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

© 2024 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI230985

Development of Integrated Data Quality Management System for Observational Medical Outcomes Partnership Common Data Model

Seol Whan OH[a,b], Soo Jeong KO[a,b], Yun Seon IM[a,b], Surin JUNG[a], Bo Yeon CHOI[a,b], Jae Yoon KIM[a,b], Sunghyeon PARK[a,b], Wona CHOI[a,1] and In Young CHOI[a]
[a] *Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*
[b] *Department of Biomedicine &Health Sciences, The Catholic University of Korea, Seoul, Republic of Korea*
ORCiD ID: Seol Whan OH https://orcid.org/0000-0002-0328-9634,Soo Jeong KO https://orcid.org/0000-0002-6550-9188,Yun Seon IM https://orcid.org/0000-0002-2510-1380, Surin JUNG https://orcid.org/0000-0002-3314-3185, Bo Yeon CHOI https://orcid.org/0000-0003-1311-2645, Jae Yoon KIM https://orcid.org/0000-0001-8847-9586, Sunghyeon PARK https://orcid.org/0000-0002-2235-4358, Wona CHOI https://orcid.org/0000-0003-0269-6374, In Young CHOI https://orcid.org/0000-0002-2860-9411

Abstract. The amount of research on the gathering and handling of healthcare data keeps growing. To support multi-center research, numerous institutions have sought to create a common data model (CDM). However, data quality issues continue to be a major obstacle in the development of CDM. To address these limitations, a data quality assessment system was created based on the representative data model OMOP CDM v5.3.1. Additionally, 2,433 advanced evaluation rules were created and incorporated into the system by mapping the rules of existing OMOP CDM quality assessment systems. The data quality of six hospitals was verified using the developed system and an overall error rate of 0.197% was confirmed. Finally, we proposed a plan for high-quality data generation and the evaluation of multi-center CDM quality.

Keywords. Data quality, common data model, data quality management system

1. Introduction

Due to the increasing importance of healthcare data, there has been a growing interest in the study of data collection and management in recent years [1,2]. However, the structure of healthcare data is different for each hospital, making it difficult to conduct multi-institutional research related to data collection [3]. Therefore, many institutions have established a common data model (CDM), thus laying the groundwork for cooperative methodological or clinical innovation [3].

[1] Corresponding Author: Wona Choi, email: choiwona@gmail.com

Data quality is one of the main issues when gathering and managing large amounts of data [1,4]. Data quality issues arising from the nature of the source data and deficiencies in the data conversion process itself can affect the actual usefulness and reliability of CDM data [1,4]. Therefore, several research institutes including the Observational Health Data Sciences and Informatics (OHDSI) have developed and distributed data quality tools for CDM. However, the implementation of this tool requires considerable database knowledge and the applied verification rules cannot be easily changed [5].

This study sought to create a more user-friendly CDM data quality assessment system for multi-center quality evaluation to address the aforementioned limitations. Additionally, advanced evaluation rules were developed using data quality evaluation rules created by several research institutes. This will allow for the evaluation of the quality of the data derived from several institutions that have formed a CDM, as well as the identification of the quality of the CDM establishment for each institution. More importantly, the proposed tool provides a novel means for high-quality data collection, thus overcoming current bottlenecks in data acquisition and management in the health sector.

2. Methods

This study created a system for multi-center CDM data quality assessment, as well as advanced evaluation rules for quality assessment. The CDM data quality assessment system, which was loaded with advanced evaluation rules, was then used to conduct multi-center CDM data quality evaluation.

2.1. Development of CDM Data Quality Assessment System

The system was developed based on the Observational Medical Outcomes Partnership CDM (OMOP CDM), a widely known and representative CDM. The system was developed by accessing the hospital's CDM database and the results were extracted through the loaded evaluation rule query. All database queries were executed using PostgreSQL. Figure 1 illustrates the architecture of the CDM data quality assessment system developed herein.

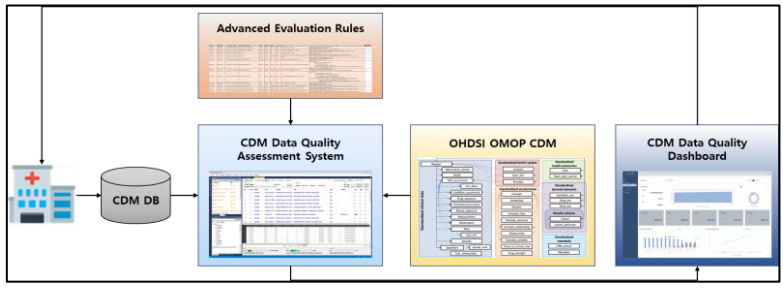

Figure 1. CDM Data Quality Assessment System Architecture.

2.2. Advanced Data Quality Evaluation Rules

The evaluation rules loaded into the system were based on DQ4HEALTH, a multi-institutional medical data quality measurement model [1]. The data quality rules were developed using five quality dimensions.

- Completeness: This rule evaluates whether null values are entered among the required items that must contain values for each column according to the table definition, such as the person number in the person table [1,2,6].
- Uniqueness: The primary key (pk) column, which is an identification value in the database, must not have duplicate values. This rule is particularly important to ensure that unique identifiers such as those in the condition occurrence ID column of the condition occurrence table are not duplicated [1,7].
- Validity: Data must have valid values and formats. For example, the number of specimens should be greater than zero and birth month must be one or two digits [1,8].
- Consistency: Some values are entered by referencing between variables in other columns or tables. This rule evaluates whether the values referenced from other tables are appropriate. For example, a drug concept ID must be a drug domain [1,8].
- Accuracy: This rule verifies whether the expression value of an object is accurately reflected. Specifically, the rule checks whether a value calculated from multiple values is correct or whether dates are in chronological order. For example, the measurement date must be between the date of birth and the date of death [1,8,9].

To optimize the evaluation rules, we compared them with other data quality rules. IQVIA, a healthcare research and service company, is supporting the construction of OMOP CDM, and the data quality rules used in this CDM were evaluated in the present study [10]. Additionally, our study also assessed the Data Quality Dashboard (DQD) evaluation rules provided by OHDSI [6].

Evaluation rules were developed by applying the DQ4HEALTH model to OMOP CDM v5.3.1, after which mapping analysis was performed with the provided IQVIA and DQD evaluation rules to obtain a set of more advanced evaluation rules (Figure 2).

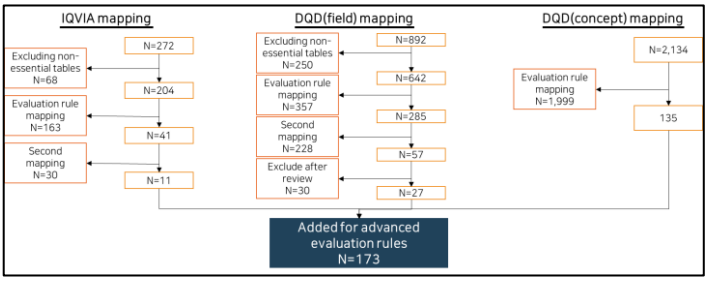

Figure 2. Data Quality Evaluation Rules Mapping Flowchart.

2.3. Verification of Multi-Center Data Quality using The System

The developed system was used to evaluate the quality of OMOP CDM data from six hospitals. Considering the development environment of each hospital, online and offline installation versions were separately developed and distributed. After quality evaluation, the results were reviewed to assess the quality of the CDM data from each hospital.

3. Results

We developed a CDM data quality evaluation system loaded with verification rule queries to evaluate the quality of multi-institutional OMOP CDM data (OMOP CDM v5.3.1). A total of 179 verification rules were developed by applying all five quality dimensions of the DQ4HEALTH model to 13 tables selected as essential tables at the time of construction.

Table 1. The Number of Advanced Verification Rules.

Table	The Number of Verification Rules	Totals
PERSON	28	
OBSERVATION_PERIOD	14	
SPECIMEN	16	
DEATH	11	
VISIT_OCCURRENCE	27	
VISIT_DETAIL	32	
PROCEDURE_OCCURRENCE	101	2,433
DRUG_EXPOSURE	33	
DEVICE_EXPOSURE	26	
CONDITION_OCCURRENCE	229	
MEASUREMENT	1,852	
NOTE	22	
OBSERVATION	42	

We loaded 2,433 optimized CDM data quality rules into our evaluation system (Table 1). The system now includes server management, verification execution, and report functions, allowing anyone to verify CDM data quality easily. It's designed for user-friendliness.

CDM data quality verification was performed by six organizations (A, B, C, D, E, F) that had built OMOP CDM v5.3.1 using a system loaded with advanced verification rules. Quality verification was conducted based on the CDM updated by each of the six hospitals as of September 2022, and the error rates were 0.208%, 0.217%, 0.218%, 0.112%, 0.228%, and 0.181%, respectively. The combined error rate of all hospitals was only 0.197%. Our results thus confirmed that the error rates were very low (Table 2).

Table 2. Evaluation results of 6 hospitals based on the CDM data quality verification system.

Hospitals	The Number of Patients	The Error Rate
A	927,997	0.207631%
B	1,951,727	0.216930%
C	1,005,002	0.218432%
D	866,168	0.112487%
E	424,752	0.228015%
F	1,159,941	0.181156%

4. Conclusions

Here, we developed a CDM data quality system for the OMOP CDM used by several hospitals. By analyzing IQVIA and DQD rules, we developed advanced quality criteria and incorporated them into the system for easy hospital assessments. With this, we evaluated six hospitals' data quality and highlighted their current CDM status and direction.

A key limitation of this study was the limited number of institutions evaluated for CDM quality. All six assessed hospitals produced high-quality CDM data as confirmed by our results. Since this study, the system has been expanded to more institutions with ongoing CDM data quality assessments. While we evaluated each hospital's data quality, a follow-up comparative study is needed after implementing improvement plans.

Despite the study's constraints, we developed a CDM data quality system that simplifies the process for institutions. A dashboard was also crafted for data quality comparison among institutions. These tools aid in evaluating data quality for OMOP CDMs, ensuring better CDM data and aiding institutions in data improvement decisions.

References

[1] Kim KH, Choi W, Ko SJ, Chang DJ, Chung YW, Chang SH, Kim JK, Kim DJ, Choi IY. Multi-center healthcare data quality measurement model and assessment using OMOP CDM. Applied Sciences. 2021 Jan;11(19):9188, doi: 10.3390/app11199188.

[2] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC). 2016 Sep;4(1):1244, doi: 10.13063/2327-9214.1244.

[3] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021 Sep;28(10):2251-7, doi: 10.1093/jamia/ocab132.

[4] Lynch KE, Deppen SA, DuVall SL, Viernes B, Cao A, Park D, Hanchrow E, Hewa K, Greaves P, Matheny ME. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform. 2019 Oct;10(5):794-803, doi: 10.1055/s-0039-1697598.

[5] Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, Staab J, Zozus MN, Kahn MG. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC). 2017 Jun;5(1):8, doi: 10.5334/egems.223.

[6] Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, Park RW, Boyce RD, Duke J, Khare R, Utidjian L, Bailey C. Multisite evaluation of a data quality tool for patient-level clinical data sets. EGEMS (Wash DC). 2016 Nov;4(1):1239, doi: 10.13063/2327-9214.1239.

[7] Amicis FD. A methodology for data quality assessment on financial data. Stud Commun Sci. 2004;4(2):115-37.

[8] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Commun ACM. 1996 Nov;39(11):86–95, 10.1145/240455.240479.

[9] Xiao Y, Bochner AF, Makunike B, Holec M, Xaba S, Tshimanga M, Chitimbire V, Barnhart S, Feldacker C. Challenges in data quality: the influence of data quality assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe. BMJ Open. 2017 Jan;7(1):e013562, doi: 10.1136/bmjopen-2016-013562

[10] Candore G, Hedenmalm K, Slattery J, Cave A, Kurz X, Arlett P. Can we rely on results from iqvia medical research data uk converted to the observational medical outcome partnership common data model?: a validation study based on prescribing codeine in children. Clin Pharmacol Ther. 2020 Apr;107(4):915-25, doi: 10.1002/cpt.1785.