National Institutes of Health
Office of Data Science Strategy
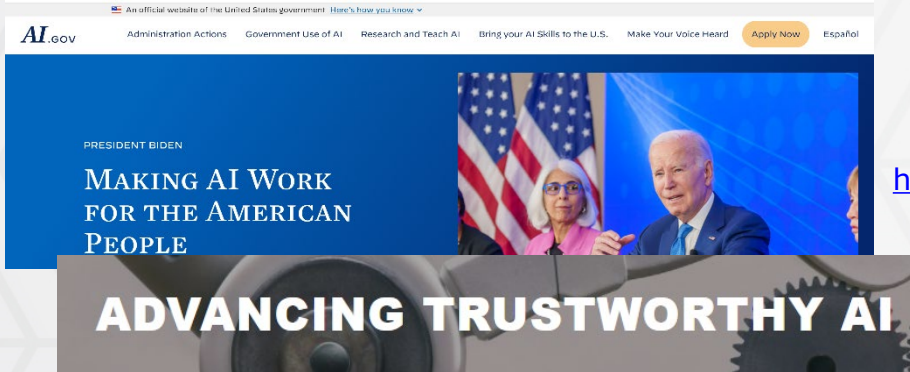
Creation of AI Working Group
Exploring NIH's Strategies, Challenges, and an AI Vision for the Future

Lawrence Tabak, DDS, PhD, *Principal Deputy Director, NIH*

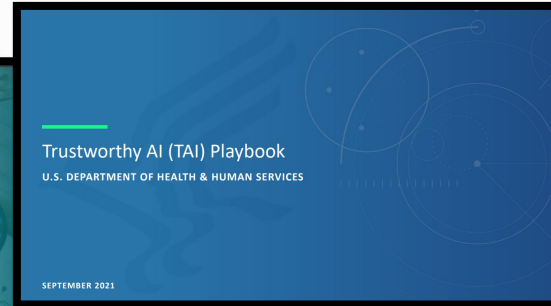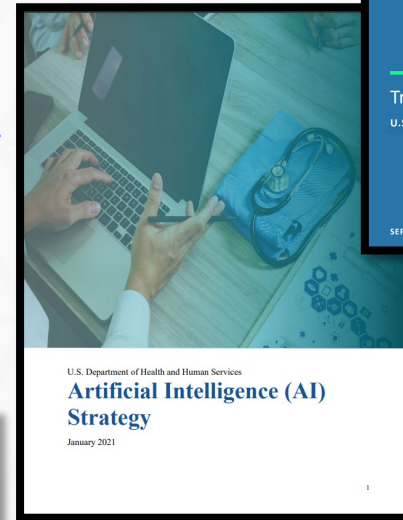Susan Gregurick, PhD, Associate Director for Data Science, NIH

June 14, 2024

Accelerating Trustworthy AI

https://ai.gov/

https://www.hhs.gov/sites/default/files/hhs-ai-strategy.pdf

https://web.archive.org/web/20230601171218/https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/

https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
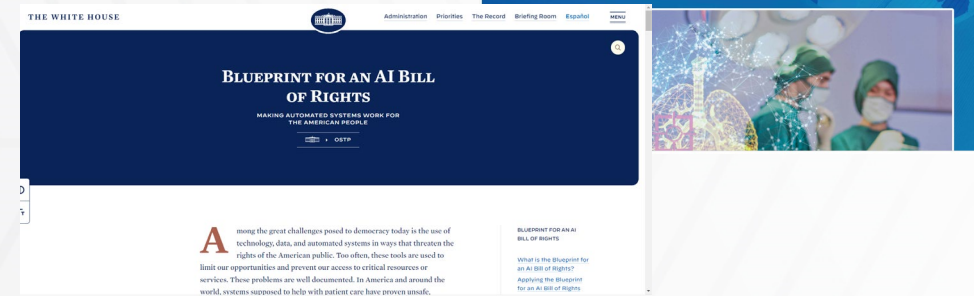
BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan

https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

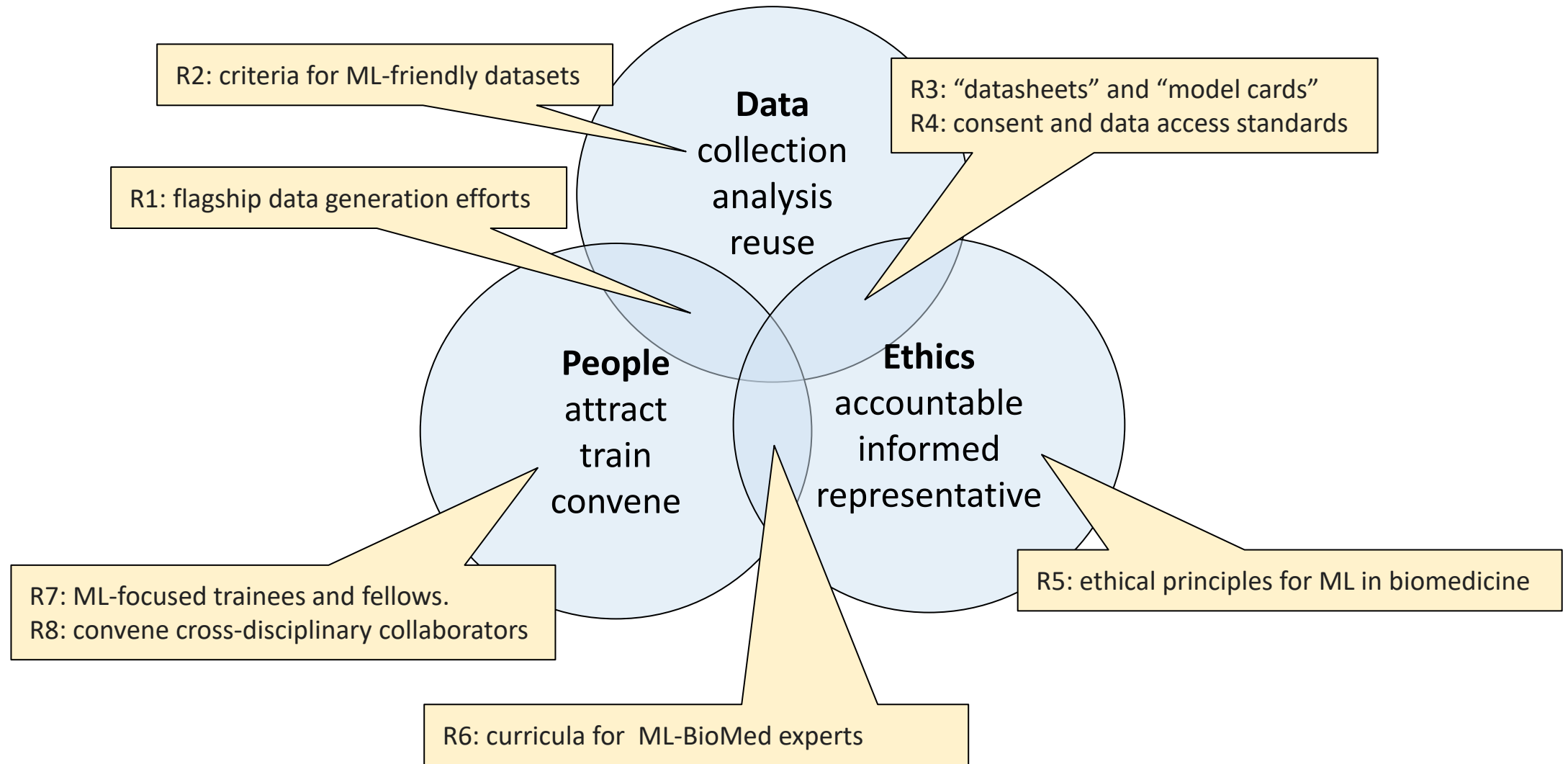https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

➢ **Investments in AI-related education, training, development, research, and capacity**

➢ Establish a program to identify and **attract top talent in AI**, and other critical and emerging technologies, at universities, research institutions, and the private sector overseas

➢ Launch a pilot program **implementing the National AI Research Resource (NAIRR)**

➢ Support **2024 Leading Edge Acceleration Project** cooperative agreement awards to improve healthcare-data quality, support the responsible development of AI tools for clinical care, real-world-evidence programs, population health, public health, and related research

➢ **Accelerate the National Institutes of Health Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD)** program
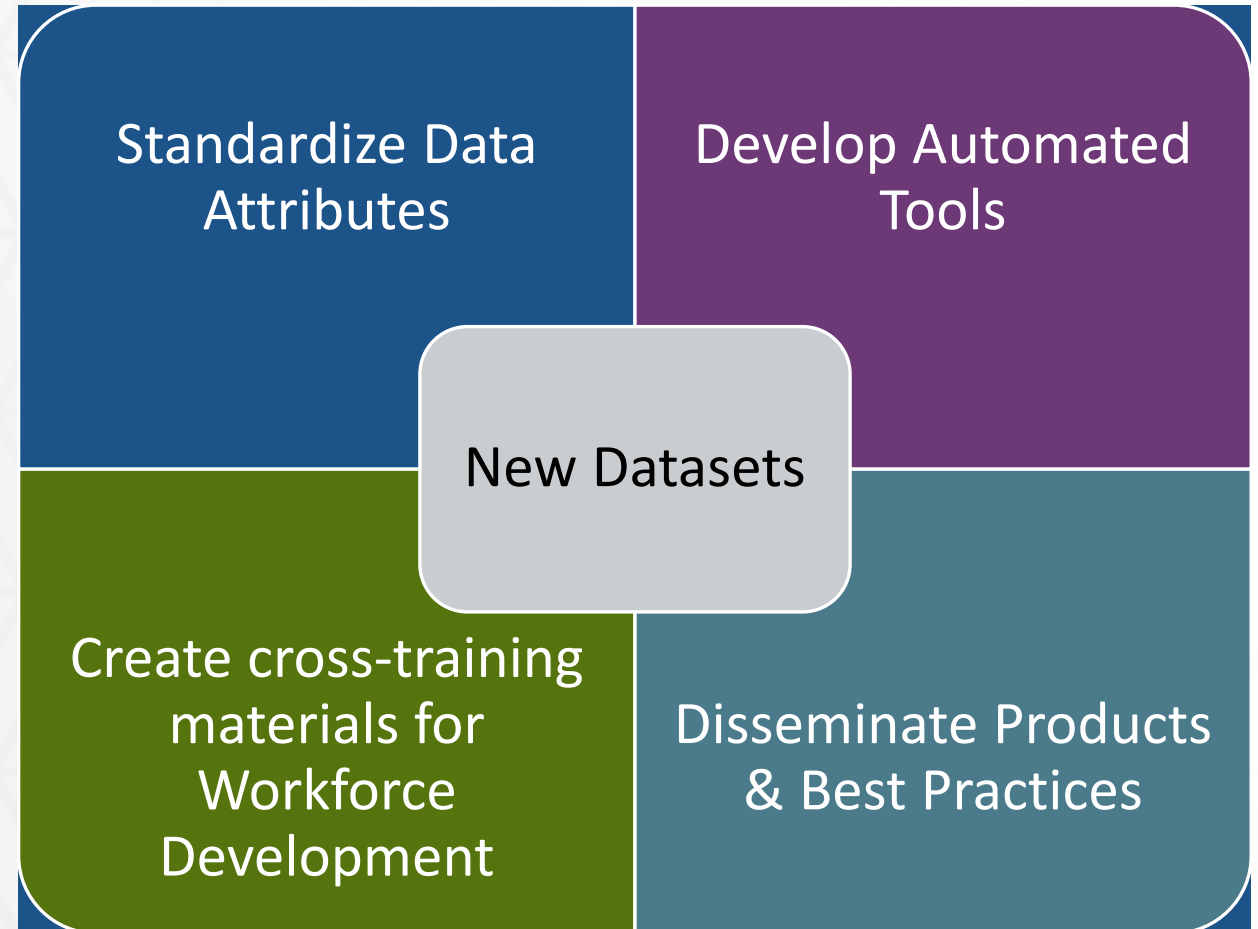
Recommendations from 2019 ACD AI Working Group

R2: criteria for ML-friendly datasets

R3: "datasheets" and "model cards"
R4: consent and data access standards

R1: flagship data generation efforts

Data
collection
analysis
reuse

People
attract
train
convene

Ethics
accountable
informed
representative

R7: ML-focused trainees and fellows.
R8: convene cross-disciplinary collaborators

R5: ethical principles for ML in biomedicine

R6: curricula for ML-BioMed experts

data	people	ELSI

Bridge2AI Program Goals

Determining how to:

- Use biomedical and behavioral research grand challenges to generate **flagship data sets**
- **Prepare** AI/ML-friendly data
- Emphasize **ethical** best practices
- Promote **diverse teams**

Standardize Data Attributes	Develop Automated Tools
Create cross-training materials for Workforce Development	Disseminate Products & Best Practices

New Datasets

Bridge2AI Milestones

The program established a public portal (https://bridge2ai.org/) for disseminating information regarding program activities and products.

Data	People	Ethics
• The four data generation projects (DGPs) have started to release data • External users worked with data in April 2024	• 5 internship programs to train researchers in both AI and biomedical research	• Paper submitted (under review) which discusses the ethics considerations as part of data and AI modeling lifecycle

About the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD)

Goals

Enhance the **participation** and **representation** of researchers and communities currently underrepresented in the development of AI

Address health disparities and inequities using AI/ML

Improve the capabilities of this emerging technology

https://aim-ahead.net/

AIM-AHEAD Impact | Year 1 & 2 Awards

AIM-AHEAD supported over **274 awards** to increase researcher diversity, address data & AI biases, engage underserved communities, and build institutional capacity

Training Programs
- Leadership Fellowship (50)
- Research Fellowship (47)

Community Engagement
- Hub Pilot Projects (35)

AI Health Equity Research
- Pilot Projects (21)
- Consortium Projects (21)

Institutional Capacity Building
- Program for Artificial Intelligence Readiness (15)
- Data and Infrastructure Capacity Building (13)

Joint training to increase researcher diversity in AI/ML by leveraging *All of Us* and N3C datasets, infrastructure, and training components.

212 Applications
25 trainees

National COVID Cohort Collaborative

120 Applications
50 trainees

NIH Office of Data Science Strategy

Impact   *as of June 5, 2024*

4,418+   Total Members

2,519   Mentees

1,170   Mentors

1,306   Institutions

AIM-AHEAD Named in White House Executive Order

Examples of AIM-AHEAD Supported AI Studies

AIM-AHEAD-supported studies have appeared in high-impact journals, including **Nature Communication, Scientific Report**, **Journal of Medical Internet Research AI**, **PLOS One**, **Journal of Clinical Oncology, Journal of Systemics, Cybernetics and Informatics, etc.**

Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach

Rachele Hendricks-Sturrup[1]; Malaika Simmons[1]; Shilo Anders[2]; Kammarauche Aneni[3]; Ellen Wright Clayton[2]; Joseph Coco[2]; Benjamin Collins[2]; Elizabeth Heitman[4]; Sajid Hussain[5]; Karuna Joshi[6]; Josh Lemieux[7]; Laurie Lovett Novak[2]; Daniel J Rubin[8]; Anil Shanker[9]; Talitha Washington[10]; Gabriella Waters[11]; Joyce Webb Harris[2]; Rui Yin[12]; Teresa Wagner[13]; Zhijun Yin[2]; Bradley Malin[2]

Teaching Health Informatics in Middle School: Experience from an NIH AIM-AHEAD pilot

Gregory TARDIEU
Alexandria City Public Schools
Alexandria, VA 22314, USA

Senait TEKLE
Biomedical Informatics Center, The George Washington University
Washington, D.C. 20037, USA

Linda ZANIN
Alexandria City Public Schools
Alexandria, VA 22314, USA

Derivation and Validation of a Clinical Risk Assessment Model for Cancer-Associated Thrombosis in Two Unique US Health Care Systems

Authors: Ang Li, MD, MS ✉; Jennifer La, PhD; Sarah B. May, MS; Danielle Guffey, MS, Wilson L. da Costa Jr, PhD; Christopher I. Amos, PhD, Raka Bandyo, MS, ... SHOW ALL ..., and Nathanael R. Fillmore, PhD | AUTHORS INFO & AFFILIATIONS

NIH wide AI Collaborative Programs

Ethics, Bias, and Transparency for People and Machines

Improving the AI-readiness of Existing, IC-supported Data

Addressing the Workforce Gap in Data Governance for AI in Biomedicine

Impact 2021-2023

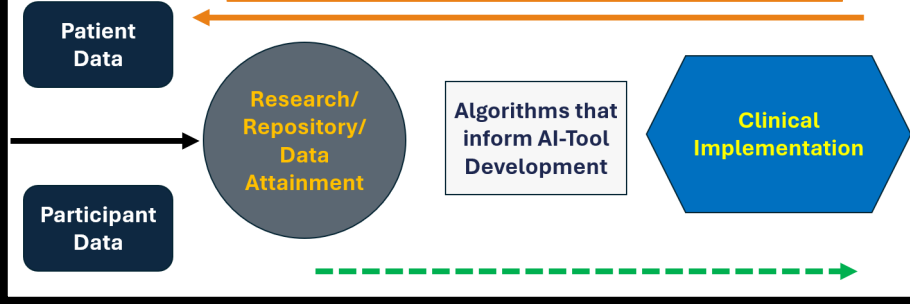|	
$54M	In funds to investigators
152	Investigators
14	Idea State Institutions
4	Workshops, Labs, PI meetings

NIH wide AI Collaborative Impacts

Potential harms & unintended consequences flow back to individuals and subgroups (discriminatory interpretations, inequity in access, confidentiality/privacy breach)

Patient Data

Participant Data

Research/ Repository/ Data Attainment

Algorithms that inform AI-Tool Development

Clinical Implementation

PI: Alex Federman, Icahn School of Medicine at Mount Sinai
TITLE (IC): Natural Language Processing and Automated Speech Recognition to Identify Older Adults with Cognitive Impairment Supplement (NIA)
HIGHLIGHTS: Qualitative Examination of Patients' and Clinicians' perspectives on AI-driven Automated Screening for Cognitive Impairment

PI: GILMORE, JOHN HORACE, UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
TITLE (IC): Rescuing Missed Longitudinal MRI Scans in the UNC Early Brain Development Study (EBDS)
HIGHLIGHTS: Developed AI methods to fill in missing brain images in cohort studies of early brain development, thereby making a complete AI-ready dataset for further analyses.
RESEARCH PRODUCTS: *Code and documentation:*
https://github.com/yoonmihong/DeepImputation

FAIR+AIML+CAPAC
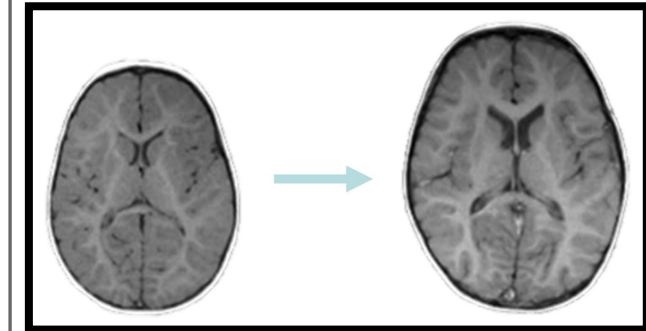
PI: Ana Patricia Ortiz, University of Puerto Rico Comprehensive Cancer Center
TITLE (IC): Preparing a workforce to apply AI/ML techniques to datasets derived from Hispanic populations to advance cancer prevention and control research (NCI)
HIGHLIGHTS: Training focuses on techniques to manipulate/pre-process cancer datasets from Hispanic populations and make them FAIR & AI-Ready and methods to create predictive models for cancer diagnosis with focus on datasets from Hispanic populations.

Multimodal AI

Embed in context of ethical, trustworthy AI practices and assessment

Expected outputs:

- New systems-level biomedical research using multimodal AI technologies
- Elucidation of the unique opportunities, risks, and challenges for applying multi-modal AI methods
- Identification of considerations for the appropriate use of multimodal AI, relative to other methodologies

Example of NIDA-sponsored AI studies in human neuroscience

Brain activity — AI-Based Model — Health outcome

Yan W et al., Biol Psychiatry. 2023

- AI can be used to predict health-related variables based on brain activity patterns

- To generate reproducible markers, extremely large samples may be needed

- We can use cutting-edge methods from AI, to translate AI-based models from big datasets to independent smaller datasets

(He et al, 2022)

Schizophrenia   ASD
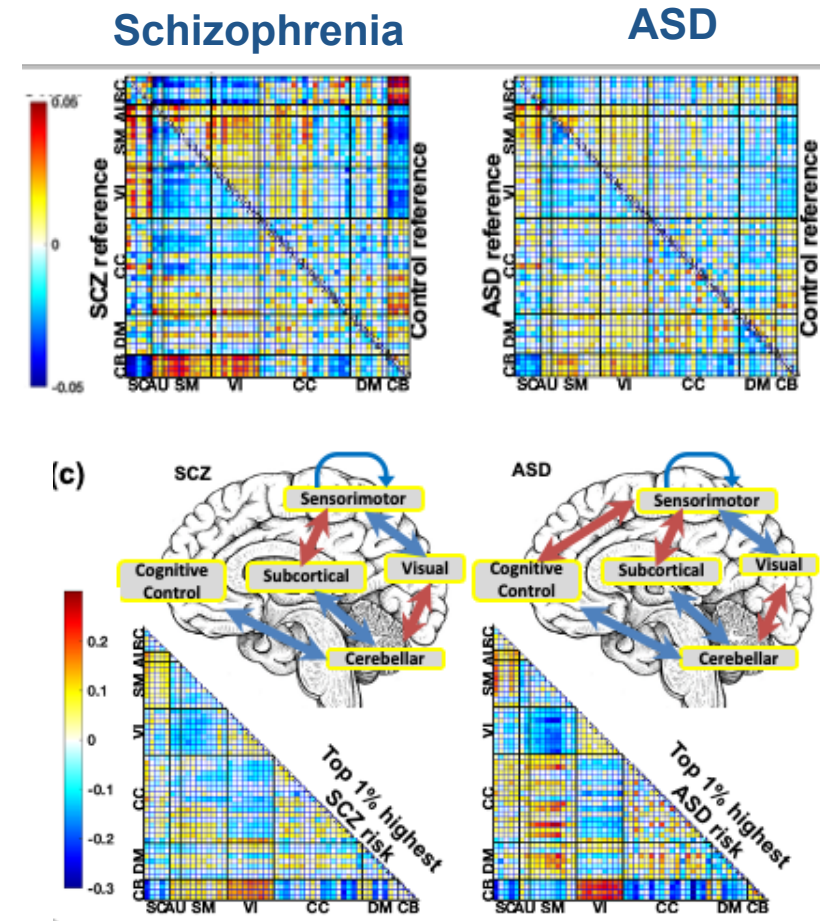
Agency Collaborations

- [Machine Learning and Artificial Intelligence NSTC Subcommittee](#)

- [National AI Research Resource (NAIRR)](#)

- [Health and Human Services AI Task Force](#)

- [NCI-DOE Collaboration for Advanced AI to end Cancer](#)

Impacts of NCI-DOE collaboration:

- **30 AI/ML publicly-available resources**

- **134 publications** since 2016

- **50+ public and private organizations** participating in innovation challenges

DBliss/NIH

Agency Collaborations: National AI Research Resource (NAIRR)

National AI Research Resource: a shared research infrastructure facilitating access to compute, software, datasets, models, training and user support for researchers and students

Objective: To strengthen and democratize the U.S. AI Innovation ecosystem in a way that protects privacy, civil rights, and civil liberties

Goals:

Spur **innovation**

Increase the **diversity** of talent in AI

Improve U.S. **capacity** for AI R&D

Advance **trustworthy AI**

NSF | NIH | DOE | NASA | NOAA

NIH Contributions to the NAIRR Pilot

DATA: NIH contributes large amounts of curated, interoperable, **AI-Ready Data**

SECURITY: NIH brings different data together in secure, privacy-preserving ways to **protect patient safety in AI**

TRUSTWORTHY AND ETHICAL AI: NIH adopts ethical practices and addresses bias, diversity, and transparency to **build trustworthy AI**

BROADENING PARTICIPATION: NIH cloud-based platforms and programs **lower the barriers to data science and grow capacity** in under-served and -resourced communities

NIH Contributions to NAIRR Pilot

Governance	• Experience developing and overseeing federated interoperability
NAIRR Open	• Integration of *ImmPort* datasets into the NAIRR • Integration of Health Equity Action Network (HEAN) datasets and ScHARE analysis tools into the NAIRR
NAIRR Secure*	• Integration of the Medical Imaging and Data Resource Center (MIDRC) and National COVID Cohort Collaborative (N3C) into NAIRR Secure
Software Stack	• Coordinate with NSF and DOE a NAIRR software stack community workshop
Classroom	• NIH Cloudlab and other platform tools leveraged in NAIRR
Outreach	• Leverage NIH networks to attract diverse users and data

* NIH and DOE jointly lead NAIRR Secure

NIH Data and Computational Infrastructure Ecosystem

NIH | *STRIDES Initiative*
STRIDES / CloudLab

IMMPORT Private Data
Your site for managing ImmPort data uploads

ScHARe

N3C: Translating Health Data Into Health Solutions — National Center for Advancing Translational Sciences

MIDRC
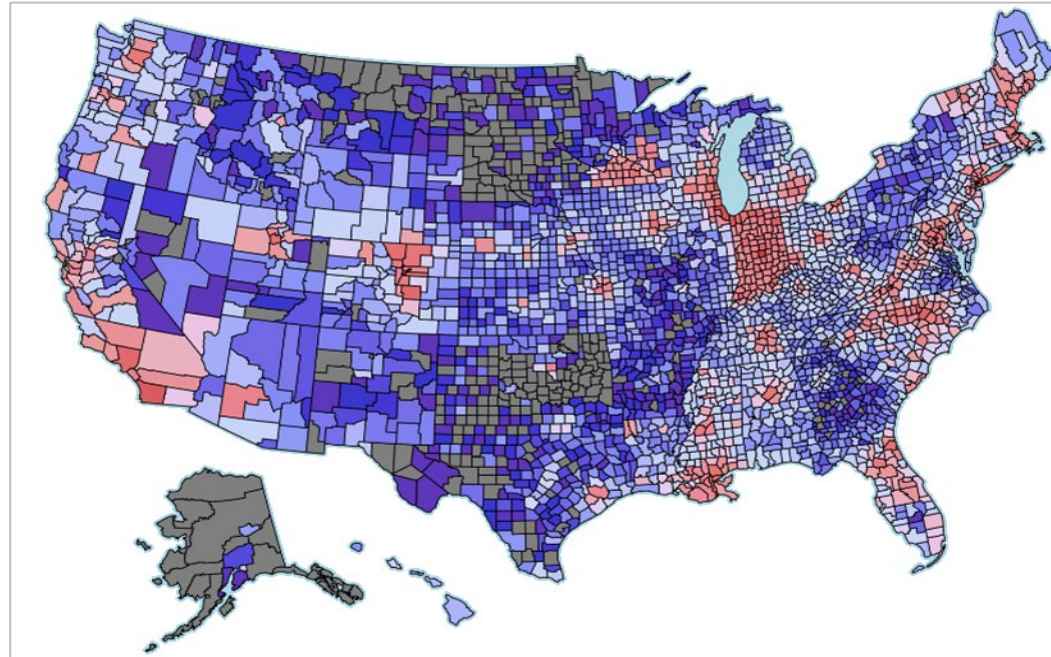MEDICAL IMAGING AND DATA RESOURCE CENTER.

AI Ready Datasets for COVID and Clinical Modeling
National COVID Cohort Collaborative (N3C)

Data Diversity, Representation, and Harmonization are foundations for robust, trustworthy AI

Geographics: 50/50 States >92% of all US Counties in USA
Representative of US population
Source: Community, Academic, FQHCs
Patient Mix: Inpatient ~20%, Outpatient ED ~80%
Longitudinal Data: 1/1/2018 to Present

https://covid.cd2h.org/

COVID+ CASES	Patients
8.9m	22.5m
Rows of Data	Contributors Health Systems
32.7b	84
Institutions Using Data	Active Investigators
391	>3900
Research Studies	Citations & H-Index.
553	3890/29

National COVID Cohort Collaborative

NIH National Center for Advancing Translational Sciences
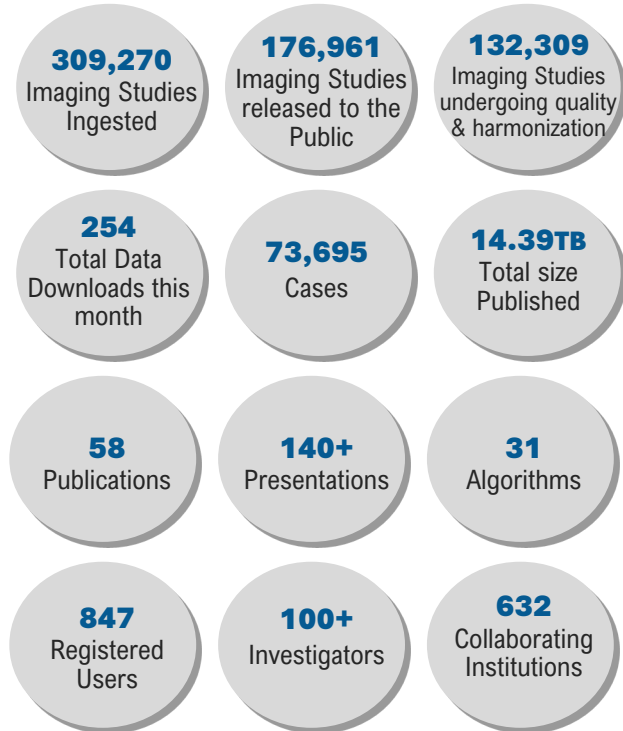
![Database icon] **AI Ready Datasets for Medical Imaging AI Applications**
Medical Imaging and Data Resource Center (MIDRC)
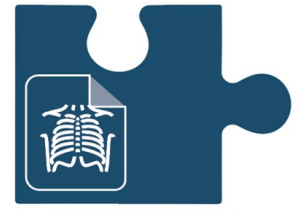
MIDRC
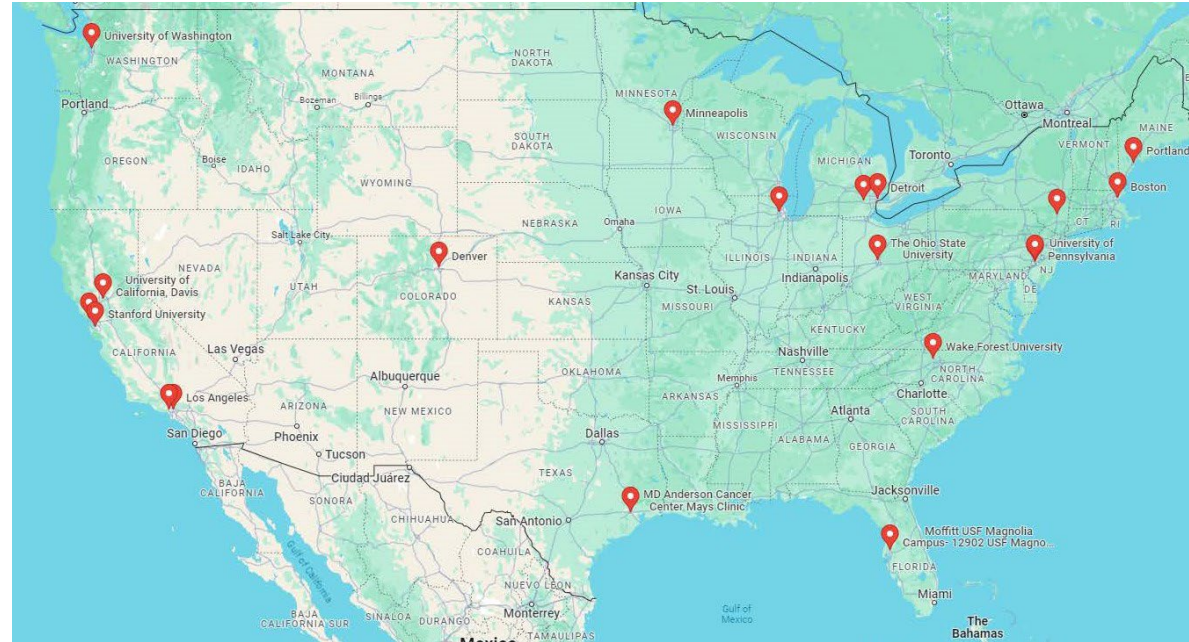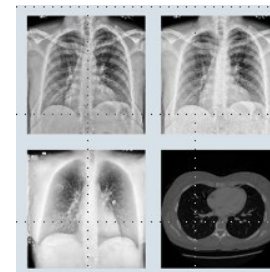### BY THE NUMBERS

Information dense medical imaging data has many applications in AI algorithm development

* Map of MIDRC Investigators
A collaboration of ~20 institutions &
> 100 investigators from academia, community practices, FDA, and others

https://www.midrc.org/

309,270 Imaging Studies Ingested

176,961 Imaging Studies released to the Public

132,309 Imaging Studies undergoing quality & harmonization

254 Total Data Downloads this month

73,695 Cases

14.39TB Total size Published

58 Publications

140+ Presentations

31 Algorithms

847 Registered Users

100+ Investigators

632 Collaborating Institutions

MIDRC NIH National Institute of Biomedical Imaging and Bioengineering
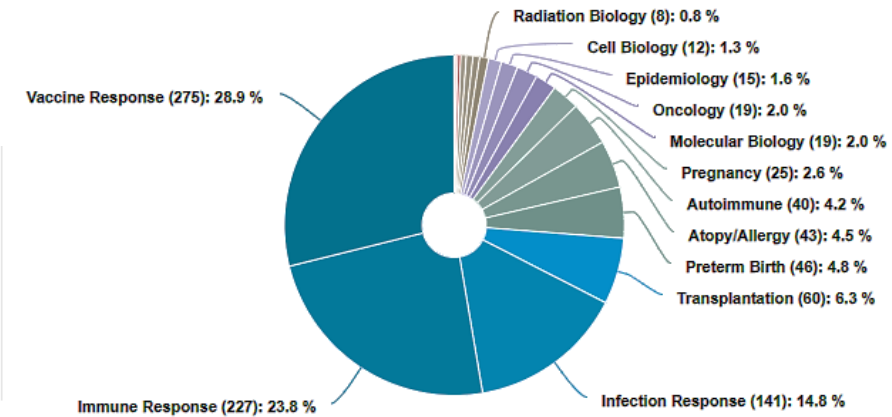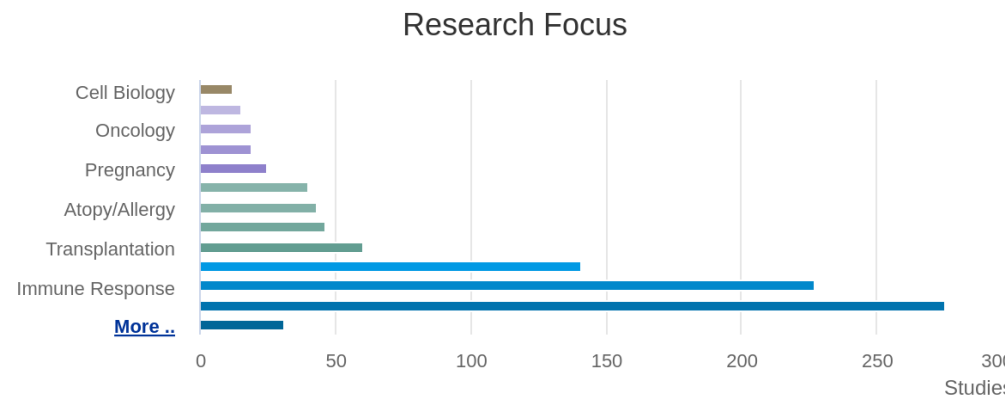MEDICAL IMAGING AND DATA RESOURCE CENTER.

AI Ready Datasets for Immunology Research
ImmPort

Immunology data supports AI development and data reuse in a broad range of applications across test and study types

Studies	**Subjects**	**Diseases**
952	101125	160
Experiments	**Total Results**	**Lab Tests**
3669	7037834	1286834

Research Focus

Radiation Biology (8): 0.8 %
Cell Biology (12): 1.3 %
Epidemiology (15): 1.6 %
Oncology (19): 2.0 %
Molecular Biology (19): 2.0 %
Pregnancy (25): 2.6 %
Autoimmune (40): 4.2 %
Atopy/Allergy (43): 4.5 %
Preterm Birth (46): 4.8 %
Transplantation (60): 6.3 %
Infection Response (141): 14.8 %
Immune Response (227): 23.8 %
Vaccine Response (275): 28.9 %

https://immport.org/shared/home

AI Ready Datasets for Social Determinants of Health (SDOH)
ScHARe Health Equity Action Network (HEAN)
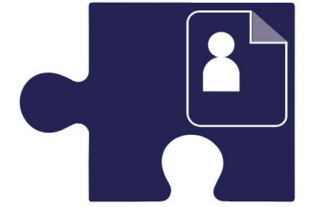
Social Determinants of Health brings the lived experience to AI algorithms and meets communities where they are

- ScHARe Cloud Platform host over 245 national federated data sets focused on population science, social determinants of health (SDoH), environmental and behavioral data

- ScHARe data repository will be available in Sept 2024 to host NIH funded project's data targeting social science and behavioral research, especially in areas of health disparities, healthcare delivery and health outcomes data

- HEAN (Health Equity Action Network)

  o Focus on chronic disease and relationship of the SDoH - 7 of 29 initial R01-level projects have transferred data to the RCC for a total of 639 records

  o Aligned data with ScHARe CDEs to enhance data integration and interoperability

https://www.nimhd.nih.gov/resources/schare/

National Institute on Minority Health and Health Disparities

AI at NIH

- AI and machine learning methods are not new to NIH

- We lack data resources at a scale necessary to optimally use AI/ML methods to improve health
  - Data relevant to decision making for individual patients, in real time
  - Data to drive innovation in technology development, new therapeutics and prevention methods, and health care delivery
  - Data to eliminate disparities: rural, elderly/adolescents, designated populations with health disparities

- We must respectfully engage people to earn trust

- Diversity is critical to avoid harm to some populations

- Current needs:
  - Dramatic increase in data collection from the clinical care environment, including all populations
  - NIH resource to facilitate optimal and especially ethical use of AI for health

Charge to the AI Working Group

*The ACD AI WG is charged with articulating a **strategic and integrated vision** for biomedical research opportunities that would benefit from developing and application of novel AI methods. In considering this charge, the AI WG's recommendations should address the following:*

- Assess progress to date and develop a framework to support strategic priorities and biomedical research opportunities in AI, particularly involving the development and application of novel methodologies (i.e. foundational models, multimodal generative AI, Edge AI, etc) for knowledge discovery and human health. This should include the necessary data and computing resources that will be required for using and scaling AI in biomedicine, allowing for interdisciplinary collaboration across fields.

- With respect to these priorities, define the potential privacy, security, ethical, policy, and cost challenges that NIH should consider in supporting and deploying AI to maximally benefit the biomedical enterprise. Consider potential approaches for mitigating these challenges, including new areas of science that could be developed.

- Recommend strategies for ensuring equitable benefits result from these strategic priorities, including equitable benefits in inclusive algorithmic development, the application of transparent and explainable AI, and collaborative training programs to enable a health learning environment using AI and AI-enabled tools.