U.S. Department of Health and Human Services (HHS)
National Institutes of Health (NIH)
Office of the Director
Division of Program Coordination, Planning, and Strategic Initiatives
Office of Data Science Strategy (ODSS)

Advancing the Use and Development of Common Data Elements (CDEs) in Research

March 6–7, 2024
Auditorium Balcony C, Natcher Building
NIH Main Campus
Bethesda, MD
and Virtual

Workshop Summary

Day 1: Wednesday, March 6, 2024

Welcoming Remarks

Susan K. Gregurick, Ph.D., Associate Director for Data Science and Director, ODSS, NIH

Dr. Susan Gregurick welcomed attendees and outlined the ODSS goals for the next 5 years regarding data science at NIH, including improving capabilities to sustain the NIH Policy for Data Management and Sharing; developing programs to enhance human-derived data for research; providing new opportunities in software, computational methods, and artificial intelligence (AI); supporting a federated biomedical research data infrastructure; and strengthening a broad data science community. She explained that CDEs are standardized, defined questions paired with specific allowable responses; if used systematically across sites, studies, or clinical trials, CDEs ensure the data collected are consistent. The workshop's aims were to assess, enhance, and broaden the development, adoption, and use of CDEs for research across diseases and conditions; demonstrate successes and discuss strategies to encourage the adoption and use of CDEs; and engage participants from diverse professional backgrounds.

Opening Keynote

Monica Bertagnolli, M.D., Director, NIH

Dr. Monica Bertagnolli outlined her background and emphasized that CDEs are essential to making science valuable, especially given the recent increase in analytic capabilities. Although electronic health records (EHRs) are promising, data from the clinical-care environment must be integrated with data from other sources and reused many times across disciplines. Dr. Bertagnolli noted health care challenges in the United States, including lower life expectancy at higher cost than comparable nations, the low quality of evidence used to make clinical decisions, lack of adequate representation across populations in research, and issues with data sharing. She emphasized that NIH's work is not finished until all people are living long and healthy lives and added that NIH scientists must take charge of all steps of this journey. Dr. Bertagnolli pointed out that standards developed by researchers are adopted across the biomedical community, so NIH support for broad access to advanced analytics and computational power is critical to advancing data science and promoting new discoveries. She also emphasized the importance of education and collaboration, as exemplified by this workshop.

Setting the Stage: Making Data Interoperable

Denise Warzel, M.Sc., National Cancer Institute (NCI), NIH

Ms. Denise Warzel emphasized that this workshop is intended to elevate participants' understanding of CDEs, including the unique characteristics that help with data integration, mapping, and transformation and that can be used to leverage CDE metadata and make data more interoperable. Governance among collaborators who agree to use certain principles to make data interoperable also is critical. Important aspects of NIH CDEs include standard terminology and structure, independent semantics, and unique identifiers. These characteristics ensure data are FAIR (findable, accessible, interoperable, and reusable). Ontologies also are useful to encode existing language in a domain-specific way. Ms. Warzel emphasized that the foundation for machine computability is meaning, and standardizing CDE concepts allows researchers to compute meaning by assessing the concepts associated with the question and allowable responses. She reiterated that NIH hopes to motivate attendees to adopt CDEs to improve data quality and consistency, support data harmonization, enhance knowledge acquisition, simplify collaboration, and reduce project startup time.

Session I: The Value of Common Data Elements

How Biomedical Ontologies Overlap With CDEs and Contribute to Harmonization

Richard Scheuermann, Ph.D., National Library of Medicine (NLM), NIH

Dr. Richard Scheuermann outlined the value of ontologies for structuring the content of data and facilitating interoperability. Ontologies—formal, explicit specifications of a shared conceptualization—capture relationships among the entities represented and add value to the information. Ontologies can be formally represented as axioms, which then can be used to infer new information from the embedded knowledge. Dr. Scheuermann provided examples of ontology use and explained that they are developed by a community of informatics professionals in collaboration with subject matter experts and involve principle development approaches. By design, ontologies promote FAIR data principles, and their underlying semantic architecture can structure the way that information is considered, allowing users to make inferences within the semantic knowledge.

Use of Clinical Data Interchange Standards Consortium Data Standards at the U.S. Food and Drug Administration (FDA)

Helena Sviglin, M.P.H., Office of Strategic Policy, Center for Drug Evaluation Research, FDA

Ms. Helena Sviglin presented on the relationship between FDA's policy framework and NIH's CDEs. FDA has the authority to require researchers to submit standardized data to enhance public health, which is FDA's primary objective. The relevant guidance includes two documents: the Study Data Technical Conformance Guide and the FDA Data Standards Catalog, which supports increased interoperability between CDEs and the data models in the catalog. She demonstrated the use of the catalog, explaining that every model has been investigated to help reviewers more efficiently and accurately make decisions about the safety and efficacy of data submitted to the agency. Ms. Sviglin also noted that an NIH-maintained vocabulary is embedded within the standards.

Garbage In, Garbage Out: Standardized Data for Better Reuse

Avinash Shanbhag, M.S., Office of Technology, Office of the National Coordinator for Health Information Technology (ONC), HHS

Mr. Avinash Shanbhag presented on the importance of standardized data for reuse, particularly ONC's United States Core Data for Interoperability (USCDI) and USCDI+ projects. ONC focuses on formulating a health information technology strategy that meets the needs of all HHS agencies. ONC built a certification program for EHR technology that now is used in more than 90 percent of hospitals, and it also ensures that governance is in place at the network level. ONC worked with industry to build a core set of CDEs with common semantics and a common format. EHRs now are required to begin exchanging USCDI version 3 elements, which expand upon earlier sets to help measure health equity and health outcomes. The USCDI+ program extends USCDI elements to meet specific data needs in different domains while ensuring the data remain interoperable. Mr. Shanbhag emphasized ONC's commitment to serving agency needs and ensuring standards are incorporated into regulations to build the foundational infrastructure that is supported by researchers' use of CDEs.

Common Data Elements: A Healthy People Perspective

Allan Uribe, Dr.P.H., M.P.H., CPH, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention

Dr. Allan Uribe outlined the Healthy People 2030 initiative, which provides a framework for prevention and wellness programs across a diverse array of users. He discussed the initiative's data-driven national objectives to improve the health and well-being of people across the United States. The current iteration of Healthy People increases the focus on social determinants of health (SDOH), which are conditions in the environments in which people live that affect a wide variety of health functioning and quality of life outcomes and risks. NCHS provides statistical advice to HHS and the Healthy People topic area workgroups, and the center compiles all data for the initiative into the Data 2030 database. NCHS also analyzes and presents the data and provides statistical expertise and technical assistance to users. Healthy People was the first initiative to apply a standard data template to national health objectives, and Data 2030 includes data from a wide variety of data systems and more than 80 sources. To be included, data must be nationally representative and publicly available, with a known population coverage and a complete set of documentation, which supports transparency and allows findings to be replicated.

Discussion

- The Data 2030 database is available online; Healthy People does not collect the data but aggregates them from publicly available sources and summarizes them.

- Data models are the intellectual property of external organizations, so grassroots initiatives within those organizations are needed to version CDEs; agencies often do not have the authority to stipulate how data are collected, and data needs of different agencies or health care providers vary.

- EHR data interoperability is complicated by the large variety of EHR systems and the fragmentation of the U.S. health care system, but EHR leadership and vendors should be invited to participate in such discussions as this meeting.

Session II: Current NIH Resources for CDEs

Overview of the NIH CDE Repository, CDE Governance Committee, and Request for Information (RFI)

Robin Taylor, M.L.I.S., NLM, NIH

Ms. Robin Taylor explained that the NIH CDE Repository is a resource for all NIH, providing access to CDEs and forms recommended or required by NIH institutes and centers (ICs) for use in research. The NIH-wide CDE Governance Committee was established to review CDEs and decide whether they merit NIH endorsement, after which they are published in the repository. The Governance Committee does not adjudicate the scientific merit of each CDE submitted but ensures that they meet the criteria set by the NIH Scientific Data Council and that appropriate evidence of validity is provided. Ms. Taylor outlined the process for CDE submission, approval, and review. The repository currently contains 171 NIH-endorsed

CDEs and is preparing to publish 500 more; another 1,000 CDEs currently are undergoing review. Ms. Taylor emphasized that NIH currently does not have any requirements for CDEs. The repository is open to anyone to use, but the primary end-user group is NIH-funded researchers looking for CDEs that are recommended or required by their NIH organization or funders; a secondary group is NIH program staff developing CDEs and surveying the landscape. Ms. Taylor also noted a currently open RFI regarding a core minimum set of CDEs for all NIH-funded research.

Q&A Panel Session
NIH CDE Governance Committee

- CDEs in the repository may be listed as "qualified," reflecting their status as recommended by an NIH IC, or "standard," but these statuses are due for review to ensure they are being applied consistently.

- The protracted governance process makes versioning difficult, but the repository governance is reviewed annually, and improvements can be made. Redundancy is a known issue in the Repository; however, sometimes there are CDEs in the repository that seem redundant but are not. For example, they may have different permissible values that distinguish them. The governance committee could begin streamlining the number of CDEs by recommending existing CDEs during the application process. The committee is actively working to reduce redundancy by encouraging use of the NIH-Endorsed CDEs in the repository.

- Strategies to increase the use of CDEs include tying them to milestones and requiring certain types of data collection as a condition of the grant.

- Usage is difficult to assess because only usage within the repository is visible to NLM.

- The committee is considering ways to incorporate large language models (LLMs) to improve CDE workflows and harmonization.

Session III: Overcoming Barriers in CDE Adoption, Mapping, and Use in Community Research

RADx® Underserved Populations—CDEs
Warren A. Kibbe, Ph.D., FACMI, Duke University

Dr. Warren A. Kibbe discussed RADx-Underserved Populations (RADx-UP) in a community context. RADx-UP was established to help understand the factors associated with COVID-19 mortality and morbidity among underserved populations. The team is using CDEs to better understand the communities that are being engaged as a part of this effort, as well as strategies to connect with the communities. The team was interested in asking standardized questions across different populations and settings. Dr. Kibbe explained that the CDEs are categorized into two tiers and various standard categories. He noted that the project has evolved over time, and he suggested setting terms and conditions regarding data sharing and use of CDEs, as well as conveying their importance to the research community (e.g., through academic publications). His group's CDEs were developed in partnership with community members and were based on "four Ds" of success: diversity, data, development, and deliverables.

Seeking Community Feedback in the Development of Autoimmune and Immune Mediated CDEs
Sirimon O'Charoen, Ph.D., Crohn's & Colitis Foundation

Dr. Sirimon O'Charoen spoke on the importance of CDEs for the IBD Plexus® program. Inflammatory bowel disease (IBD) represents a complex collection of diseases with unknown origin that present in various heterogeneous manners, with overlapping pathogeneses among other immune-mediated diseases.

IBD Plexus is a research ecosystem that aims to accelerate progress toward precision medicine by providing researchers integrated clinical patient-reported data, with the goals of accelerating precision medicine and finding a cure for IBD. The Crohn's & Colitis Foundation is mandating the use of CDEs and other forms of structured data for integration. Current challenges include a lack of standard disease characteristics and interoperability across studies. Additionally, Crohn's & Colitis Foundation patient symptoms, outcomes, experiences, and abilities can also vary significantly as a result of their diverse experiences. Dr. O'Charoen underscored the importance of actively engaging both investigators and patients and of providing CDEs early in the research process. Important considerations include language and readability, cultural competency, acceptability, and diversity. She concluded by emphasizing the importance of a collaborative approach among researchers, advocates, and patients.

Social Determinants of Health Common Data Elements in the NIMHD Health Equity Action Network

Stuart Gansky, Dr.P.H., M.S., University of California (UC), San Francisco

Dr. Stuart Gansky presented on the use of CDEs within the National Institute on Minority Health and Health Disparities (NIMHD) Health Equity Action Network, which is studying comorbid chronic diseases in communities with health disparities and focusing on disease prevention, treatment, and management, as well as studying a wide variety of health outcomes. Centers within the network developed their own sets of integrated CDEs across their projects, with a focus on SDOH, comorbidities, and quality of life. The network's Common Data Element and Data Harmonization Working Group was charged with compiling a list of CDEs and discussing data transfer and data harmonization procedures. For this effort, the team modified existing toolkits and collections and added specific elements from individual centers or studies. This approach allowed them to develop validated items that were comparable across studies and centers but were not overly lengthy. This list was also modified in response to community feedback. Dr. Gansky also discussed the National Artificial Intelligence Research Resource pilot, which involved preparing a data warehouse to be AI ready for the NIH Science Collaborative for Health Disparities and Artificial Intelligence Bias Reduction (ScHARe) data repository. The team is working with community advisory board members to develop guidelines for involving community members in data collection and data reuse.

Discussion

- Long COVID is a complex disease that affects multiple organ systems, and physicians have noted that questions related to the disease can be overwhelming for patients. Simplified language can help address this barrier.

- By adapting existing CDEs, researchers can generate CDEs that are more relevant to the communities; however, loss of information is a concern. Community validation is crucial for ensuring that relevant information is being collected. Data that relate to action and improvement are often the most effective.

- Researchers should assure community members that their data are being protected. Several approaches for data protection can be employed, but more work is needed to determine which approach is likely to be most effective. The NIMHD ScHARe data repository is compiling a data warehouse of protected data, which researchers can request to access.

- CDEs can accelerate the path of discovery and the use of the data across the life cycle of the patient cohort. CDEs also can be used to develop clinical research networks.

- Tribal sovereignty and data sovereignty are critical considerations in the context of CDEs, and engagement with Tribal Nations in developing data agreements is crucial. NIMHD has funded a Tribal data repository that is focused on this topic.

- Patient groups are likely to change over time, and communities represent dynamic experiences. Continued engagement with these communities is important for capturing these dynamics, and trust building is crucial. In-person meetings can help foster meaningful conversations. Some repositories include capabilities for version control as communities change over time.

Day 2: Thursday, March 7, 2024

Session IV: Technical Implementation Aspects of Mapping, Transformation, and Harmonization

The National Institute of Allergy and Infectious Diseases Food Allergy Data Dictionary and CDEs
Shruti Sehgal, M.D. (Hom.), M.S., Northwestern Feinberg School of Medicine

Dr. Shruti Sehgal discussed efforts to standardize the collection of food allergy clinical and research data. The Center for Food Allergy and Asthma Research at Northwestern University has been at the forefront of epidemiological research and prevention strategies related to food allergies, which are highly complex. Dr. Sehgal explained that none of the existing coding systems for clinical terminologies adequately capture the food allergy concepts, which are determined by domain experts. The food allergy domain lacks a common base of terminology, which prevents data exchange among institutions. The Food Allergy Data Dictionary was developed to address these challenges and provide a usable community resource. This effort involves domain experts and is vetted by core team members, academic centers, and industry partners. The dictionary has helped enable other initiatives, such as standardized note templates for food allergies, new partnerships, and wireframe designs. The dictionary is helpful for comparing data across trials. The team is creating specific CDEs, rather than recreating common CDEs. Dr. Sehgal stated that her team's long-term vision is to develop a centralized resource to support data aggregation, distribution, and analysis.

Making the "Common" in CDEs More Common
Anne E. Thessen, Ph.D., University of Colorado Anschutz Medical Campus

Dr. Anne E. Thessen presented her work on promoting data integration and interoperability. She explained the importance of CDE mapping, which requires data–model alignment and value set alignment. She highlighted the National COVID Cohort Collaborative as an example of data interoperability and explained how LinkML, a modeling language, allows researchers to compare data in different formats. Dr. Thessen presented a proposed workflow for making CDEs more comparable at scale via LinkML and CurateGPT. She explained that CDEs are not as computable as researchers would like, which reduces interoperability. CDEs can be updated to address this issue. Recent efforts have focused on developing mapping standards and LLM-based tools. Ultimately, these tools can be used to formulate a curation strategy, which will provide numerous benefits for data interoperability.

National Marrow Donor Program: Harmonizing Data Through CDEs
Kathleen Malum, B.A., NMDP

Ms. Kathleen Malum spoke on harmonizing data using CDEs. She highlighted the Center for International Blood and Marrow Transplant Research (CIBMTR), which collects and maintains outcome data for clinical research. Two of CIBMTR's strategic priorities are data centralization and transformation. Most of CIBMTR's data are submitted via FormsNet, a comprehensive electronic data submission system containing more than 250 forms related to the capturing of hematopoietic cell transplantation outcomes for both donors and recipients. FormsNet does not allow for direct transmission of data from one transplant center's database, but A Growable Network Information System allows for electronic data exchange between the transplant center databases and FormsNet. CDEs are used to represent the metadata and allow unambiguous data interpretation. The team uses CDEs that contain standardized terminology concepts that define the meaning of the data. With this approach, data can be standardized across forms aligning to FAIR principles. Ms. Malum emphasized that using CDEs at the NMDP is important for CIBMTR's collection, storage, transmission, and use of data. Streamlining data collection and facilitating data sharing generates new knowledge and ultimately leads to better clinical outcomes.

Discussion

- NIH has established definitions for CDEs and ontologies, and a glossary might be useful for vocabulary alignment. Bridging knowledge gaps remains an ongoing challenge, and ontologies for the entire CDE space would be beneficial.

- Words can have different meanings based on context. It is helpful for researchers to use identifiers with a logical structure because it allows semantic similarity analysis to determine the relationships among the identifiers. Value sets that are used for the CDEs can facilitate semantic mapping.

- When developing CDEs, it is valuable to understand the mechanism of the disease, and this mechanism can be factored into CDE collections.

- Researchers working with data face two problems: preserving the data that have already been collected and optimizing future data collection. Workflows will depend on the nature and structure of the data, and a clear understanding of the data and infrastructure is essential.

- Collaboration is essential for data dictionary development, and domain experts can provide insights into best practices.

- Developing CDEs for food allergy is a critical effort in the context of the broader immune-mediated disease condition community. The data could be extended to this area in the future.

- Publicly accessible CDEs can help facilitate interoperability across repositories.

- NIH can play a role in facilitating CDE mapping and tool implementation. Automated tools and guidance would help further progress in this area. An interface could help users track CDEs that have been established across various topics.

Session V: Approaches to Improve Interoperability

Leveraging REDCap for Finding and Reusing CDEs for Making Data Interoperable
Paul Harris, Ph.D., FACMI, FIAHSI, Vanderbilt University Medical Center

Dr. Paul Harris shared an overview of REDCap, a software platform for designing research databases to support diverse clinical and translational studies. The software is shared with members of the REDCap Consortium at no cost. Consortium administrators convey feedback from users about requested updates and improvements, which are built, tested, and disseminated by the REDCap engineering team on a monthly basis. REDCap has been used to support large research programs. Dr. Harris highlighted the REDCap CDE workflow and current state. He noted that REDCap's rate of adoption has been linear. Dr. Harris also spoke on efforts toward instrument assessment and validation. He underscored the importance of engaging with researchers on new developments. Additionally, he noted that generative AI modeling can help predict CDEs of interest.

Implementation of National Institute of Neurological Disorders and Stroke (NINDS) CDEs: Lessons Learned

Jocelyn Craven, M.P.H., Medical University of South Carolina, and Sara Meyer, Medical University of South Carolina

Ms. Jocelyn Craven and Ms. Sara Meyer discussed lessons learned in incorporating the CDEs into their case report forms through StrokeNet and the Strategies to Innovate Emergency Care Clinical Trials Network (SIREN). Both networks use the NINDS CDEs in different repositories. Ms. Craven and Ms. Meyer briefly outlined their process for CDE implementation. They explained that they draft case report forms from approved protocols and via standardized forms that are used across the networks. The CDEs are added to the forms after approval is obtained from the trial principal investigator. They underscored the importance of allocating adequate time from the first request to approval. The CDEs are then incorporated into a clinical trial management system. They explained that the CDE reviews are incorporated into the electronic database change request approval and implementation process. The final step is the creation of the public-use data set. They also briefly discussed both direct and indirect mapping during the process. They highlighted a summary of submitted suggestions, which relate to usability and security.

Motivations, Challenges, and Benefits of Building and Deploying the NCI Clinical Trials Evaluation Program (CTEP) CDEs

Ginger Riley, Data Solutions Sector, Westat

Ms. Ginger Riley discussed considerations related to standardization of NCI CTEP CDEs. She explained that her team has adopted a standardized clinical data management system to help continue fostering the development of standards, as well as a standard means to optimize data standardization and harmonization. Primary areas of interest include stakeholder engagement, a lessons-learned approach, and governance. She briefly described her team's efforts in standardizing, implementing, identifying, and curating CDEs, standard CDEs, and standard case report forms. She underscored the importance of identifying and adopting standards within the areas of standardized CDEs, identifying standard case report forms, and identifying processes to support use of those standard CDEs and case report forms. Considerations include a common methodology for data capture, as well as stakeholder collaboration. Challenges have included previously established local codes, lack of existing standardization, and disparate systems. Ms. Riley outlined her team's roadmap for developing a set of harmonized case report forms for the CTEP community.

Using CDEs to Enable Harmonization of Common Data Models

Kenneth Gersing, M.D., National Center for Advancing Translational Sciences, NIH

Dr. Kenneth Gersing spoke on common data model harmonization. He highlighted the Common Data Model Harmonization project, a collaborative effort that was established to improve interoperability among models in different formats used by different communities. Previously, lack of harmonization among models led to widespread issues following model updates. Harmonization bypasses the need for remapping on a semantic basis when models change. Dr. Gersing explained that this system was important for mapping data collected during the COVID-19 pandemic, as well as EHR data. He emphasized that these efforts help bridge health care and research and increase the overall value of the data. Dr. Gersing discussed current system developments, which include developing tools for data transformation and harmonizing legacy models.

Discussion

- The time and effort needed to implement CDEs can depend on the nature of the protocol.
- Data dictionaries can serve as exchange mechanisms and currencies across repositories. The metadata are computable and can be stored as needed. Additionally, case report forms can be processed and clustered for real-world applications. The REDCap data dictionary has been stable from the beginning of the project; metadata fields have been added and are backward compatible.
- Leveraging expertise is crucial for understanding the relevance of existing content. The user community can inform development of upcoming standards, building on their own local codes and values.
- REDCap includes functionalities to help people who are building local field banks elevate them to a master field bank that could be reused with other institutional studies. It includes the ability for top-down curation and can accommodate sharing of clusters of field banks or fields. Additionally, study templates can be generated as needed.
- Metadata are important for understanding differences among patients, and facilitating patient engagement is an ongoing challenge. Additionally, communication between clinician researchers and patients is critical.

Session VI: Use Cases for Preparing and Applying CDEs for Intelligent Technologies

ScHARe AI Use Cases

Deborah Duran, Ph.D., NIMHD, NIH

Dr. Deborah Duran provided an overview of ScHARe, which was built to bring underrepresented people, including women and people of color, into data science and improve the use of big data in social science and health disparities research. NIMHD prioritizes bias mitigation efforts, and ScHARe includes a focus on community colleges and low-resource, minority-serving institutions. The ScHARe platform is a data ecosystem with federated data sets and shared CDEs focused on population science, and the program runs a number of events and initiatives focused on increasing representation and reducing barriers for people who are underrepresented in data science. Dr. Duran emphasized the importance of collaboration among smaller ICs to increase capacity. She explained that ScHARe has developed a core set of 20 CDEs, drawn from a broad review of other sources and databases, that will be required to deposit data in an upcoming repository. Dr. Duran outlined how CDEs can be mapped to existing standards to increase the representation of social sciences and health disparities data in repositories.

Defining and Developing CDEs for Use in AI/Machine Learning (ML) Applications for Clinical Research With Electronic Health Record Data

Katherine P. Liao, M.D., M.P.H., Harvard Medical School

Dr. Katherine Liao provided a clinical investigator’s perspective on CDEs, outlining an AI/ML project to study treatment response in rheumatoid arthritis. Patients often cycle through multiple treatments before clinicians can determine which drug is most effective, but most clinical trials are not designed to answer broad questions about treatment response. Dr. Liao’s team has access to a large database of primary patient data across several institutions, but several CDE-related challenges have become apparent. An algorithm is needed to define rheumatoid arthritis in the EHR because the diagnosis codes are not standardized, and treatment response is not recorded with codes. Any algorithms developed also must be harmonized across institutions. CDEs identified across institutions can be used to build a knowledge network to leverage and standardize patient data with AI/ML strategies. Dr. Liao noted that the CDEs used to evaluate data will evolve over time as technology evolves.

The Possibilities and Challenges of Using CDEs in Making Data AI Ready and the Future of AI in Achieving Data Interoperability

Sally Baxter, M.D., M.Sc., UC San Diego School of Medicine

Dr. Sally Baxter commented on how to use CDEs to make data AI ready, noting the example provided by Dr. Liao of the use of CDEs in AI algorithms. The ophthalmology field uses a significant amount of imaging data, but standardization and interoperability have been difficult to develop because many different devices and proprietary systems are used. Harmonization is critical to analyze large and diverse sample sizes. Dr. Baxter explained several efforts in the ophthalmology field to develop CDEs and data standards with multidisciplinary teams and diverse types of data. Using AI to develop CDEs is not a mature area of research in the field, and much of the work is manual and labor intensive. LLMs are being explored, but transparency is a major issue. Other challenges in the field include gaps in representation, lack of standard terminology in eye exam findings, and use of complicated data elements. Dr. Baxter pointed out that many efforts are in progress that will address data and health disparities, especially using standardized tools and definitions.

Discussion

- Local codes can be mapped manually or with natural language processing, but there is no single best system. CDEs are important because the best system will always depend on the needs of a particular project. EHR vendor engagement is necessary to improve interoperability across systems.

- Natural language processing could be used to assess whether the gender of patients and providers affects treatment, but data must be interoperable before such questions can be addressed.

- Data validated in some clinical trials may not be in the ideal format to use for a different study, so trial findings might need to be recapitulated.

- The ScHARe repository that will open soon will be based on CDEs and focus on common interoperability, but other data can be included.

- Imaging-based phenotypes can be generated from segmentation algorithms, but many imaging metrics in ophthalmology are not represented in standardized terminologies, so the lack of visibility in many algorithms may complicate the results. This area is in its infancy and is likely to evolve.

Closing Keynote

Victoria Shanmugam, M.B.B.S., MRCP, FACR, CCD, Office of Autoimmune Disease Research, NIH

Dr. Victoria Shanmugam explained that autoimmune diseases affect many people in the United States, and many of the affected are women, but longitudinal data repositories that could be studied to explain this discrepancy are lacking in the United States. Recent data suggests that Xist, a molecule responsible for inactivation of the additional X chromosome in women, may play a role in autoimmunity. Immune activation can release Xist from the chromosome, and this seems to be associated with autoimmunity but The number of X chromosomes is correlated with immune response; in some circumstances, the inactivation of one X chromosome that occurs early in development can be lost in immune cells, but the cause is unknown. Time data are needed to be able to assess how autoimmunity changes lead to these conditions. Several existing registries and databases focus on specific autoimmune conditions, but answers to these questions may require integrating information about the genome, microbiome, exposome, and immunome.

Closing Remarks and Adjournment

Steve Tsang, Ph.D., ODSS, NIH

Dr. Steve Tsang summarized the conference and commended the broad participation from organizers and attendees representing many disciplines. Crosscutting themes included the need to be aware of the community when developing CDEs, the need for better collaboration and communication and platforms through which to do so, and the importance of emerging technologies. Dr. Tsang thanked attendees and invited them to provide input on the RFI and the future of CDEs.