How to Improve the Reuse of Clinical Data-- openEHR and OMOP CDM

Bei Li¹,²,* and Rich Tsui²

¹Department of Biomedical Informatics, Life Science School, Central South University, No. 172, Tongzi Road, Changsha 410013, China
²Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, 2716 South Street 19104, Philadelphia, Pennsylvania, United States

*Corresponding author email: 802599@csu.edu.cn

Abstract. All medical big data reuse projects are faced with the challenging problem of collecting and transforming heterogeneous data from different sources in a distributed research network. Both openEHR and OMOP CDM are open source tools for medical data. In this paper, the principles, feasibility, implementation, and characteristics of the two main clinical data secondary use methods are compared and discussed. We analyzed two data conversion frameworks in the medical data secondary utilization project conducted in China and the United States, and summarized the experience of designing the data ETL process, and compared the principles, implementation, characteristics between openEHR-based data acquisition system and reusing medical data approach based on Common Data Model with literature. OpenEHR from the Scandinavian countries is one of promising two-level modeling approach to extract data from various medical databases. It separates the operations of medical experts and software engineers, and changes in medical knowledge can be embedded in the new prototypes without affecting the EHR system. However, some shortcomings overshadow its advantages, such as poor compatibility with medical data other than EHR, difficulties in defining prototypes, steep learning curve, and the lack of mature development tools and guidelines. We adopted a minimalist data transformation model in Xiangya medical big data acquisition system based on openEHR to solve the large-scale data exchange problem faced by the distributed clinical data center. Many experimental projects have proved the feasibility and utility of OMOP CDM for multiple, disparate health databases. This is why it is widely used for the model framework of patient-level prediction and safety surveillance, including a transformation from source data into standard vocabulary, which solves semantic interoperability; technology neutrality that does not rely on special computer technology; open community, open resources, free tools; generating aggregated analysis results directly from desensitized data, etc. Some issues should be under consideration in the use. Not all source data encodings can be converted to standard vocabulary, and there will be a loss of semantics, and concepts matching requires a lot of time and effort. The model and vocabulary were originally developed and designed for pharmaceutical safety research and clinical observation data, while the development of vocabularies in other fields is limited. In conclusion, both openEHR and CDMs are designed for exporting and reusing data from a distributed clinical database. The former is suitable for collecting data from distributed EHR systems and building medical big data warehouses, while the latter is a better model for sharing data in some decentralized medical database.

Keywords: openEHR; OMOP CDM; Clinical data reuse.
1. Introduction
The value of medical big data lies in the reuse of data. However, with the rapid growth of data, it becomes more and more difficult to find useful information and knowledge from massive medical data. With the universal application of EHR (electronic medical records) and the continuous expansion of hospital information systems, the secondary use of clinical data faces more challenges. Recognizing the multiple potential benefits and numerous difficulties of clinical data reuse, several institutions have proposed recommendations. The American Medical Informatics Association has published a white-paper listing recommendations for national framework and the European initiative proposed similar recommendations for the trustworthy reuse of health data [1]. In 2016, China outlined in the "Healthy China 2030 Plan" that medical big data should not only be limited to the diagnosis and treatment process of hospitals, but also be integrated into the whole human health field, including drug development, clinical diagnosis and treatment, insurance payment and commercial insurance design, healthcare management and public health services, and it is suggested to strengthen the integration and sharing of clinical and scientific data resources to improve the efficiency of scientific research applications [2-3]. Unfortunately, unsatisfied data quality issues in multisite environments can caused by variation in data capture, differences and changes in medical coding, local clinical workflows and so on. Even if the platforms and repositories make clinic data sharing technically possible, the practices of data reuse are not in place [4].

2. Barriers Against Reuse of Clinical Data

2.1. Massive and Ugly Clinical Data
Clinical data is the original electronic record of medical documents, which is the original ecology without processing. The integrated data classified according to medical documents is usually the information occurred and recorded in the process of hospital’s internal medical business. This kind of observational data centered on a single patient's longitudinal clinical electronic file is mainly used by clinical doctors to collect, share, and consult. The problems of these data include: (1) fragment patient records and data losing; (2) data entry errors; (3) biased and underspecified diagnostic and procedure codes; (4) Clinical data stored in multiple application systems (LIS, PACS, ECG, etc.) including many unstructured data [5-6].

2.2. A Higher Standard Research Data
Obviously, the protocol-centric research data has a higher standard for data sharing than disease-centric clinical data [7]. Research data capture information in much greater depth and in maximally structured form than clinical data. There are the following characteristics: (1) the data is traceable, and the quality of the data is rigorous; (2) structured and tabular data are required; (3) a CRF (case report form) is established for each topic to collect data separately, and the process will terminate till the end of the project. Research results must be drawn from many disparate data sources and compared and contrasted to understand the effect of potential capture bias. Although there are huge expectations for the processes of reusing data, there are also important challenges. The secondary use of clinical data faces multiple difficulties from integration, interoperability and semantic sharing issues when integrating heterogeneous and different dimensions of information into a common platform, data warehouse or network [8]. This raises a question: what kind of approach is suitable for secondary use from clinical data to scientific data?

3. Methods
The author has participated in two medical big data projects in China and the United States, so we can compare the two different ways of secondary application of clinical data at Central South University in China and Children's Hospital of Philadelphia in the United States and analyze the path and characteristics of two models. In 2015, Central South University established a medical big data platform with clinical data form four affiliated hospitals. According to OpenEHR, we adopted the two-level modeling method to design an acquisition system to collect clinic data. First, to collect the data related to 100 specific diseases into a
big data platform. The medical data acquisition system has been developed and is in service. A total of 93,353 data items and 6,017 categories were created for 285 specific diseases.

In 2019, the Department of Biomedical and Health Informatics in Children’s Hospital of Philadelphia plans to convert clinical data into OMOP CDM. The data includes birth_certificate, death_certificate, delivery, person_alias, diagnosis, drug, encounter, lab, pharmacy, procedure, etc. We design the following data reuse steps.

- White Rabbit scans the local data tables, fields, and content, then creates a detailed report containing necessary information on the tables, fields, and values that appear in a field.
- Rabbit-in-a-hat reads and displays a White Rabbit scan document, we are ready to begin designing and writing the rule to map the clinical data to CDM.
- To extract the mapping from non-standard source codes to standard source codes, we use the SQL to connect different tables. Source records of the patient are modified and gathered in a temporary data structure before transforming it and writing to the related data fields in the CDM. Python scripts and SQL queries are developed to the ETL process.

4. Result and Discussion

Regardless of openEHR and CDMs are considered as data models [9], data standardized frameworks [10], data harmonization tools[11], open-source tools for medical data [12], semantic web applications [13], data interchange methods and protocols[14], and so on, we prefer to define them as pathways to achieve secondary use of medical data in this paper. Even if the existence of multiple platforms is inevitable, thinking carefully about the characteristic of various tools should be encouraged. We can learn more from the comparison of two different methods of reusing medical big data.

4.1. OpenEHR

OpenEHR from the Scandinavian countries is one of promising two-level modeling approach to extract data from various medical databases [15]. In the two-level modeling, the RM and AM play different roles. We define the basic logic structures and attributes needed for clinic data in the RM, including data types, data structures, and components of an EHR [16]. We define domain concepts in the archetype which means the archetypes can be modified without changing the RM [17]. It separates the operations of medical experts and software engineers, and changes in medical knowledge can be embedded in new prototypes without affecting the EHR system.

However, some shortcomings overshadow its advantages, such as poor compatibility with medical data other than EHR [18], difficulties in defining prototypes [19], steep learning curve, and the lack of mature development tools and guidelines [20]. So, we design a new data transformation method based on the openEHR. After clinical data were integrated on the medical big data platform, research groups extracted the required data from the platform and converted the clinical data into scientific research data[Figure 1]. Future research areas include extending the model that incorporate additional reference vocabularies into standard terminology dictionary and further explore better editor. This effort provides a reference for other similar large-scale, multi-center medical data utilization [21].
Figure 1. The data acquisition system based on OpenEHR in Central South University Medical Big Data[17].

4.2. OMOP CDM

Common data models (CDMs) are often used in research when there is a need to exchange or share a set of data for some particular use [22]. Sharing data via a CDM decreases some obstacles of data reuse and increases the possibility that the data will be used to answer additional scientific questions. The concept behind this approach is to transform the data contained in disparate databases into a common format (data model), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format [23]. The use of a standard terminology dictionary is a critical component in the development of a CDM. Multiple CDMs have been developed to support secondary use of healthcare data in research, while the predominant public CDMs are i2b2, OMOP CDM, Sentinel, the PCORnet CDM [24].

OMOP CDM is one of the most famous CDMs developed in recent years. OMOP (Observational Medical Outcomes Partnership) was a public-private partnership, chaired by the US Food and Drug Administration, administered by the Foundation for the National Institutes of Health, and funded by a consortium of pharmaceutical companies. The companies, in collaboration with academic researchers and health data partners, has set up a research program that sought to advance the science of active medical product safety surveillance using observational healthcare data [25].

The general process of ETL in medical data is provided by OMOP, which is the most fundamental and core step is to map and transform the source data into the OMOP CDM. The process is divided into the following steps: vocabulary mapping, data tables mapping, transform and load the local data into the CDM, validation of integrity, and equivalence to the local source data. The available standardized terminology, including 81 vocabularies used in CDM, ensures syntactic and semantic interoperability, which makes studies comparable and analysis methods reusable. Figure 2 shows the mapping relationship between the clinic data table in CHOP and the OMOP CDM schema, which is the basic step of ETL.
When building the ETL, the data were first categorized with an open-source tool, and all tables, fields, and distinct values in those fields were listed. Then you can document each ETL with a tool, such as WhiteRabbit and Rabbit-in-a-hat, which allows the users to connect data tables and columns from the raw dataset to where they will map into the OMOP CDM dataset. Then, a CDM Builder program was developed to transform raw data into the CDM. The development of the ETL makes it transparent to codify and document the raw data issues and apply consistent decisions about which data should be available for researchers. Different research groups may make different decisions during their CDM implementations, but the process of designing and implementing an ETL specification allows those decisions to be exposed to the broader research community [26].

Many experimental projects have proved the feasibility and utility of OMOP CDM for multiple, disparate health databases [27-28]. This is why it is widely used for the model framework of patient-level prediction and safety surveillance, including a transformation from source data into standard vocabulary, which solves semantic interoperability; technology neutrality that does not rely on special computer technology; open community, open resources, free tools; generating aggregate analysis results directly from desensitized data, etc. However, some issues should be under consideration in the use [29]. Not all source data encoding can be converted to standard vocabulary, and there will be a loss of semantics, and concepts matching requires a lot of time and effort. The model and vocabulary were originally developed and designed for pharmaceutical safety research and clinical observation data, while development of vocabularies in other fields is limited [30-31].

5. Conclusion
Reuse of clinical data is essential to fulfill the promises for high-quality health care, improved health care management, effective medical research, but is complex, especially in distributed data networks. For a clinical practice using leverage research, there are many different options for data reuse. Existing clinic reuse methods were developed for specific uses, and their fitness for other uses depends on how closely the method matches the planned use. We compared two main methods of secondary clinical data use based on our medical big data project experience and literature in this paper. Both openEHR and
CDMs are designed for exporting and reusing data from a distributed clinical databases, with their advantages and disadvantages, and they are not available if there is nobody at the site who is familiar with the technology. The first one has been used in the data warehouse construction to collect and exchange EHR data between EHR systems, or between EHR systems and a centralized EHR data repository. The second trend is to connect the EHR with external analytical tools through a complex ETL process that involves both data mapping and normalization, such as using CDMs. OMOP CDM is often used in research to generate high quality and reproducible evidence directly from EHR systems without rebuilding a centralized database [32]. This approach is to transform data contained in disparate databases to a common format (data model), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format. The demonstration of transforming EHR database into this CDM offers a promise that data from other health claims, clinical systems could also be implemented.

Routine EHR systems are not desired in the research database. The process of transforming data in a data warehouse architecture implies the modification of both the source data structure and the data storage framework. Based on our experience in building medical big data projects, the former is suitable for collecting data from distributed EHR systems and building medical big data warehouses, while the latter is a better model for sharing data in some decentralized medical database. Our experience and recommendations are not comprehensive but interesting, and invite further discussion in this field.

Acknowledgments

This study was funded by China Scholar Council (No.201906375054).

References

[1] Safran, Charles, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association 14.1 (2007): 1-9.

[2] Outline of the Healthy China 2030 Plan. http://www.gov.cn/zhengce/2016-10/25/content_5124174.htm (accessed 20 Jan 2020). (in Chinese)

[3] Guidance to Promoting and Regulating the Application and Development of Big Data on Health Care 2016. http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm (accessed 5 Dec 2019) (in Chinese).

[4] Brown, Jeffrey, Michael Kahn, and Sengwee Toh. Data quality assessment for comparative effectiveness research in distributed data networks. Medical care 51.803 (2013): S22.

[5] Meystre, S. M., et al. Clinical data reuse or secondary use: current status and potential future progress. Yearbook of medical informatics 26.01 (2017): 38-52.

[6] Gold, Sigfried, et al. Clinical Concept Value Sets and Interoperability in Health Data Analytics. AMIA Annual Symposium Proceedings. Vol. 2018. American Medical Informatics Association, 2018.

[7] Richeson, Rachel L., and Prakash Nadkarni. Data standards for clinical research data collection forms: current status and challenges. Journal of the American Medical Informatics Association 18.3 (2011): 341-346.

[8] Evans, R. S. Electronic health records: then, now, and in the future. Yearbook of medical informatics 25.S 01 (2016): S48-S61.

[9] Klann, Jeffrey G., et al. Data model harmonization for the All Of Us Research Program: transforming i2b2 data into the OMOP common data model. PloS one 14.2 (2019).

[10] Reps, Jenna M., et al. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association 25.8 (2018): 969-975.

[11] Overview of available data. https://everlasting-project.eu/wp-content/uploads/2017/09/EVERLASTING_D8.2_final_20170428_DMP_Version1.pdf (accessed 29 Jan 2020).

[12] Genomic and medical data. https://en.wikipedia.org/wiki/Genomic_and_Medical_Data (accessed 25 Jan 2020).
[13] Semantic web application and tools for life science. http://swat4ls.blogspot.com/2018/11/(accessed 29 Jan 2020)

[14] Lima, Daniel M., et al. Transforming two decades of ePR data to OMOP CDM for clinical research. Stud Health Technol Inform 264 (2019): 233-237

[15] openEHR. https://www.openehr.org/(accessed 20 Dec 2019).

[16] Min, Lingtong, et al. Modeling EHR with the openEHR approach: an exploratory study in China. BMC medical informatics and decision making 18.1 (2018): 75.

[17] Li, Bei, et al. Experiences of building a medical data acquisition system based on two-level modeling. International journal of medical informatics 112 (2018): 114-122.

[18] L.T. Cavalini, T.W. Cook, Knowledge engineering of healthcare applications based on minimalist multilevel models, 2012 IEEE 14th International Conference on IEEE e-Health Networking, Applications and Services (Healthcom) (2012) 431–434.

[19] B. Christensen, G. Ellingsen, Evaluating model-driven development for large-scale EHRs through the openEHR approach, Int. J. Med. Inform. 89 (2016) 43–54.

[20] M.B. Späth, J. Grimson, Applying the archetype approach to the database of a biobank information management system, Int. J. Med. Inform. 80 (3) (2011) 205–226.

[21] Li, Bei, et al. Experience and reflection from China's Xiangya medical big data project. Journal of biomedical informatics 93 (2019): 103149.

[22] Garza, Maryam, et al. Evaluating common data models for use with a longitudinal community registry. Journal of biomedical informatics 64 (2016): 333-341

[23] OMOP Common Data Model. https://www.ohdsi.org/data-standardization/the-common-data-model/ (accessed 25 Nov 2019)

[24] Rosenbloom, S. T., et al. Representing knowledge consistently across health systems. " Yearbook of medical informatics 26.01 (2017): 139-147.

[25] Observational Health Data Sciences and Informatics (OHDSI) Website. https://www.ohdsi.org/web/wiki/doku.php (accessed 10 Dec 2019)

[26] Voss, Erica A., et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association 22.3 (2015): 553-564

[27] Rinner, Christoph, et al. A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data. eHealth. 2018.

[28] You, Seng Chan, et al. Conversion of National Health Insurance Service-National Sample Cohort (NHIS-NSC) Database into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). Studies in health technology and informatics 245 (2017): 467-470.

[29] Hripcsak, George, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics216 (2015): 574.

[30] Reisinger, Stephanie J., et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. Journal of the American Medical Informatics Association 17.6 (2010): 652-662.

[31] Shin, Seo Jeong, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. Journal of medical Internet research 21.3 (2019): e13249.

[32] Monsen, Karen A., et al. Rigor in electronic health record knowledge representation: lessons learned from a SNOMED CT clinical content encoding exercise. Informatics for Health and Social Care 41.2 (2016): 97-111