DataJoint

July 15, 2025

Jayanta Bhattacharya, M.D., Ph.D.
Director, National Institutes of Health
Via email to: ai-rfi@nih.gov

Re: Request for Information Response — NIH AI Strategy [NOT-OD-25-117](NOT-OD-25-117)

DataJoint is pleased to provide comments in response to the NIH's Request for Information (NOT-OD-25-117) concerning its forthcoming institute-wide Artificial Intelligence (AI) strategy and initial one-year action plan.

DataJoint Inc. (“DataJoint”) is a U.S.-based company that assists life sciences research teams, laboratories, and institutions in attaining the highest standards of scientific integrity. Our software operates some of the most complex, data-intensive biology research in over 100 labs worldwide. NIH has funded both the dissemination of our open-source framework (Grant NIH U24NS116470) and the development of our commercial platform (Grant NIH R44NS129492). Our core technology manages all scientifically relevant information about a study, unifying its data, code, and computation and managing their relations over time. In effect, DataJoint creates a living record of a study in a common language that can be readily understood by human researchers and AI co-scientist agents. Today, DataJoint helps customers across research institutions, medical schools, hospital systems, and NIH itself to move up to 6x faster and produce results that are reproducible and verifiable.

Our comments focus specifically on the application of AI within experimental biology, focusing on the collection and aggregation of primary data, as well as the subsequent analysis, hypothesis testing, and pattern detection in complex multimodal data. Recognizing that good AI fundamentally requires good data, we are pleased to offer the NIH our perspective, drawn from many years developing and implementing robust data management and workflow orchestration systems, on the foundational needs for high-quality, reproducible AI in biomedical research.

The Imperative for Rigor in Biomedical AI

AI is poised to transform every facet of scientific discovery, dramatically accelerating research and progress towards understanding and solving complex biomedical problems. Future breakthroughs will depend on the application of AI to interrelated biological systems, with massive data streams across multiple modalities, by collaborative teams that blend human and artificial intelligences.

However, current haphazard approaches to AI, particularly those combining widely used large models, are **rife with possibilities for error, data loss, security risks, and the breakdown of attribution and ownership** for data, code, and algorithms. Unlike many other domains, the stakes in biomedical research are exceptionally high. Flawed AI, built on unreliable data, will inevitably lead to wasted resources, incorrect conclusions, and adverse impacts on human health.

Therefore, the reliable and efficient application of AI to biomedical research imposes **unique and challenging requirements for rigor**. While the necessary levels of rigor—in data management, computational processes, and scientific understanding—will vary depending on the research phase, task type, and potential impact, these requirements must be applied proportionally and holistically across the entire ecosystem of data, computation, and AI. This means that AI and data strategies are inextricably linked.

The Application of AI in Research: A Multi-Tiered Strategy

The integration of AI into biomedical research is complex, operating on multiple levels of scientific abstraction. A single, undifferentiated approach across these levels is likely to be too opaque and undecipherable, hindering trust, collaboration, and the effective governance of AI-driven scientific discovery. Attempting to apply one set of requirements for rigor, precision, transparency, security, and ownership tracking across all AI applications in research would prove unwieldy and ultimately ineffective.

Therefore, the adoption of a multi-tiered AI strategy will become necessary in biomedical research operations. Such a tiered framework is essential to achieve a clear separation of concerns across different AI functionalities and to provide greater control to labs, institutions, and individual researchers over their AI-augmented workflows and intellectual assets. Each tier within this framework would address distinct needs and necessitate tailored approaches to rigor, precision, transparency, security, and ownership tracking:

- **Level 0: Foundational Layer for Rigor & Precision:** This tier focuses on tasks that ensure absolute **data integrity, perfect reproducibility, and precision** in all foundational data operations. It involves translating natural language into precise, validated queries and orchestrating complex analysis workflows with mathematical and logical accuracy. Furthermore, Level 0 AI applications impose stringent requirements for **data security, access control, data residency,** and **immutable provenance** to safeguard the integrity of the scientific process and secure the intellectual property residing in the system. This layer is analogous to an **"AI Co-Pilot"** or a **"Coding Assistance and Data Interpreter Layer,"** ensuring reliable, AI-assisted data access.

- **Level 1: Collaborative Layer for Creativity & Exploration:** Building upon the impeccable data foundation managed by the Level 0 foundation layer, Level 1 acts as an intellectual partner in the scientific discovery process. A Level 1 AI would assist with tasks such as generating novel hypotheses, synthesizing literature, and preparing manuscripts and other scholarly communications, potentially using an iterative, closed-loop discovery process. While still requiring traceability, the requirements at Level 1 shift to **transparency of generative outputs, reproducibility of discovery workflows, and robust credit tracking for collaborative insights and contributions.** This layer is conceptualized as an **"AI Co-Scientist"** and various moonshot projects are underway to develop such systems (e.g. NovelSeek, Google Co-Scientist, FutureHouse.org, etc.). However, these existing initiatives **lack integration with foundational Level 0 data governance and deployment mechanisms** for systematic adoption within diverse research labs.

- **Level 2: Strategic Layer for Scientific Vision:** A Level 2 AI helps guide the long-term research trajectory, identify funding and partnership opportunities, and predict the strategic implications of scientific discoveries. It involves analyzing curated insights and connecting to external data sources like patent databases and grant solicitations. The requirements of Level 2 include secure access to sensitive strategic information, transparent strategic analysis, and clear governance over how strategic insights are shared and acted upon within a research team or consortium. This layer functions as an "AI Co-Visionary" and streamlines grant qualification and the formation of large-scale collaborations.

This multi-tiered approach acknowledges that a single AI model or set of governance rules could not simultaneously embody the unwavering precision required for data management, the unbounded creativity needed for discovery, and the nuanced understanding essential for strategic planning. By systematically addressing the distinct needs of each research phase, from foundational data interaction to creative hypothesis generation and high-level strategic planning, such an AI architecture can empower scientists and enhance collaboration through precisely governed data and workflows.

The SciOps Maturity Model: An Incremental Roadmap to Operational Excellence and AI Readiness

The ambitious vision for AI in biomedical science confronts a significant operational reality: the vast majority of research labs—both intramural and extramural—are **not currently AI-ready.** These labs, often operating with traditional, fragmented data practices, lack the integrated data management, computational rigor, and robust operational frameworks necessary to reliably leverage AI at scale. Integrating AI agents into such ad-hoc, unprincipled operations will likely lead to greater confusion and a profound loss of traceability and integrity, potentially even more so than operating without AI.

Consequently, producing the high-quality, structured, and fully traceable datasets that AI models demand for effective, trustworthy, and scalable performance remains a major hurdle. Addressing this is not a criticism of individual scientists, but an acknowledgment of a systemic challenge requiring a new paradigm for scientific operations. This paradigm shift requires a systematic framework that comprehensively addresses:

- **Experimental Data and Computation:** Ensuring the integrity and traceability of primary research data, from initial acquisition through analysis, and the computational pipelines used to process it. This includes robust data schemas, version control, and provenance tracking. Critically, it must also include stringent requirements for data access control and security, enabling granular permissions and protecting sensitive information. Furthermore, data residency requirements (e.g., ensuring data remains within specific geographic or sovereign boundaries) must be met, especially for human subject data.
- **Operational Processes (DataOps, MLOps, AI/LLM Ops):** Establishing robust, standardized, and automated workflows for managing data, machine learning models, and AI/Large Language Model operations throughout their lifecycle. These processes must also incorporate mechanisms for credit tracking of scientific contributions, ensuring proper attribution for data generation, analysis, and model development.
- **AI Algorithms:** Implementing rigorous validation protocols, ensuring interoperability with existing systems, and facilitating the ethical and compliant deployment of AI algorithms in research settings.

- **AI Training:** Providing for the rigorous management of data, code, and processes used in training AI models, ensuring transparency and reproducibility of the training environment and outputs.

Given the current state of research operations and the challenges outlined above, it's clear that research labs cannot simply be ordered into immediate compliance with the strict, rigorous processes necessary for effective and worthwhile AI integration. The complex, dynamic nature of scientific inquiry demands a more strategic, incremental approach to enhancing operational excellence.

The transition to AI readiness cannot be abrupt; it requires an **incremental roadmap for operational maturity** to allow teams to systematically enhance their capabilities for AI integration. We advocate for a **SciOps (Scientific Operations) path to operational excellence**.[1] SciOps is a transformative approach to scientific operations, inspired by DevOps, DataOps, and MLOps, that integrates computation, lab automation, and AI across the entire research cycle—from experimental design and data collection to analysis and dissemination, ultimately leading to closed-loop discovery. It provides a framework for systematic, standardized, and automated management of scientific data and computational workflows. This is the critical link that ensures scientific rigor translates into reliable AI applications, thereby enhancing reproducibility, accelerating discovery, and optimizing resource utilization.

Our SciOps Capability Maturity Model, developed in collaboration with leading academic and industry collaborators (as detailed in our preprint SciOps: Achieving Productivity and Reliability in Data-Intensive Research), describes five levels of operational maturity in research projects, from small-scale exploratory studies to large-scale, multi-disciplinary endeavors:

- **Level 1: Initial** - Characterized by ad hoc, customized processes and manual data management. The vast majority of neuroscience teams currently operate at this level.
- **Level 2: Managed** - Focuses on establishing lab-wide standard, repeatable processes, defined roles, and continuous quality controls to enhance consistency and predictability.
- **Level 3: Defined** - Teams embrace robust collaborations through community standards, adopting open-source ecosystems, and adhering to FAIR (Findable, Accessible, Interoperable, and Reusable) principles for data and workflows across laboratories and disciplines. Funding policies and publisher mandates are driving progress towards this level.
- **Level 4: Scalable** - Research operations adopt technology-enabled methods to streamline and scale collaborative efforts through semi-automated SciOps pipelines in collaborative research environments, enabling continuous project operation and efficient team workflows.

  Full-scale AI integration, particularly for advanced applications, truly begins at this level. While often more achievable in larger institutions, smaller teams can also implement these practices by leveraging advanced tools.
- **Level 5: Optimizing** - Represents the pinnacle of operational maturity, involving closed-loop discovery with the assistance of AI to accelerate breakthroughs, integrating AI with human cognition to refine experimental design, enhance knowledge synthesis,

[1] Johnson (2024), "SciOps: Achieving Productivity and Reliability in Data-Intensive Research." arXiv:2401.00077

and support continuous learning. No team has fully achieved this level, but some projects demonstrate aspects through machine learning-optimized experimental conditions.

This incremental roadmap highlights that the foundation for advanced AI applications lies in progressing through these levels of operational maturity. Implementing SciOps principles directly addresses the challenges of fragmented data, inconsistent workflows, and the lack of transparent computational provenance that currently hinder AI adoption in many research settings. Our SciOps maturity model can serve as one such guide, but it needs to be continuously improved through ongoing community input, real-world application in pilot programs, and the development of shared benchmarks to foster widespread adoption and refinement.

DataJoint's Role in Ensuring Rigor and Operational Excellence

DataJoint's technology directly addresses these challenges. We provide a declarative framework for defining and managing scientific data schemas, automating data pipelines, and linking computations to their originating acquisition instruments and data sources. This ensures an immutable audit trail for all research outputs, from raw data to derived insights and AI model training sets.

Beyond scientific rigor, implementing a robust AI strategy offers an unprecedented opportunity for NIH to enhance operational efficiency and maximize the return on taxpayer investment in biomedical research. Inefficient data handling, lack of reproducibility, and fragmented workflows lead to significant waste of resources—time, funding, and scientific effort.

We believe the NIH AI Strategic Plan should focus on:

- **Streamlining Research Workflows:** Standardizing data management and computation reduces redundancies, accelerates research cycles, and allows scientists to focus on discovery.
- **Reducing Waste and Accelerating Discovery:** Investing in robust data and computational infrastructure upfront prevents costly downstream errors and unproductive research paths, translating into greater mission effectiveness.
- **Enabling Strategic Resource Management:** Comprehensive systems for managing scientific assets allow research institutions to make more informed decisions about resource allocation and collaborative opportunities.

Recommendations for the NIH AI Strategic Plan

The NIH AI Strategic Plan presents a crucial opportunity to align with President Trump's Executive Order (EO) 14303, **"Restoring Gold Standard Science,"** and the accompanying **OSTP guidance.** This guidance calls upon agencies to implement the nine tenets of Gold Standard Science—including reproducibility, transparency, and collaborative approaches—to strengthen scientific inquiry, rebuild public trust, and ensure U.S. global leadership in rigorous, evidence-based science. Critically, the guidance encourages the use of AI and other advanced technologies to achieve these goals while minimizing administrative burdens.

We propose a two-pronged approach for the NIH AI Strategic Plan, focusing on immediate, actionable steps and a longer-term strategic vision, both centered on providing a systematic framework to lead labs up the ladder of operational excellence and rigor.

Immediate One-Year Action Plan

For its initial one-year action plan, we recommend NIH prioritize foundational steps that establish a robust operational framework for AI, directly supporting the tenets of Gold Standard Science:

1. **Develop a Multi-Stakeholder Framework for Biomedical AI & Data Integrity:**

   Establish and drive the development of a comprehensive, multi-stakeholder framework for biomedical AI and data integrity, drawing inspiration from NIST's successful Risk Management Frameworks (RMFs)[2]. This framework would:

   a. Be **iterative and agile**, similar to software versioning (e.g., v1.0, v2.0), allowing for rapid adaptation to evolving AI technologies and scientific needs.
   b. Focus on developing **best practices and norms for data governance, quality, AI validation, and reproducibility** specifically tailored to biomedical research, directly addressing the "Reproducible" and "Transparent" tenets of Gold Standard Science.
   c. Be collaboratively developed with input from **NIH intramural and extramural researchers, technology providers (like DataJoint), scientific professional societies, and publishers** to jointly define, develop, and learn what works in practice. This inclusive approach would provide technically informed guidelines for data management plans and AI integration across NIH-funded research, thereby facilitating collaborative and interdisciplinary science

2. **Launch Demonstrative Pilot Programs on SciOps and Reproducibility:**

   NIH should initiate and publicly document pilot projects within its intramural labs (and potentially with extramural partners) to test and showcase advanced data management and AI validation technologies. These pilots would:

   a. Focus on **SciOps implementation, data management rigor, and reproducibility** across diverse research domains, aiming to advance labs toward Level 4 (Scalable) operational maturity, where full-scale AI integration truly begins.
   b. Provide **tangible evidence of improved efficiency and cost savings** in complex research endeavors, such as replicating studies or scaling large-scale data initiatives, aligning with the goal of minimizing administrative burdens while enhancing scientific outcomes.
   c. Contribute to **developing test suites and benchmarks** for AI readiness and performance evaluation, similar to professional evaluation methodologies for advanced AI systems. These efforts can help quantify uncertainty and promote skepticism of findings, aligning with Gold Standard Science tenets.

Long-Term AI Strategic Plan

The lessons learned and frameworks developed during the one-year action plan should form the bedrock of the broader NIH AI Strategic Plan. This plan should largely focus on applying these proven principles and best practices across the entire NIH ecosystem, reinforcing Gold Standard Science tenets:

---

[2] https://csrc.nist.gov/projects/risk-management/

- **Scalable Infrastructure:** Develop and promote scalable data and computational infrastructure that supports enterprise-level SciOps, capable of handling the vast and complex datasets required for advanced AI. Emphasis should be placed on digital platforms that elevate scientific operations without extensive investments in engineering expertise, thereby supporting reproducible and transparent science.
- **Training and Education:** Invest in programs to upskill the biomedical research workforce in data management best practices, computational rigor, and AI literacy. This directly addresses the need for providing training and resources to ensure adherence to Gold Standard Science tenets.
- **Incentivizing Rigor:** Revisit and strengthen grant requirements around data management plans, potentially incorporating elements developed by the collaborative working group. This will incentivize the adoption of AI-ready practices and foster a transition to Level 3 (Defined) and higher maturity, aligning with requirements for data sharing plans and robust statistical methods.
- **Continuous Improvement:** Establish mechanisms for ongoing assessment and refinement of AI policies and technical guidelines based on emerging research and technological advancements, supporting the aspirational goal of Level 5 (Optimizing) operations and closed-loop discovery. This mirrors the OSTP's call for continual improvement and assessment of adherence to tenets.

Conclusion

The promise of Artificial Intelligence to revolutionize biomedical discovery is immense, but its full realization hinges on a **foundational commitment to scientific rigor, data integrity, and operational excellence**. As detailed herein, integrating AI effectively requires a nuanced, multi-tiered strategy that addresses distinct levels of AI application, supported by an **incremental roadmap for operational maturity** like the SciOps Capability Maturity Model. The current state of many research labs necessitates a systematic approach to bridge the gap to AI readiness, ensuring that advanced technologies amplify, rather than compromise, scientific trustworthiness.

By embracing these principles and proactively implementing the recommended one-year and long-term actions, NIH can not only meet but exceed the mandates of Gold Standard Science. DataJoint, with its proven technology and expertise in building robust scientific data ecosystems, stands ready to collaborate with NIH and the broader scientific community in this transformative endeavor. We believe that investing in these foundational elements will solidify the trustworthiness of U.S. biomedical research, maximize the impact of federal funding, and ensure continued global leadership in an AI-powered future.

Sincerely,

DataJoint, Inc.
Dimitri Yatsenko, Ph.D., Founder - dimitri@datajoint.com
Monty Kosma, J.D., Co-Founder - monty@datajoint.com