Biomedical Data for AI

Lead: Jihad Obeid (MUSC)

Despite the exponential growth in biomedical data, the provision of AI-ready data has several challenges. Real-world data from the electronic health record (EHR) and other data sources, such as omics, environmental, social determinants of health, and remote monitoring devices, is inherently complex and requires significant expertise to analyze and prepare for research. Moreover, the quality of the data is particularly challenging since the source systems are optimized for health care but not research or AI.

Researchers must comply with privacy and confidentiality guidelines under ethical and regulatory mandates when using patient data. Investigators must also navigate IRB approvals when appropriate and comply with HIPAA requirements when using protected health information.
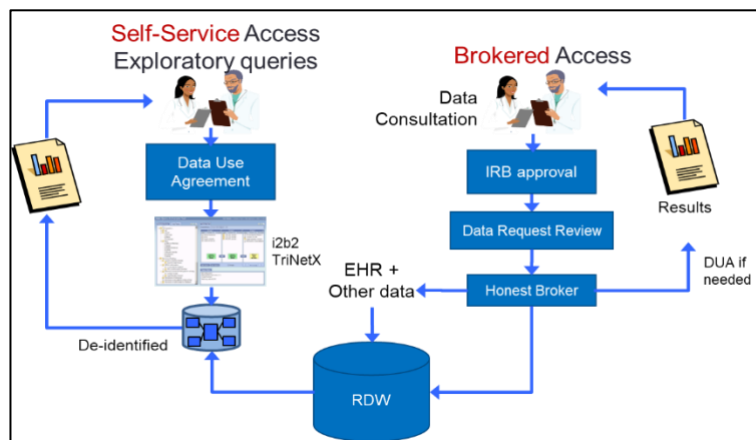

Figure 3. Data governance overview (RDW is research data warehouse)

Overview, Status, and Key Results: This activity addresses biomedical data readiness for AI (e.g., data acquisition, processing, curation, tools, standards, and preparation for AI models). We are leveraging the EHR research infrastructure and data governance processes at the CRUs as shown in Figure 3 to establish services and expertise for data delivery. As part of this effort, we are training researchers on best practices for AI-ready data acquisition.

The first project-wide half-day biomedical data training session was held on 11/29/2023. The agenda included: EHR data infrastructure for research, MUSC data governance, the data request process, IRB requirements and IRB reliance across the state, regulatory compliance, data sharing and data use agreements across institutions, and a panel discussion ending with a Q&A session.

One-on-one data services and training will be provided on demand and tailored to specific research projects. We are also currently compiling an inventory spreadsheet of publicly available clinical and biomedical data sets for educational purposes. These data sets will be made available to students in AI across the state.

Relevance/Significance of work: Biomedical data is the fuel that enables AI research for the advancement of personalized and transformative healthcare in South Carolina. However, researchers must comply with privacy and confidentiality guidelines under ethical and regulatory mandates when using patient data. Expanding best practices across the state will democratize access to clinical data essential for training and testing AI models for biomedical devices while ensuring data security and protection of patient privacy and confidentiality.