National Institutes of Health
National Institute of Diabetes and Digestive and Kidney Diseases
Data Management and Sharing Webinar Series
Virtual Meeting

Session 4: The "R" in FAIR: Data Reuse
July 13, 2023

SUMMARY

Welcome and Introductions

Shannon Bell, Ph.D., RTI International

Dr. Shannon Bell, Senior Bioinformatician, RTI International, welcomed the participants to the fourth session of the Data Management and Sharing (DMS) Webinar Series. Recordings of the previous three webinars and associated information are available on the NIDDK DMS website.

Advancing Knowledge Through Secondary Data Use

Vivian Ota Wang, Ph.D., Office of Data Science Strategy (ODSS), National Institutes of Health (NIH)

Dr. Vivian Ota Wang, Lead, Policy and COVID, ODSS, explained that many of the leading causes of death have associated modifiable behaviors, and data analysis efforts aim to understand broad causes of diseases. The current profusion of types and amounts of data is moving the field toward the democratization of knowledge, aligned with the human rights idea that everyone has the right to information and knowledge. A study by the National Academies of Sciences, Engineering, and Medicine outlined how to design purposefully open science, emphasizing innovation in data sharing as a path to making data available to all members of society. The movement toward open science will support development of the next generation of research tools, increase the statistical power of studies, and improve research quality through validation and replication of data.

This kind of data use is fundamental to the development of precision medicine, which considers not only biological factors related to health and disease but also the social, cultural, and psychological factors that influence the underlying biology. Precision medicine will lead to advancements in methods, technology, and management of many types of data. Early advances already are changing how diseases are classified and treated. At a fundamental level, the open science movement is increasing the scientific value of data by allowing researchers to explore, combine, and analyze data from many sources. This movement will increase the scale of studies and the types of scientists who can participate because secondary data analysis tends to be multidisciplinary and includes researchers from many fields.

Data begin as unstructured information and must be organized before researchers can identify trends. The ecosystem of data comprises many levels, from individual genomic and molecular data to national-level population data. Scientific studies collect many types of data, including real-world data, self-reported data, data from electronic health records, and exposure data. Data now are generated very rapidly, and researchers must determine how to verify, store, and use data at a lower cost. Until recently, only large institutions had the ability to store and use data securely. Current technology, such as cloud computing, allows greater democratization of data access and sharing. Cloud computing is cost-effective and scalable, with computational capacity for many large data sets.

As environments and repositories are evolving, use of data is becoming more social, requiring increased focus on communication and equity to support the work of multidisciplinary teams with diverse expertise. Research has become a data-intensive enterprise—data are rapidly becoming pervasive and extending beyond laboratories and clinics into homes and every aspect of today's world. These data advances are illuminating the functional roles of genes across tissues and diseases, the interrelatedness of diseases and conditions, and the relationships between behaviors and health. The large amounts of data collected across studies can be used for secondary analyses to identify trends and improve research quality. This paradigm shift in science—from a hypothesis confirmation mindset to a hypothesis generation mindset—is elaborated on in the *NIH Strategic Plan for Data Science*. Remaining challenges include issues related to privacy and confidentiality, and researchers must keep equity and disparities issues at the forefront of their considerations. Dr. Ota Wang emphasized that data and information are not neutral and are prone to exacerbating stigma, decreasing inclusion, and perpetuating disparities and injustices. She challenged listeners to design studies that address equity in innovative ways.

Best Practices for Secondary Data Use
Harold Lehmann, M.D., Ph.D., Johns Hopkins University School of Medicine

Dr. Harold Lehmann, Professor, Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, emphasized that investigators must unlearn some of their habits as they begin to work with new types of data. In hypothesis-driven science, researchers know which elements of their study are definitive and can plan around established rules to minimize biases; when real-world data are used to generate hypotheses, the definitive aspects of the study are unknown, and collecting data to confirm hypotheses often maximizes biases.

Some data are generated with structures that must be adjusted to fit the desired analytic variables, but data constructs and analytic variables influence each other. Data constructs may include a concept identification for a standard code, a set of synonymous codes, a standard formula, the time frame in which the data were collected, the phenotypes defined, and the structure of the cohort. Data cleaning is required to transform raw data into the variables needed for analysis. Many checklists are available to help researchers use best practices to report real-world evidence, but most research based on electronic health records does not use sufficiently cleaned data.

Several large open-science networks are engaging in the work of secondary data analysis, including the National Patient-Centered Clinical Research Network (PCORnet), the Observational Health Data Sciences and Informatics (OHDSI) program, and the National COVID Cohort Collective (N3C). Such networks can conduct massively parallel analyses, which make many simultaneous comparisons among the data at a fraction of the cost of standard analysis. Bespoke analyses remain useful, especially for new treatments, and the rules of massively parallel analyses remain unknown; however, researchers need to resist the impulse to avoid these efficient methods.

One area still lacking innovation in this new environment is communications. Scientific communication continues to be conducted mainly through the publication of articles, but researchers must identify how the power of computers can be used to communicate. Several efforts are in progress that build on the foundation of Fast Healthcare Interoperability Resources (or FHIR) systems. Dr. Lehmann emphasized the need for such systems to include all the elements that comprise evidence—data, method, and context.

Generalist Repositories for Sharing and Finding Data
Ana Van Gulick, Ph.D., Figshare

Dr. Ana Van Gulick, Government and Funder Lead, Head of Data Review, Figshare, pointed out that domain- or institution-specific repositories within the data sharing system should be researchers' first choice for hosting their data; however, specific repositories are not available for all types of research outputs. In addition, institutional repositories often are not open to researchers outside that institution and may not meet the sharing requirements of the NIH DMS policy. In such cases, a generalist repository may be an option to consider. Figshare runs two public-facing generalist repositories: Figshare.com, which is available free to researchers around the world to share scholarly outputs of any type, and Figshare+, a new repository for larger data sets.

Figshare strives to provide services that meet NIH's list of desirable characteristics for data repositories. It offers flexibility, integrates existing tools to support researchers' current workflows, uses best practices, provides full open access to published material, and offers ways to track the impact of the work. Dr. Van Gulick noted that generalist repositories should be used for data that are not sensitive in nature, have been fully de-identified, and can be made publicly available.[1] Figshare+ also helps support data curation and metadata completeness during deposit.

The Generalist Repository Ecosystem Initiative (GREI), supported by ODSS, brings together seven generalist repositories, including Figshare, to enhance support for NIH data sharing and discovery under FAIR (findable, accessible, interoperable, reusable) principles. The GREI repositories all provide digital object identifiers (DOIs) and structured metadata to make the data findable and discoverable, and standardized metadata are used to link among systems and provide interoperability. The repositories are fully accessible, including through programmatic access, and the clear, machine-readable licenses and metadata provide context for the data, ensuring that anyone who finds the data can make sense of them.

Generalist repositories also can be tools for finding and reusing data. They offer the flexibility to share any research output or file type, and they support discoverability and open access. Generalist repositories also use open application programming interfaces (APIs), provide DOIs for all records and other persistent identifiers (PIDs) as needed, and support structured metadata. Supporting these data sharing practices can make data more reusable in the future. Because generalist repositories cannot set precise requirements for most metadata, researchers must understand the best practices of their fields—as well as data sharing generally—and take the initiative to add descriptive information to their metadata to help others navigate reuse of the data.

Figshare and other generalist repositories can be used to create FAIR records for all types of research outputs, and researchers can use different repositories for different outputs as appropriate, then link the data via the metadata. Dr. Van Gulick showed many of the functions available on Figshare, including sharing, creating programmatic records, integrating with other tools and PID systems, creating DOIs for every research output, using PIDs to link related materials, and linking to funding resources. Dr. Van Gulick noted that discoverability is key for FAIR outputs, and outputs that are well described have a much better chance of being found and reused.

Question and Answer Session

- Dr. Lehmann asked Dr. Ota Wang how she distinguishes between reuse of data collected for a specific research purpose and use of real-world evidence. Dr. Ota Wang pointed out that the most important consideration is how to adhere to the informed consents that participants signed when they provided their data. Regardless of the data's initial purpose, researchers must respect the conditions to which the participants agreed, especially because participants may not be aware of the nuances of reuse.
- Dr. Van Gulick reminded attendees that many resources are available at institutions and repositories and encouraged them to use these resources to share their data as much as possible.

[1] For more information on data characteristics for repositories participating in the GREI, visit https://www.rd-alliance.org/sites/default/files/Generalist%20Repository%20Comparison%20Chart_version%203.0.pdf.

Adjournment

Dr. Bell thanked the panelists and attendees and noted that all the materials will be available on the NIDDK DMS website.