The NCI Office of Data Sharing presents the Annual Data Sharing Symposium:

Driving Cancer Advances through Impactful Research

October 15 – October 16, 2024

Meeting Purpose and Goals

The National Cancer Institute (NCI) Office of Data Sharing (ODS) Annual Data Sharing Symposium convened the cancer community stakeholders to work together on critical data sharing efforts; discuss respective successes and challenges; identify knowledge gaps and evolving needs across the field; and explore solutions that align community efforts in ways that benefit the majority. High-level goals for the 2024 symposium included:

- Discuss best practices and key challenges of data sharing in a learning health care system, focusing on technology, legal considerations, and collaborative strategies to improve patient care.
- Showcase exemplary research performed using shared data and highlight features that facilitated its reuse, focusing on key exemplars over the past decade (e.g., Kids First Data Resource Center, Cancer Research Data Commons).
- Outline leadership views on NCI data strategy, how technology can be leveraged to empower discovery and innovation that drives understanding of disease biology and optimal therapeutic strategies, and the overall impact on the cancer research community.
- Highlight how artificial intelligence (AI) is being implemented to support data sharing efforts and how policy guides the development and use of this technology.
- Share perspectives on challenges and success and tackle big questions in data sharing so that ODS can grow and improve with the cancer community.

Key themes of presentations and discussions reflected session topics, including honoring contributions of research participants, developing a learning health system for cancer, highlighting impact of data sharing and reuse, impact of AI on data use and sharing, lessons learned from building cancer data ecosystems, and exploring infrastructure and resources that support data sharing and reuse.

Day 1: Tuesday October 15

Welcome

ODS Director Dr. Jaime Guidry Auvil welcomed participants, reviewed the symposium agenda, provided a brief history of the first ODS symposium, and described key ODS achievements for the past year. These accomplishments included expanding the Index of NCI Studies, ongoing activities of the Childhood Cancer Data Initiative, and engaging with key NCI programs through workshops focused on clinical and phenotypic data types. She described the National Data Ecosystem for Cancer that integrates cancer research data—from basic research to clinical care to population level studies—in a meaningful way to enhance understanding of disease, guide optimal therapeutic strategies, and evaluate how those strategies evolve over time in individuals and across cohorts.

Keynote Address: ODS & Kids First—10 Years In

Dr. Adam Resnick, Director, Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, highlighted the work of the Gabriella Miller Kids First Pediatric Research Program (Kids First), which focuses on pediatric cancer and congenital birth defects. Emphasizing the value of partnership and collaboration in achieving success in data sharing, he described the role of the ODS team in supporting the program, which succeeded in releasing data within 6 months of generation. Dr. Resnick noted the importance of interoperability and data harmonization in meeting the challenges of integrating data from multiple sources to enable comprehensive analysis and discovery; for example, integration of imaging data into the Kids First portal allows researchers to combine imaging and genomic data. Dr. Resnick outlined the development of the participant index to link clinical trials, research data, and real-world data (RWD), with its potential to inform clinical trial design and improve patient outcomes. He described the innovative Real-time Analysis and Discovery in Integrated And Networked Technologies (RADIANT) Project, funded by an award from the Advanced Research Project Agency for Health (ARPA-H) and aims to leverage RWD to drive personalized care and improve patient outcomes.

Data Sharing in a Learning Health System

An expert panel discussed best practices and key challenges of data sharing in a learning health system, with a focus on technology, legal considerations, and collaborative strategies to improve patient care. Dr. Catherine Lerro, U.S. Food & Drug Administration (FDA), presented on the role of FDA in RWD and real-world evidence, highlighting the importance of ensuring data quality and relevance to support regulatory decision-making and describing collaborative research efforts of the Alliance for Clinical Trials in Oncology and the University of Maryland to enhance oncology source data. She emphasized the need for IT support and clinical alignment to implement structured data collection methods in electronic health records (EHRs). Ms. Reda Wilson, Centers for Disease Control and Prevention (CDC), described the agency's role in cancer surveillance and data sharing as well as providing support for cancer registries and technical assistance to stakeholders. She outlined challenges of data collection and integration and the need for collaboration and standardization to improve data quality and support public health decisions. Dr. Patricia Bright, FDA, presented background on the FDA Sentinel System, a medical product safety surveillance system designed for postmarket active risk identification and analysis that incorporates healthcare data from a distributed network of public and private data partners. Sentinel standardizes RWD from disparate sources to make them query ready and generates RW evidence about safety and effectiveness of medical products using data produced by health systems. Health systems apply RW evidence to improve healthcare delivery.

Data Sharing and Reuse: Exemplars and Challenges

Well structured, accessible data not only enhances rigor and reproducibility of research but also plays a vital role in secondary research. Reusing shared data saves time and resources, leads to increased productivity, minimizes duplication, and accelerates innovation. Four presenters showcased exemplary research performed using shared data and highlighted features that facilitated its reuse. Dr. Matt Wyczalkowski, Washington University, presented data sharing and reuse within the Clinical Proteomic Tumor Analysis Consortium (CPTAC3). These genomic data are derived from over 1,500 cancer patients across 10 cancer types and include the following data types: matched tumor and normal whole genome, whole exome, bulk RNA-Seq, and methylation. Examples of data sharing included CPTAC3 genomic data processing and pathogenic germline variant discovery. Dr. Katie Campbell, Broad Institute of MIT and Harvard, described reuse of integrated deep phenotype and screening data to link cell line models to patient tumors. She noted that large resources efforts to harmonize data generation and data provenance optimize data reuse for analyses well beyond the vision of a single project. Recognizing that harmonizing bioinformatic pipeline processing is not always feasible, she advised participants not to let “the perfect be the enemy of the good” (enough). Dr. Larry Kushi, Kaiser Permanente, presented data sharing and reuse in epidemiologic studies conducted in the Kaiser healthcare delivery system. He shared an example of a “waiver of consent” disparities study in ovaria cancer treatment and survival and a prospective cohort study “with consent” of women with breast cancer. Dr. Hai Hu, Chan Soon-Shiong Institute of Molecular Medicine, presented three examples of using public clinicopathologic and molecular data in research. He described development of the TCGA Clinical Data Resource, a cancer outcome comparison between military health system beneficiaries and the general population, and a proteogenomic analysis of breast cancer in young women.

Breakout Session Report Outs: Equity, Access, Infrastructure

During the afternoon, participants had the opportunity to engage with ODS staff and other attendees in small group discussions around critical topics in data sharing. Symposium participants attended breakout sessions in person or virtually to discuss how the culture of biomedical data sharing can shift to improve secondary use through the perspective of equity, access, or infrastructure.

Equity. Participants in this session noted the multiple aspects of equity relevant to data sharing and analysis—equitable access, representation of study populations, and capability to use data. They identified barriers, gaps, and opportunities for improvement.

Access. Participants in this breakout session discussed what is working well, barriers, gaps, and opportunities for improvement. Hurdles include lengthy time from requesting data access to obtaining access; inability to confirm that the data needed is included in the dataset until after downloading it; and unclear guidance on some forms. Access could be improved through streamlined application processes, centralized study registration, and making summary-level information available outside of the firewall.

Infrastructure. Participants in this session highlighted the value of existing collaborations that help share data and discussed infrastructure hurdles such as lack of funding for data curation or deidentification and challenges related to tracking data use. Available infrastructure might be improved by cataloging datasets to facilitate findability; broadening infrastructure to support sharing of multiple data types, and defining data retention policies.

Day 2: Wednesday October 16, 2024

Artificial Intelligence in Data Use and Sharing

Five panelists highlighted policy considerations for AI development and use in research as well as AI-related activities, resources, and initiatives at FDA and NIH. Topics included policy considerations, safeguarding data and privacy, mitigating biases in biomedical AI, and the future of AI in data sharing. Dr. Juli Klemm, NCI Center for Strategic Scientific Initiatives, emphasized the need for coordination of AI in cancer research across NCI and highlighted the trans-NCI AI working group, web pages that communicate about AI opportunities, collaborations with the Department of Energy, and the Cancer AI Conversations webinar series. Dr. Ellen Wann, NIH Division of Scientific Data Sharing Policy, emphasized the importance of data management and sharing policies in the context of AI and complexities around sharing AI models, and she described the NIH centralized resource for guidance on safe and responsible use of AI in research. Dr. Ravi Samala, FDA Center for Devices and Radiological Health (CDRH), discussed the increasing amount of FDA-authorized AI software as medical devices and related regulatory challenges; he noted that the CDRH tools catalog for public use includes AI tools. Dr. Sean Mooney, NIH Center for Information Technology, described the NCI SRIDES cloud program and its potential for integration and harmonization of data; he emphasized the need for scalable, easy-to-use, standardized technology for AI. Dr. Veerasamy “Ravi” Ravichandran, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), described a pilot study that uses secure large language models (LLM) for review and evaluation of NIH required Data Management and Sharing (DMS) plans; currently, program officers review these plans manually. LLM generates a DMS plan derived from the research strategy in the grant application and compares it with the DMS submitted with the grant application.

Fireside Chat: Leadership Views on NCI Data Strategy and Its Impact on the Cancer Research Community

Dr. Warren Kibbe, NCI Deputy Director for Data Science and Strategy, and Ms. Amanda Haddock, President, Dragon Master Initiative, shared their personal connections to cancer. Dr. Kibbe emphasized the importance of community and the goal of making data available for research uses in real time. He described the impact of TCGA and the NCI RAS Initiative. He highlighted the importance of making data accessible, intuitive, and simple to use. Dr. Haddock discussed the value of connecting data across different lines of research and the impact of shared data on research.

A Message from NCI Director Dr. Kimryn Rathmell

Dr. Rathmell thanked symposium participants for their energy and passion for cancer data access and sharing. Their dedication has led to incredible advances in technology and accelerated the pace of discovery and innovation in cancer research. She described the wealth of available RWD and the opportunities—particularly through emerging AI solutions, LLMs, and informatics—for wholesale changes in matching patients to clinical studies, assessing side effects, and monitoring disease progression. She said it is imperative that we make data collection as simple as possible, make data interoperable, and ensure that everyone has the chance to contribute their data. The right group of people have gathered at this symposium to help drive this forward. Doing so will require action from all of us to break down silos, collaborate in bigger, bolder ways, work with data and people in fields very different from our own, engage communities and people who have not traditionally been included, build a scientific workforce that can pursue the right questions for all populations, and empower young investigators to be our teachers and not just learners.

Exploring Technology in Data Sharing:

During the "Exploring Technology" session, representatives from several NIH-supported resources, demoed tools and platforms, showcasing the valuable data and services available through these repositories. In a collaborative afternoon session, celebrating 10 years of the “NCI Cancer Research Data Commons (CRDC)”, CRDC colleagues reviewed how CRDC supports the research community. They highlighted the CRDC’s role across the entire NCI data lifecycle — from developing data management and sharing plans to data submission, access, use (analysis and tools), and retention. The session concluded with a panel addressing how CRDC meets the "desired characteristics" of repositories as outlined in NIH's guidance on selecting a repository for sharing data from NIH-supported research.

The symposium highlighted the critical role of collaboration, innovation, and technology in advancing biomedical data sharing, with a strong emphasis on equity, access, and infrastructure.