The Evolving Cyberinfrastructure at the National Institutes of Health to Support Data and AI in Biomedical Research*

Ojas A. Ramwala
Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington, 98195, USA
Email: ramwala@uw.edu

Nick Weber
Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA
Email: nick.weber@nih.gov

Sean D. Mooney†
Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA
Email: sean.mooney@nih.gov

Technological advancements have made biomedicine rich in data. With the generation of enormous volumes of biomedical and clinical data, it has become imperative to support biomedical computing investigators to utilize this wealth of biologically meaningful information. Moreover, advancements in Artificial Intelligence (AI) techniques, in conjunction with improved capabilities in implementing large-scale data processing pipelines, have led to the development of robust computational techniques and algorithms to solve complex biological problems. However, there are many challenges associated with providing researchers with secure systems for accessing biomedical data and computational resources that must be addressed. Establishing and maintaining an impactful data and AI ecosystem to support efforts in advancing biomedical research requires effective, scalable, and standardized information technology solutions, funding programs, and technical guidance that facilitate researchers in utilizing the state-of-the-art. The U.S. National Institutes of Health (NIH) has established novel initiatives to implement a cyberinfrastructure that democratizes secure access to large biomedical datasets and cloud-based computing resources, equipping biocomputing scientists to pursue pioneering research. This workshop will highlight the major issues restraining researchers' access to biomedical datasets and computing infrastructures and will cover the key components of the NIH's cyberinfrastructure aimed at advancing data science and AI research for biomedical applications.

Keywords: Artificial Intelligence; Biomedicine, Cloud Computing, Cyberinfrastructure, Data Science

* This work is supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, and the Office of Scientific Computing Services, Center for Information Technology (CIT), National Institutes of Health.
† Corresponding author.

1. Introduction

The rapidly evolving landscape of acquiring, sharing, and processing biological data has fueled the burgeoning volume of omics, clinical, imaging, and electronic health record (EHR) datasets available to biomedical researchers. While this has immense potential to enhance healthcare outcomes, intensive efforts are necessary to address challenges associated with big biomedical data. These challenges include, but are not limited to, data heterogeneity, patient data privacy and information security concerns, limited standardization and interoperability, and siloed data systems, which cause a lack of reproducibility in biomedical research. This is exacerbated by challenges in establishing secure and reliable high-performance computing infrastructures for analyzing large biomedical datasets.

A paradigm shift in the biomedical data-sharing and analysis workflow has emerged[1–3]. Traditional methods involve a resource-intensive process of curating de-identified data and sharing it with individual researchers, all under legal agreements aimed at safeguarding the privacy of biomedical data. Recently, improved data-sharing techniques involving centralized cloud-based infrastructures that host biomedical data have been developed to provide controlled access to data while addressing the limitations of traditional approaches. To operate effectively in this environment, researchers require efficient data access mechanisms that enable them to build computational workflows that facilitate the advancement of biological discoveries.

The growth of biomedical data implies a growing demand for computational resources to develop reusable interactive data science workflows and to build, validate, and deploy robust AI algorithms for diverse applications, including predicting diagnostic outcomes, proposing patient-specific treatment strategies, discovering novel therapeutics, and analyzing complex biological relationships, among many others. Cloud-based solutions have demonstrated the potential to aid researchers in rapidly acquiring scalable computational resources, often in a cost-effective fashion.

However, investigators and research institutions can leverage cloud technologies, but emergent tools in the cloud have been evolving somewhat slowly. Reasons for this often include structuring acquisition approaches, forecasting and budgeting for usage, negotiating costs, training and/or hiring staff, and implementing techniques to scale and secure prototype systems into robust infrastructures. Aligning with its Strategic Plan for Data Science and its Digital Strategy[4], the NIH has been developing a novel set of tools that provide secure biomedical data sharing mechanisms, affordable access to cloud services, and secure data analytics workspaces to enable the biomedical research community to achieve the potential of the emerging digital ecosystem. This workshop will be of interest to computational researchers who want to learn about and discuss cloud architecture, as well as explore successful approaches that leverage tools supporting the biodata community. Opportunities to leverage these platforms will be presented.

2. Workshop Goals and Organization

The workshop on **The Evolving Cyberinfrastructure at the National Institutes of Health to Support Data and AI in Biomedical Research** at the 2026 Pacific Symposium on Biocomputing is aimed at addressing the key technical factors enabling AI and Data Science research to advance biomedical and clinical medicine and healthcare. This workshop will highlight some of the major

challenges associated with storing and analyzing biomedical data, as well as developing robust algorithms. We will present impactful programs within the NIH digital ecosystem to tackle these challenges. Speakers will present a vision for reusable cyberinfrastructure "building blocks" and discuss how the biomedical research community can better understand, leverage, enhance, and extend NIH resources within an open ecosystem. The workshop will conclude with a moderated panel discussion designed to stimulate conversations and foster a deeper understanding of the technological needs of biomedical scientists, ultimately shaping AI research in biocomputing.

2.1. *Enhancing Access to Cloud Resources to Deploy Standardized and Scalable Infrastructures*

A multifaceted set of challenges acts as an obstacle to the adoption of cloud-based solutions in biomedical research workflows. The NIH STRIDES (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability)[5] program is designed to enhance biomedical discovery and improve efficiency through new digital data management strategies that contribute to NIH's efforts to develop and sustain a modern biomedical data ecosystem. Through partnerships with commercial cloud service providers (CSPs), STRIDES provides a cost-effective way for biomedical researchers at NIH and NIH-funded institutions to access advanced computational infrastructure, tools, and services provided by CSPs. The overall goal of the initiative is to accelerate biomedical advances by reducing economic and technological barriers to data and resources.

Today, STRIDES supports more than 2,500 research programs and, in aggregate, manages more than one-third of an exabyte of data (~370,000,000 gigabytes). Resources built within STRIDES include the All of Us Research Program; the Helping to End Addiction Long-term (HEAL) initiative; the Rapid Acceleration of Diagnostics (RADx) initiative for COVID research; the Accelerating Medicines Programs for Alzheimer's and Parkinson's Diseases (AMP AD / AMP PD); the Cancer Research Data Commons (CRDC) and Cancer Moonshot programs; the Gabriella Miller Kids First Pediatric Research Program; and the NIH INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE (INCLUDE), to name a few. STRIDES aims to facilitate researchers' access to and use of high-value NIH research data that are currently stored in environments; to support researchers' transition to commercial cloud technologies through a cost-efficient framework; to provide NIH researchers access to and training on new and emerging cloud-based tools and services to enhance cutting-edge research; and ultimately to contribute to the formation of an interconnected ecosystem that breaks down silos that inhibit generating, analyzing, and sharing research data. STRIDES provides the following to the biomedical research community:

- **Discounts on Cloud CSP Services**: Favorable pricing on computing, storage, and related cloud services for NIH and NIH-funded institutions.
- **Professional Services**: Access to professional service consultations and technical support from CSP partners.
- **Training**: Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- **Potential Collaborative Engagements**: Opportunities to partner with CSPs to explore novel

methods and approaches to advance NIH's biomedical research objectives.

- **Access to Cloud Marketplaces**: Ability to find, purchase, and deploy cloud-based products and services from a wide variety of third-party CSP partners to jumpstart research and discovery.

2.2. *Effective Mechanisms for Secure Biomedical Data Sharing*

A modernized biomedical data ecosystem based on FAIR (i.e., Findable, Accessible, Interoperable, and Reusable) data principles is necessary to manage the plethora of health-related datasets generated for AI and data science research. However, risks associated with the privacy of patient or study participant information while sharing healthcare and other personal data can stymie the efforts to develop and deploy robust algorithms for applications in bioinformatics and clinical medicine. Thus, an advanced paradigm of data governance mechanisms is required to protect sensitive information in environments designed to enable greater access to and sharing of data.

NIH's Researcher Auth Service (RAS)[6] program has established a novel, cloud-based mechanism to authenticate and authorize researchers to controlled datasets. Based on Global Alliance for Genomics and Health (GA4GH)[7] standards, this initiative is designed to simplify and secure login processes — for example, by federating with trusted institutional identity providers — to allow researchers to safely access sensitive data using a model similar to "passports" (authentication) and "visas" (authorization) for international travel. This model serves to dynamically control access to biomedical data in a fine-grained fashion to align with consent agreements provided by patients/participants while implementing robust cybersecurity controls to minimize risk. Moreover, RAS features interoperable resources aimed at streamlining data sharing and analyses across biomedical data repositories and analytical platforms. With comprehensive auditing and a growing community of adopters, RAS and related approaches have the potential to greatly simplify and enhance computational research in healthcare by supporting secure mechanisms for sharing and accessing sensitive biomedical data.

2.3. *Secure Data Analytics Workspaces to Improve Collaboration and Advance AI*

Hundreds of domain-specific as well as generalist data repositories support the biomedical research enterprise. Significant challenges exist in bringing together data from multiple repositories; in meeting security and compliance requirements, especially with controlled-access data; and in providing common tools and interfaces to enable accessible, reproducible analysis to researchers regardless of local or institutional resources. The cyberinfrastructure ecosystem at the NIH is expanding to include secure, centralized, low-cost, cloud-based instances of workspace environments (i.e., "Secure Workspaces") that align with NIH expectations for interoperability within an open data ecosystem.

These workspaces are being designed to allow multiple researchers to study and interpret large-scale datasets from diverse data repositories and across cloud service providers to bring robust analysis tools and AI models to tackle complex biological questions. Common technical and data standards such as RAS, the Global Alliance for Genomics & Health (GA4GH) Data Repository Service (DRS)[8], the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)[9], and the Common Workflow Language[10] are being incorporated into these computational environments

to securely provide controlled data access to users and to enhance interoperability, data harmonization, and reproducibility of analyses. Moreover, within these environments, biomedical researchers can leverage popular informatics tools to perform interactive analyses and develop novel bioinformatics pipelines without having to download or manage large and complete datasets; set up, secure, manage, and support local IT infrastructure; or be an expert software developer.

Combined, NIH's STRIDES, RAS, and Secure Workspaces programs provide the biomedical research community with a stable foundation to build upon and modular, interoperable components to build with. By utilizing these cyberinfrastructure "building blocks", researchers can worry less about the technical aspects of authentication, authorization, computation, and data management and standardization, and instead, they can focus on the collaboration and innovation that drive biomedical discovery.

3. Workshop Topics and Speakers

3.1. *Building the tools to support the biomedical research digital ecosystem*

Sean D. Mooney, PhD (National Institutes of Health)

Dr. Sean Mooney, Ph.D., serves as the Director of the NIH Center for Information Technology and the NIH Associate Director for Information Technology, Cyberinfrastructure and Cybersecurity (AD ITCC). He oversees an approximately $400 million portfolio that includes a world-renowned supercomputer, a state-of-the-art network, cloud-based services, and the latest collaboration tools.

3.2. *From Infrastructure to Insight: Enabling the Next Generation of Biomedical Science*

Nick Weber, MBA (National Institutes of Health)

Nick Weber is the Acting Director of the Office of Scientific Computing Services at the NIH Center for Information Technology (CIT). He has been supporting cloud computing, high-performance computing, and other scientific infrastructure activities to enable research within the NIH for 17 years. For the past few years, he has been focused on the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative.

3.3. *Establishing Cyberinfrastructures for Trustworthy AI in Biomedicine*

Ojas A. Ramwala, B.Tech (University of Washington)

Ojas A. Ramwala is a final-year Ph.D. candidate at the University of Washington, Seattle, in the Department of Biomedical Informatics and Medical Education, School of Medicine. His research focuses on enhancing the clinical translation of deep learning algorithms for breast cancer screening.

3.4. *The Biomedical Data Sustainability Paradox*

Philip E. Bourne, PhD, FACMI (University of Virginia)

Philip E. Bourne, PhD, FACMI, is the Stephenson Founding Dean of the School of Data Science, Professor of Data Science and Biomedical Engineering at the University of Virginia, USA. Prior to that, he was the Associate Director for Data Science (ADDS) for the US National Institutes of Health (NIH) and a Senior Investigator at the National Center for Biotechnology Information (NCBI).

3.5. *Ten Tips for Effective AI Analysis of Biomedical Data*

Jason H. Moore, PhD, FACMI, FIAHSI, FASA (Cedars-Sinai Medical Center)
Dr. Moore completed a Ph.D. in Human Genetics and an M.A. in Statistics at the University of Michigan in 1999. He joined Cedars-Sinai Medical Center in 2021 as the founding Chair of the Department of Computational Biomedicine and Director of the Center for Artificial Intelligence Research and Education.

3.6. *Advancing Dementia Research with AI and Informatics: Mining Multidimensional Biobank Data*

Li Shen, PhD, FAIMBE, FACMI, FAMIA (University of Pennsylvania)
Li Shen, Ph.D., is Professor of Informatics and Radiology at the Perelman School of Medicine, University of Pennsylvania, where he serves as Associate Director for Bioinformatics at the Institute for Biomedical Informatics and Co-Director of the Penn Center for AI and Data Science for Integrated Diagnostics.

The workshop will also feature a talk from the ClinPGx team and conclude with a moderated panel discussion focusing on the following topics:

- Lessons learned from AI-driven biomedical discoveries
- Current challenges limiting access to high-quality biomedical datasets
- Community requirements for cost-effective and robust computing infrastructures
- Development of novel solutions to support the emerging AI and data ecosystem

References

1. Mooney, S. D. Technology Platforms and Approaches for Building and Evaluating Machine Learning Methods in Healthcare. *J. Appl. Lab. Med.* **8**, 194–202 (2023).
2. Ramwala, O. A. *et al.* ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **30**, 215–228 (2025).
3. Ramwala, O. A. *et al.* Establishing a Validation Infrastructure for Imaging-Based Artificial Intelligence Algorithms Before Clinical Implementation. *J. Am. Coll. Radiol.* **21**, 1569–1574 (2024).
4. Digital NIH: Innovation, Technology, and Computation for the Future of NIH FY2023 - FY2028.
5. About STRIDES. https://cloud.nih.gov/about-strides/.
6. Researcher Auth Service Initiative | Data Science at NIH. https://datascience.nih.gov/researcher-auth-service-initiative.
7. Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* **1**, 100029 (2021).
8. Data Repository Service (DRS). https://www.ga4gh.org/product/data-repository-service-drs/.
9. Index - FHIR v5.0.0. https://www.hl7.org/fhir/.
10. Language (CWL), C. W. Home. *Common Workflow Language (CWL)* https://www.commonwl.org/.