Research on Data Governance in Guangdong Province of Healthcare Big Data Platform Based on Common Data Model

ZHANG Hongzheng, LI Lin, LIU Mimi, CHENG Yinlin, ZHOU Yi
Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

Background

In order to promote the application of personalized healthcare big data and artificial intelligence in clinical diagnosis and treatment and health management, explore the mechanism of healthcare data sharing, openness and payment, create an environment for data circulation and innovative application, drive the new business form of digital economy, and promote the digital transformation of healthcare field with big data, this study established a clinical big data research and analysis platform. Based on the common data model, this platform standardizing multi-center data with complex types, different standards and uneven quality through data extraction and cleaning, text data structuring and terminology mapping, provides high-quality data for clinical research across institutions and departments, and realizes data interconnection, sharing and utilization. It also provides experience and reference for multi-center healthcare big data governance.

Methods

The platform obtains basic information, medical information, diagnosis information, medication information, test information, surgery information, text information and other data from different information systems in each data center(Shenzhen Luohu hospital, Sun Yat-sen memorial hospital, the sixth affiliated hospital of Sun Yat-sen University), and performs data encryption and data desensitization. The data obtained included structured data and text data. The structured data were directly extracted, cleaned, and quality checked, while the text data were transformed into structured data using natural language processing technology for entity identification and relationship extraction. After preprocessing the structured data and text data, the diagnostic, surgical, pharmaceutical, laboratory and other data were referred to the Standard terminology set to formulate the Standard Operation Procedure (SOP) of terminology mapping, and the mapped data were reviewed by medical experts. Mapping qualified data is common data model data. Data governance processing is showed in figure 1.

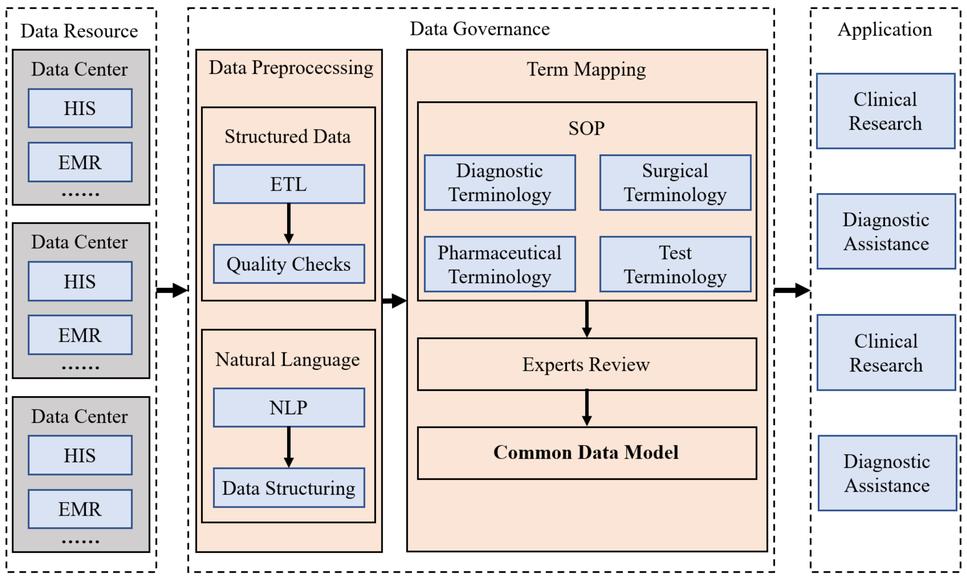

Figure 1: Data governance processing based on common data model

Results

Through data governance based on the common data model, the platform(showed in figure 2) gathers health and medical data from three medical institutions, including 1.31 million patient data, including about 120 million inpatients, 1.17 million outpatients, 90 million surgical patients, and about 30 million examination data. The platform has functional modules such as data overview, exploration and discovery, cohort discovery, and scientific research management, which can support researchers to efficiently and conveniently study, count, manage and analyze patient data, improve research efficiency and expand the scope of research. In terms of data overview, it supports the statistics and visualization of the full data of the platform and the number of patients in a specific cohort, the number of inpatients, the number of outpatients, the number of surgeries, the number of examinations, gender, age, geographical distribution and other data. The data are presented in a variety of charts and charts, and researchers can quickly understand the overall situation of the data.

Figure 2: Platform interface

Conclusions

In the data governance practice of the platform, the real-world multi-center healthcare data standardization and quality improvement have been achieved. Through the development of different data governance standard operating procedures, the health care data of different medical institutions with different data quality and data structure are converted into a common data model format, so as to provide high-quality and reliable data support for real-world multi-center health care research. However, in the process of data governance, there are still problems such as ambiguity of Chinese text data and incomplete data, which are also common problems in the current real-world multi-center health care big data governance research.