Leveraging Foundation and Large Language Models for Health Workshop Report

Prepared By

Digital Health Research and Development
Interagency Working Group

of the

Networking and Information Technology
Research and Development Program

Co-Chairs
Wendy J. Nilsen (NSF) and Dana L. Wolff-Hughes (NIH)

January 2025

Subcommittee on Networking and Information Technology Research and Development (NITRD) Co-Chairs

Joydip Kundu
U.S. National Science Foundation (NSF)

Kirk Dohne
NITRD National Coordination Office (NCO)

Digital Health Research and Development (DHRD) Interagency Working Group (IWG)
Co-Chairs

Wendy Nilsen, Deputy Division Director, Information and Intelligent Systems, Computer & Information Science & Engineering Directorate
U.S. National Science Foundation (NSF)

Dana Wolff-Hughes, Program Officer, Risk Factor Assessment Branch, Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences
National Cancer Institute, National Institutes of Health (NIH)

Technical Coordinator

Olachi Onyewu, NITRD National Coordination Office (NCO)

Copyright Information

Note: Reference in this report to any product, service, enterprise, or individual, including any written works (i.e., books, articles, papers), is not an endorsement and does not imply official government sanction or endorsement of those entities or their views. Published in the United States of America, 2025.

Table of Contents

Executive Summary

The Networking and Information Technology Research and Development (NITRD) Program Digital Health Research and Development (DHRD) Interagency Working Group (IWG) held a federal-only workshop, “Leveraging Foundation and Large Language Models for Health”, on July 22, 2024. Over 35 participants from six federal agencies took part in the workshop to explore artificial intelligence (AI), especially foundation and large language models (LLMs), in biomedical, public health, and healthcare research.

The goal of the workshop was to identify the opportunities and challenges for these AI models in health-related research and development[1] (R&D) applications. AI has the power to be a transformative tool that can revolutionize health-related R&D if used properly; however, there are issues that need to be addressed to safely implement current foundation and LLMs in health-related R&D. This report identifies target areas, challenges, and opportunities that should be addressed in future digital health R&D and summarizes the key takeaways from the workshop.

Introduction

The NITRD Program’s DHRD IWG brought together federal staff for a one-day workshop on “Leveraging Foundation and Large Language Models for Health”. The workshop brought together researchers in foundation and large language models, collectively called generative AI,[2] to explore the opportunities and challenges facing AI. This meeting focused on foundation models, which are AI general purpose systems that have been trained on large data sets, and large language models, which are also trained on large, diverse datasets, but focus on language-based tasks, in health-related R&D.

The goal of the workshop was to identify the opportunities and challenges for generative AI in health-related research and development. AI has the power to be a transformative tool that can revolutionize research and clinical applications if used properly, however, there are several challenges that need to be understood to safely implement generative AI models in health-related R&D.

The workshop included six speakers from government and academia. The sessions included:

- Introduction to Large Language and Foundational Models, Michael Littman, Ph.D., U.S. National Science Foundation
- Large Language Models for Accelerating Biomedical Information and Knowledge Management, Hongfang Liu, Ph.D., University of Texas Health, Houston
- Transforming Health through AI, Predictive, and Behavioral Technologies, Ramesh Jain, Ph.D., University of California, Irvine
- Generative AI for Disease Modeling: A Roadmap, Fei Wang, Ph.D., Cornell University
- Hypothesis Generation with Large Language Models, Aidong Zhang, Ph.D., University of Virginia
- Ethical, Legal, and Social Implications of AI, Camille Nebeker, Ph.D., University of California, San Diego

The talks provided background on the mechanisms of generative AI, select use cases, and an overview of some of the ethical, legal, and social issues raised by these models to identify some

[1] In this document, health-related research and development (R&D) includes activities focused on improving health and well-being of individuals and communities. Application areas include clinical research, public health, healthcare, health services research, and health information technology.

[2] For convenience, large language and foundation models are collectively referred to as generative AI.

of the strengths and limits of the approach. Speakers noted that the term artificial intelligence (AI) includes a range of analytic methods (such as machine learning, neural networks, and computer vision) that address a range of mathematical problems. In recent years, AI has gained attention because of the capabilities exhibited by generative AI models. With increasing availability of data and computational capacity, generative AI builds on previous AI methods. While previous AI research worked to predict the next variable in a sequence; that is, pixel in an image, or word in a sentence, generative AI adds context by bringing in millions or billions of parameters that define how much processing can happen before the model reaches an answer. These generative AI models can be trained in a variety of data formats, including text, images, and raw data, and are capable of continuously learning by modifying the prompts and adding new data to the training dataset. That said, training a generative model requires access to the data and expensive computational resources. Thus, retraining a model cannot be done frequently.

Generative AI models developed in the last few years have gained attention because of the ability of single models to create complex responses across a range of disciplines and problems. Developers start with a so-called foundation model trained on massive amounts of general data (pre-train), and then use that as a starting place for solving specific problems (fine tune). These new models hold the potential to revolutionize various facets of health-related R&D, from exploring diagnostics to personalized medicine to operational efficiency, as well as enhancing clinical decision making, accelerating drug discovery, and creating personalized care plans that are reflective of individual variability.

However, the integration of generative AI into clinical and biomedical applications requires caution. First, most commercially available models lack transparency; that is, the data used in the training models are unclear, the code is not public, and how models are calibrated is opaque. These factors make it challenging for health researchers, clinical providers, and patients to trust the output. Further, generative AI has been known to create responses, called hallucinations, that are generated by AI systems and are incorrect and/or out of context. Because the data from which a model is trained is not usually public, it is also hard to identify biases that may have been introduced. Finally, research on the output quality of most generative AI models has either not been done or has not been published. Thus, in the high-stakes world of health-related R&D, the validity of any response needs to be checked by humans before use.

Furthermore, there are ethical, legal, and social implications to be considered for deploying generative AI in health-related R&D that cannot be overlooked. Concerns about data ownership, patient privacy, data security, and the potential for misuse of AI models and products raise critical questions about the responsible use of these technologies. There is also a need to ensure that AI-driven health solutions, derived from the input data used to train the foundation models, are equitable and do not perpetuate and/or exacerbate existing disparities in health-related R&D.

Despite all these issues, generative AI has created considerable enthusiasm in the health-related R&D community. Generative AI has revealed new opportunities from hypothesis generation to providing new insights into rare diseases[3] to creating user-friendly text from highly technical data. In addition, there is interest within the scientific community in building new generative AI models, using high-quality data, whose code and tuning are publicly available. These opportunities and possible future directions highlight the potential of generative AI in health-related R&D.

The speakers' comments set the stage for discussion of these areas by members of the workshop. The focus of the workshop was not specific to any one outcome, sector, or population, but served

[3] Rare diseases are most commonly defined in reference to the Orphan Drug Act (https://www.ecfr.gov/current/title-21/chapter-I/subchapter-D/part-316), which identifies a rare disease as a disease or condition that affects less than 200,000 people in the United States.

as a platform for multidisciplinary collaboration on the rapidly evolving topic by focusing on general target areas for advancement and highlighting areas of greatest risk and concern.

For the first breakout session, each group was asked to address four questions:

- How to evaluate systems that continue to learn (i.e., continuously adding new training data)?
- Is there an optimal strategy for federal funding of foundation and large pre-trained model development?
- Who needs to be at the table for the research to succeed? What kind of teams do we need?
- What are the next steps for exploring new ideas and establishing partnerships?

After the second session, the group was asked to address four questions:

- What kind of teams do we need to succeed with next-generation foundation models?
- How can we safeguard privacy and ethics in generative AI?
- What is the right approach for handling proprietary data in AI?
- What are the next steps for exploring new ideas and establishing partnerships?

Opportunities and Challenges of Using Generative AI in Health-related Research and Development

We are at a pivotal moment where the convergence of multimodal data, systems biology, quantitative physiology, and AI offer the potential to transform how we approach human health and medicine. This synergy may allow us to move from the traditional models of R&D that are expensive and time-intensive to new approaches and rapid hypothesis generation and simulation. It could also change our understanding of human biology, treatment and prevention from a reactive, episodic, pathogenic model of health to a more balanced, proactive, anticipatory, salutogenic model capable of more personalized approaches. By leveraging connected and continuous data from a wide range of sources, considerable innovation can occur in this area. Trusted, accessible, and unbiased generative AI models that are reliable, explainable, robust, and secure will be central to achieving this goal.

Generative AI models, particularly in the case of health-related R&D, are uniquely positioned to serve as powerful supporting tools. They are not intended to and should not replace human-led decision making, but they do have immense potential to augment and improve many aspects of the scientific process. The strengths of large language and foundation models lie in the adaptability of their massive data and computing capacity, which can be used on an ever-growing range of tasks to enhance scientific discovery. From clinical trial design to drug discovery, generative AI can automate tasks such as the formatting of clinical documentation, extracting and compiling information, and conducting early-stage trials by predicting biological or clinical outcomes without trials or wet labs. In this capacity, AI can serve as an essential partner for scientists and developers, improving efficiency and productivity in many applications.

An example of this partnership is the “expert-in-the-loop” model where AI performs many tasks that are guided and verified by human experts who provide and refine the questions and evaluate responses. This collaborative dynamic can streamline the R&D process by assisting with feasibility assessments, automating eligibility screening, content analysis, and outcome dissemination. Similarly, this partnership can be used to accelerate drug discovery, particularly when used in a multi-omics application. By integrating genomic, proteomic, and metabolomic data, generative AI could dramatically accelerate drug discovery processes, enhancing our understanding of disease development and progression in ways previously thought unattainable. The ability to model complex biological processes could reveal new therapeutic targets or optimize existing treatment protocols with unprecedented speed and efficiency even in cases with limited data, such as rare diseases. Generative AI models may also be able to simulate an individual's unique biology and physiology through detailed medical history records and could enable individualized treatment strategies, improving outcomes, patient experience, and reducing trial-and-error approaches in medicine.

Generative AI also holds promise for disease modeling, enabling better mapping of disease progression and the identification of disease phenotypes that could lead to optimized treatments. Using techniques such as retrieval augmented generation, in-context learning, and chain-of-thought reasoning can greatly improve the fact-based responses increasing the reliability of generative AI.

However, despite these promising opportunities for advancement, there remains uncertainty around practical issues, particularly regarding the requirements for foundation models in health-related R&D. Questions remain regarding the minimum requirements for a foundation model, such as the training data size, volume, and diversity needed. Additionally, there is no current common framework or guidance on how such models would be developed and maintained by the federal government. This is challenging for important areas, such as rare diseases or small, underrepresented populations, where data pools are inherently limited, but which may be augmented by a well-developed generative AI model (i.e., where training includes a broader sample of the whole population). Other systems for evaluation of these models will need to be developed to assess their accuracy and how responses change over time. Attendees agreed that the health-related R&D community must strike a careful balance that ensures generative AI use mitigates, rather than exacerbates, existing disparities in health-related R&D. Furthermore, issues such as model accuracy, explainability, and robustness remain significant hurdles, as do many ethical, legal, and social implications.

Ethically, developers and users of generative AI must address data ownership, privacy, confidentiality, and biases. In health-related R&D, where data are often highly sensitive, use of commercial generative AI systems may be inappropriate because de-identification measures may not be enough, given the potential for re-identification. Biases in data can exacerbate existing health disparities, making it critical to ensure that the datasets we use are diverse and representative. Additionally, the boundaries of informed consent are vague when it comes to data use, so clearly defining this aspect for those providing data and emphasizing data literacy will be crucial to building and maintaining public trust.

Legally, AI researchers and developers must navigate the complex regulatory environment for data protection, including compliance with the United States' Health Insurance Portability and Accountability Act (HIPAA)[4] and the European Union's General Data Protection Regulation (GDPR).[5] Intellectual property rights are another concern, as it is unclear who owns the data, models, and algorithms generated by the many proprietary AI systems and pipelines. Establishing clear frameworks for liability and accountability will be crucial. Independent data and systems governance policies will also be necessary to ensure the quality, security, and integrity of the data used to train AI, as well as the outputs generated by AI.

[4] Department of Health and Human Services. *Health Information Privacy*. https://www.hhs.gov/hipaa/index.html

[5] Intersoft Consulting. *General Data Protection Regulation*. https://gdpr-info.eu/

Socially, public trust will be essential for the successful integration of generative AI into health-related R&D. Individuals need to trust that their data will be used responsibly and that AI tools will not perpetuate health disparities. This will require concerted efforts to evaluate tools built using AI to ensure that the outcomes are equitable, accessible, and culturally appropriate for diverse populations. Additionally, the evaluation of these tools should be disseminated to support public trust in the output. As digital health tools become more ubiquitous, addressing the digital divide will be increasingly important, ensuring that everyone can benefit from AI-related advancements in health, regardless of their level of access or literacy.

Federal Engagement with AI

Much of the federal-only breakout session discussion centered on strategies to accomplish the successful use and integration of generative AI in health and to identify the resources and personnel to achieve success. This will be a paradigm shift for many and there are a lot of unknowns so it will take cooperation and collaboration to forge ahead. This pursuit should remain a focus for the DHRD IWG.

A considerable challenge lies in establishing evaluation criteria for continuously learning AI systems, including the metrics for success. Previous work has started tackling this issue in specific use cases, such as with chatbots in healthcare,[6] but these use cases do not create an evaluation framework for the many possible uses of these systems. Building on the Food and Drug Administration's framework for the Predetermined Change Control Plan (PCCP[7]) could provide a starting point for adapting an evaluation framework suited to generative AI systems that are continuously updated. This structured approach to evaluation will require data ecosystem building, data standardization and harmonization, and data linkages with federated architecture being preferred.

Breakout sessions also underscored the need for large, multidisciplinary teams of stakeholders. Despite the development of some federal foundation models, it was not clear whether the federal government is positioned to support the development of high-quality, open-source foundation models for health. To address this, the community will need to work in public-private partnerships to develop templates that leverage resources, streamline collaboration, and foster innovation. There should also be careful consideration for the role and continued investment at the federal level by gauging the return on investment when many private entities are actively pursuing major development in this sector.

Another major consideration is the need to reduce redundancies across agencies, an activity in which NITRD may play a pivotal role. Aligning efforts across different agencies and sectors will not only accelerate progress but also optimize the use of federal resources. The group emphasized the importance of determining a sustainable funding model, including how funding allocation and the administration of funded projects would occur across agencies.

[6] Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by Generative AI. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01074-z

[7] Food and Drug Administration. (2024, August). *Predetermined Change Control Plans for Medical Devices*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devices

Lastly, the federal government already has some level of investment in efforts to utilize AI through initiatives like NIH's Bridge2AI, the Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI), and NSF's National AI Research Institutes program, although only the last effort includes work in developing generative AI models. The NIH-funded Bridge2AI seeks to foster collaborations across technological, biomedical, social science, and humanities disciplines with the goal of creating ethically sourced datasets optimized for AI analysis. A key challenge in health-related R&D is that while data are abundant, many datasets are incomplete or inconsistent, making them unsuitable for robust modeling. The Bridge2AI program is working to address this by producing well-defined, trustworthy, FAIR (Findable, Accessible, Interoperable, and Reuseable) datasets, while simultaneously developing the tools and standards necessary to standardize and harmonize these data. To scale these efforts, establishing a federated learning network could help ensure secure and efficient training of AI models across agencies without compromising data privacy.

Other federal investments, like the National Artificial Intelligence Research Resource (NAIRR),[8] provide access to generative models, computational capacity, data and AI tools and training. The NAIRR brings together a public-private partnership to democratize AI in a wide range of domains and problems.

Conclusion

In conclusion, this workshop highlighted the undeniable potential for generative AI in health-related R&D, but with the knowledge that the full integration into existing systems requires careful consideration. Despite the promise generative AI holds for transforming health as the research community now knows it, we must acknowledge the considerable technical and ethical issues involved are complicated and cannot be regulated or evaluated using traditional frameworks and processes. The guidance in place for other technologies may not be sufficient for generative AI and we may need to imagine and develop new ways to assess the impact, especially when data are constantly evolving and there are many unknown factors.

As new ways to apply AI are explored, we must ensure safeguards are in place to reduce the likelihood of misuse and to find balance to minimize potential individual and societal harm when weighed against the potential benefits. This will require some degree of trust in the process and people responsible for oversight as well as a robust implementation system that both protects privacy and holds researchers and developers to a high standard. This will be accomplished by better transparency and access to existing models and/or development of new trusted models, as well as quantifying the accuracy, explainability, accountability, and ethics of the work.

Importantly, the revolutionary and transformative capability of AI should be leveraged rather than feared. While AI represents a paradigm shift, we can collectively discover ways to adapt, regulate, and ensure its proper use for the betterment of health at a societal level. This workshop and discussion surrounding generative AI for health applications is only the beginning. With a commitment to innovation and a strong ethical and legal framework, advances can be made in this area. With the right measures, AI can help achieve more equitable, effective, and accessible health for all.

[8] NSF. *The National Artificial Intelligence Research Resource (NAIRR) Pilot*. www.nairrpilot.org

List of Abbreviations and Acronyms

Acronym	Definition
AHRQ	Agency for Healthcare Research and Quality
AI	Artificial intelligence
DHRD	Digital Health Research and Development
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
IWG	Interagency Working Group
LLM	Large language model
mHealth	mobile health
NCO	National Coordination Office
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development
NSF	U.S. National Science Foundation
R&D	Research and development

About NITRD and the Digital Health R&D Interagency Working Group

The NITRD Program is the Nation's primary source of federally funded work on pioneering information technologies in computing, networking, and software. The NITRD Subcommittee of the National Science and Technology Council guides the multiagency NITRD Program in its work to provide the R&D foundations for ensuring continued U.S. technological leadership and that meets the Nation's advanced IT needs. The National Coordination Office (NCO) supports the NITRD Subcommittee and the Interagency Working Groups (IWGs) and teams that report to it. The NITRD Subcommittee's Co-Chairs are Kirk Dohne, Acting NCO Director, and Joydip Kundu, Deputy Assistant Director of the NSF Directorate for Computer and Information Science and Engineering. More information about NITRD is available online at https://www.nitrd.gov/.

NITRD's Digital Health R&D IWG focuses on more efficient and effective biomedical, healthcare, and public health R&D through digital health technologies that support effective health monitoring; individualized screening, diagnosis, and treatment; improved disease prevention and disaster and emergency response; and broad and inclusive access to health and healthcare information and resources. The IWG's activities also contribute to building and sustaining a vibrant community of professional digital health researchers and practitioners. More information is available online at https://www.nitrd.gov/coordination-areas/dhrd/.

Acknowledgements

The NITRD Digital Health R&D IWG gratefully acknowledges the workshop planning committee members Caitlin Bailey (NIH), Tom Martin (NSF), Virginia Moore (NCO), Wendy Nilsen (NSF), Christopher Yang (NSF), David Zahavi (NIH), Samuel LaMunion (NIH), and Dana Wolff-Hughes (NIH), who helped to plan the workshop and to write and review this report; the workshop panelists for their valuable insights; and the attendees for their contributions to the workshop discussions.