

Data Sharing in Digital Age: Harnessing the Power of AI for Cancer Research

*Emily Boja, Ph.D.
Branch Chief
Office of Data Sharing
National Cancer Institute*

Sharing the Right Data for the Right Model

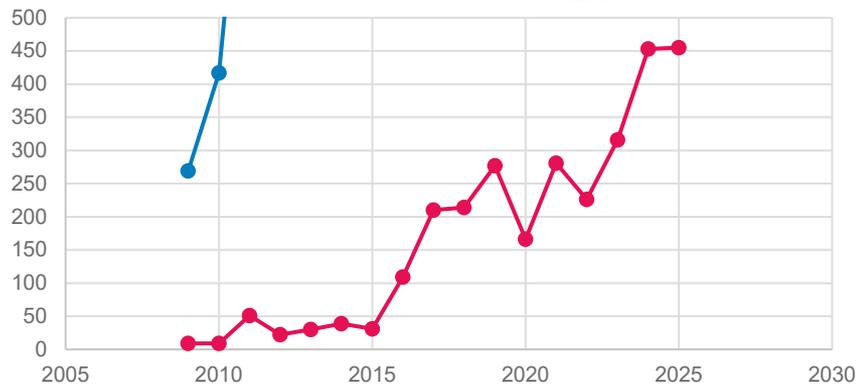
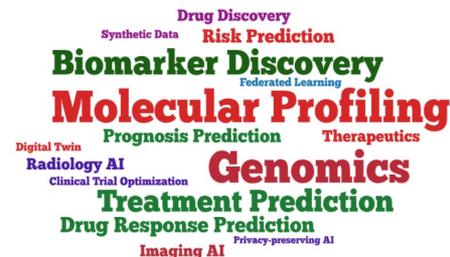


AI in Research: Policy Considerations and Guidance (NIH Office of Science Policy)



How to Request Controlled-Access Data

GenAI Models Trained on Controlled-Access Genomic Data (pre-decisional)



● Total #DARs Approved ● Total #AI/ML DARs Approved

NIH Robust Process to Protect Controlled Access Data

Protecting Human Genomic Data when Developing Generative Artificial Intelligence Tools and Applications

Notice Number:
NOT-OD-25-081

Key Dates



GenAI models including model parameters, SNPs & imputed data trained on controlled access data from NIH-designated repositories constitute Data Derivatives (DUC).

- Approved users may **not** share the model, including model parameters, except with collaborators who are also Approved Users.
- May **not** retain GenAI models & model parameters upon closeout of the project.
- Sharing controlled-access data (GDS Policy) with public generative AI tools via prompts or other user interfaces is **not** allowed.

Enhancing Data Utility for AI Applications in Oncology

CCDI
BUILDING A COMMUNITY CENTERED AROUND CHILDHOOD CANCER CARE AND RESEARCH DATA

DATA TYPES:

- CLINICAL
- TREATMENT
- OUTCOME
- MOLECULAR
- BIOSPECIMEN
- LONGITUDINAL
- POPULATION

LEARN FROM EVERY CHILD

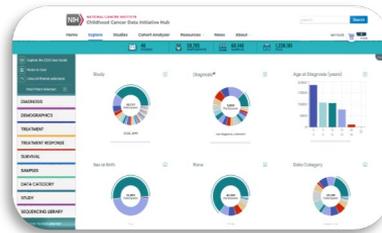
Improving the quality, consistency, and accessibility of data to make it easier for researchers to develop new and better treatments for children with cancer.

THE WHITE HOUSE

UNLOCKING CURES FOR PEDIATRIC CANCER WITH ARTIFICIAL INTELLIGENCE

EXECUTIVE ORDERS SEPTEMBER 30, 2025

ARPA-H
Biomedical Data Fabric Toolbox



Enhancing Childhood Cancer Data Sharing and Utility

NCI Office of Data Sharing proudly presents its inaugural data jamboree

September 29 - 30, 2025

United States and European Union Artificial Intelligence Administrative Arrangement

ITCR
Informatics Technology for Cancer Research

Cancer

The **USCDI+ Cancer** Program will define real-world data (RWD) elements to further cancer prevention, diagnosis, treatment, research, and care. Enhanced data exchange for research purposes and clinical care contributes to the U.S. government's support of persons with cancer. This is an area of mutual interest and responsibility for agencies across HHS. This work is collaboratively managed by NCI and ONC with input from CMS, CDC and FDA and includes focus on the following use cases:

- Clinical Trials Matching (CTM)
- Immune Related Adverse Event (iRAE) tracking in Immunotherapy trials
- Enhance the efficiency and timeliness of collection of cancer registry data
- Enhancing Oncology Model (EOM) alignment

The aim of USCDI+ Cancer is to improve the underlying data quality issues and improve the reproducibility of methods.

AI

NCI-DOE Collaboration

- Foster a growing predictive oncology community
- Provide open access to FAIR AI/ML resources, including datasets and computational models

NCI-DOE Projects

MOSSAIC
Modeling Outcomes Using Surveillance Data and Scalable Artificial Intelligence for Cancer

ADMIRRAL
AI-Driven Multiscale Investigation of the RAS/RAF Activation Lifecycle

IMPROVE
Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

ATOM
Accelerating Therapeutics for Opportunities in Medicine

↑ ↑ ↑

Infrastructure

CANDLE
CAnCER Distributed Learning Environment

Computational Resources for Cancer Research

MoDaC
NCI Predictive Oncology Model and Data Clearinghouse

NIH NATIONAL CANCER INSTITUTE

How to Unlock AI Potential for Effective & Ethical Use?



Engage with us:

<https://datasharing.cancer.gov/>

nciofficeofdatasharing@mail.nih.gov

- Other data modalities (e.g., imaging, EHR)?
- Risk assessment of multi-modal data?
- Benchmarks and reference datasets?
- Scalable governance framework for data access and downstream research (e.g., federated learning, data linkage)?
- AI-specific attributes to consider:
 - Sample size bias
 - Label noise
 - Performance drift monitoring
- Develop strategies for AI/ML-related training and workforce development?
- Cancer research specific applications?
- How to align AI application types with sensitive data: Biomedical Impact vs. Risks?

Thank You



**NATIONAL
CANCER
INSTITUTE**

[cancer.gov](https://www.cancer.gov)

[cancer.gov/espanol](https://www.cancer.gov/espanol)