The Alcoholic Hepatitis Network Research Data Commons (ARDaC): Design and Development

Jing Su, PhD[1], Nanxin Jin, MS[2], Zuotian Li, MS[3], Carla D. Kettler, BS[1], Bruce Barton, PhD[4], Greg L. Puetz, BS, Chi Mai Nguyen, PhD, Donna McGrath, MS[4], Yingjie Chen, PhD[3], Baijian Yang, PhD[2], Vijay Shah, MD[5], Svetlana Radaeva, PhD[6], Samer Gawrieh, MD[7], Wanzhu Tu, PhD[1]

[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana, USA; [2]Department of Computer and Information Technology, Purdue University, Indiana, USA; [3]Department of Computer Graphics Technology, Purdue University, Indiana, USA; [4]Department of Population and Quantitative Health Sciences, University of Massachusetts, Worcester, Massachusetts, USA; [5]Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA; [6]National Institute on Alcohol Abuse and Alcoholism, Maryland, USA; [7]Division of Gastroenterology and Hepatology, Department of Medicine, Indiana University, Indiana, USA

Abstract

The Indiana University Data Coordinating Center is developing ARDaC, the Alcoholic Hepatitis Network (AlcHepNet) Research Data Commons, to facilitate the effective use of the rich and complex AlcHepNet multimodal data and synergize efforts and expertise within the consortium and beyond. ARDaC provides a comprehensive solution for representing, querying, visualizing, and analyzing clinical, biological, molecular, multiomics, and behavioral data for alcoholic hepatitis research. ARDaC is the central data hub across AlcHepNet clinical and translational teams. It functions as an engine to drive AlcHepNet research projects, the data interface between AlcHepNet consortium and other research data commons, and the research nexus to ignite new research and collaborations.

Introduction

Alcoholic hepatitis is a leading cause of liver-related morbidity and mortality in the United States. The AlcHepNet project, sponsored by the National Institute on Alcohol Abuse and Alcoholism (NIAAA), aims at improving the treatment of severe alcoholic hepatitis. The AlcHepNet consortium is composed of 8 clinical sites. It conducts a randomized clinical trial and a large observational study. It also provides data and biosamples to 10 translational research projects, including basic science and preclinical studies. The network is recruiting over 1,700 participants for clinical studies, following participants for up to 180 days, collecting more than 24,000 blood, urine, saliva, and liver biopsy biosamples, capturing demographic and behavioral information, clinical conditions, laboratory tests, treatments, and treatment outcomes, as well as generating 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 the essential research infrastructure, including experimental design, study implementation, data management, and statistical analysis in support of the two primary studies within the consortium as well as the translational projects that utilize the biospecimens collected by the two primary studies. To facilitate the effective research use of the rich and complex data generated by the AlcHepNet, we are developing ARDaC, the Alcoholic Hepatitis Network Research Data Commons, as the central data hub and research nexus.

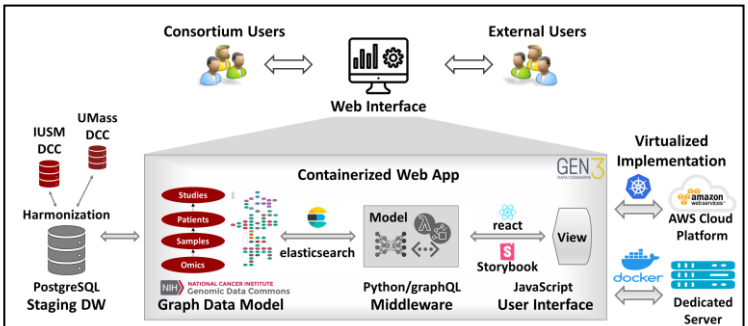

Figure 1. Architecture of ARDaC.

Design of ARDaC

The ARDaC design is in full accordance with the guiding principles of FAIR (Findable, Accessible, Interoperable, Reusable). The architecture of the ARDaC system is demonstrated in Figure 1, with the following components:

1) The ARDaC Data Warehouse (Figure 2). The heterogeneous information about clinical features, biospecimens, and omics data is extracted from the IU DCC and UMass DCC, standardized according to the ARDaC Data Standard, harmonized according to the ARDaC Common Data Model (CDM), and hosted in a central ARDaC Data Warehouse. Specifically, the novel ARDaC CDM is derived from and compatible with the Genomics Data Common (GDC)[1] Data Model and is compliant with the FAIR Principles[2]. The ARDaC Data Warehouse is the data source for the ARDaC web application open to the public as well as for regular reporting and customized services within the AlcHepNet consortium. A graph-based provenance model supports comprehensive data

dependency and version control so that the ARDaC digital entities, including the standards, data model, data and metadata, and scripts, are attributable, trackable, and reproducible.

2) The ARDaC web application. The ARDaC system uses the Gen3 data common framework, which has been widely used in NIH-sponsored projects. At the data layer, the standardized and harmonized data is extracted from the ARDaC Data Warehouse (Figure 2) and injected into the ARDaC Staging Data Warehouse (Figure 1) according to the ARDaC Graph Data Model. In the middleware layer, based on the user's input of the filtering criteria, the graph-based data is queried using graphQL through the elastic search engine, analyzed with Python, and delivered interactively to users using the JavaScript-based React libraries and Storybook. The ARDaC web application is containerized as a series of images, each providing a specific service. The ARDaC system can be and deployed to the AWS cloud services through the Kubernetes platform or to a dedicated server through the Docker platform.

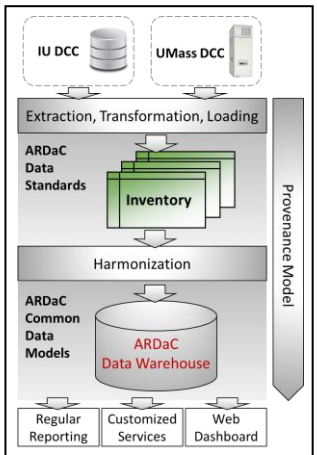

Figure 2. Data standardization and harmonization.

By leveraging the GDC common data model and the Gen3 data commons ecosystem, ARDaC enables the data integration with other NIH-funded data commons and delivers broad impact of AlcHepNet research and data to other research communities. The new common data elements (CDEs) introduced by ARDaC address the research needs in similar studies.

Essential functionalities.

The ARDaC system supports the representation of behavioral and pathologic data unique to alcoholic hepatitis, facilitates data querying, visualization, and exploring which are specific to the AlcHepNet clinical studies (Figure 3). A key novelty factor is that the system is designed for the unique needs of alcoholic hepatitis research and similar studies. As a disease stemming from human behavior, ARDaC incorporates not only biological data, but also human behavioral data such as various measures of alcohol consumption are used to characterize the history, current use, abstinence, and relapse. Besides the general data query and visualization functions, ARDaC allows data exploration using the study-related criteria such as the study cohorts or arms, alcohol use history, alcoholic hepatitis treatments, prognosis information such as mortality and liver transplantations, liver functions such as MELD's scores, omics data availability and biospecimens availability, and omics-derived features such as differentially expressed genes and enriched signaling pathways. ARDaC system also provides GraphQL query interface, cloud-based workspace, and R and Python programming environments for in-depth data analysis. As a research nexus, if researchers are interested in proposing new data generation plan, ARDaC allows users to explore sample availability, to visualize and evaluate the synergy of their data generation plans with existing data and funded projects, and to plan for new data generation. The ARDaC system is available at github.com/jing-su/ardac.

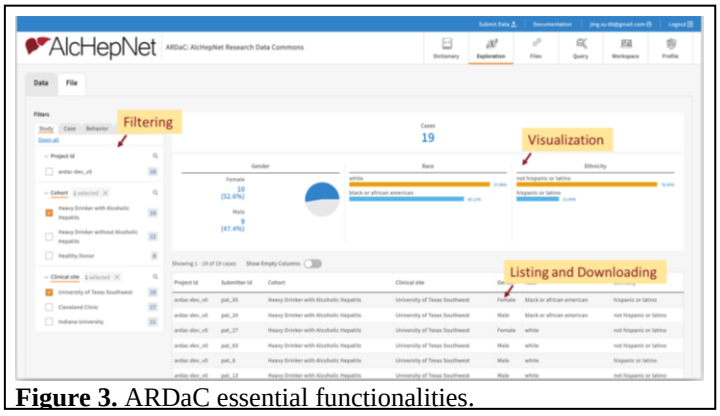

Figure 3. ARDaC essential functionalities.

Conclusion

In summary, ARDaC is the central data hub connecting data of multiple modalities across clinical and translational teams, the engine to drive AlcHepNet research projects, the data interface between AlcHepNet consortium and research other data commons, and the research nexus to ignite new research and collaborations. While the data commons is designed for alcoholic hepatitis researchers, the basic design and functionality could be extended to other disease areas and potentially be used by a broader group of researchers.

References

1. Heath AP, Ferretti V, Agrawal S, et al. The NCI Genomic Data Commons. Nat Genet 2021;**53**(3):257-62 doi: 10.1038/s41588-021-00791-5[published Online First: Epub Date]|.
2. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review. JMIR Res Protoc 2021;**10**(2):e22505 doi: 10.2196/22505[published Online First: Epub Date]|.