modeling and simulations with data-driven capabilities.
- **Promote** opportunities to expand the software development community by making these resources accessible to more people.

OBJECTIVE 3-3

Support FAIR Software Sustainability

Software is an integral component of biomedical and behavioral research due partly to the speed and growth of new technology innovations in software and computing, including AI, computer transistors, and microchips. NIH collaborates across ICOs to support developing and enhancing software tools for open science[57] by fostering new collaborations between biomedical and clinical scientists and software engineers. For example, significant progress has been made in developing software 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 achieved over 275 million compute hours for data analysis in the cloud. Yet challenges remain in creating FAIR software.[58] 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 larger biomedical research communities sustain research software over time. NIH recently issued best practices for software sharing[59] that align with the FAIR software principles.

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 non-profits are tackling software sustainability head-on, including the NSF program on Cyberinfrastructure for Sustained Scientific Innovation,[60] the recent NIH supplements to support enhancement of software tools for open science, the Schmidt Futures Virtual Institutes for Scientific Software,[61] and the Chan Zuckerberg Initiative's program for Essential Open-Source Software for Science.[62] A long-standing program at NIH is NCI's Information Technology for Cancer Research (ITCR)[63] program. The ITCR program serves the informatics needs of the cancer research continuum and supports informatics resources across the development lifecycle (see textbox for ITCR).

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, optimize research software for robustness, and ultimately increase software sustainability.

Information Technology for Cancer Research (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**
- **Prototype and Hardening of Software**
- **Enhancement and Dissemination of Software**
- **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 the 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

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 various scientific domains, and integration of tools from other domains into biomedical research. Additional measures of progress include improved auditing AI of models, development of 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 like the Research Software Alliance.[64]

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Goal 4

Support for a Federated Biomedical Research Data Infrastructure

Throughout the last five years, NIH has seen remarkable growth in supporting and using biomedical data repositories and platforms for biomedical research. More and more of these data infrastructures are 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 is to provide greater connections across NIH cloud-based data platforms for easier access to multiple datasets, streamlining similar functions, and enabling more facile analytics.

Current cloud-based data platforms include the NHLBI's BioData Catalyst®,[65] 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)[66] which provides secure access to a large, comprehensive, and expanding collection of cancer research data; the Kids First Data Resource[67] which houses data on 44 childhood cancer and structural birth defects cohorts; the National Human Genome Research Institute (NHGRI)'s Genomic Data Science Analysis, Visualization and Informatics Lab-space (AnVIL)[68] genomic data sharing and analysis platform; *All of Us*[69] which has collected data from over 500,000 participants; the NIH database of Genotypes and Phenotypes (dbGaP)[70] 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.

NIH ICOs are also developing data ecosystems, including the NIH Cloud Platform Interoperability initiative;[71] the NIBIB's MIDRC[72], 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 (NIAID) Data Ecosystem (see textbox for NIAID Data Ecosystem); the Common Fund Data Ecosystem;[73] and NCBI's Comparative Genomic Resource,[74] 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 is responsible for integrating 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' 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 the 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 [NIEHS] Strategic Plan)[75] 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 discoveries.

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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

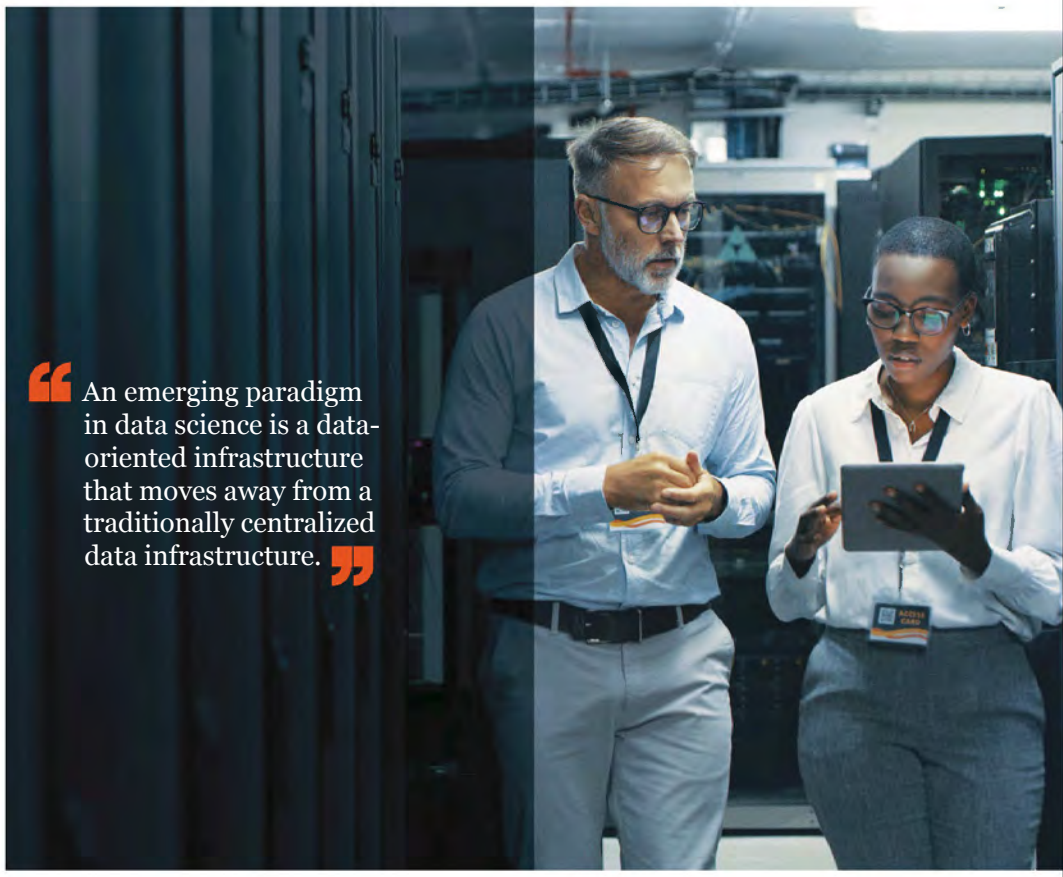
OBJECTIVE 4-1

Develop, Test, Validate, and Implement Ways to Federate NIH Data and Infrastructure

As the National Institute on Aging (NIA) Strategic plan articulated,[76] NIH needs to develop comparable databases on health outcomes, risk factors, and lifestyles. An emerging paradigm in data science is a 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 maximum interoperability and economies of scale. 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 catalogs, 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 data and adopt standard processes and interfaces, analysis tools, and services that can be used broadly for biomedicine and behavioral research. In line with this vision, this objective aims to improve efficiencies and maximize researchers' ability to find, access, and use data 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 for Research Auth Service [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. As a result, researchers could demonstrate data access across multiple cloud-based NIH data repositories and perform combined analysis with meaningful results.[77] These initial piloted efforts are the genesis of an NIH-wide federated data ecosystem and are articulated as priorities in the Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap[78] priorities.

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., standard search capabilities, APIs, identity 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 generated from federally funded research and stored in cloud-based data repositories.

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 FY2023 Federal Cybersecurity R&D Strategic Plan Implementation Roadmap to protect systems and ensure confidentially, integrity, as well as availability and privacy of data. The RAS initiative is advancing data infrastructure and ecosystem goals as defined in the 2018 NIH Strategic Plan for Data Science.



> An emerging paradigm in data science is a data-oriented infrastructure that moves away from a traditionally centralized data infrastructure.

Implementation Tactics

- **Enhance** the utilization of cloud and hybrid computing architectures and provide opportunities for low-resource institutions to access and utilize NIH-supported cloud capabilities.
- **Support** efficiency and sharable technologies across NIH data platforms, including increasing new and existing technology industry partnerships.
- **Expand** RAS to increase researchers' access to data and ensure accountability for privacy protection and cybersecurity of systems.
- **Ensure** a robust and connected data resource ecosystem that supports linkages and interoperability across NIH-supported cloud platforms for curation, analysis, and sharing of data and metadata.
- **Develop** new data search and discovery capabilities by enhancing metadata standards and indexing techniques and improving data interoperability and harmonization.
- **Explore** new paradigms in computing for biomedical and behavioral applications.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

GOAL 4

Partnerships and Measuring Progress

For this goal, the potential measures of progress should advance biomedical research through an integrated infrastructure and include ease of findability of datasets and the ability of researchers to integrate NIH data and increase the 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. This work will require partnership and collaboration across multiple NIH ICOs and scientific organizations such as GA4GH, RDA, and Research Software Alliance.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Goal 5

Strengthen a Broad Community in Data Science

Data science has become essential to progress in biomedical and health research. NIH is committed to building a broad and collaborative community that ensures all members – learners, researchers, and professionals – can effectively use data science skills and knowledge to advance research.

Meeting the full spectrum of research needs requires talent at varying levels of data expertise, including:

- **Data science literate**: Comfortable reading and understanding outcomes from data science approaches.
- **Data science savvy**: Data science literate and able to actively use data science approaches in designing research projects and collaborating with data scientists.
- **Data science proficient**: Experts in bioinformatics, AI, 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 research training and education programs in data science. These collective efforts will continue to grow, focusing on increasing the data science community to take full advantage of the talent nationwide. A growing community will help NIH identify new research questions, facilitate the translation of scientific data and findings to the public, and build the public's trust in data-driven biomedical research. The time has never been better for all biomedical and behavioral researchers to take full advantage of data science, including innovations from cloud computing, the availability of significant amounts of biomedical and behavioral data, and new advances in AI. To ensure that data science advances in biomedical research can benefit all, NIH will help create a vibrant, innovative, and inclusive data science community.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

› OBJECTIVE 5-1

Increase Training Opportunities in Data Science

Based on the network of existing extramural training programs, NIH will coordinate across the ICOs to promote data science training and education. 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. Therefore, early intervention strategies in research education are 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 retaining 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.
- **Increase** pairing 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.
- **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, NIH has made significant progress in enhancing its administrative and programmatic data science workforce to meet the growing needs of biomedical and health research. For example, in 2020, NIH launched the Data and Technology Advancement National Service Scholars (DATA Scholars) Program,[79] which recruited 31 highly skilled data scientists from academia, industry, and government to lead transformative, high-impact projects across 17 ICOs. These scholars spend one to two years at NIH applying advanced data science methods, including AI, cloud analytics, and interoperability frameworks, to improve transparency, reproducibility, and operational efficiency. In collaboration with the non-profit organization Coding it Forward, NIH has also implemented the Civic Digital Fellowship[80] program 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 of data science approaches for established investigators, strengthening data science skills among clinician-scientists, and expanding the role of data science in funded research nationwide.

Implementation Tactics

- **Expand** the number of data science investigators and broaden data science's reach 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 skills.
- **Facilitate** recruitment and retention of innovative data science talents at the NIH.
- **Develop** a pathway for early-career data scientists to join the NIH.

> "These efforts strengthen the data science workforce within NIH and provide a strong foundation for continued growth."

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

OBJECTIVE 5-3

Enhance 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 broaden the recruitment of data science trainees to the Intramural Research Program (IRP).[81] The NIH IRP is the largest biomedical research institution and conducts long-term, high-impact science that would otherwise be difficult to undertake. Moreover, NIH supports Biowulf,[82] 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 foster a steady pipeline of intramural data science students and researchers, develop a supportive network for data science trainees in the IRP, and enhance intramural computational capabilities to realize new opportunities and partnerships across NIH, 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, and postdoctoral fellows.
- **Support** cross-institute intramural data science projects and enhance the 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 take advantage of new opportunities in cloud computing, AI, and other data science and computing initiatives.

OBJECTIVE 5-4

Broaden and Champion Capacity Building and Community Engagement Efforts

Developing and sustaining a biomedical and behavioral research workforce that reflects the communities being served and supported in an environment that nurtures their success is essential to advancing health research. However, many researchers and institutions continue to encounter challenges in leveraging data science, including easy access to cloud computing environments, sufficient training and mentoring in data science, and opportunities to apply their unique expertise to biomedical challenges. NIH is committed to broadening participation in data science by engaging a wider range of institutions and investing in efforts to grow human capital, enhance institutional infrastructure, and build partnerships.

Following the first strategic plan for data science, new programs have resulted from data science partnerships with the National Institute of General Medical Sciences (NIGMS), including enhancement to the IDeA Networks of Biomedical Research Excellence (INBRE) program to support new data science cores and the development of cloud-based learning modules for the NIH CloudLab. Institutional data science capacity has been enhanced through administrative supplements supported in collaboration with several ICOs. Training activities have included workshops and professional certificates. At the same time, new computing infrastructure continues to support data science practices, and partnerships with industry and federal laboratories provide access to real-world applications. NIH also supported efforts in the Native American Research Centers for Health (NARCH) program (see box for NARCH) to advance data science training, hire experts, develop early-stage investigator careers, and coordinate tribally managed health organizations. 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.

The Native American Research Centers for Health (NARCH) program supports federally recognized American Indian and Alaska Native (AI/AN) Tribes and tribally based organizations in research, capacity building, and career development opportunities. The program fosters partnerships between tribes and research institutions, supports health research, provides AI/AN researchers training, and strengthens community research infrastructure to improve health outcomes.

NARCH projects have actively focused on data science. Highlights from NARCH project teams include establishing data repositories and governance frameworks, organizing data science conferences, and offering training courses on grant writing, data management, biostatistics, and AI.

Implementation Tactics

- **Collaborate** with existing NIH programs to develop and expand programs to enhance data science capacity in nationwide institutions.
- **Leverage** datasets in NIH-supported data repositories and data platforms as training resources.
- **Build** synergies across government, academic, non-profit, international, and industry stakeholders.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

GOAL 5

Partnerships and Measuring Progress

For this goal, potential measures of progress include an increase in the number of data science trainees and the number of trainees leveraging NIH-supported data platforms, as well as an increase in the number of trainees who matriculate to data science careers, data scientists recruited to the NIH, and 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.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

Endnotes

1 https://sharing.nih.gov/data-management-and-sharing-policy
2 FAIR data denotes Findable, Accessible, Interoperable, and Reusable datasets.
3 https://datascience.nih.gov/strategicplan
4 NIH defines metadata as information intended to make scientific data citable, interpretable, and reusable.
5 www.hl7.org/
6 https://datascience.nih.gov/clinical-
7 https://allofus.nih.gov/
8 https://ncats.nih.gov/n3c
9 www.nih.gov/research-training/medical-research-initiatives/radx
10 www.nibib.nih.gov/covid-19/radx-tech-program/mars
11 commonfund.nih.gov/bridge2ai
12 www.gida-global.org/care
13 www.nimhd.nih.gov/resources/schare/
14 ncats.nih.gov/funding/challenges/bias-detection-tools-in-health-care
15 https://datascience.nih.gov/strides
16 https://datascience.nih.gov/sites/default/files/NIH%20Search%20Workshop%20Summary%20Final.pdf
17 https://ocio.nih.gov/Documents/Digital%20NIH%20Strategy_2023.02.06_Final_508C.pdf
18 www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf
19 Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020).
20 https://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 https://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.
25 https://cde.nlm.nih.gov/home
26 https://evs.nci.nih.gov/
27 https://datascience.cancer.gov/resources/metadata
28 https://heal.nih.gov/
29 https://datacommons.cancer.gov/cancer-data-aggregator
30 www.facebase.org/
31 https://health.mitre.org/mcode
32 https://iamhrf.org/projects/driving-adoption-common-measures
33 https://datascience.nih.gov/news/nih-joins-datacite-consortium
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 https://osf.io
39 https://datacite.org
40 https://wellcome.org
41 https://globalbiodata.org/
42 www.insdc.org/
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
47 www.nlm.nih.gov/research/umls/index.html
48 www.niddk.nih.gov/about-niddk/strategic-plans-reports/niddk-strategic-plan-for-research
49 www.ninds.nih.gov/modules/custom/ninds/assets/files/NINDS_Strategic_Plan_2021-2026_Final_508C.pdf
50 AI includes knowledge representation, machine learning, natural language processing, computer vision and perception, deep learning, and language models.
51 www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1218
52 www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf
53 https://doi.org/10.6028/NIST.AI.100-1
54 www.nichd.nih.gov/sites/default/files/2019-09/NICHD_Strategic_Plan.pdf
55 www.nimh.nih.gov/sites/default/files/documents/about/strategic-planning-reports/NIMH%20Strategic%20Plan%20for%20Research_2022_0.pdf
56 www.nhlbi.nih.gov/about/strategic-vision
57 www.nhlbi.nih.gov/sites/default/files/2017-11/NHLBI-Strategic-Vision-2016_FF.pdf
58 https://datascience.nih.gov/tools-and-analytics/administrative-supplements-to-support-enhancement-of-software-tools-for-open-science
59 Barker, M., Chue Hong, N.P., Katz, D.S. *et al.* Introducing the FAIR Principles for research software. *Sci Data* 9, 622 (2022).
60 datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq
61 https://beta.nsf.gov/funding/opportunities/cyberinfrastructure-sustained-scientific
62 www.schmidtfutures.com/our-work/virtual-institute-for-scientific-software
63 https://chanzuckerberg.com/eoss
64 https://itcr.cancer.gov/about-itcr
65 https://www.researchsoft.org
66 https://biodatacatalyst.nhlbi.nih.gov
67 https://datascience.cancer.gov/data-commons
68 https://kidsfirstdrc.org
69 https://anvilproject.org
70 https://allofus.nih.gov
71 www.ncbi.nlm.nih.gov/gap/
72 https://datascience.nih.gov/nih-cloud-platform-interoperability-effort
73 www.midrc.org
74 https://commonfund.nih.gov/dataecosystem
75 www.ncbi.nlm.nih.gov/comparative-genomics-resource
76 www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf
77 www.nia.nih.gov/sites/default/files/2020-05/nia-strategic-directions-2020-2025.pdf
78 Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) (biorxiv.org)
79 www.nitrd.gov/pubs/FACE-SP-FY22-Implementation-Roadmap.pdf
80 https://datascience.nih.gov/data-scholars-2022-closed#overview
81 www.codingitforward.com/fellowships
82 irp.nih.gov/
83 hpc.nih.gov/

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

APPENDIX 1

Accomplishments

From the first NIH Strategic Plan for Data Science

GOAL 1

Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

- NIH's Partnership with Google Cloud Services, Amazon Web Services, and Microsoft Azure through the STRIDES program has resulted in over 200 petabytes of biomedical data on the cloud, 320 compute hours, 5,000 researchers trained, 1,300 program workers working in the cloud, and the development of the NIH CloudLab.
- The development of the RAS for single sign-on and efficient data access across NIH data platforms includes integrating over 30 data programs into RAS and partnering 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, resulting from single sign-on and cross-platform data analysis.

GOAL 2

Promote Modernization of the Data-Resources Ecosystem

- NIH has supported funding opportunities for data resources (databases and knowledgebases), which have resulted in 17 new awards across 7 NIH Institutes and Centers. NIH has also supported supplemental funding to existing databases to align with the OSTP characteristics for FAIR data repositories, which has resulted in 21 awards across 12 NIH Institutes and Centers.
- NIH has also launched the GREI, partnering with 7 generalist repositories to establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure across them. This initiative conducted several 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 ongoing 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 in leveraging FHIR® for clinical data platforms. NIH partnered with the RDA and HL7® to develop and publish a FHIR® implementation guide with 6 real-world use cases, 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 Smart and Connected Health for AI and data science, which resulted in 17 awards from 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 Scholar Program to bring together data experts, computer scientists, and engineers to tackle challenging biomedical data problems with the potential for substantial public health impact. The program has resulted in 31 DATA Scholars across 17 NIH Institutes and Centers. Of the 23 Scholars who have completed the program, 11 chose to continue in longer-term positions at the NIH.
- NIH partnered with the Civic Digital Fellows program to bring 80 Coding-it-Forward fellows to NIH over four consecutive summers.
- NIH has supported code-a-thons to encourage broad participation in data science, including forming coding partnerships with the African Society for Bioinformatics and Computational Biology. NIH funded 24 new awards supporting initiatives that advanced research training, education, and capacity building in data science.

GOAL 5

Enact Appropriate Policies to Promote Stewardship and Sustainability

NIH published the 2023 NIH Data Management and Sharing Policy, which includes new training and infrastructure support for its implementation.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

APPENDIX 2
List of Abbreviations

AI
Artificial Intelligence

AI/AN
American Indians and Alaska Natives

AnVIL
Genomic Analysis, Visualization and Informatics Lab-space

API
Application Programming Interface

ASPIRE
A Specialized Platform for Innovative Research Exploration

ASTP
Assistant Secretary for Technology Policy

AWS
Amazon Web Services

Bridge2AI
Bridge to Artificial Intelligence

caDSR
cancer Data Standards Registry and Repository

CARE
Collective benefit, Authority to control, Responsibility, and Ethics

CADRs
Controlled-Access Data Repositories

CDE
Common Data Element

CMS
Centers for Medicare & Medicaid Services

CRDC
Cancer Research Data Commons

DATA Scholars
Data and Technology Advancement National Service Scholars

dbGAP
database of Genotypes and Phenotypes

DMS Policy
Policy for Data Management and Sharing

DOE
Department of Energy

DSPC
Data Science Policy Council

EHRs
Electronic Health Records

EVS
Enterprise Vocabulary Services

FAIR
Findable, Accessible, Interoperable, and Reusable

FAQs
Frequently Asked Questions

FHIR®
Fast Healthcare Interoperability Resources

GA4GH
Global Alliance for Genomics and Health

GREI
Generalist Repository Ecosystem Initiative

HEAL
Helping End Addiction Long-Term

HHS
Health and Human Services

HL7®
Health Level Seven International

IAM
Identify and Access Management

ICs
Institutes and Centers

ICOs
Institutes, Centers, and Offices

INBRE
IDeA Networks of Biomedical Research Excellence

IRP
Intramural Research Program

IT
Information Technology

ITCR
Information Technology for Cancer Research

LLM
Large Language Model

LOINC
Logical Observation Identifiers Names and Codes

MARS
Mobile At-Home Reporting through Standards

mCODE
Minimal Common Oncology Data Elements

MIDRC
Medical Imaging and Data Resource Center

ML
Machine Learning

MOST
Maintainable, Observing, Securing, and Timing

N3C
National Clinical Cohort Collaborative

NARCH
Native American Research Centers for Health

NAIRR
National AI Research Resource

NCATS
National Center for Advancing Translational Sciences

NCI
National Cancer Institute

NEI
National Eye Institute

NHGRI
National Human Genome Research Institute

NHLBI
National Heart, Lung, and Blood Institute

NIA
National Institute on Aging

NIAID
National Institute of Allergy and Infectious Diseases

NIBIB
National Institute of Biomedical Imaging and Bioengineering

NICHD
National Institute of Child Health and Human Development

NIDCR
National Institute of Dental and Craniofacial Research

NIDDK
National Institute of Diabetes and Digestive and Kidney Diseases

NIEHS
National Institute of Environmental Health Sciences

NIGMS
National Institute of General Medical Sciences

NIH
National Institutes of Health

NIMH
National Institute of Mental Health

NINDS
National Institute of Neurological Disorders and Stroke

NIST
National Institute of Standards and Technology

NLM
National Library of Medicine

NNLM
Network of the National Library of Medicine

NSF
National Science Foundation

NSTC
National Science and Technology Council

ODSS
Office of Data Science Strategy

OER
Office of Extramural Research

OIDC
OpenID Connect

OMOP
Observational Medical Outcomes Partnership

OSP
Office of Science Policy

OSTP
Office of Science, Technology, and Policy

PROs
Patient-Reported Outcomes

RAS
Researcher Auth Service

RDA
Research Data Alliance

RDCRN
Rare Diseases Clinical Research Network

RMF
Risk Management Framework

RWD
Real-World Data

SDC
Scientific Data Council

SNOMED
Systematized Nomenclature of Medicine

STRIDES
Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability

TEFCA
Trusted Exchange Framework and Common Agreement

TRUST
Transparency, Responsibility, User focus, Sustainability, and Technology

UMLS
Unified Medical Language System

USCDI+
United States Core Data for Interoperability Plus

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

> This new strategic plan is built on more than five years of investment into the creation of a strong foundation for a modern biomedical data ecosystem. As we look boldly to the future, we will build on that foundation to advance new, sustainable, trustworthy capabilities and enable a strong workforce community that leverages data science to inform discovery and care for all.

Dr. Susan Gregurick

Associate Director for Data Science and the Director of ODSS

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012

National Institutes of Health

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00725
Progenics EX2012