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

Session 3: Metadata and Data Standards for National Institute of Diabetes and Digestive and Kidney Diseases Research Data

June 27, 2023

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) has developed resources to support researchers in addressing metadata and data standards, including institute-specific guidance on integrating metadata and data standards during the study design phase and how to describe these efforts in the data management and sharing (DMS) plan. A Data and Metadata Standards Examples tool includes decision steps to help guide thinking and a table of standards and example repositories by data type. Four example DMS plans provided by NIDDK (as well as additional sample plans from the National Institutes of Health) highlight the integrated use of metadata and data dictionaries (described in Element 1C) and the use of data standards and common data elements (Element 3) that should be incorporated in DMS plans. For more information, please review the guidance and resources provided by NIDDK on the NIDDK Guidance for Writing a DMS Plan and NIDDK DMS Tools & Examples pages.

SUMMARY

Welcome and Introductions

Jeran Stratford, RTI International

Dr. Jeran Stratford, Bioinformatician and Data Scientist, RTI International, welcomed the participants to the third session of the Data Management and Sharing (DMS) Webinar Series, hosted by NIDDK to catalyze understanding within the NIDDK scientific community about DMS topics. The previous two webinars covered planning for DMS and selecting an appropriate repository for scientific data.

Metadata and Data Standards—What and Why

Matthew Schu, RTI International

Dr. Matthew Schu, Director of ‘Omics, Epidemiology, and Analytics Program, RTI International, explained that strong metadata are needed to validate research findings, strengthen analysis through combining data sets, allow reuse of data, open new discovery frontiers, and foster trust in publicly funded research by ensuring that the findings are reproducible. Metadata are “data about data” and provide the information required for researchers to understand details of data collection and explore deeper questions about the data.

Metadata can be considered on several levels. Study- or data set–level metadata include information about the purpose of collecting the data, including information about the study. Variable-level metadata include information about the individual variables measured in a study or data set; metadata at this level often are stored in a data dictionary, which usually is a separate table that describes fields contained in a larger table. File-level metadata include information about the file, such as location, format, and access procedures. Most researchers already collect metadata, which often are required for manuscript or repository submission; although these procedures now are standard, the quality of metadata must be ensured to facilitate data sharing. Many commonly shared data types now have metadata standards, which outline the minimal amount of information that must be shared with data to allow its reuse by others.

Dr. Schu outlined the Nutrition for Precision Health (NPH) Initiative, a study powered by *All of Us* to deeply explore the relationship between nutrition and health. NPH is aiming to enroll 10,000 participants across 14 U.S. sites and to gather details about the participants' diets, environments, microbiomes, and metabolomic signatures and then integrate those data to advance the scientific understanding of how each individual responds uniquely to food. All data collected in this initiative will be shared on the *All of Us* Researcher Workbench and integrated with data collected in other *All of Us* initiatives, so data and metadata standards are particularly important.

The microbiome is the collection of bacteria living on the skin or in the gut. NPH researchers use sequencing techniques to detect the flora in participants' guts. Metadata included with the sequencing files include the sample name, type, and qualities; physical specimen collection data; and country. One unique type of metadata included with these data is the sample's rating on the Bristol stool scale, which provides information on what the sample looks like.

Consistent data standards are needed to ensure that study data can be interoperable and harmonizable with data from other studies, as well as to minimize lost or missing data. Several existing repositories have strong data type–specific data standards; researchers are encouraged to use such repositories and standards, when possible, to allow potential downstream data integration. Data standards and metadata are important for processing and automation, allowing high-throughput data analysis.

Establishing Data Structure to Increase Accessibility

M. Todd Valerius, Brigham and Women's Hospital, Harvard Medical School

Dr. Todd Valerius, Associate Biologist and Instructor of Medicine, Brigham and Women's Hospital and Harvard Medical School, explained that the GenitoUrinary Development Molecular Anatomy Project (GUDMAP) and ReBuilding a Kidney (RBK) programs are working to combine their data through the Analysis, Technology, Leadership, Administration, and Science—Data to Knowledge (ATLAS-D2K) Center with an overarching goal of building the "go-to" open-access resource for the research community of mouse and human renal and genitourinary development and disease. ATLAS-D2K acts as the data repository and data discovery hub of the GUDMAP and RBK consortia and ensures that the complex data are in a form accessible to the community, establishes connections between kidney and lower urinary tract data in these and related consortia, and enables researchers with various levels of experience to engage with the generated data by providing data-interaction tools.

GUDMAP and RBK have slightly different objectives; however, many investigators are involved in each. New technology is added frequently, and data from several species (human and animal models) are recorded. To form an atlas, adult human data must be correlated to data from other species and developmental stages. ATLAS-D2K also aims to serve a broad range of investigators interacting with and using the data. ATLAS-D2K intends to make the data maximally useful to people beyond the original study, thereby accelerating science. ATLAS-D2K develops example queries and graphical tools, integrates molecular and imaging data, establishes reference data sets, and harmonizes data across consortia.

Ontologies and controlled vocabularies have become important for DMS activities. "Data dictionaries" is the term most often used in the clinical space, and researchers more commonly refer to "ontologies," but the meaning is the same. Ontologies are structured relationships of terms that ensure that data can be entered with a limited vocabulary to facilitate search and use of the data. Researchers can search for scored expression data and tie annotations of structures in histological images to ontological terms used in sequencing data. Dr. Valerius recommended that researchers select a source of anatomical and cell-type terms that best fit their research and ensure that the terms are used consistently.

Dr. Valerius commented on several types of data formats. Data shared in a raw sequence format have privacy concerns, but the computational analysis and alignment of such formats are collected well by current systems. Metadata for biosamples must be captured at the time of the study. Many types of image formats are available, but data can become locked in one format if technology changes. When data are standardized to a common format, open-source software can be used. Dr. Valerius emphasized that consortia need to be able to adapt to new technologies, and he recommended that users develop a plan to capture biosample metadata and protocols and use consistent image formats. He pointed out that libraries at many institutions can be helpful in this area. The privacy issues inherent in raw sequence data formats can be avoided by adding additional layers of sequencing and focusing on the processed data. Researchers can use standardized approaches and provide data dictionaries to inform future users of how the data were generated.

Realignment may be necessary when references change, but with good metadata, researchers can redo the analysis on a reference genome and update older data to allow new users to perform new visualizations. Dr. Valerius recommended that users consider asking their cores to use programmatically published base processing pipelines, publish the pipelines and the analysis code through a coding repository, consider what intermediate layers should be protected, and publish analysis intermediates.

Role of Data Standards in Quality and Harmonization

Sanjay Jain, Washington University in St. Louis

Dr. Sanjay Jain, Professor of Neurology, Pathology, and Pediatrics and Director of the Kidney Translational Research Center, Washington University School of Medicine in St. Louis, explained how quality control and harmonization in data standards are used to understand the human kidney. The Human Kidney Atlas aims to create a high-resolution integrated single-cell and spatial multimodal atlas of the human kidney in health and disease across an individual's life span. Atlas projects are collections of spatially resolved morphological and molecular maps that help researchers understand the vital functions of the kidney and will generate knowledge that provides insights to prevent kidney injury and promote recovery.

The atlas involves many assays conducted across numerous institutions, so quality assurance and control are ensured by following the tissue pipeline. Assay and data harmonization processes are used to bridge data types and ensure rigor, and examples of cross-species integration are used to outline cellular diversity and injury time course. At each step of single-cell technologies, choices are made that affect the data generated. Without quality control, the resulting data will be unusable.

Quality control processes are used to generate data that are reproducible and rigorous. These processes minimize technical variations, allowing investigators to uncover biological variations. Quality control processes begin at the participant level; capturing metadata related to the subject is critical because such factors can affect how results are interpreted. Minimizing confounders helps researchers avoid making conclusions that may be inaccurate.

At the specimen level, metadata include pre-analytical variables about the procurement, preservation, and storage of the samples. Some analytes may be sensitive to these factors, and researchers will not know in advance which will be affected, so recording as much data as possible is key. Analytical components should be captured along with quality control cutoffs. Many researchers are using human samples from multiple studies, so providing accurate registration and mapping for the sample's source is critical. Clinical samples often are very small, and data are generated at a single time point. When samples are registered, researchers can understand variability and interpret the data properly.

Assay harmonization is also important. In sequencing, including genomic and imaging, use of the reference genomes or sequencing attributes must be documented. When ‘omic assays are used, researchers must ensure that variations do not appear over time because of software or analytical pipeline change. Standardized tissues are used in each batch to ensure that the assay is performing well; this helps researchers gauge any deviations and identify any problems.

Nomenclature also must be harmonized—data sets are generated by different laboratories, in different anatomical contexts, and using different cell types, but researchers must ensure that they can communicate about the same factors. Ontologies are used at various levels to confirm that structures and biomarkers under consideration are the same.

Dr. Jain pointed out that in the Human Kidney Atlas, the definitions for healthy and nonhealthy states are supported by the literature, which helps the community interpret the information. Each of the five technologies used in the project creates its own maps, but unification of names across technologies allows researchers to combine them into an atlas, as well as translate across species. When all cell types are defined, the data can be applied to human clinical observations to better understand the biology of kidney injury. Many of these studies are possible only because of harmonized formats.

Dr. Jain noted cross-collaborative tools available for the project. An Azimuth reference is available on the Human BioMolecular Atlas Program (HuBMAP), which allows users to identify cell types in their own data set. Cell types and states are available through CELLxGENE, allowing users to explore genes and cell types and correlate them with clinical data. An atlas explorer available through the Kidney Precision Medicine Project shows gene expression using multiple technologies, and Anatomical Structures, Cell Types, plus Biomarkers (ASCT+B) tables are available on HuBMAP to standardize cell-type nomenclature and markers.

Implementation of Data and Metadata Standards—The Added Value

Kenneth Young, University of South Florida

Dr. Kenneth Young, Assistant Professor and Chief Information Officer, Health Informatics Institute, University of South Florida, provided an overview of The Environmental Determinants of Diabetes in the Young (TEDDY) Study, a cohort of more than 8,000 children in the United States and Europe who were enrolled before 4.5 months of age and followed for up to 15 years to identify genetic and environmental triggers of type 1 diabetes. Children from the general population and children with a parent or sibling with type 1 diabetes are included. The study investigates potential gene–environment interactions both before birth and during childhood to explore the development of pre-diabetes, autoimmunity, and type 1 diabetes. Study participants are followed for 15 years to track the appearance of beta cell auto antibodies and diabetes with documentation of early childhood diet, reported and measured infections, vaccinations, and psychosocial stressors.

A variety of data types are collected, including clinical metadata and laboratory test results across various ‘omics analytes. Data are integrated at the Data Coordinating Center and provisioned for analysis by investigators. The TEDDY website provides many types of clinical metadata and documentation, including a clinical metadata overview and collection summary. TEDDY also collects ‘omics analytes and provides an ‘omics overview and data summary. Use of data standards enables reuse of data elements and their metadata, which reduces redundancy between systems, improves readability, and reduces costs.

To improve interoperability, TEDDY implemented several biomedical ontologies for adverse events, diagnostic information, medications, and other standardized data sets. To improve the quality and reusability of the data, electronic case report forms (eCRFs) were designed to capture certain data standards directly. In addition to standard ontologies, TEDDY has unique “TEDDY codes” that are used to capture clinical data in a standardized manner for the study. Collecting codes limits the use of free-text fields, improving data accuracy and implementing consistency across reported values by restricting abbreviations, synonyms, and misspellings.

Each clinical data set TEDDY shares is accompanied by a data dictionary, which contains metadata and additional information to make the scientific data interpretable and reusable. Dr. Young pointed out that providing sufficient and well-structured metadata is a key component of abiding by FAIR (findable, accessible, interoperable, reusable) principles. Clinical data dictionaries can be provided in multiple formats, and dictionaries for ‘omics data vary by data type.

Direct-to-investigator data releases also receive release notes describing the data freeze date, population, data sets provided, and any relevant notes for investigators. The release notes metadata summarize the information about data releases that can make tracking or working with specific data easier. The TEDDY ‘omics metadata also are shared with data repositories, including the database of Genotypes and Phenotypes (dbGaP) and the Metabolomics Workbench.

Another key component to TEDDY standardization is the use of eCRFs that are annotated with the data set name and variable name. These forms have been shared with the NIDDK Central Repository as a searchable PDF file, allowing investigators to find data of interest, understand how they were collected, and identify related variables. The TEDDY public website has information to help investigators find documents detailing the data collection procedures, data availability, and data sharing policies. This information is updated periodically when more data are made available.

The TEDDY study has adopted policies and procedures in support of its commitments to sharing data with the scientific community while also protecting the privacy of participants. Data releases have been submitted at different time points and to various repositories depending on the requirements and nature of the data; each submission is treated as an independent release. If researchers want to combine independent releases, the repository can provide identifier mapping materials once investigators have received approval to access the data. Data are shared to dbGaP, the Metabolomics Workbench, and the NIDDK Central Repository.

Question and Answer Session

- When asked about the best time during the study lifecycle to consider data standards and metadata collection to maximize the value of the study and minimize the effort required, Dr. Young suggested that the best time to start planning data standards is during study design, even though aspects may change as the study proceeds. Dr. Jain agreed that data standards should be considered as early as possible. With a plan in mind, researchers can identify potential collaborating sites or individuals with relevant expertise to contact.
- In response to a question about how to respond to a lack of appropriate standards, Dr. Valerius suggested that researchers select standards that work for their study from the available resources and then work to improve ontologies. He pointed out that understanding metadata can be a significant challenge when standards are lacking, so any efforts to improve metadata will improve the research results.

Adjournment

Dr. Stratford thanked the panelists and attendees and noted that a recording of the webinar will be available on the NIDDK DMS website. NIDDK also will post a Q&A document with responses to questions received during this webinar, including those that were not addressed live. He invited attendees to return for the next webinar in the series, focused on data reuse, which is scheduled for July 13, 2023, from 12:00 p.m. to 1:00 p.m. EDT.