

NIH Collaborative AI Assurance Research Laboratory: Justification and Impact Report

**Approved for Public Release;
Distribution Unlimited.
Public Release Case Number 26-0299**

Prepared for:

National Institutes of Health (NIH)

MITRE | Center for
Transforming Health

Operator of the CMS Alliance to Modernize Healthcare (The Health FFRDC)

A Federally Funded Research and Development Center

March 2026

Notice

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

This technical data report was produced for the U.S. Government under Contract Number 75FCMC23D0004, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data-General.

No other use other than that granted to the U.S. Government, or to those acting on behalf of the U.S. Government under that Clause is authorized without the express written permission of The MITRE Corporation.

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.

© 2026 The MITRE Corporation.

Authors

Ariel Kapusta, Ph.D.	The MITRE Corporation
Matthew McQueen, Sc.D.	The MITRE Corporation
Patrick Minot, M.S.	The MITRE Corporation
Laura Freeman, Ph.D.	Virginia Tech National Security Institute
John Halamka, M.D., M.S.	Mayo Clinic College of Medicine and Science
Heidi Hanson, Ph.D., M.S.	Oak Ridge National Laboratory
Stephen Sodeke, Ph.D.	Tuskegee University
Michael Young, M.D., M.Phil.	Massachusetts General Hospital

Acknowledgements

We acknowledge Christine Cutillo, Samson Gebreab, Chun-Ju Hsiao, and Larry Holt from NIH/ODSS for expertise, guidance and direction on this work.

Executive Summary

The National Institutes of Health (NIH) is increasingly adopting Artificial Intelligence (AI) to advance biomedical and health research. To **assess the current state of AI assurance in research and identify solutions to key challenges**, NIH partnered with MITRE, operator of the Health Federally Funded Research and Development Center (Health FFRDC). Guided by an advisory board of AI and health experts in addition to NIH AI Program Officers and cybersecurity subject matter experts, MITRE conducted exploratory pilots using real-world AI use cases. The pilots evaluated barriers and gathered insights from the biomedical and health research community to provide recommendations for practical solutions to accelerate responsible AI development and adoption. The pilots uncovered gaps in the AI assurance resources available to the NIH research community (e.g., curated playbooks, formalized benchmarks, testing and evaluation methods, and standardized tools) that are slowing advancements in the field.

Challenges for the biomedical and health research community include:

- Absence of clear, standardized guidance and best practices for AI development and deployment.
- Inconsistent processes for aligning AI workflows with established ethical, technical, and assurance standards.
- Reliance on inefficient, custom-built tools and processes due to a lack of standardized and accessible AI resources.
- Unscalable and resource-intensive efforts required to develop and maintain AI systems throughout the research lifecycle.

Recommendation

To help the biomedical and health research community overcome these challenges, NIH in partnership with MITRE recommends establishing the **Collaborative Research for the Advancement of Artificial Intelligence Laboratory (CORAL)**. CORAL would serve as a trusted resource to accelerate innovation and overcome barriers to AI adoption in biomedical and health research. Through collaborative research engagements exploring real-world AI use cases, CORAL would provide the biomedical and health research community with tailored lessons learned, playbooks, and other assurance resources to conduct responsible AI-enabled research.

CORAL would operate using an iterative framework that emphasizes collaboration, continuous improvement, and validation through real-world evidence. This framework enables CORAL to discover emerging AI assurance challenges, create solutions, and

tailor those solutions to practical biomedical and health research settings. By fostering interdisciplinary partnerships and streamlining AI workflows in research, CORAL would accelerate the adoption of AI across NIH research initiatives, ultimately advancing scientific discovery, enabling precision medicine, and improving public health.

The Collaborative Research for the Advancement of Artificial Intelligence Laboratory (CORAL)

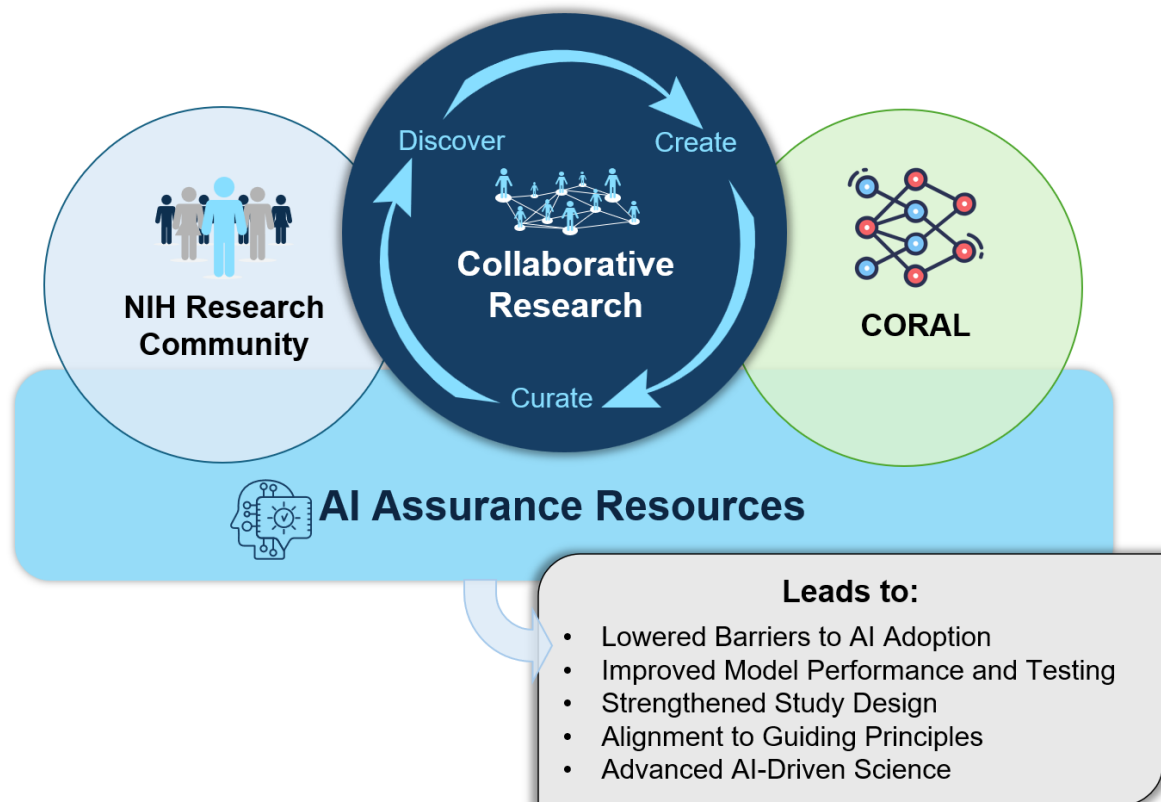


Table of Contents

Executive Summary	ii
1 Introduction	1
1.1 The Need for Federal Leadership in AI Assurance.....	1
1.2 Establishing AI Assurance Resources.....	3
1.3 Building The Initial AI Assurance Framework.....	4
1.4 Key Terms & Definitions	5
1.5 Related AI Initiatives & Frameworks.....	6
2 Approach Overview	15
2.1 Exploratory Pilots	17
2.1.1 Exploratory Pilots Background Details.....	17
2.1.2 Exploratory Pilot Documentation	19
3 Findings	21
3.1 Challenges	21
3.2 Preliminary Biomedical and Health AI Assurance Practices.....	22
4 Recommendations & Next Steps	25
4.1 Lab Design Overview	26
4.2 Phased Lab Implementation.....	27
4.2.1 Option Year (planned: June 2025–June 2026)	28
4.2.2 Future Years (planned: June 2026–June 2027).....	30
5 Conclusion	33
6 References	34
Appendix	37
A-1 The National Center for Biotechnology Information (NCBI)	37
A-2 Current State of AI R&D In Health	41
A-3 Additional Tables and Figures	43

1 Introduction

As the largest public funder of biomedical and health research in the world, the National Institutes of Health (NIH) facilitates transformative and innovative approaches to advance our understanding of human health and disease. Research funded through NIH generates vast amounts of biomedical and health data and information across many substantive areas from cells to populations. Recent advances in Artificial Intelligence (AI) have been rapidly integrated into NIH projects and health data ecosystems (Le et al., 2026). For example, in fiscal year (FY) 2018, fewer than 100 projects were funded by NIH that included “artificial intelligence” in the project title, project abstract, and/or project terms. By FY 2025, that number grew to over 1,700 funded projects.¹ Additionally, in FY 2025, nearly every NIH Institute/Center funded a project utilizing AI at some level. This highlights the broad adoption of this emerging technology across different health domains.

Rapid adoption of AI impacts the entire NIH community. As AI-enabled systems are increasingly developed by NIH-funded researchers, there will be greater demand and reliance on peer review of multidisciplinary experts to ensure that published research leveraging AI is high quality, transparent, and reproducible. The NIH research community is undergoing a paradigm shift as AI applications become increasingly integrated into the research process. This underscores the urgent need for AI assurance resources, such as playbooks and standard methodologies, to accelerate discovery and innovation, break down barriers to adoption, and ensure consistent quality of biomedical and health research using AI (Foote et al., 2025).

1.1 The Need for Federal Leadership in AI Assurance

Given the rapid integration of AI into biomedical and health research and emerging federal priorities, there is a critical need for NIH to accelerate the development of AI assurance resources that ensure responsible AI development and use. As outlined in the memorandum titled, “Accelerating Federal Use of AI through Innovation, Governance, and Public Trust” (M-25-21; April 3, 2025),² there are three key priorities

¹ NIH RePORTER. Retrieved January 2026, from <https://reporter.nih.gov/>

² Executive Office of the President, Office of Science and Technology Policy, Accelerating Federal Use of AI through Innovation, Governance, and Public Trust. (Washington DC: White House, April 3, 2025). <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

for federal agencies for AI:

1. **AI Innovation:** Remove barriers to innovation and accelerate responsible AI adoption.
2. **Governance:** Establish clear leadership structures for AI deployment.
3. **Public Trust:** Ensure that AI is used responsibly and with appropriate safeguards that include human oversight and training.

It is of critical importance that NIH is well positioned to meet the increasing demand for AI leadership from the biomedical and health research community while aligning with the emerging federal priorities for AI innovation, governance, and public trust.

Additionally, as outlined in the White House's "America's AI Action Plan" (July 23, 2025), "While much work has been done to advance the field of AI Assurance, promoting resilient and secure AI development and deployment should be a core activity of the U.S. Government."³

An example of NIH leadership in the context of emerging revolutionary technology started in 1988, when the National Center for Biotechnology Information (NCBI) within the National Library of Medicine (NLM) was established by NIH in response to the "oncoming genetic revolution in science." (Smith, 2013). NCBI's role following the completion of the Human Genome Project in April 2003 represents one of the most successful federal initiatives in modern biomedical science. By supporting the development of foundational resources across the international community including standardized databases, common file formats, consistent analysis tools, and shared reference resources, NIH helped to accelerate emerging high-throughput genomics. The field changed from a fragmented system dominated by siloed and inefficient efforts to a coherent scientific ecosystem that accelerated discovery and democratized innovation. This example offers a powerful template for addressing the current challenges and opportunities around the development, testing, and implementation of AI in biomedical and health research.

Private industry plays a key role in commercializing biomedical and health AI applications. However, competitive market pressures and emphasis on differentiation make it less suited to creating open-source AI assurance resources essential for the NIH research community. Even in the context of open-source AI standards that may be

³ Executive Office of the President, Winning the Race: America's AI Action Plan (Washington DC: White House, July 2025). <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

developed through entities such as the International Organization for Standardization (ISO), there are still challenges in the direct applicability of these standards to the biomedical and health research community. Further, companies are inherently incentivized to establish their platforms as dominant, leading to competing standards that further fragment the field. This pattern has played out repeatedly across technology sectors as industry faces strong incentives to develop proprietary data formats, closed algorithms, and unique validation approaches that differentiate their products and lock customers and users into siloed ecosystems. While there have been open-source solutions provided through established technology entities that may be relevant to general solutions in AI system implementation (e.g., Google's TensorFlow,⁴ Meta's PyTorch,⁵ and Apache's MXNet⁶), the responsible use of these frameworks in biomedical and health research requires further advancements not within their scope.

1.2 Establishing AI Assurance Resources

By creating a common foundation of resources, an NIH collaborative AI assurance research lab would enable researchers to focus on developing novel applications rather than reinventing common infrastructure.

An NIH collaborative AI assurance research lab would serve as a trusted authority to develop AI assurance resources for the biomedical and health research community. These resources include playbooks, toolkits, benchmarks, and workflow standards, and to enhance, accelerate, and safeguard AI development, adoption, and reuse. Like NCBI's initiative in biotechnology, NIH could address fragmentation by fostering consensus on biomedical AI assurance standards, playbooks, and providing shared datasets and tools across the biomedical and health research community. The NIH is uniquely positioned to ensure reference datasets, frameworks, benchmarks and other resources are accessible and curated for the biomedical and health research community. Just as the reference human genome provided a foundation for genomic research, NIH-established reference datasets could ensure that biomedical and health AI solutions are reliable, ethical, reproducible and broadly support and interact with NIH AI-initiatives

⁴ TensorFlow developed by Google. <https://www.tensorflow.org>

⁵ PyTorch developed by Facebook AI Research team (now known as Meta AI). <https://pytorch.org>

⁶ Apache MXNet developed by Amazon Web Services (AWS). <https://mxnet.apache.org/versions/1.9.1/>

such as Bridge to Artificial Intelligence (Bridge2AI)⁷, Precision Medicine with AI: Integrating Imaging with Multimodal Data (PRIMED-AI)⁸ and AIM-AHEAD.⁹ By creating a common set of assurance resources, an NIH collaborative AI assurance research lab would enable researchers to focus on developing novel applications rather than reinventing common infrastructure or searching for tools, guidance, and other resources, and thereby accelerate progress across competing tools and methodologies.

The NIH has demonstrated its capacity to maintain and continuously improve biomedical and health resources over decades. This long-term commitment is essential for the development of AI systems that can be trusted in biomedical and health research as well as clinical settings, where transparency, stability, and consistency are paramount.

1.3 Building The Initial AI Assurance Framework

MITRE, in collaboration with NIH, conducted exploratory pilots using real-world use cases of AI applications to identify the distinct AI assurance challenges and opportunities in biomedical and health research. These pilots revealed that while investigators recognize AI as a powerful tool for enabling deeper insights and improving research efficiency, its adoption remains daunting and bottlenecked due to resource-intensive development, rapidly evolving technology, and the need for alignment with established technical, ethical, and assurance principles. Investigators often resort to developing in-house solutions and spending significant time searching for costly, dense, or irrelevant AI assurance resources that are not tailored to biomedical and health research.

To address these challenges, MITRE captured critical insights to inform the design of a collaborative AI assurance research lab. This lab would provide AI assurance resources to design, develop, evaluate, validate, and monitor AI-enabled systems aligned with NIH's mission. Key strategic components of the framework include:

- **Research Lifecycle Integration:** Discover challenges for integrating responsible AI seamlessly into existing research workflows, including Institutional Review

⁷ NIH Common Fund, Bridge to Artificial Intelligence (Bridge2AI). <https://bridge2ai.org>

⁸ NIH Division of Program Coordination Planning and Strategic Initiatives (DPCPS), Precision Medicine with AI: Integrating Imaging with Multimodal Data (PRIMED-AI). <https://commonfund.nih.gov/primed-ai>

⁹ NIH ODSS, AIM-AHEAD. <https://datascience.nih.gov/artificial-intelligence/aim-ahead>

Board (IRB) processes, grant proposals, experimental design, and dissemination of findings.

- **Collaborative Development:** Create practical solutions through real-world AI use cases in partnership with the NIH research community.
- **Curation & Standardization:** Curate findings to establish a shared foundation of AI assurance resources to standardize workflows across wide-ranging biomedical and health research contexts and computational environments.

To help the biomedical and health research community overcome these challenges, MITRE recommends NIH establish a collaborative AI assurance research lab. By developing AI assurance resources, this lab would empower the biomedical and health research community to harness AI effectively and efficiently, advancing NIH's mission to enhance health, extend life, and reduce illness and disability.¹⁰

1.4 Key Terms & Definitions

The following are definitions of key terms that will be used throughout this report.

- **AI Assurance:** “A process for discovering, assessing, and managing risk throughout the life cycle of an AI-enabled system so that it operates effectively to the benefit of its stakeholders.” (Robins et al., 2024)
- **AI-enabled System:** “An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.” (Tabassi, 2023)
- **AI Use Case:** “The description of a particular application of an AI system for an intended goal or solution.” (International Organization for Standardization/International Electrotechnical Commission, 2024)
- **Exploratory Pilot:** An AI use-case driven experiment aimed at obtaining discoveries on specific aspects of the AI assurance process.
- **Intended Use:** “Use for which a product, process or service is intended according to the specifications, instructions and information provided by the manufacturer.” (International Organization for Standardization, 2019)

¹⁰ NIH, Mission and Goals, accessed January 2026. <https://www.nih.gov/about-nih/what-we-do/mission-goals>

-
- **Process:** “Set of interrelated or interacting activities that use inputs to deliver an intended result.” (International Organization for Standardization, 2019)
 - **System:** “Integrated composite consisting of one or more of the processes, hardware, software, facilities, and people, that provides a capability to satisfy a stated need or objective.” (Jordan, 2006)

1.5 Related AI Initiatives & Frameworks

The landscape of AI in biomedical and health research is rapidly evolving, with numerous initiatives and frameworks being developed across industry, academia, and government. These efforts aim to establish standardized best practices, processes, and tools that aim to accelerate AI innovation while ensuring ethical and effective application in biomedical and health research. However, these efforts face challenges with widespread adoption due to issues such as competing interests, proprietary constraints, accessibility, and applicability (The Lancet, 2025; Cao et al., 2026). Additionally, surveys of patients and clinicians make the case for risk-based frameworks depending on the clinical stakes (i.e., there is a high degree of comfort leveraging AI transcription services, but a low degree of comfort with an AI chatbot interacting directly with patients regarding direct clinical care). Below is an abridged overview of existing AI initiatives and frameworks. Further details on related works can be found in the Appendix in Section A-3, Table 4. Inventory of AI Initiatives and Frameworks. While there has been significant progress in developing formal AI guidelines and frameworks, many of them focus on high-level principles (e.g., ethics, fairness, transparency) rather than providing detailed, domain-specific, standardized instructions for practical implementation.

Industry Standards Organizations

Organizations such as the International Organization for Standardization (ISO)¹¹ and the International Electrotechnical Commission (IEC)¹² are developing global standards to ensure the reliability and interoperability of AI technologies. The Coalition for Health AI (CHAI)¹³ is focusing on guidelines specific to healthcare applications. These industry organizations and others like them are leading efforts to create globally recognized standards for AI development, focusing on reliability, interoperability, and ethical

¹¹ International Organization for Standardization (ISO). <https://www.iso.org/home.html>

¹² International Electrotechnical Commission (IEC). <https://www.iec.ch/homepage>

¹³ Coalition for Health AI (CHAI). <https://chai.org/>

considerations. These initiatives aim to align AI technologies with regulatory requirements and ensure their safe and effective use in healthcare applications primarily focused on AI-enabled systems in commercial use.

ISO/IEC is actively developing standards for AI, including concepts and terminology for AI (International Organization for Standardization/International Electrotechnical Commission, 2022) and AI management systems (International Organization for Standardization/International Electrotechnical Commission, 2023). These standards provide foundational frameworks to ensure the reliability, accountability, and ethical deployment of AI systems across industries, including healthcare applications. CHAI is a multi-stakeholder organization that focuses on ethical guidelines and assurance standards specific to healthcare applications of AI. CHAI's initiatives emphasize trustworthy AI implementation for healthcare, aiming to bridge the gap between abstract principles and practical implementation (Coalition for Health AI, 2023, 2024a, 2024b).

The development of global standards for AI technologies is essential for ensuring reliability, interoperability, and ethical alignment. AI standards can help ensure that AI systems perform consistently and accurately, reducing errors or unintended consequences, which is particularly critical in healthcare. Despite these efforts, the standardization of AI technologies faces several challenges. The rapid pace of technological advancements often outpaces the development of standards, while global variability in regulatory environments, cultural values, and healthcare systems complicates the creation of universally applicable guidelines. Additionally, the complexity of AI systems, including their reliance on large datasets and intricate algorithms, makes it difficult to address all potential risks and ethical concerns comprehensively (Kapusta et al., 2025). A collaborative AI assurance research lab would address these challenges by serving as a trusted hub to translate global standards like those developed by ISO, IEC, and CHAI, tailoring them to the unique needs of the research and development community in biomedicine and health. By focusing on specific needs of the NIH research community, like rigorous validation, ethical alignment, clinical relevance, and interoperability, an assurance lab would bridge the gap between abstract principles and practical application, ensuring AI systems are reliable, transparent, and effective in advancing healthcare outcomes.

Academic and Not-for-Profit Organizations

Academic institutions are central to advancing AI standardization, policy development, and ethical frameworks, particularly in biomedical and health research. Through interdisciplinary research and collaborative initiatives, academia is shaping assurance guidelines and methodologies that ensure AI technologies are ethical, responsible, and

reproducible. These efforts focus on addressing societal, technical, and regulatory challenges while emphasizing transparency, accountability, inclusivity, and representativeness.

The Partnership on AI (PAI)¹⁴ is a leading initiative that promotes responsible AI practices, particularly in healthcare applications. PAI's work includes developing ethical guidelines that emphasize fairness and inclusivity, conducting research on algorithmic fairness and explainability, and fostering collaboration between academia, industry, and non-profits. Similarly, the Stanford Institute for Human-Centered Artificial Intelligence (HAI)¹⁵ focuses on ensuring AI technologies are designed with human values at their core (Jiang et al., 2025; Liu et al., 2026). HAI develops frameworks to address bias, privacy, and accountability in healthcare AI systems while collaborating with policymakers to shape regulations that promote transparency and fairness.

Other academic efforts, such as the AI4Health Consortium¹⁶ and the Alan Turing Institute,¹⁷ focus on standardization and reproducibility in AI research. The AI4Health Consortium develops open datasets, benchmarking tools, and reproducibility guidelines to ensure reliable and scalable AI applications in healthcare. The Alan Turing Institute, the UK's national institute for data science and AI, is advancing ethical and technical standards for AI, including research on algorithmic fairness and privacy-preserving techniques like federated learning. Their collaboration with the UK National Health Service (NHS) ensures that AI applications in public health adhere to ethical and technical standards.

Institutions such as MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL)¹⁸ and the University of Toronto's Vector Institute¹⁹ further contribute to AI governance and policy development. CSAIL focuses on creating interpretable AI models for healthcare decision-making and collaborates with policymakers to promote responsible AI use. The Vector Institute develops standardized methodologies for AI applications in drug discovery, genomics, and medical imaging while engaging with

¹⁴ Partnership on AI (PAI). <https://partnershiponai.org/>

¹⁵ Stanford Institute for Human Centered AI (HAI). <https://hai.stanford.edu/>

¹⁶ AI4 Health Consortium. <https://ai4healthcro.eu/>

¹⁷ Alan Turing Institute. <https://www.turing.ac.uk/>

¹⁸ MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). <https://www.csail.mit.edu/>

¹⁹ University of Toronto, Vector Institute. <https://vectorinstitute.ai/>

government agencies to shape ethical AI policies.

There is also a collection of guidelines from a group called Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) that has provided guidance on developing, reporting, implementing, and validating predictive models in clinical applications (Collins et al., 2015). This work has been extended to include guidance for reporting clinical prediction models that use regression or machine learning (ML) methods (Collins et al., 2024) as well as large language models (Gallifant et al., 2025). Collectively, the guidelines aim to improve the quality, transparency, and reproducibility of AI research, particularly in biomedical and health domains.

Researchers and developers are encouraged to use these frameworks to ensure that their studies and models meet high standards of rigor and reliability. Depending on the type of study (e.g., clinical trials, diagnostic accuracy, prediction models), different guidelines may be more appropriate.

Through these initiatives, academic institutions are driving progress in creating assurance frameworks and methodologies that benefit healthcare providers, patients, and researchers alike. A collaborative AI assurance research lab would serve as a hub for efficient dissemination of, and access to, valuable assurance insights and best practices across academia that are curated for the NIH research community.

Government Initiatives by Entity

Government organizations play a critical role in advancing AI standardization, policy development, and ethical governance.

Government organizations play a critical role in advancing AI standardization, policy development, and ethical governance. These entities are tasked with ensuring that AI technologies are safe, reliable, and transparent while fostering innovation and addressing societal challenges. Through regulatory frameworks, funding initiatives, and interagency collaborations, U.S.

government agencies are shaping the landscape of AI development and deployment in biomedical and health research as well as healthcare.

The National AI Research Resource (NAIRR) Pilot²⁰ is an initiative aimed at democratizing access to AI research infrastructure across the U.S. Proposed by the National AI Initiative Office and supported by agencies such as the National Science Foundation (NSF) and the White House Office of Science and Technology Policy

²⁰ National Artificial Intelligence Research Resource (NAIRR) Pilot. <https://nairrpilot.org/>

(OSTP), NAIRR seeks to provide researchers with access to high-performance computing resources, curated datasets, and tools necessary for advancing AI innovation. By lowering barriers to entry, NAIRR aims to empower researchers to contribute to AI advancements. NAIRR's focus is on how access to data and computing resources could advance research, discovery and innovation.

The National Institute of Standards and Technology (NIST)²¹ is a leader in developing comprehensive frameworks for AI governance and risk management. NIST's AI Risk Management Framework provides organizations with guidelines to assess and mitigate risks associated with AI systems, emphasizing transparency, accountability, and reliability. In healthcare, this framework is particularly relevant for ensuring that AI tools meet safety and ethical standards. NIST also works on standardizing AI-related terminology and methodologies, fostering interoperability and reproducibility across sectors. By engaging with stakeholders from academia, industry, and government, NIST ensures that its frameworks are adaptable to the evolving needs of healthcare AI.

The Assistant Secretary for Technology Policy (ASTP)²² focuses on integrating AI into health IT systems while ensuring interoperability and patient privacy. ASTP's efforts include promoting the use of standards like HL7® FHIR® (Fast Healthcare Interoperability Resources) to enable seamless data exchange between AI systems and electronic health records (EHRs). By fostering interoperability, ASTP ensures that AI technologies can be effectively integrated into clinical workflows. Additionally, ASTP is actively involved in addressing ethical concerns, such as bias in AI algorithms and equitable access to AI-driven healthcare solutions, particularly for underserved populations.

The U.S. Food and Drug Administration (FDA)²³ plays a critical regulatory role in overseeing AI-based medical devices and software. FDA has developed guidelines for Software as a Medical Device (SaMD), providing a regulatory pathway for AI innovations in healthcare. These guidelines emphasize algorithmic transparency, real-world performance monitoring, and patient safety. FDA also collaborates with industry and academic stakeholders to refine its regulatory frameworks for adaptive AI systems, ensuring that these technologies can evolve while maintaining compliance with safety

²¹ National Institute of Standards and Technology (NIST). <https://www.nist.gov/>

²² Assistant Secretary for Technology Policy. <https://www.healthit.gov/>

²³ U.S. Food and Drug Administration (FDA). <https://www.fda.gov/>

standards. Through initiatives like the Digital Health Center of Excellence,²⁴ FDA is fostering innovation in AI-driven healthcare while prioritizing ethical and safety considerations.

U.S. government agencies are also contributing to AI standardization and governance. NIH is actively funding AI research through initiatives like Bridge2AI,²⁵ which focuses on creating high-quality, ethically sourced datasets for AI model training. NIH also collaborates with other agencies to address challenges in areas such as AI ethics, reproducibility, and transparency. The Department of War (DoW), through its Chief Digital and Artificial Intelligence Office (CDAO),²⁶ is exploring AI applications for military use, emphasizing ethical use and operational safety. Department of Energy (DOE), through its Office of Science, is also working in this space implementing transparency initiatives across their component data, models, and analysis groups to ensure that outputs are reproducible, transparent, and communicative of error and uncertainty.²⁷

Interagency collaborations further strengthen the U.S. government's efforts in AI governance. For example, the National Artificial Intelligence Initiative Office coordinates AI research and policy across agencies, ensuring alignment with national priorities. The Global Partnership on Artificial Intelligence (GPAI),²⁸ which includes U.S. participation, fosters international collaboration on AI ethics, governance, and standardization. These efforts demonstrate the U.S. government's commitment to advancing AI technologies responsibly while addressing critical challenges in biomedical and health research.

Through these initiatives, governmental organizations are driving progress in AI standardization, policy development, and governance. By fostering collaboration across sectors and prioritizing transparency, accountability, and representativeness, the U.S. government is ensuring that AI technologies benefit patients, providers, and researchers while safeguarding public trust. While U.S. government agencies are making significant strides in advancing AI research, governance, and deployment, efforts can often become fragmented due to the variable missions, priorities, and operational frameworks

²⁴ U.S. Food and Drug Administration (FDA), Digital Health Center of Excellence. <https://www.fda.gov/medical-devices/digital-health-center-excellence>

²⁵ NIH Common Fund, Bridge to Artificial Intelligence (Bridge2AI). <https://bridge2ai.org>

²⁶ Department of War, Chief Digital and Artificial Intelligence Office (CDAO). <https://www.ai.mil/>

²⁷ Department of Energy, Restoring Gold Standard Science DOE Implementation Plan. <https://www.energy.gov/gold-standard-science/>

²⁸ Global Partnership on Artificial Intelligence (GPAI). <https://gpai.ai/>

of individual organizations. Without a collective strategy, AI resources may fail to meet the specific needs of their intended use cases, resulting in suboptimal outcomes or barriers to adoption. NIH, through a collaborative AI assurance research lab, is ideally positioned to engage with biomedical and health researchers to understand the domain specific challenges to AI development and adoption, thereby developing AI assurance resources tailored to those specific needs.

Industry Frameworks and Platforms

Software frameworks and platforms are essential for the development, testing, and deployment of AI systems in healthcare. These tools provide researchers and developers with the infrastructure needed to build robust, scalable, and reproducible AI models. By offering specialized libraries, pre-trained models, and streamlined workflows, these frameworks accelerate innovation while ensuring consistency and reliability in AI applications. From general-purpose ML frameworks to healthcare-specific platforms, these tools are driving progress in areas such as medical imaging, genomics, natural language processing (NLP), and predictive analytics.

One healthcare-specific framework making significant strides is MONAI, or Medical Open Network for AI.²⁹ MONAI is an open-source platform, built on PyTorch, tailored for medical imaging applications. It provides tools for data preprocessing, model training, and evaluation, enabling researchers to develop reproducible workflows for tasks such as segmentation, classification, and anomaly detection in radiology and pathology. MONAI's modular design and support for distributed computing make it a powerful resource for scaling AI research in medical imaging, ensuring that models can be standardized and deployed effectively across disparate clinical settings.

General-purpose frameworks like PyTorch and TensorFlow are widely used across healthcare AI applications due to their flexibility and scalability. PyTorch, developed by Meta, with its dynamic computation graphs and intuitive interface, is ideal for rapid prototyping and experimentation, making it a popular choice for researchers working on medical imaging, genomics, and predictive analytics. TensorFlow, developed by Google, provides tools for building and deploying AI models at scale, with support for distributed computing and multi-platform deployment. TensorFlow's ecosystem enables researchers to deploy AI models in variable healthcare environments, from hospitals to remote clinics.

²⁹ Medical Open Network for AI (MONAI). <https://monai.io/>

Platforms specializing in NLP, such as Hugging Face,³⁰ are reshaping how AI is applied to clinical text analysis and patient communication. Hugging Face's Transformers library provides pre-trained models for tasks like text classification, question answering, and language translation, making it an invaluable resource for healthcare applications such as automated medical coding, clinical documentation, and patient sentiment analysis. Its repository of models allows researchers to quickly adapt state-of-the-art AI techniques to healthcare-specific challenges. Hugging Face also integrates seamlessly with frameworks like PyTorch and TensorFlow, enabling streamlined workflows for AI development.

Other platforms, such as NVIDIA Clara³¹ and OpenMined,³² address unique challenges in healthcare AI. NVIDIA Clara provides tools for medical imaging, genomics, and computational drug discovery, with a strong emphasis on federated learning capabilities that preserve patient privacy while enabling collaborative AI research across institutions. OpenMined, through its PySyft framework, specializes in privacy-preserving AI, allowing researchers to train models on decentralized datasets without compromising data security. These platforms are critical for ensuring that AI systems adhere to ethical standards while advancing innovation in sensitive domains like healthcare.

These frameworks and platforms collectively empower the healthcare AI community to develop solutions that improve patient outcomes, enhance clinical workflows, and advance biomedical and health research. By leveraging open-source tools and healthcare-specific platforms, researchers can address critical challenges in AI ethics, privacy, and interoperability while accelerating the development of scalable and impactful AI models. These technologies are not only driving innovation but also ensuring that AI systems are accessible, equitable, and trustworthy across the healthcare ecosystem.

While private sector contributions to open-access AI resources for biomedical and health research have been influential, this sector often prioritizes scalability, marketability, and competitive advantage, which can conflict with the goals of open science, equitable and sustained access, and reproducibility. These challenges can hinder the widespread adoption and utility of AI in biomedical and health research and illustrate the need for federal leadership to establish evidence-based AI assurance

³⁰ Hugging Face. <https://huggingface.co/>

³¹ NVIDIA Clara. <https://www.nvidia.com/en-us/clara/>

³² OpenMined. <https://openmined.org/>

resources to drive innovation through collaborative research and engagement. An NIH collaborative AI assurance research lab would have the capacity to help researchers identify assurance resources aligned with the goals mentioned above and provide guidance on how to best employ industry generated tools.

Challenges and Opportunities

The rapid evolution of AI in biomedical and health research presents both significant challenges and unique opportunities for NIH. While industry, academia, and government agencies have made strides in developing AI assurance standards, frameworks, and tools, the fragmented nature of these efforts could lead to competing interests, proprietary constraints, and limited accessibility. For NIH, these challenges underscore the need for a trusted and holistic AI assurance effort serving the unique needs of the NIH research community.

A collaborative AI assurance research lab could provide researchers with assurance resources for high quality AI development that align to established technical, ethical, and assurance principles. These assurance resources can help researchers identify the best practices in performing their AI research which can include pointing to reference datasets, scalable infrastructure, and specialized AI resources designed for biomedical and health research applications. By building on existing initiatives and frameworks, such as MONAI or Hugging Face, NIH would empower its researchers to advance AI innovation while maintaining rigorous standards for reproducibility, privacy, and performance.

2 Approach Overview

MITRE, in collaboration with NIH's Office of Data Science Strategy (ODSS), worked with representatives from industry, academia, and government to collaboratively explore AI assurance, implementation challenges, and solutions for biomedical and health research. Figure 1 provides an overview of the stakeholder and leadership groups that were considered at various levels, with the primary focus of understanding the needs and challenges of the biomedical and health research community. This cross-sector approach fostered rich perspectives and expertise, enabling a broad examination of challenges and opportunities in AI development, adoption, and assurance. Participants included specialists in biomedical and health research, computer science, AI development, digital health, AI policy, and bioethics, as well as representatives from NIH offices, institutes, and initiatives such as Bridge2AI and AIM-AHEAD.

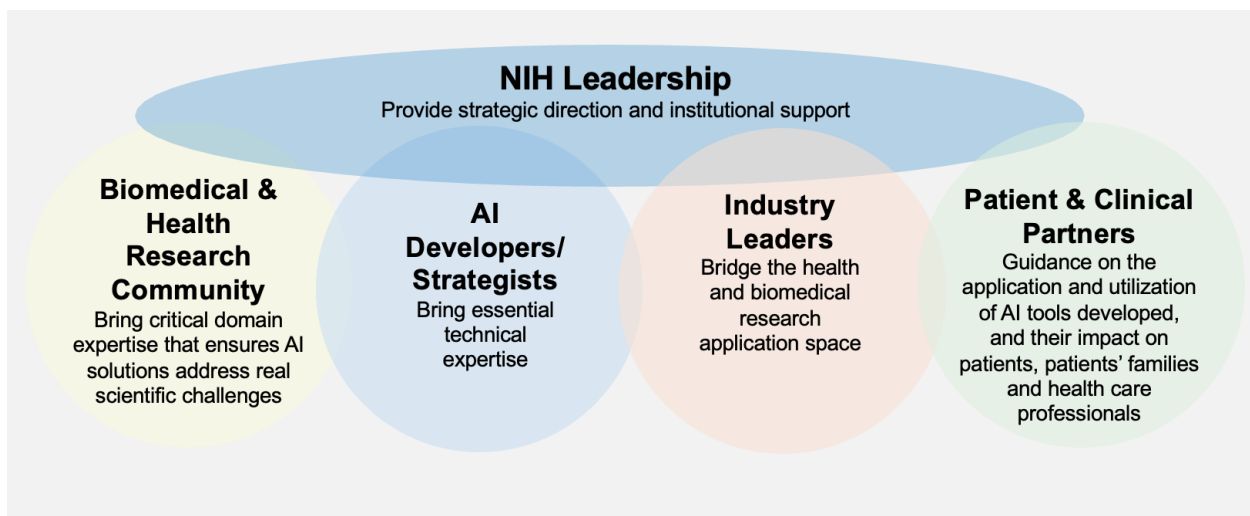


Figure 1. Overview of Stakeholders and Leadership to Understand AI Challenges

To understand the needs of the research community, MITRE conducted a series of engagements with select biomedical and health researchers to gather insights on AI assurance methodologies, challenges, and potential solutions.

The engagements throughout the base period included initial explorations with exploratory pilot teams and relevant subject matter experts such as Bridge2AI investigator teams, AIM-AHEAD investigators, and other connections through the NIH research community. Additionally, monthly meetings were held with MITRE's advisory board to provide strategic guidance, feedback on ongoing work, and facilitate connections. The advisory board comprised representatives from multiple organizations, including the Mayo Clinic, the NIST Center for AI Standards and Innovation, and the

Virginia Tech National Security Institute. This group brought expertise spanning healthcare innovation, academic research, government policy, and bioethics.

The core effort for this project was hosting regular technical meetings with the three selected exploratory pilot teams. The pilot engagement provided a platform for identifying critical needs, such as improving the integration of responsible AI practices into research workflows, addressing risks associated with resource gaps, and iterating on preliminary approaches to align with assurance, ethical, and technical frameworks. This in turn, set the foundation for an initial set of assurance practices for AI in biomedical and health research.

Further, the collaborative efforts across all the engagements were instrumental in shaping the work's direction. Collectively, the representatives highlighted the importance of addressing challenges such as data curation and methodologies, reproducibility, scalability, and ethical considerations while fostering innovation in AI assurance methodologies. Table 1 below summarizes the representative composition of the engagement activities.

Table 1. Representative Composition of Engagement

Engagement	Institution
Leadership	NIH Office of Data Science Strategy (ODSS)
Leadership	MITRE
Expertise	NIH Common Fund's Bridge to Artificial Intelligence (Bridge2AI)
Expertise	NIH ODSS AIM-AHEAD
Exploratory Pilot	NIH National Institute on Minority Health and Health Disparities (NIMHD)
Exploratory Pilot	NIH National Cancer Institute (NCI)
Exploratory Pilot	Oak Ridge National Laboratory (ORNL)
Exploratory Pilot	Massachusetts General Hospital (MGH)
Advisory Board	Mayo Clinic
Advisory Board	Virginia Tech National Security Institute

Engagement	Institution
Advisory Board	UMass Chan Medical School
Advisory Board	Tuskegee University National Center for Bioethics in Research
Advisory Board	NIST Center for AI Standards and Innovation

2.1 Exploratory Pilots

Three exploratory pilots were conducted to investigate different concepts related to AI assurance within NIH research. The pilots focused on explainable AI (XAI) methods, model calibration techniques, and designing human-AI interaction studies for novel applications. These pilots provided valuable insights into the challenges, best practices, and methodologies that shape AI development, assurance, and use in biomedical and health research workflows.

2.1.1 Exploratory Pilots Background Details

The exploratory pilots were chosen to study a variety of AI use cases to solve real-world biomedical and health problems. The pilots largely represented three stages of the AI lifecycle (model testing, deployment, and operation), though spanned aspects from across the AI lifecycle. Figure 2 summarizes the lifecycle phases and functional groups explored with each pilot.

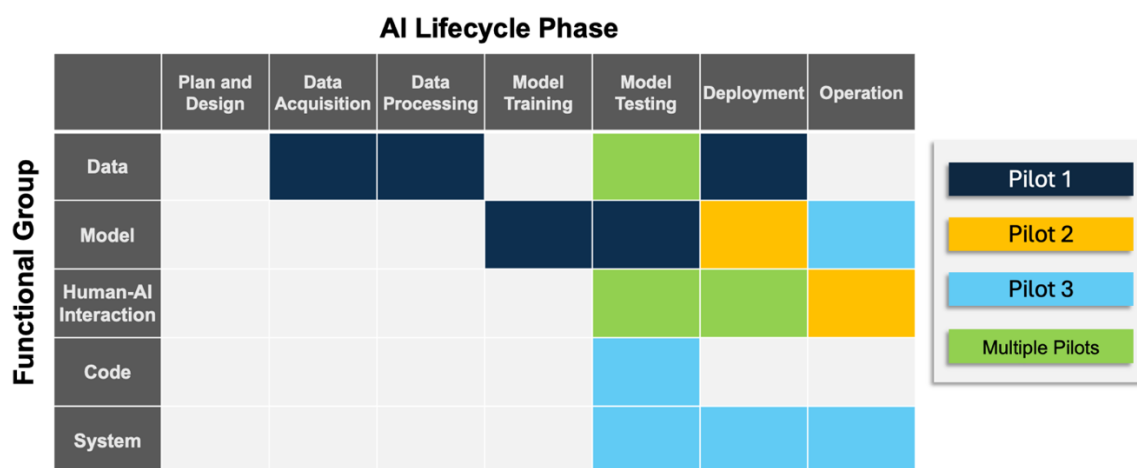


Figure 2. Aspects of AI Assurance Explored

Exploratory Pilot 1 (Model Testing Phase)

Description: Intramural investigator team leveraging *All of Us* Research Program dataset to identify and understand factors contributing to diabetes.

Objective: Investigate the utility of XAI tools, such as SHAP (SHapley Additive exPlanations), in identifying features in a diabetes detection model and correlating them with real-world determinants of diabetes.

Area of Investigation: Role of AI explainability methods for deriving factors contributing to chronic disease.

Approach:

- Validate the features identified by XAI methods to ensure alignment with research objectives.
- Address the lack of established guidance for interpreting real-world data using XAI tools.

Exploratory Pilot 2 (Deployment Phase)

Description: Extramural clinical research team developing coma prognosis tool.

Objective: Develop an AI system to assist clinicians in coma recovery prognosis and inform care decisions.

Area of Investigation: Role of human-AI interactions in interpretation of model predictions and contributing factors, using an AI-enabled UI to present key contributing factors to clinicians to support coma prognosis.

Approach:

- Ensure system performance in high-risk clinical settings.
- Help physicians interpret model outputs effectively.
- Establish evaluation methods for novel applications with limited benchmarks, especially around human-AI interactions.

Exploratory Pilot 3 (Operation Phase)

Description: Large-scale platform automating cancer registry tasks.

Objective: Reduce errors related to uncertainty in predictions by leveraging uncertainty estimation methods such as deep abstention classifiers for large-scale production-level platform using multimodal architecture to analyze patient data and automate cancer registry tasks.

Area of Investigation: Post-processing ML techniques for predictions informing cancer registry data, using large scale multimodal ML to accelerate efficient and more accurate cancer registries.

Approach:

- Develop a post-processing methodology to ensure reliability, such as model confidence calibration, to lower errors and improve system performance.
- Implement mechanisms for continuous monitoring and adaptation in real-world environments.

2.1.2 Exploratory Pilot Documentation

MITRE reviewed the AI use case documentation and preliminary approaches for each pilot, including details on the development process, data sources, model architecture, workflows, decision-making criteria, and user requirements. This documentation provided a detailed view into the role AI plays in researchers' broader work.

Capture & Collection of Current AI Assurance Approaches

Each pilot had developed their own, in-house, AI assurance approaches, many of which were not formalized or documented. To gain insights into current concepts and methods, MITRE gave pilots an initial structure for documenting existing AI assurance approaches. An entry had the following components: a method or approach, desired outcome using that approach, and the origin of how the approach was derived. The current approaches across pilots highlighted an emphasis on the following areas:

- **Data Curation and Management:** Strategies for handling data bias, preprocessing, and ensuring data quality.
- **Model Training, Validation, and Testing:** Techniques for optimizing model performance, ensuring reliability, appropriate use, and minimizing risks such as bias in predictions.
- **Clinical vs. Technical Interpretation:** Approaches for bridging the gap between technical outputs and clinical relevance, ensuring results are interpretable and actionable for healthcare professionals.
- **Continuous Monitoring and Maintenance:** Methods for integrating AI systems into existing workflows and continuously monitoring and maintaining their performance over time.

There was broad consensus that the exercise identified a need for an evidence-based and widely accepted playbook for AI assurance in biomedical and health research. A total of nine approaches to assurance challenges were identified across the pilots and evaluated to analyze their methodology, rationale, and derivation process. These evaluations served as a foundation for understanding AI assurance challenges and as a precursor to establishing an initial knowledge base.

The exploratory pilots and other collaborative engagements provided critical insights into the challenges and opportunities of integrating AI into NIH research. By documenting use cases, assessing current approaches, and aligning with technical, ethical, and assurance principles, this initiative creates an initial playbook that establishes scalable assurance practices and tools to support AI adoption throughout the biomedical and health research lifecycle.

3 Findings

3.1 Challenges

The exploratory pilots and broader engagements revealed several substantial challenges about the current state of AI-related activities within the biomedical and health research community. These challenges highlight the unique barriers investigators face with the integration of AI into research workflows. Below are the challenges identified from the exploratory pilots.

Lack of Curated Assurance Resources and Standardized Framework

The biomedical and health research community urgently needs curated AI assurance resources, including a formalized playbook, standardized processes, and frameworks. This challenge was shared across all exploratory pilots and the absence of these resources led to fragmented approaches, slowed progress, and limited opportunities for knowledge sharing. For example, Pilot 1 investigators delayed publication due to concerns about alignment with AI assurance practices with their use of XAI methods, and Pilot 3 required an extensive amount of time and resources for AI development and deployment due to lack of standardized processes, often gathering disparate resources from a variety of sources (e.g., journal check lists, medical center internal resources). This gap can create uncertainty, particularly with novel applications, where mistakes could have serious implications for patient safety, public health, and scientific integrity.

Challenges in Aligning with Guiding Principles

Researchers recognized the importance of aligning AI workflows with established technical, ethical, and assurance principles. However, this alignment was often challenging due to the lack of clear, tailored guidance and resources specific to biomedical and health applications. Every exploratory pilot recognized the importance of developing responsible AI, yet they had limited ability to substantiate such claims. For Pilot 1, this meant confirming their usage of XAI methods was meeting ethical and technical principles, and for Pilot 2, how to implement their technology ethically with limited benchmarks for human-AI interactions in neurocritical care.

Varying Perspectives and Approaches to AI Development & Assurance

Stakeholders across academia, industry, and government hold varying perceptions of what constitutes effective AI research, development, and assurance. Some view the development of AI as primarily technical, focused on model training, testing, validation, explainability, and performance metrics, while others emphasize investigating ethical

considerations, such as safety, data privacy, fairness, and accountability. This heterogeneity across the various engagements including pilots, subject matter experts, and the advisory board, reflects the interdisciplinary challenges of AI in biomedical and health research, where technical, scientific, ethical, legal, and regulatory dimensions intersect and are required.

Unclear Guidance for Deployment and Evaluation of Novel Applications

Despite the potential of AI in NIH research, there is a lack of clear guidance on how to evaluate and deploy AI systems for novel applications, such as Exploratory Pilot 2, prediction of coma recovery. These applications often involve complex ethical and technical considerations, such as ensuring patient safety, usability, and validating AI systems clinical utility in real-world clinical settings. Without robust assurance frameworks and practices to provide guidance, researchers may hesitate to pursue ambitious AI-driven projects or struggle to navigate regulatory and ethical challenges. For Pilot 2, limited resources were available to design studies for novel AI applications in neurocritical care, where inherent uncertainty (such as incomplete ground truth labels for predicting coma recovery) poses a challenge. For Pilot 3, insufficient guidance on AI deployment led investigators to develop their own evaluation methods, which included assessing the performance in out-of-distribution data over time, but did not include systematic monitoring of the system over time. They also ran into issues when the workflows leveraged APIs, which were not originally designed for that specific application space, resulting in erroneous AI outputs and reduced end-user trust in the system.

3.2 Preliminary Biomedical and Health AI Assurance Practices

Below are a preliminary set of AI assurance practices in biomedical and health research. These stemmed from the pilot documentation and current approaches to assurance challenges that each pilot initially utilized, iterated on via the weekly technical meetings to align with technical, ethical, and assurance principles, and then drafted into the below assurance practices. They emphasize transparency, reliability, fairness, and clinical utility. As AI systems increasingly influence critical decisions in healthcare, these preliminary assurance practices sketch the unique challenges posed by the NIH community, such as ensuring fair and representative outcomes across populations and mitigating high risks associated with uncalibrated predictions. Though preliminary, these are included to provide examples of the type of outputs that would be enabled by a collaborative AI assurance research lab moving forward.

Assurance Practice 1: Towards Reliability in AI Explainability

Explainable AI (XAI) methods can provide valuable insights into model behavior, but they are often insufficient to fully ensure model transparency and reliability. In some cases, XAI findings may oversimplify complex model behavior, fail to capture biases, or provide explanations that are inconsistent with the model's true decision-making process, leading to misleading or incorrect conclusions. To address these challenges, like those experienced by the team from Pilot 1, XAI outputs can be supplemented with metrics such as predictive parity, which evaluates fairness across groups, and uncertainty quantification, which assesses confidence in predictions. Tools like subgroup performance analysis can further help identify disparities in model performance across different populations. By incorporating these additional steps, practitioners can better evaluate fairness, reliability, and confidence in XAI outputs, promoting transparency, accountability, and ethical use of AI.

Assurance Practice 2: Path From Model Performance to Clinical Utility

In AI applications that influence life-and-death decisions, the absence of ground truth on outcomes introduces significant ethical and practical risks. To address these challenges, researchers and clinicians need an assurance toolkit (i.e., a structured "workbench") to guide the design of human-centered AI experimentation that prioritizes safety, efficacy, and dignity. Such a toolkit should include tools for managing uncertainty, evaluating fairness, and ensuring accountability in decision-making. Beyond technical metrics like accuracy and precision, it is essential to assess the system's interpretability and usability, enabling clinicians and end-users to understand and trust the AI's predictions. Incorporating domain-specific metrics, such as clinical utility and impact on patient outcomes, ensures alignment with real-world healthcare goals and supports evidence-based decision-making. Ultimately, designing effective and safe AI systems requires balancing technical performance with ethical considerations and practical usability, ensuring that these systems enhance patient care while respecting human dignity. This was critical for Pilot 2, where the team wanted to implement potentially lifechanging technologies within neurocritical care that are desperately needed, but require resources and a path forward that aligns AI assurance practices with the complexities of neurocritical care.

Assurance Practice 3: Confidence is not Correctness

Selecting the appropriate metrics for a given AI use case is crucial for evaluating performance and requires a deep understanding of the intended use and associated risks. In high-risk domains like healthcare, traditional metrics such as accuracy,

precision, and recall may provide a misleading picture of performance, as uncalibrated models can exhibit high confidence in incorrect predictions, potentially leading to dangerous clinical decisions. Pilot 3 utilized a deep abstaining classifier (DAC), which incorporates an “abstain” or “reject” class – i.e., the model learns to automatically weigh the risk of making an incorrect prediction against the cost of abstaining. The trade-off, however, is that the model may abstain on a large percentage of samples. For Pilot 3, this resulted in higher abstention rates for certain multiclass classification tasks for tumor characteristics (morphology and stage) surveillance, despite having a highly performant model. Calibration metrics like Expected Calibration Error (ECE) and post-processing techniques such as temperature scaling can help align confidence scores with actual probabilities, improving model reliability and trustworthiness. Identifying which approach has the most utility under a range of scenarios would help NIH researchers optimize their algorithms for specific applications. Beyond calibration, fairness and bias metrics are essential to ensure equitable performance across heterogeneous demographic groups. Metrics such as demographic parity, which ensures equal outcomes across groups, and equalized odds, which ensures consistent error rates across groups, can help mitigate disparities in an AI system.

While these preliminary AI assurance practices illustrate a step forward in promoting responsible AI, additional real-world evidence with further use cases will be needed to substantiate claims. They also raise important considerations for the use of AI in NIH research. By addressing key issues such as explainability, calibration, fairness, and clinical utility, refinement of these AI assurance practices into a validated playbook could provide a foundational knowledge base for accelerating the adoption of responsible AI in NIH research.

4 Recommendations & Next Steps

As noted throughout this document, MITRE recommends NIH establish a collaborative AI assurance research lab that extends initial work with the exploratory pilots to begin developing AI assurance resources, including practices/playbooks, benchmarks, evaluation methods, and standardized tools. As illustrated in Figure 3, the **Collaborative Research for the Advancement of Artificial Intelligence Laboratory (CORAL)** is envisioned to foster collaborative assurance research, focused on real-world applications, leveraging the expertise and needs of the biomedical and health research community. CORAL could be structured around the following mission, vision, and strategic components, each addressing the critical challenges, opportunities, and needs identified.

Mission: To provide AI assurance resources that accelerate the adoption and impact of AI in biomedical and health research while ensuring alignment to established technical, ethical, and assurance principles.

Vision: Enable responsible and scalable AI applications in biomedical and health research.

Strategic Components:

- **Research Lifecycle Integration:** Discover challenges for integrating responsible AI seamlessly into existing research workflows.
- **Collaborative Development:** Create practical assurance solutions through real-world AI use cases in partnership with the NIH research community.
- **Curation & Standardization:** Curate findings to establish a shared foundation of AI assurance resources to standardize workflows across biomedical and health research contexts and computational environments.

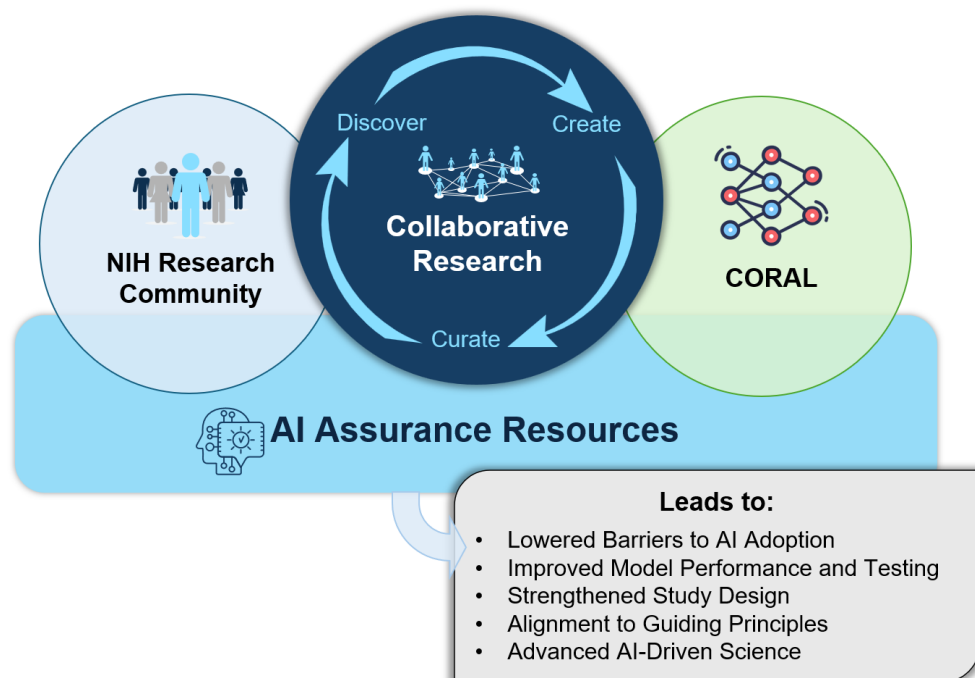


Figure 3. NIH Collaborative Research for the Advancement of AI Lab (CORAL)

4.1 Lab Design Overview

By prioritizing collaboration, CORAL would seek to create an environment where investigators can jointly discover, create, and curate AI assurance methodologies and applications that directly address pressing research questions – building upon the existing body of knowledge around AI assurance and translating those real-world insights into actionable AI assurance practices for the NIH research community. Due to the collaborative and dynamic nature of the lab, much of the initial infrastructure resources, such as compute and storage, would reside with AI use case owners. Once the lab has established its initial AI assurance knowledgebase, further assessment of infrastructure needs would be conducted.

The lab's workflow would be structured to deliver immediate and tangible value to collaborators while simultaneously contributing to the broader scientific community. Through each collaboration, CORAL would aim to enhance, expand, and add to collective AI knowledge and resources in meaningful ways. Investigations would be designed to produce actionable outcomes, aligned with the solutions identified above, that improve research workflows and overall impact. Examples of these outcomes include:

Lowered Barriers to AI Adoption

Access to curated AI assurance resources from a trusted source such as domain-specific pre-trained models, standardized codebases, and researcher-friendly tools will lower AI adoption barriers and accelerate investigators' use of AI to conduct robust and reproducible research more efficiently and effectively.

Improved Model Performance and Testing

Testing infrastructure with tailored benchmarks, guidance on metric selection, and methods for training and testing enable researchers to refine AI models, ensuring they are optimized for accuracy, reliability, and scalability. This includes rigorous testing and validation to address challenges with appropriate use, transparency, efficacy, and robustness in AI systems.

Strengthened Study Design

Collaborative investigations provide opportunities to advance methods of AI evaluation as part of study design processes, enabling researchers to develop more efficient, innovative, and reproducible approaches to data collection, analysis, and interpretation.

Alignment to Guiding Principles

A focus of the lab is ensuring that AI-driven research adheres to principles including accessibility, reproducibility, and transparency. By developing validated and standardized frameworks and tools, the lab supports investigators in producing results that can be reliably trusted to generate appropriate results and replicated across different studies and contexts.

Advanced AI-Driven Science

Through novel insights and iterative collaboration, the lab tackles pressing research challenges while fostering communal knowledge. This approach not only addresses current needs but also establishes a robust foundation for future breakthroughs in AI-driven biomedical and health research.

4.2 Phased Lab Implementation

Successful implementation of CORAL requires a clear and actionable roadmap that aligns with the needs of the NIH research community. This section outlines the proposed phased approach to rolling out the lab, emphasizing milestones that span short-term and long-term activities related to the lab's mission and vision.

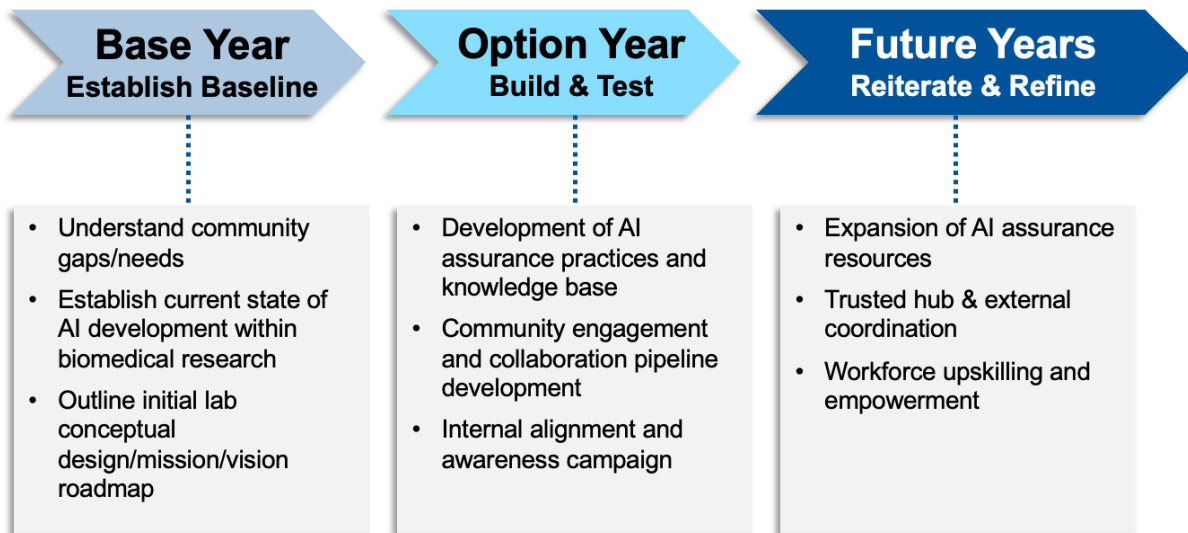


Figure 4. Overview of CORAL Roadmap

4.2.1 Option Year (planned: June 2025–June 2026)

CORAL's second phase, build and test, is centered on three critical focus areas designed to establish a strong foundation for its activities and long-term impact. First, the *Development of AI Assurance Playbook and Knowledge Base* aims to advance AI reliability, trustworthiness, reproducibility, and scalability by documenting standardized methodologies and creating a prototype knowledge base aligned with established technical, ethical, and assurance principles. This work will be aimed at testing and refining the lab's approach to conducting real-world investigations. Second, *Community Engagement and Collaboration Pipeline Development* focuses on understanding more broadly the needs of biomedical and health researchers and AI developers through workshops, user personas, and town halls, while building a pipeline for future collaborations. Finally, the *Internal Alignment and Awareness Campaign* seeks to align the lab's objectives with NIH priorities and initiatives, leveraging existing resources and fostering communication across NIH institutes and centers. Together, these efforts aim to ensure the lab is well-positioned to support impactful and responsible AI research.

1. Development of AI Assurance Playbook and Knowledge Base

Activities:

- Conduct two collaborative research efforts to identify, document, and validate assurance practices on areas such as AI reliability, reproducibility, and scalability in real-world biomedical and health research contexts.

-
- Develop an initial approach for capturing and providing access to AI assurance playbook that support established technical, ethical, and assurance principles.
 - Coordinate with other initiatives across NIH to assess infrastructure needs to support knowledge base, such as cloud compute, network, and storage.

Milestone(s):

- Publish two case studies from collaborations to highlight the impact of lab activities.
- Create a prototype knowledge base and have collaborators contribute five assurance practices for the playbook.

2. Community Engagement and Collaboration Pipeline Development

Activities:

- Conduct workshops with user groups to map out workflows for AI integration into research lifecycles.
- Develop detailed user personas using methods like empathy maps to represent the variable needs of biomedical and health researchers and AI developers.
- Refine and prioritize needs through user group workshops to identify critical requirements for AI assurance resources.
- Build a collaboration pipeline based on community feedback.

Milestone(s):

- Identify five AI use cases for future engagement.
- Obtain input from the biomedical and health research community at three town halls regarding their needs for AI assurance resources, workflows, and tools.
- Create an initial set of demonstrations and mockups of lab resources, tools, and interfaces based on feedback from stakeholders.

3. Internal Alignment and Awareness Campaign

Activities:

- Launch an internal awareness campaign to inform NIH staff and researchers about the lab's mission, vision, and goals.
- Conduct an internal review to inventory existing NIH initiatives, resources, tools,

datasets, and infrastructure that can support CORAL.

- Use feedback from the awareness campaign to refine lab objectives and ensure alignment with NIH priorities and ongoing NIH initiatives, such as Bridge2AI and AIM-AHEAD.

Milestone(s):

- Receive feedback from three other NIH AI-related initiatives.
- Establish a coordination and communication strategy working with external stakeholders so the lab aligns with NIH priorities and ongoing AI initiatives.

These activities are critical for laying the foundation of CORAL's long-term success. Researchers need robust AI methodologies that are effective, ethical, and scalable across the unique biomedical and health use cases. By focusing on playbooks, community engagement, and internal alignment, CORAL would address urgent gaps in AI development, assurance, and standardization. Such efforts will accelerate the research community toward AI systems can be trusted, reproducible, responsible, and interoperable, ultimately advancing scientific discovery and improving public health outcomes.

4.2.2 Future Years (planned: June 2026–June 2027)

Future years would focus on ways to scale CORAL's impact, foster innovation, and establish a trusted hub for AI-driven biomedical and health research. By building upon the initial set of AI assurance resources, creating a trusted hub to connect across internal and external resources, and advancing workforce upskilling, CORAL would aim to address systemic challenges in AI adoption and integration while driving advancements in research workflows.

1. Expansion of AI Assurance Resources

Objective: Build high-quality, innovative AI resources aligned to established technical, ethical, and assurance principles tailored to biomedical and health research workflows.

Activities:

- Expand on AI assurance resources, such as benchmarking, testing and evaluation tools, an AI-ready dataset registry, pre-trained model repository, and reusable AI workflow frameworks, designed to ensure reproducibility, scalability, and reliability.

-
- Aggregate findings from investigations, user workshops, and early lab efforts to refine the lab's design and knowledgebase functionality.

Milestone(s):

- Launch an initial suite of AI assurance resources.
- Publish playbook and standardized formats for AI development and assurance.
- Validate the lab's knowledge base functionality through aggregated findings and benchmarking efforts.

2. Trusted Hub & External Coordination

Objective: Provide an access point for NIH AI-related assurance resources while fostering collaboration and capacity building.

Activities:

- Create a trusted hub for reference resources inside and outside of NIH, including code, documentation, training programs, tutorials, and playbook.
- Develop an intuitive user-friendly interface to make resources accessible to non-experts.
- Host events to demonstrate the hub concept, receive community feedback and build partnerships with government initiatives like NAIRR, industry collaborators, and academic institutions to expand access and foster innovation and collaboration.

Milestone(s):

- Establish a trusted hub for expanded AI assurance resources, ensuring accessibility and usability for researchers across the NIH community.
- Achieve 1,000 monthly visits 12 months after launch of lab's digital hub.
- Host five events aimed at obtaining community feedback and building partnerships.
- Form three strategic partnerships with other federal agencies to work on common, persistent challenges across agencies.

3. Workforce Upskilling and Empowerment

Objective: Equip interested researchers with the skills and tools needed to adopt and

integrate responsible AI methodologies into their workflows while advancing experimental design and workflow standardization.

Activities:

- Host training workshops to provide orientation on using AI assurance resources.
- Scale workforce upskilling initiatives to ensure researchers can adopt AI assurance methodologies effectively and responsibly.

Milestone(s):

- Develop a responsible AI development tutorial for demonstrating how to use AI assurance resources.
- Host two training workshops on responsible AI development for 100 attendees.

The long-term phase is critical for ensuring the lab's work has a lasting impact on biomedical and health research. By focusing on AI assurance resources, a trusted hub, and workforce upskilling, the lab would address systemic challenges in AI adoption, standardization, and scalability. This phased implementation ensures that CORAL is developed in a way that capabilities can be built, tested, and refined iteratively, ensuring they continue to meet the needs of the NIH research community as they evolve.

5 Conclusion

Investigators recognize AI as a powerful tool for enabling deeper insights and improving research efficiency, yet its adoption remains challenging. The exploratory pilots, conducted through real-world NIH use cases, underscored how resource-intensive development, rapidly evolving technology, and inconsistent alignment with ethical, technical, and assurance standards hinder scalable deployment. In the absence of clear, standardized guidance and best practices, investigators often rely on custom, in-house tools and fragmented workflows, spending significant time on costly, dense, duplicative, or poorly tailored resources that slow progress and limit the adoption, safety, and impact of AI in biomedical and health research.

To address these challenges, NIH in partnership with MITRE recommends establishing the Collaborative Research for the Advancement of Artificial Intelligence Laboratory (**CORAL**) as a trusted resource for the biomedical and health research community. CORAL would leverage collaborative research engagements using real-world AI use cases to generate tailored lessons learned, curated playbooks, benchmarks, testing and evaluation methods, and other assurance resources that directly support responsible AI-enabled research. Operating through an iterative framework centered on collaboration, continuous improvement, and validation with real-world evidence, CORAL would identify emerging assurance gaps, develop solutions, and adapt them to practical biomedical and health research settings. By fostering interdisciplinary partnerships, streamlining AI workflows, and setting new benchmarks for ethical and effective AI, CORAL would accelerate AI adoption across NIH initiatives, ultimately advancing scientific discovery, enabling precision medicine, and improving public health for the benefit of both the scientific community and society at large.

6 References

- Cao, W., Zhang, Q., Liu, J., & Liu, S. (2026). From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era. *Journal of medical Internet research*, 28, e86550. <https://doi.org/10.2196/86550>
- Coalition for Health AI (2023). *Blueprint for trustworthy AI implementation guide and assurance for healthcare*. <https://www.chai.org/workgroup/responsible-ai/blueprint-for-trustworthy-ai>
- Coalition for Health AI (2024a). *Responsible AI Guide (RAIG)*. <https://www.chai.org/workgroup/responsible-ai/responsible-ai-guide-raig-and-raig-executive-summary>
- Coalition for Health AI (2024b). *Responsible AI Checklist (RAIC)*. <https://www.chai.org/workgroup/responsible-ai/responsible-ai-checklists-raic>
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ (Clinical research ed.)*, 350, g7594. <https://doi.org/10.1136/bmj.g7594>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., Hoffman, M. M., ... Logullo, P. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ (Clinical research ed.)*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- Foote, H., Hong, C., Anwar, M., ... Lindsell, C.J. (2025). Embracing Generative Artificial Intelligence in Clinical Research and Beyond: Opportunities, Challenges, and Solutions. *JACC Adv*, Mar, 4(3). <https://doi.org/10.1016/j.jacadv.2025.101593>
- Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., Umeton, R., ... Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature medicine*, 31(1), 60–69. <https://doi.org/10.1038/s41591-024-03425-5>
- Institute of Electrical and Electronics Engineers (IEEE)/Electronic Industries Association (EIA) (2008). *Standard for Information Technology - Software Life Cycle Processes* (IEEE/EIA 12207-2008). <https://ieeexplore.ieee.org/document/707581>
- International Organization for Standardization (ISO) (2019). *Medical Devices - Application to risk management to medical devices* (ISO 14971:2019). <https://www.iso.org/standard/72704.html>

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2022). *Artificial Intelligence concepts and terminology* (ISO/IEC TR 22989:2022). <https://www.iso.org/standard/74296.html>

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2023). *Information Technology - Artificial Intelligence - Management System* (ISO/IEC 42001:2023). <https://www.iso.org/standard/42001>

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2024). *Information Technology - Artificial Intelligence - Use Cases* (ISO/IEC TR 24030:2024). <https://www.iso.org/standard/84144.html>

Jiang, Y., Black, K.C., Geng, G., Park, D., Zou, J., Ng, A.Y., & Chen, J.H. (2025). MedAgent Bench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2. <https://doi.org/10.1056/Aldbp2500144>

Jordan, P. (2006). Standard IEC 62304 - Medical Device Software - Software Lifecycle Processes. IET Seminar on Software for Medical Devices (pp. 41-47). <https://doi.org/10.1049/ic:20060141>

Kapusta, A.S., Jin, D., Teague, P.M., Houston, R.A., Elliott, J.B., Park, G.Y., & Holdren, S.S. (2025). A framework for the assurance of AI-enabled systems. *Proc. SPIE 13476, Assurance and Security for AI-enabled Systems 2025*. <https://doi.org/10.1117/12.3056719>

Le, J. P., Morrison, J., Malhotra, A., Nemati, S., Wardi, G., & Ford, J. S. (2026). National Institutes of Health-Funded Artificial Intelligence and Machine Learning Research, 2019-2023: Cross-Sectional Study. *Journal of medical Internet research*, 28, e84861. <https://doi.org/10.2196/84861>

Liu, Y., Carrero, Z. I., Jiang, X., Ferber, D., Wölflein, G., Zhang, L., Jayabalan, S., Lenz, T., Hui, Z., & Kather, J. N. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. *NPJ digital medicine*, 10.1038/s41746-026-02443-6. Advance online publication. <https://doi.org/10.1038/s41746-026-02443-6>

Robbins, D., Eris, O., Kapusta, L., Booker, L., & Ward, P. (2024). AI assurance: A repeatable process for assuring AI-enabled systems. *The MITRE Corporation*. <https://www.mitre.org/sites/default/files/2024-06/PR-24-1768-AI-Assurance-A-Repeatable-Process-Assuring-AI-Enabled-Systems.pdf>

Smith K. A Brief History of NCBI's Formation and Growth. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK148949>

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100>

The Lancet (2025). Reclaiming care in the age of AI. *Lancet (London, England)*, 406(10512), 1535. [https://doi.org/10.1016/S0140-6736\(25\)02062-8](https://doi.org/10.1016/S0140-6736(25)02062-8)

Appendix

A-1 The National Center for Biotechnology Information (NCBI)

Overview

The National Center for Biotechnology Information (NCBI), operating under the National Library of Medicine (NLM) at the National Institutes of Health (NIH) played a crucial role in establishing best practices, standards, and attainable workflows after the Human Genome Project. Examples include:

- **GenBank and Data Repositories:** NCBI created standardized databases like GenBank that provided uniform formats for storing and accessing genetic sequence data.
- **Sequence Identification Standards:** Development of accession numbers and unique identifiers for genetic sequences enabled consistent referencing across research globally.
- **BLAST and Standardized Analysis Tools:** Implementing bioinformatics tools like the Basic Local Alignment Search Tool (BLAST) provided researchers with consistent methods to analyze genetic data.
- **Reference Genome Assembly:** NCBI maintained reference genome assemblies that served as standardized workflow templates for comparative analysis.
- **Standardized Ontologies:** The Gene Ontology (GO) project established shared vocabularies for gene functions across organisms.
- **Data Sharing Policies:** NIH implemented policies requiring researchers to deposit data in standardized repositories, promoting access and reuse.

Impact of NCBI on Innovation

The establishment of best practices, standards, and attainable workflows by NCBI and NIH created a robust ecosystem for innovation across several areas as outlined below.

Democratization of Genomic Research

- **Lowered Technical Barriers:** Prior to these standards, only well-funded labs with significant computational resources could effectively analyze genomic data.

NCBI's standardized databases and analysis tools allowed smaller labs and even individual researchers to participate in genomic research.

- **Global Participation:** Researchers from developing countries gained access to the same high-quality genomic data and tools as those in well-funded institutions, expanding the global research community.
- **Educational Impact:** The standardization of formats and tools made it easier to teach genomics, creating a pipeline of trained researchers who could further drive innovation.

Acceleration of Discovery Timeline

- **Reduced Redundant Efforts:** Before NCBI best practices, standards, and attainable workflows, research groups often reinvented analysis methods that accommodated reformatted data. Common workflows and standards eliminated this duplicative work.
- **Enabled Large-Scale Analyses:** Common formats allowed researchers to combine raw data (mega-analysis) as well as results from multiple studies (meta-analysis), leading to insights that would otherwise be limited through individual studies.
- **Automated Processing Pipelines:** Best practices resources enabled the development of automated analysis workflows, reducing analysis time.

Commercial Ecosystem Development

- **Sequencing Technology Companies:** Companies (e.g., Illumina) could focus innovative improvements to sequencing technologies rather than data workflows and standards, driving both sequencing speed and cost reduction.
- **Bioinformatics Tools:** A market emerged for specialized analysis software built on standard data formats, creating an entire industry sector.
- **Diagnostic Applications:** Clinical genomics companies could develop diagnostic tests based on standard references, accelerating the translation of genomic discoveries to patient care.

Cross-Disciplinary Integration

- **Integration with Clinical Data:** Standard genomic formats allowed for easier connection with electronic health records and clinical databases.

-
- **Multi-omics Integration:** The success of genomic standards provided a template for standardizing other “-omics” data (e.g., proteomics, metabolomics), enabling more comprehensive etiological studies.
 - **Computational Biology Growth:** Best practices and workflows made it easier for computer scientists to apply advanced computational methods to biological questions without deep domain expertise in molecular biology.

Long-Term Scientific Culture Change

- **Data Sharing Norms:** The expectation that genomic data would be shared in standard repositories before publication became firmly established, creating a more collaborative research environment.
- **Reproducible Research:** Standardization improved reproducibility, building confidence in genomic findings and reducing wasted research effort.
- **Precompetitive Collaboration:** Even competing pharmaceutical companies recognized the value of collaborating on standard formats and references, leading to novel consortia approaches to drug discovery.

The NCBI/NIH standardization following the Human Genome Project demonstrates that thoughtfully implemented best practices, standards, and attainable workflows do not constrain innovation but rather create fertile ground for it to flourish more rapidly and in unexpected directions. This example offers a compelling model for how similar approaches could accelerate innovation in AI for biomedical and health research.

A Model for AI Innovation in Biomedical & Health Research

The NCBI/NIH approaches that catalyzed post-Human Genome Project innovation offer valuable frameworks that could be adapted to support AI innovation in biomedical and health research across the following domains.

Open Data Infrastructure with Appropriate Guardrails

- Creating managed access repositories for de-identified clinical data with standardized formats and robust privacy protections.
- Developing synthetic datasets that preserve statistical properties while protecting patient privacy.
- Establishing tiered access systems where increasingly sensitive data requires additional verification.

Example: An “AI BioBank” could provide access to curated, de-identified biomedical data with standardized formats for imaging data, clinical data, and multi-omics data.

Foundational Reference Resources

- Developing reference benchmark datasets for specific biomedical and/or disease areas.
- Creating standardized patient archetypes that represent all populations.
- Establishing baseline model performance expectations for different clinical tasks.

Example: An “AI Reference Collection” could provide well-characterized medical imaging datasets with expert annotations across different modalities and disease states, serving as the standard evaluation resource for new algorithms.

Common Analysis Tools with Open Standards

- Developing open-source implementations of AI assurance architectures.
- Creating standardized preprocessing pipelines for biomedical data.
- Establishing common evaluation frameworks for model performance.

Example: An “AI BioToolkit” could provide standardized methods for data preprocessing, model training, and evaluation specific to biomedical applications while supporting established technical, ethical, and assurance principles.

Metadata and Annotation Standards

- Extending biomedical ontologies to be more AI-compatible.
- Standardizing how AI models are documented (e.g., model cards, datasheets).
- Creating common frameworks for describing model limitations and appropriate use cases and/or exploratory pilots.

Example: An “AI Biomedical Ontology” could standardize how concepts are represented for machine learning (ML), making models more interoperable and interpretable.

Federated Approaches to Data Access

- Developing federated learning infrastructures that allow models to train across institutions and laboratories without data sharing.
- Creating secure compute environments for sensitive data analysis.

-
- Establishing data use agreements that protect privacy while enabling innovation.

Example: A “Federated AI Research Network” could allow algorithms to train across multiple systems without transferring sensitive data.

Public-Private Partnership Models

- Facilitating joint academic-industry collaborations on biomedical AI standards.
- Developing shared validation frameworks that benefit both sectors.
- Establishing IP frameworks that protect innovation while encouraging standardization.

Example: An “AI Biomedical Consortium” could bring together technology companies, pharmaceutical companies, medical centers, and academic researchers to develop shared resources for algorithm development and validation.

A-2 Current State of AI R&D In Health

Evolving Health AI Policy and Public Protection

AI systems in healthcare must adhere to rigorous safety and reliability standards. Health AI policies and protections have advanced over time with the development of key standards including the following. The IEEE Software Life Cycle Process standard, introduced in 1998, laid the groundwork for structured software development processes, emphasizing the importance of systematic methodologies.¹ This was further refined by the IEC 62304 standard in 2006, which specifically addressed medical device software, underscoring the need for comprehensive life cycle processes to ensure safety and efficacy.² In 2016, the U.S. Food and Drug Administration (FDA) established the National Evaluation System for health Technology (NEST) to enhance the evaluation

¹ Institute of Electrical and Electronics Engineers (IEEE)/Electronic Industries Association (EIA) (2008). Standard for Information Technology - Software Life Cycle Processes (IEEE/EIA 12207-2008). <https://ieeexplore.ieee.org/document/707581>

² Jordan, P. (2006). Standard IEC 62304 - Medical Device Software - Software Lifecycle Processes. IET Seminar on Software for Medical Devices (pp. 41-47). <https://doi.org/10.1049/ic:20060141>

and safety assurance of medical devices, aiming to create a robust ecosystem for continuous monitoring and improvement.³

Emerging AI R&D Standards for Biomedical & Health Research

International technical standards are necessary for fostering innovation, global competition, collaboration, and interoperability. They ensure safety and reliability, address ethical considerations, and support regulatory compliance. The International Organization for Standardization (ISO)⁴ together with the International Electrotechnical Commission (IEC)⁵ are leading the development of emerging AI standards, including those pertinent to AI assurance. This ISO/IEC partnership brings together experts to create technical standards that make technology products safer, better, and more interoperable. The ISO/IEC Joint Technical Committee 1 (JTC1) Subcommittee 42 (SC42) on Artificial Intelligence is principally responsible for developing and maintaining international AI technical standards.⁶

In December 2023, SC42 published ISO/IEC 42001:2023, titled “Information technology - Artificial intelligence - Management system”.⁷ This is the world’s first AI management system (AIMS) standard, providing a structured approach to managing AI-related risks and opportunities, balancing innovation with governance. It outlines requirements for establishing, implementing, maintaining, and continually improving an AIMS within an organization. Designed for those providing or using AI-based products or services, it ensures the responsible development and use of AI systems. The standard includes a set of controls for risk treatment based on a risk assessment, covering categories such as policy, resources, data, lifecycle, and use. SC42 is also developing additional standards to support AI assurance, including ISO/IEC 42005 on AI system impact assessment and ISO/IEC 42006 on requirements for bodies providing audit and

³ U.S Food and Drug Administration, National Evaluation System for Health Technology. <https://www.fda.gov/about-fda/cdrh-reports/national-evaluation-system-health-technology-nest>

⁴ International Organization for Standardization (ISO). <https://www.iso.org/>

⁵ International Electrotechnical Commission (IEC). <https://www.iec.ch/>

⁶ International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) Joint Technical Committee 1 (JTC1) Subcommittee 42 (SC42). Artificial Intelligence. <https://www.iso.org/committee/6794475.html>

⁷ International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2023). Information Technology - Artificial Intelligence - Management System (ISO/IEC 42001:2023). <https://www.iso.org/standard/42001/>

certification of AI management systems. Additionally, SC42 has a subgroup focused on testing AI-based systems, with related standards in development.

While the applicability of these emerging standards extends beyond the scope of this paper, which focuses on risk identification and characterization, they will become increasingly relevant as this body of work expands across the full AI implementation process.

A-3 Additional Tables and Figures

Table 2. Advisory Board Composition

Name	Primary Affiliation	Domain Expertise
Laura Freeman, PhD	Virginia Tech National Security Institute	AI Assurance National Security
Stevie Bergman	NIST Center for AI Standards and Innovation	Government Responsible AI
John Halamka, MD, MS	Mayo Clinic	Digital Health/AI Data Standards in Health Care
David McManus, MD	UMass Chan Medical School	Clinical AI Research
Stephen Sodeke, PhD	Tuskegee University National Center for Bioethics in Research and Health Care	Bioethics Biomedical Research

Table 3. AI Use Case Composition

Use Case	Lifecycle Stage	Area of Investigation
(1) An investigator team based in the NIH Intramural Research Program (internal research program of the NIH) leveraging <i>All of Us</i> dataset for diabetes detection model – how AI can be used to identify patterns of root cause of	Model Testing	Role of AI explainability methods for deriving factors contributing to chronic disease

Use Case	Lifecycle Stage	Area of Investigation
chronic diseases using AI explainability methods.		
(2) An extramurally funded clinical research team based at an academic institution developing coma prognosis tool -- using an AI-enabled User Interface (UI) to present key contributing factors to clinicians to support coma prognosis	Deployment	Human-AI interactions result in an interpretation of model predictions and contributing factors
(3) Large-scale platform using multimodal models for cancer registry -- using a largescale, multimodal Machine Learning (ML) to accelerate efficient and more accurate cancer registries	Operation	Post processing ML techniques for predictions informing cancer registry data

Table 4. Inventory of AI Initiatives and Frameworks

Entity	Initiative	Type	Initiative Details
The Assistant Secretary for Technology Policy (ASTP) / The Office of the National Coordinator for Health Information Technology (ONC)	AI in health IT	Government	Outlines a strategic framework for integrating AI into health IT systems, focusing on interoperability, data standards, and ethical considerations.
The Assistant Secretary for Technology Policy (ASTP) / The Office of the National Coordinator for Health Information Technology (ONC)	2020–2025 Federal Health IT Strategic Plan	Government	Includes initiatives related to the adoption of AI technologies in health IT, emphasizing data interoperability and patient privacy.

Entity	Initiative	Type	Initiative Details
National Institute of Standards and Technology (NIST)	NIST AI Risk Management Framework	Government	Provides guidelines for managing risks associated with AI technologies, focusing on trustworthiness, transparency, and accountability.
NIST	NIST Special Publication 1270: A Proposal for Identifying and Managing Bias in Artificial Intelligence	Government	Addresses challenges of bias in AI systems and offers strategies for identifying and mitigating bias, crucial for ethical AI applications in healthcare.
Food and Drug Administration (FDA)	Proposed Regulatory Framework for Modifications to AI/ML-Based Software as a Medical Device (SaMD)	Government	Outlines the FDA's approach to regulating AI/ML-based medical devices, focusing on safety, effectiveness, and iterative improvements.
FDA	Artificial Intelligence and Machine Learning in Software as a Medical Device: An FDA Discussion Paper	Government	Provides insights into the FDA's perspective on AI/ML in medical devices, highlighting the need for transparency, real-world performance monitoring, and regulatory oversight.
Office of Science and Technology Policy (OSTP)	The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update	Government	Outlines federal priorities for AI research and development, including initiatives to promote ethical AI use in healthcare and other sectors.
OSTP	Guidance for Regulation of Artificial Intelligence Applications	Government	Provides principles for the regulation of AI technologies, emphasizing a balanced approach that fosters innovation while protecting public welfare.
International Organization for Standardization (ISO)/ International Electrotechnical Commission (IEC)	ISO/IEC 22989:2022	Industry Standards	Provides a comprehensive framework for understanding AI concepts and terminology, applicable across various domains, including healthcare.

Entity	Initiative	Type	Initiative Details
ISO/IEC	ISO/IEC 23053:2022	Industry Standards	Outlines a framework for AI systems using ML, focusing on interoperability and reliability.
ISO/IEC	ISO/IEC TR 24027:2021	Industry Standards	Addresses the assessment and mitigation of bias in AI systems, crucial for ethical AI deployment in healthcare.
ISO/IEC	ISO/IEC TR 24028:2020	Industry Standards	Provides an overview of trustworthiness in AI, emphasizing transparency, accountability, and ethical considerations.
ISO/IEC	ISO/IEC 23894:2023	Industry Standards	Focuses on managing risks associated with AI systems, ensuring their safe and effective deployment.
IEC	IEC 62304:2006/Amd 1:2015	Industry Standards	Relevant for AI-integrated medical device software, focusing on safety and life cycle management.
IEC	IEC 81001-1:2021	Industry Standards	Outlines principles for ensuring the safety and security of health software, including AI applications.
Coalition for Health AI (CHAI)	Health AI Ethical Guidelines	Industry Standards	Developing guidelines that address ethical considerations in the deployment of AI in healthcare, focusing on data privacy, bias, and transparency.
CHAI	Collaborative Reports and White Papers	Industry Standards	Collaborates with academic and industry partners to produce reports highlighting best practices and ethical considerations for AI in healthcare.
Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)	TRIPOD	Academic Standards	Provides the minimum reporting recommendations for studies developing or evaluating the performance of a prediction model.

Entity	Initiative	Type	Initiative Details
TRIPOD	TRIPOD+AI	Academic Standards	Extends the TRIPOD guidelines to AI-based prediction models, providing a standardized framework for reporting and validating AI-driven diagnostic and prognostic tools in biomedical research.
TRIPOD	TRIPOD-LLM	Academic Standards	Extension of TRIPOD+AI statement, extended to large language models in biomedical applications.
Standards for Reporting of Diagnostic Accuracy (STARD)	STARD-AI	Academic Standards	Provides guidelines for studies evaluating the diagnostic accuracy of AI-based tests (in preparation).
CLAIM (Checklist for Artificial Intelligence in Medical Imaging)	CLAIM	Academic Standards	Offers reporting guidelines for medical imaging studies using AI.
Developmental and Exploratory Clinical Investigations of DEcision (DECIDE)	DECIDE-AI	Academic Standards	Provides a framework for early-stage clinical evaluation of AI-driven decision support systems, including safety and human factors evaluation.
Consolidated Health Economic Evaluation Reporting Standards (CHEERS)	CHEERS-AI	Academic Standards	Focuses on health economic evaluations to estimate the cost-effectiveness and value for money of AI interventions in healthcare.
Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)	SPIRIT-AI	Academic Standards	Provides protocols for clinical trials evaluating interventions with AI components.
Consolidated Standards of Reporting Trials (CONSORT)	CONSORT-AI	Academic Standards	Offers guidelines for reporting clinical trials evaluating interventions with AI components.

Entity	Initiative	Type	Initiative Details
Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)	PRISMA-AI	Academic Standards	Provides systematic review and meta-analysis guidelines for AI interventions (in preparation).
Scikit-learn	Scikit-learn Framework	Industry Framework	A widely used open-source ML library for Python, offering tools for predictive modeling, data preprocessing, and validation, applicable across domains, including healthcare.
Keras	Keras Framework	Industry Framework	A high-level neural network API that simplifies the development of deep learning models, widely used for healthcare applications such as medical imaging and disease prediction.
Apache MXNet	Apache MXNet Framework	Industry Framework	A scalable deep learning framework optimized for distributed computing, used for building AI models in healthcare and other industries.
FastAI	FastAI Framework	Industry Framework	A library built on PyTorch that simplifies the development of deep learning models, with applications in medical imaging and other healthcare domains.
H2O.ai	H2O.ai Framework	Industry Framework	An open-source platform for AI, offering tools for predictive analytics and model deployment.
PyTorch	PyTorch Framework	Industry Framework	An open-source ML library widely used for developing AI models across various applications, including healthcare.
TensorFlow	TensorFlow Framework	Industry Framework	An open-source ML library that supports the development of AI models, offering tools for healthcare research and applications.

Entity	Initiative	Type	Initiative Details
Medical Open Network for AI (MONAI)	MONAI Framework	Industry Framework	Provides tools optimized for medical imaging, supporting healthcare research with domain-specific capabilities.
Observational Health Data Sciences and Informatics (OHDSI)	OHDSI Framework	Industry Framework	Offers resources for analyzing large-scale health data, facilitating the integration of AI in epidemiological studies and clinical research.
MLflow	MLflow MLOps Framework	Industry Framework	Provides tools for tracking experiments and deploying models, ensuring reproducibility and scalability in AI model lifecycle management.
Kubeflow	Kubeflow MLOps Framework	Industry Framework	Facilitates the deployment of ML workflows on Kubernetes, supporting robust and scalable AI applications.
TensorFlow Extended (TFX)	TFX MLOps Framework	Industry Framework	Offers components for building production-ready ML pipelines, enhancing the efficiency and reliability of AI systems in healthcare.
Amazon SageMaker	Amazon SageMaker MLOps Platform	Industry Framework	Integrates MLOps capabilities with cloud services, streamlining the deployment and monitoring of AI models, incorporating DevOps and AutoML principles.
National Institutes of Health (NIH)	Bridge2AI	Government	NIH initiative focused on creating high-quality, ethically sourced datasets and tools to enable AI applications in biomedical and health research.
NIH	NIH Common Fund Data Ecosystem	Government	Provides infrastructure for managing large-scale biomedical datasets, enabling FAIR-aligned AI research and collaboration.

Entity	Initiative	Type	Initiative Details
NIH	AIM-AHEAD	Government	Supports researchers across the US in the development of AI/ML models and enhances the capabilities of AI/ML technologies.
World Health Organization (WHO)	WHO Guidance on Ethics and Governance of AI for Health	International Guidelines	Provides ethical principles and governance frameworks for the use of AI in healthcare, emphasizing equity, accountability, and transparency.
WHO	WHO Collaborating Centers for AI in Health	International Collaboration	Facilitates global partnerships to advance AI applications in public health and healthcare systems.
European Commission	Ethics Guidelines for Trustworthy AI	International Guidelines	Provides a framework for developing and deploying AI systems that are ethical, transparent, and trustworthy, applicable to healthcare and other domains.
Global Partnership on AI (GPAI)	Responsible AI Working Group	International Collaboration	Focuses on promoting responsible AI development and deployment, including applications in healthcare, through global collaboration and shared best practices.