Executive Summary

Inaugural NCI Office of Data Sharing Symposium

Meeting Dates: October 16th and 17th, 2023

The National Cancer Institute's (NCI) Office of Data Sharing (ODS) held its inaugural Data Sharing symposium on October 16th and 17th 2023, bringing the full cancer community (biomedical, advocacy, and industry) together to kick off what is intended to be a continuous, open discussion on the current and future states of scientific data sharing for NCI-supported care and research. This meeting was intentionally divided into one full day, that focused primarily on more immediate data sharing issues and practical ways to meet those needs, and another half day to explore aspirational goals moving forward towards more Open Science implementations as an overall ideal for government funded research. The ODS and NCI are committed to ensuring that the full cancer community can benefit from all the data collected in NCI studies in ways that accelerate novel discovery and innovation for understanding biology and optimally treating each patient with cancer.

Day One

Welcome & Introduction:

Dr. Jaime M. Guidry Auvil (founding Director of ODS)

Dr. Jaime M. Guidry Auvil, founding Director of ODS, provided opening remarks, reflecting on the history and policies of data sharing at the NCI and National Institutes of Health (NIH) at large, but more importantly offered a high-level view of the comprehensive ODS strategic vision to enable broad and equitable data sharing for NCI that fosters collaboration, facilitates innovation, accelerates scientific discovery, and enhances rigor and reproducibility. Dr. Guidry Auvil began with a powerful metaphor, using a geographical view of Cape Town, South Africa, to explain how individual snapshots of the highly diverse city are grossly inadequate to fully delineate a complete and accurate image of the landscape, let alone connect those disparate pieces of information to understand the fabric of what makes Cape Town so special. In the same way, individual data sets, collected in a variety of ways and shared in formats with different or no standard descriptions or processing, can never fully reveal the complexity of a single disease, and can often lead to inaccurate conclusions. Capturing the essence of the meeting goals and the ODS vision, she set the meeting tone with the idea that thoughtful generation and alignment of the many separate pieces of information (aka views of the city, or data sets of a disease) is critical for building a complete and accurate depiction needed to allow for the advancement of scientific inquiry.

Keynote:

Mr. Adam Hayden

The meeting kicked off with patient advocate, Adam Hayden, describing his journey with cancer and his perspective on data sharing from a patient's point of view. He was diagnosed with glioblastoma in 2016, while juggling academic pursuits and a burgeoning family. In his remarks, Mr. Hayden discussed how the integration of health information technology (HIT), interoperability, Health Data Utilization (HDU), and patient data access could significantly benefit researchers, clinicians, and patients alike. For researchers

and clinicians, these advancements promise enhanced healthcare quality and efficiency, facilitate clinical research by empowering patients in managing their health, and helps bridge gaps in access and quality care, particularly regarding social determinants of health and social services referrals. It was emphasized that patients stand to gain significantly from improved healthcare delivery, transparent information about their health, better participation in clinical trials, and an understanding of the evolving complexities in modern healthcare systems. He shared the view that patients "are the data providers" and went on to note that "It's also about contributing your information that will benefit the next generation of patients." His powerful and moving presentation offered a much-needed perspective to the researchers and clinicians at the symposium and set the stage for the discussions to follow.

"Sharing" Data Experiences:

Dr. Henry Rodriguez (Director, NCI Office of Cancer Clinical Proteomics Research/Division of Cancer Treatment and Diagnosis)
Dr. David Chambers (Deputy Director for Implementation Science, NCI Division of Cancer Control and Population Sciences)
Dr. Sean Hanlon (Deputy Director, NCI Center for Strategic Scientific Initiatives)
Dr. Margaret Mooney (Associate Director, NCI Cancer Therapy Evaluation Program, Division of Cancer Treatment and Diagnosis)
Dr. Eric Miller (NCI Division of Cancer Prevention)
Dr. Andrea Ramirez (Chief Data Officer All of Us Research Program)

The first panel of the meeting, titled "Sharing Data Experiences", gathered leaders from NCI distinguished programs to share and discuss successes, challenges and lessons learned in collecting, managing, and sharing different types of research data. Dr. Guidry Auvil opened the panel discussion with a comprehensive overview of current NIH and NCI data sharing policies, most notably the recently activated NIH Data Management and Sharing (DMS) policy. The panelists then each provided an overview of the key programs they are leading while highlighting the various considerations for each data sharing element expected to be addressed in all DMS planning moving forward (data type, tools and software, data standards, data preservation and access, data distribution and reuse, and oversight). The participating programs have shown multiple data sharing successes in their respective fields via international collaboration, harmonization of data standards, strategies of prioritizing high value data, public engagement, and cloud tool/service advancements. Programs are taking different approaches to address challenges relating to the nature of diverse science and data types. The panelists also pointed out that increasing sources of data, including electronic medical records, administrative data, cancer registries, etc., are contributing to the complexity of data governance and data security, as well as other complicating factors like the large amount of data accumulated over time, especially notable in variations of the consenting process, data user agreements, regulations and policies, and languages used. The panel discussion demonstrated remarkable success in data sharing in NCI programs, though more discussions on collaboratively working on the challenges are clearly needed.

Summary of Afternoon Breakout Sessions:

During the afternoon, participants had the opportunity to engage with ODS staff and other attendees in small group discussions and brainstorming around critical topics in data sharing. This included a networking lunch, open office hours with the ODS Director, and interactive whiteboarding on some focused questions. The afternoon finished with breakout sessions that explored key topics of interest in

data sharing among our various stakeholders. Participants, virtual or in person, were able to attend two out of three available breakout discussions to lend their insights.

Sharing throughout the Scientific Data Lifecycle:

Dr. Veena Gopalkrishnan (NCI Division of Cancer Treatment and Diagnosis)
Mr. Ishwar Chandramouliswaran (Lead, FAIR Data & Resources, NIH Office of Data Science Strategy)
Dr. Erika Kim (NCI Center for Bioinformatics and Information Technology/Cancer Research Data Commons)

Various high-level topics focused on planning data management and sharing into all NCI-funded research were discussed. These included planning for Data Management and Sharing in accordance with the DMS policy: outlining supplemental guidance, describing elements of a DMS plan, and providing helpful tips for developing such a plan, along with repository selection and available repository options. NCI biospecimen and data resources were highlighted, including the Cancer Moonshot Biobank and IT infrastructure, as well as an overview of Cancer Research Data Commons (CRDC) and associated Cloud Resources to access and analyze datasets in the CRDC. Datasets accessible through CRDC, data submission processes, and data analysis capabilities were described. Attendees also heard about the NIH Generalist Repository Ecosystem Initiative (GREI) and its role, task areas, and accomplishments. NCI's data access tiers, controlled data access, and expedited review procedures were discussed for reference, and NCI staff were guided through how to obtain access to controlled data. These topics collectively provided insights into data management, sharing, access and analysis within the scientific data lifecycle.

Measuring Impact of Secondary Data Sharing:

Dr. Tanja Davidsen (Data Ecosystems Branch Chief, NCI Center for Bioinformatics and Information Technology/Cancer Research Data Commons)
Dr. Matthew McAuliffe (Chief, NIH Scientific Application Services)
Dr. Lisa Federer (NIH National Library of Medicine)

There were several key points discussed in the Breakout Session focusing on metrics for data sharing and reuse. First, the use of digital object identifiers (DOIs, which are a string of unique characters that can identify an object such as a dataset) for publications and datasets were reviewed, emphasizing how they can greatly enhance tracking data reuse. Secondary use of controlled-access tier data, those requiring approval by an authorizing body through dbGaP, are easier to monitor compared to open access tier data. This is, in part, due to completion of data use certificates that are signed by each approved user and their Institutional Signing Official during the request process, whereas tracking openly available data often relies on altmetrics like web page views and downloads. Second, there was a suggestion to develop a system similar to the h-index, an author-level metric that measures both the productivity and citation impact of the publications, to incentivize data sharing. However, concerns about potential biases were raised (for a review of the h-index and its limitations, see Dinis-Oliveira, DOI: 10.2174/2589977511021911111141801), and there was discussion about Goodhart's Law, which loosely states that when a measure becomes a target, it ceases to be a good measure. Third, measuring how meaningful an impact data sharing has on a scientific field depends on stakeholders' perspectives, which can vary greatly. For example, repository staff may focus on publications and number of data downloads, yet academics and funders have more interest in knowing how their data are contributing to others' research and how data sharing could advance their career, while clinicians tend to prioritize data's role in

specifically advancing treatment options. The field of data reuse metrics is still emerging with further discussions needed on emphasizing the importance of consistent, useful metrics and identifying various stakeholders who stand to benefit.

Institutional Approaches Breakout Summary:

Dr. Stuart Levine (Massachusetts Institute of Technology)
Dr. Caroline Shamu (Harvard University)
Dr. Michael Townsend (The Ohio State University Comprehensive Cancer Center)
Dr. Carla Williams (Howard University)
Dr. Joanna Groden (University of Illinois, Chicago)
Dr. Valeria Mezzano (New York University School of Medicine)
Dr. Ricardo Richardson (North Carolina Central University)

During the Institutional Approaches Breakout, it was evident that while large and well-resourced academic institutions are actively pursuing strategies to promote data sharing and management in line with NCI and NIH policies: such as using electronic notebooks and educating researchers on data standards, smaller institutions have experienced challenges in providing support to faculty to develop DMS plans and noted a lack of data sharing infrastructure for cross-institutional partnerships. Panelists noted challenges in sharing data in a public repository due to lack of defined or consistent expectations for standardized data, limited institutional resources to support data management activities, trust and bias issues in community-based participatory research that result from a lack of consultation with the community from which data is generated or collected, and a lack of data management infrastructure for cross-institutional collaborations. To address trust and bias concerns, Dr. Carla Williams provided an innovative vision to propose a "TRUST 2.0" framework, based on TRUST principles, that emphasized building capacity, readiness, unbiased data (representation, context and meaning), sharing, and trustworthiness to ensure that communities benefit from their contributions to research, shifting the perspective of these research participants from "subjects" to "contributors." This comprehensive approach aims to enhance data management and sharing practices, emphasizing health equity and community involvement. Recommendations for improvement shared by panelists included incorporating data sharing activities alongside publications in the NIH Biosketch template, addressing institutional support and building capacity for long-term data storage within institutions, integrating data management training in onboarding and offboarding processes at each institution (by NIH or each Institution), and promoting data annotation efforts.

Day Two

Year of Open Science:

Dr. Maryam Zaringhalam (White House Office of Science and Technology Policy)
Dr. Lisa Federer (NIH - National Library of Medicine)
Dr. Paige Martin (National Aeronautics and Space Administration)
Ms. Kristen Ratan (Stratos/Incentivizing Collaborative Open Research)

The second day of the Symposium focused more on aspirational goals of data sharing, including the opportunity to highlight Open Science. The White House Office of Science and Technology Policy (OSTP) is using 2023 to celebrate a “Year of Open Science”, which is an opportunity to advance actions and policies promoting broad and equitable research that enhances public trust and provide access to publicly funded data to accelerate discovery and innovation. Panelists began by providing remarks focusing on their work which included:

- A discussion of the benefits and impacts of open science, emphasizing increased accessibility, reproducibility, and diversification of research, along with showcasing examples of how shared data can lead to unexpected discoveries
- The various open science initiatives and their influence on scientific research, collaboration, and accessibility, including NASA TOPS, the White House Year of Open Science, ICOR, ASAP, Moonshot 2.0, and Plan S/Coalition S
- Means to enhance equitable inclusion and access to scientific research, specifically addressing under-resourced countries and institutions' access to data
- Highlights of the NIH Public Access Policy's response to the 2022 OSTP Memorandum, emphasizing equity in publication opportunities, data access, and transparency through persistent identifiers

Panelists discussed the importance of stakeholder engagement and collaboration in open science and research initiatives, acknowledging the unique potential role of citizen scientists in advancing discoveries when access to data is not limited to certain scientific fields. Throughout the session a central theme emerged – a needed cultural shift within the scientific community valuing data sets more for career advancement and grant funding than the current model of publication in high-impact biomedical journals. Some panelists suggested possible solutions include improving open access for science that specifically focuses on treating shared data sets on par with publications, reporting them on CVs and grant applications, and tracking their impact over time.

Fireside Chat:

Dr. Monica Bertagnolli (former NCI Director, current NIH Director)
Dr. Susan Gregurick (NIH Associate Director for Data Science)

The inaugural NCI Data Sharing Symposium closed out with a recorded message of the Biden administration’s support of our efforts surrounding data sharing and reuse from the White House Cancer Moonshot office, as well as a Fireside Chat with NCI and NIH leadership. Dr. Guidry Auvil sat down with Drs. Monica Bertagnolli and Susan Gregurick to discuss the NCI and NIH vision of data sharing that benefits cancer patients, survivors, and their families. As a research surgeon and breast cancer patient, Dr. Bertagnolli emphasized the importance of data in clinical strategies and supported Dr. Gregurick’s iteration of the need to incentivize data reuse and ensure interoperability. They highlighted initiatives like the NCI Childhood Cancer Data Initiative (CCDI), and NIH All of Us and NIH Cloud Platform Interoperability (NCPI) efforts to connect NCI and other NIH data platforms; all of which aim to create communities around data and its analysis universally. Some challenges of data sharing were acknowledged, including data formatting and the variety of roles within the research community. They discussed the critical need for responsible use of artificial intelligence (AI) and patient involvement in research, emphasizing the positive impact of involving patients and restoring trust in science. Opportunities to use COVID real-world data, cooperation among government agencies, and the importance of partnerships across the research community were also mentioned, along with employment of data standardization and use of core data elements. In the future, they envisioned scaling data science for wider accessibility and communities focused on critical public health questions that serve everyone. Dr. Guidry Auvil ended the informative and engaging conversation with some important messages:

- Gratitude messages for the ODS team who organized the meeting and all the incredible speakers and panelists that provided critical insight

- Reminders of current data sharing expectations, NCI approaches that are serving as exemplars in this space, and goals for moving data sharing forward in equitable ways
- Reiterating the NCI commitment to ensuring that all stakeholders (scientists, clinicians, patients, advocates, and industry) can work together to optimize how we collectively leverage the power of research data to benefit the whole cancer community
  Invitation to all attendees to remain engaged and keep the dialogue open throughout the year in preparation for the next annual NCI Data Sharing Symposium. Anyone interested is encouraged to subscribe to email listserv for future updates.