MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230916

ARDaC Common Data Model Facilitates Data Dissemination and Enables Data Commons for Modern Clinical Studies

Nanxin JIN[a,b], Zuotian LI[a,c], Carla KETTLER[b], Baijian YANG[b], Wanzhu TU[a] and Jing SU[a,1]

[a]*Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana, USA*

[b]*Computer and Information Technology, Purdue University, Indiana, USA*

[c]*Computer Graphics Technology, Purdue University, Indiana, USA*

ORCiD ID: Nanxin Jin https://orcid.org/0000-0003-3609-374X, Baijian Yang https://orcid.org/0000-0003-4440-3701, Jing Su https://orcid.org/0000-0003-4917-6173

Abstract. Modern clinical studies collect longitudinal and multimodal data about participants, treatments and responses, biospecimens, and molecular and multiomics data. Such rich and complex data requires new common data models (CDM) to support data dissemination and research collaboration. We have developed the ARDaC CDM for the Alcoholic Hepatitis Network (AlcHepNet) Research Data Commons (ARDaC) to support clinical studies and translational research in the national AlcHepNet consortium. The ARDaC CDM bridges the gap between the data models used by the AlcHepNet electronic data capture platform (REDCap) and the Genomic Data Commons (GDC) data model used by the Gen3 data commons framework. It extends the GDC data model for clinical studies; facilitates the harmonization of research data across consortia and programs; and supports the development of the ARDaC. ARDaC CDM is designed as a general and extensible CDM for addressing the needs of modern clinical studies. The ARDaC CDM is available at https://dev.ardac.org/DD.

Keywords. Common data model, data commons, clinical study

1. Introduction

Modern clinical studies such as randomized clinical trials (RCT) and observation studies acquire longitudinal clinical information about participants; follow up participants for clinical events or their responses to interventions; collect and biobank biospecimens and generate multiomics data from biospecimens to reveal the molecular underpinnings of the observed intervention responses and clinical events. One example is the RCT and the observational study conducted by the Alcoholic Hepatitis Network (AlcHepNet) consortium sponsored by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and aiming to improve the treatment of severe alcoholic hepatitis. The two

[1] Corresponding Author: Jing Su, email: email: su1@iu.edu.

AlcHepNet studies recruited 1,700 participants from 8 clinical sites, followed up participants for up to 180 days. The studies are capturing demographic and behavioral information, clinical conditions, laboratory tests, treatments, and outcomes. Over 24,000 blood, urine, saliva, and liver biopsy biospecimens are collected and sent to 10 translational research projects to generate multiomics data including microbiome, immunologic, proteomic, metabolomic, lipidomic, RNA-seq, and ChIP-seq data. The Indiana University Data Coordinating Center (IU DCC) and the University of Massachusetts Data Coordinating Center (UMass DCC) collaboratively provide essential research infrastructure such as data management and statistical analysis.

The rich and complex data generated in modern clinical studies requires novel common data models and data modeling tools to support data standardization, harmonization, query, retrieval, and interoperable reuse following the FAIR principles [1]. Three data domains involved in such studies are: clinical data about the participants, metadata about biospecimens, and molecular and multiomics data generated from biospecimens. The National Cancer Institute's Genomic Data Commons (GDC) [2,3] CDM provides a graph data model to represent the relations of data elements belonging to these three data domains. However, the GDC CDM is not ready to support modern clinical studies, including clinical study designs such as randomized trials, case-control studies, and cohort studies, and longitudinal biospecimen collections in follow-up visits, and trajectory analysis of multiomics profiles. Data modeling tools that can bridge the gap between the data models used in data coordinating centers and the GDC CDM are significantly underdeveloped.

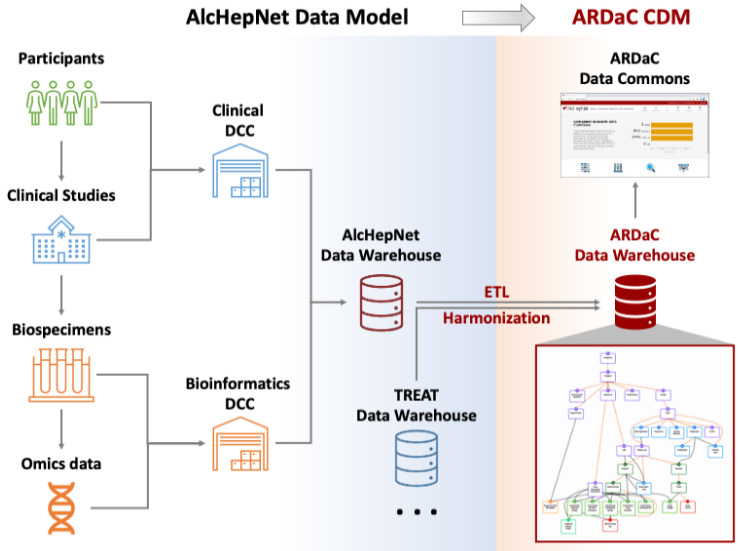

Figure 1. Data flow of ARDaC ETL and harmonization.

To facilitate the effective research use of the rich and complex data generated by modern clinical studies, we developed a novel CDM to enable the development of the Alcoholic Hepatitis Network Research Data Commons (ARDaC), which serves as the central data hub and research nexus for the AlcHepNet consortium. The ARDaC CDM supports common clinical study designs and longitudinal biospecimen collection. It accommodates rich clinical and behavioral features as well as clinical events and outcomes such as acute kidney injury, sepsis, severe adverse events, liver transplantation, and death. The ARDaC CDM is compatible with GDC CDM and allows the development of research data commons. We also developed the ARDaC CDM ETL toolkit to streamline the data extract, transform, load (ETL), and harmonization from different sources to the ARDaC CDM. As a general and extensible data model, the ARDaC CDM supports the data dissemination for modern clinical trials. We implemented ARDaC CDM on the AlcHepNet RCT and observational study as well as the Translational Research and Evolving Alcoholic Hepatitis Treatment (TREAT) observational study to demonstrate its functions and performance.

2. Methods

2.1. Data Flow

The overall data flow is illustrated in Figure. 1. Briefly, participants are recruited into clinical studies at a clinical site, with clinical and behavioral information, treatment, responses, laboratory results, and clinical events collected at baseline and follow-up visits into REDCap and hosted at IU Clinical DCC. Meanwhile, biospecimens are collected at baseline and during follow-up visits as planned and distributed to translational labs to perform molecular tests or to generate multiomics data. Information about biospecimens as well as the generated data are sent to UMass Bioinformatics DCC. The data from the two DCCs are then merged into the AlcHepNet Data Warehouse for downstream data analysis. The REDCap-based AlcHepNet Data Model, which is specialized for the operation of clinical studies, is used for data capture and management. The ARDaC ETL and Harmonization tool is developed to bridge the gap between the AlcHepNet Data Model and the ARDaC CDM. Similarly, other NIAAA-funded clinical studies such as TREAT are ETL'ed to the ARDaC CMD. The harmonized data are then stored in the ARDaC Data Warehouse and disseminated through the ARDaC research data commons. Table 1 summarizes the current data size of both the AlcHepNet Data Warehouse and the ARDaC Data Warehouse. As both the AlcHepNet clinical studies and the incorporation of previous clinical study data are ongoing, the ARDaC Data Warehouse is quickly growing. In this work we include the AlcHepNet RCT (NCT04072822) and observational study (NCT03850899) as well as the TREAT observational study (NCT02172898).

2.2. ARDaC CDM

The ARDaC CDM is compatibly extended from the GDC data model by introducing new entities and remodeling the graph topology (Figure 2 A). Briefly, new nodes such as AUDIT for drinking behaviors, Lab for AlcHepNet translational projects, and Study for clinical studies are introduced. The major topology change is to move nodes Aliquot and Molecular Test under the node Follow Up, which enables longitudinal collections of biospecimens and data generation. Also introduced are clinical study design features such as treatment arms in RCTs and cohorts in observational studies; lifestyles such as drinking behaviors; clinical events such as acute kidney injuries and sepsis; key liver disease indicators such as MELD Score, Child-Pugh Score, Lille Score, and Maddrey's Discriminant Function Core; and socioeconomic determinants of health such as employ status and education level. New controlled vocabularies about aliquot types, molecular tests of liver functions, kidney functions, and cardiovascular risk indicators, sepsis indicators such as blood culture results, as well as new omics data types such as lipidomics, metabolomics, microbiomics, and ChIP-Seq are also included. The NIH Common Data Elements are used for new properties and controlled vocabularies when applicable.

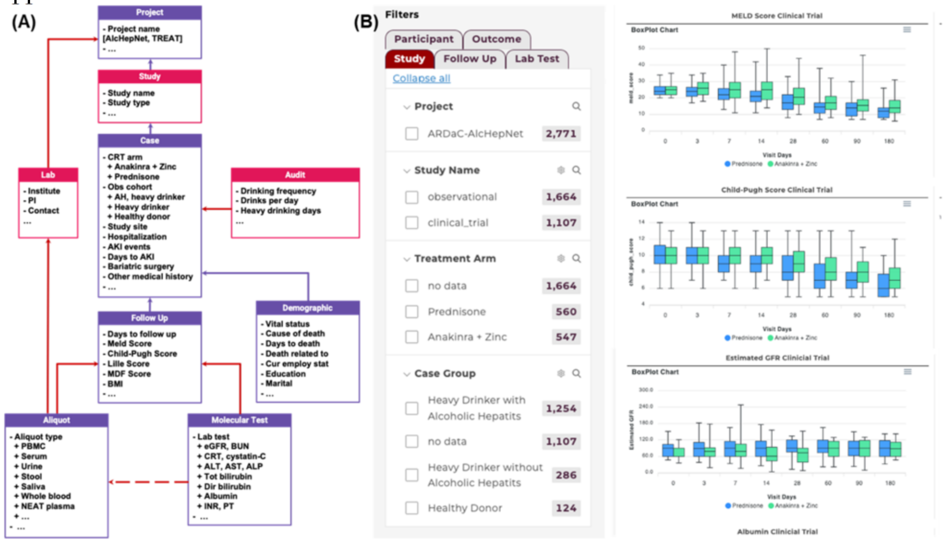

Figure 2. (A) Major modifications in ARDaC CDM. (B) ARDaC CDM enables study design features.

The data model and de-identified data is hosted on Amazon AWS. The data harmonization transfers the data to meet the data model requirement. All datapoint failures are reported to Quality Control (QC) group. The latest version of the ARDaC CDM is available at https://github.com/jing-su/ardac-cdm.

3. Results

Table 1. Data entities in AlcHepNet Data Warehouse and ARDaC Data Warehouse.

Data entities	AlcHepNet Data Warehouse	ARDaC Data Warehouse
New entities	9 tables 781 columns 68,122 rows	10 nodes 8 revised edges 232 properties 109 controlled vocabularies
Total entities	39 tables 2,507 columns 104,309 rows	33 nodes 49 edges 648 properties 217 controlled vocabularies 72,480 rows

The new data entities as well as the overall summarization of the AlcHepNet data and the harmonized ARDaC data are provided in Table 1, with the major changes highlighted in Figure 2 (A). The ARDaC enables the functionalities of query and visualization of study designs (Figure 2 B, left and middle panels) and the trajectory comparison of clinical indicators across different RCT arms or observational study cohorts (Figure 2 B, right panel of the longitudinal liver disease scores and kidney function indicator between the two RCT arms) on the ARDaC data commons.

4. Discussion

4.1. Novelty and Technical Significance

This is the first CDM that supports modern clinical trials. The incorporation of clinical study designs, the support of longitudinal biospecimen collection, and wide coverage of rich clinical and longitudinal information are accomplished by the backward-compatible extension of the GDC CDM and the topological modification of the graph structure of the original schema. This allows non-destructive conversion of GDC CDM to ARDaC CDM, and the interactions with other Gen3-based data commons to ARDaC.

4.2. Significance in Data Dissemination of Clinical Studies

The ARDaC CDM enables the development of ARDaC for harmonizing various NIAAA-sponsored clinical studies for data dissemination and the incorporation of other clinical studies, paving the way toward clinical study data commons.

4.3. Perspectives

We are incorporate data elements from the Observational Medical Outcomes Partnership (OMOP) CDM to support the use of electronic medical records in AlcHepNet studies. We are also extend the current ARDaC CDM to support the data dissemination of the multidomain data collected and generated during regular clinical care such as the electronic medical records of patients, the biospecimens in local tissue banks, and the multiomics data generated from such biospecimens.

5. Conclusions

The ARDaC CDM meets the urgent needs of the data dissemination in modern clinical studies; bridges the gap between the data models used in clinical studies; and enables the development of research data commons for clinical studies.

Acknowledgements

The work is sponsored by NIH grants (U24AA026969, R01LM013771, and P30CA082709) and the Indiana University Precision Health Initiative.

References

[1] Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, Pétavy F, Galvez J, Becnel LB, Zhou FL, Harmon N. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform. 2020 Jul;107:103421, doi: 10.1016/j.jbi.2020.103421.

[2] Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, Bajari R, Baqar B, Barnowski JH, Burt J, Catton A. The NCI genomic data commons. Nat Gen. 2021 Mar;53(3):257-62, doi: 10.1038/s41588-021-00791-5.

[3] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017 Jul;130(4):453-9, doi: 10.1182/blood-2017-03-735654.