Multi-Center Healthcare Data Quality Measurement Model and Assessment Using OMOP CDM

Ki-Hoon Kim 1,2, Wona Choi 1,2, Soo-Jeong Ko 1,2, Dong-Jin Chang 3, Yeon-Woog Chung 4, Se-Hyun Chang 5, Jae-Kwon Kim 2, Dai-Jin Kim 5 and In-Young Choi 2,*

1 Department of Biomedicine & Health Sciences, The Catholic University of Korea, Seoul 06591, Korea; kkh9525@naver.com (K.-H.K.); choiwona@gmail.com (W.C.); tn9934@icloud.com (S.-J.K.)
2 Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; tpgus6525@naver.com (S.-H.C.); jaekwonkorea@naver.com (J.-K.K.)
3 Department of Ophthalmology, Yeouido St. Mary’s Hospital, The Catholic University of Korea, Seoul 06591, Korea; hpalways@catholic.ac.kr
4 Department of Ophthalmology and Visual Science, St. Vincent’s Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; forgetmenot21c@gmail.com
5 Department of Psychiatry, Seoul St. Mary’s Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; kdj922@catholic.ac.kr
* Correspondence: iychoi@catholic.ac.kr

Abstract: Healthcare data has economic value and is evaluated as such. Therefore, it attracted global attention from observational and clinical studies alike. Recently, the importance of data quality research emerged in healthcare data research. Various studies are being conducted on this topic. In this study, we propose a DQ4HEALTH model that can be applied to healthcare when reviewing existing data quality literature. The model includes 5 dimensions and 415 validation rules. The four evaluation indicators include the net pass rate (NPR), weighted pass rate (WPR), net dimensional pass rate (NDPR), and weighted dimensional pass rate (WDPR). They were used to evaluate the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) at three medical institutions. These indicators identify differences in data quality between the institutions. The NPRs of the three institutions (A, B, and C) were 96.58%, 90.08%, and 90.87%, respectively, and the WPR was 98.52%, 94.26%, and 94.81%, respectively. In the quality evaluation of the dimensions, the consistency was 70.06% of the total error data. The WDPRs were 98.22%, 94.74%, and 95.05% for institutions A, B, and C, respectively. This study presented indices for comparing quality evaluation models and quality in the healthcare field. Using these indices, medical institutions can evaluate the quality of their data and suggest practical directions for decreasing errors.

Keywords: healthcare data; OMOP CDM; multisite study; data quality assessment

1. Introduction

Healthcare data is evaluated as data with economic value; subsequently, it attracts global attention from observational studies and clinical studies alike [1–3]. Healthcare data can be utilized remarkably rapidly, due to its large data set, continuity over time, and timely availability. Despite this potential, it remains difficult to analyze and integrate multicenter data due to skepticism among medical centers and different data structures of electronic health record (EHR) systems [4–11]. To overcome this, the recent introduction of a large-scale distributed research network (DRN) is expected to provide answers in areas that cannot be studied through conventional controlled clinical trials [12–15]. Notably, the importance of the equivalent level of data quality is increasing in these multicenter studies [9,16–19].

Data quality studies continue to use tools and assessment approaches to improve the quality of EHR data [15–30]. Lynch conducted a quality improvement study on errors in
the OMOP CDM V4.0 conversion process in the clinical data warehouse (CDW) of a hospital [14]. Additionally, the scalability of research on the detailed issues of the established common standard model is emphasized. Wang proposed a rule-template schematic model. This study emphasizes that quality evaluation is carried out by constructing a rule-based model and deriving dimensions through an expert group composed of IT experts and clinicians [30]. When reviewing healthcare research, Savitz points out that different institutions have different methodologies for data quality [11]. Institutions can work closely with multiple stakeholders to address healthcare data quality issues. Some of these stakeholders may include clinicians, administrators, and IT professionals. Despite the importance of the quality of healthcare data, there were few studies on multicenter quality management indices involving research.

Data quality assessment is vital in multicenter studies. The Observational Health Data Sciences and Informatics (OHDSI), a representative multicenter research network, is a network in which multiple stakeholders collaborate by analyzing large-scale medical data converted into the Common Data Model (CDM) format [12,16]. The CDM enables the integration of heterogeneous data from healthcare institutions by presenting a standardized data model from other healthcare institutions. In addition, it is possible to simplify the research network between institutions by converting it into analytical data.

This study presents a new conceptual model that can be applied to the quality of healthcare data through the review of existing data quality literature. In addition, this study selected medical institutions that established a large-scale cohort based on OHDSI’s CDM. A data quality evaluation was then performed. This study contributes to the improvement of healthcare data quality by checking the difference in quality for each medical institution after a quality evaluation.

2. Materials and Methods

The development and evaluation of clinical epidemiological information linked to human materials quality model involves the following four steps:

1. Develop a model for healthcare data quality evaluation;
2. Define the quality evaluation rules to be applied to quality evaluation;
3. Define the evaluation method using the quality evaluation model;
4. Verify the model by using it to evaluate the CDM data of hospitals.

2.1. Healthcare Data Quality Conceptual Model Development

This study used Google Scholar to compare and review eight data quality papers from 1990 to 2020. This allowed us to build a conceptual model of healthcare data quality. Based on the literature on information system data quality dimensions, we selected five dimensions as the evaluation criteria [9,17,20,30–34]. Specifically, these dimensions included completeness, validity, accuracy, uniqueness, and consistency (Appendix A Table A1).

We conducted a systematic review to examine 17,800 pieces of related literature and then reviewed eight studies that suggested an original data quality measurement method for research purposes. The studies did this by collecting expert opinions and conducting in-depth reviews. We dimensioned a total of 5 DQ4HEALTH.

We excluded 5251 articles that were either related to data quality for unstructured data or contained a database for the purpose of operating an information system, not a database for research purposes. Additionally, some data quality measurement methods were applicable only to a specific clinical domain, and cases in which it was difficult to use as a generalized evaluation item. This caused us to exclude an additional 9033 cases. Finally, 3506 documents were excluded from the data quality measurement method as they were difficult to derive by applying survey and analysis methods.

2.1.1. Completeness

Completeness is a criterion for evaluating whether data are missing in the process of expressing real-world data as a system. For example, if a patient visits a hospital and
undergoes an examination, the patient, examination, and visit information should not be omitted from the data’s point of view.

2.1.2. Validity

Validity is a criterion for assessing whether the data included in the system complies with the acceptable range and format. For example, January to December can be accepted for the month of birth, and the format must be an integer.

2.1.3. Accuracy

Accuracy evaluates whether real-world data are accurately represented by the system. For example, (i) the end date of a patient’s prescription medication must be equal to one day subtracted from the start date of the medication prescription, and (ii) the medication cannot be prescribed after the patient’s date of death.

2.1.4. Uniqueness

Uniqueness is a dimension that examines a unique data value according to the characteristics of the data and evaluates the word for the data value. For example, a unique prescription number should be assigned to each prescription of medication, measurement, or procedure.

2.1.5. Consistency

Consistency assesses the relationship between the inside and outside of the system by evaluating whether the data are consistent according to the system structure. For example, if a patient visits a hospital as an outpatient and receives a medication prescription, the value of their assigned patient number should exist in the visit information and medication information. This rule examines whether different data are linked to the foreground key in the database.

The importance of validation rules for seven clinical experts, data managers, and data quality experts with more than three years of experience in the above five dimensions were evaluated. The weights developed using these are reflected in the evaluation index (Table 1).

| Dimension       | Sub-Dimension | Definition                                                                 | Importance | Weight |
|-----------------|---------------|----------------------------------------------------------------------------|------------|--------|
| Completeness    |               | This rule verifies that there are no missing required columns.              | 9.6        | 0.23   |
| Validity        | Range         | This rule checks whether a data value is within a given range.             | 7.5        | 0.18   |
|                 | Format        | This rule checks whether a data value conforms to the data type.           | 7.5        | 0.18   |
| Accuracy        | Calculation   | This rule verifies whether the values of different columns are correct.   | 7.7        | 0.18   |
|                 | Timeline      | This rule verifies the precedence of time.                                | 7.7        | 0.18   |
|                 | Business Rule | This rule verifies the hospital business rules.                            | 7.7        | 0.18   |
| Uniqueness      |               | This rule verifies the value corresponding to the primary key.             | 9.0        | 0.22   |
| Consistency     | Standard      | If an international standard code is used, this rule verifies the standard code. | 7.9        | 0.19   |
|                 | Relationship  | If there is a referential relationship between tables, this rule verifies the referential integrity. | 7.9        | 0.19   |

Table 1. DQ4HEALTH dimensions of specialist group review.
2.2. Data Quality Assessment Rule Development

Five dimensions were applied to OMOP CDM V5.3.1 to develop a total of 415 evaluation rules that allowed us to assess the quality of medical data (Appendix B Table A2). The evaluation rules were divided into errors and warnings based on their importance (Table 2). An error must be cleansed and corrected when it is confirmed in the validation rule. A warning is an error that does not need to be cleansed, even if it is confirmed in the validation rule. The developed rule evaluates its importance with expert advice, and the weight developed using it is reflected in the evaluation index.

Table 2. Type definition of specialist group review.

| Type   | Definition                                                                 | Importance | Weight |
|--------|---------------------------------------------------------------------------|------------|--------|
| Error  | This error is one that must be cleansed and corrected once it is identified in the validation rule. | 9.3        | 0.64   |
| Warning| This error is one that was identified in the validation rule but does not need to be corrected. | 5.3        | 0.36   |

2.3. Data Quality Assessment Method Development

The evaluation results were quantified when the data quality was evaluated. For this purpose, the result is expressed as the ratio of error data with the data that passed or did not pass the rule, and the error rate that reflects the weight of each data rule evaluated by experts. Four indices were developed as an evaluation index: NPR, WPR, NDPR, and WDPR.

- NPR: this is a data quality evaluation index that does not reflect the weights for data errors. The NPR is calculated by subtracting 100 from the total error rate, which is the result of the data quality evaluation for each institution, and by adding the error rates of error and warning;
- WPR: this is a level evaluation index of data in which weights are applied to errors or warnings among data errors;
- NDPR: this is a data level evaluation index that does not reflect the weight of each index for each data error after quality is evaluated by the five dimensions. The NDPR is calculated by subtracting the error rate for each total dimension error rate, which is the data quality evaluation result for each dimension, 100;
- WDPR: this is a data level evaluation index that reflects the weights for data errors in each of the five dimensions. The WDPR evaluates the importance of each dimension by experts and sets the weight after finding the average value. Thereafter, the value obtained by multiplying the total dimension error rate by each dimension of each weight was calculated as the WDPR.

2.4. Multicenter Data Quality Assessment

To study data quality research using the OMOP CDM, we selected three medical institutions that allowed access to their data to build the OMOP CDM V5.3.1 model (Table 3). The institutions that agreed to the evaluation were medical institutions located in the metropolitan areas of Korea that serve more than two million people. In addition, the analysis was performed using the chi-square method to determine whether the difference between the data quality evaluation results of the three medical institutions was significant.

Table 3. Characteristics of A, B, C hospital.

| Center | Provider Type     | Region     | The Number of Bed Hospitals | The Number of OMOP CDM Person |
|--------|-------------------|------------|-----------------------------|-------------------------------|
| A      | Tertiary Hospital | Seoul      | approximately 1400          | 3,598,955                     |
| B      | General Hospital  | Gyeonggido | approximately 900           | 2,279,292                     |
| C      | General Hospital  | Seoul      | approximately 400           | 2,077,837                     |
3. Results

3.1. Multicenter OMOP CDM Data Quality Assessment Results

DQ4HEALTH was applied and evaluated for three medical institutions and built by OMOP CDM V5.3.1.

3.1.1. NPR and WPR

Without weighting the data error, the NPR was 96.58%, 90.08%, and 90.87% for institutions A, B, and C, respectively (Figure 1 and Table 4). The WPR result, which is an evaluation index that experts employed to weigh error and warning, was 98.52%, 94.26%, and 94.81% for institutions A, B, and C, respectively. Compared to that of the NPR, the WPR scores were higher; however, the difference in scores between the institutions was similar. The difference in quality between institutions is due to the influence of weights reflecting expert evaluation of verification rules, as classified into “error” and “warning.”

For example, regarding the quality of patient information, the Patient ID rule that has a value of not null is an error. As this is a rule that has an important influence on quality, experts evaluated it with a weight of 0.64. However, the patient’s racial classification was evaluated with a weight of 0.36, as it was identified as a rule that did not affect data quality. We confirm that the importance can be different even within tables that collect the same information.

It was possible to verify the overall errors of healthcare data quality, and the following five types of errors were identified:

![Figure 1. Comparison of NPR and WPR according to error and warning weights.](image)

|       | Total Error Rate | Error Rate | Warning Rate | NPR    | WPR    |
|-------|------------------|------------|--------------|--------|--------|
| A     | 3.42%            | 0.89%      | 2.53%        | 96.58% | 98.52% |
| B     | 9.92%            | 7.73%      | 2.19%        | 90.08% | 94.26% |
| C     | 9.13%            | 6.79%      | 2.34%        | 90.87% | 94.81% |

The difference in quality between institutions is due to the influence of weights reflecting expert evaluation of verification rules, as classified into “error” and “warning.” For example, regarding the quality of patient information, the Patient ID rule that has a value of not null is an error. As this is a rule that has an important influence on quality, experts evaluated it with a weight of 0.64. However, the patient’s racial classification was evaluated with a weight of 0.36, as it was identified as a rule that did not affect data quality. We confirm that the importance can be different even within tables that collect the same information.

It was possible to verify the overall errors of healthcare data quality, and the following five types of errors were identified:
1. We identified a type of error in which the inspection result value of the inspection information table is not an integer greater than 0, but a negative number. Obviously, there cannot be a case where a test result value exists as a negative number. As a result of tracing the source data, it was confirmed that the unmeasurable value was defined as $-9999$, owing to an error in the inspection machine.

2. A type of error was revealed that is caused by a problem with the source data value (source_value), and it is a type of error that includes missing spelling such as “Test Name (“88888_Drug Name”, mi(missing spelling) minor salivary gland”). This type of error suggests that meaningless data can be loaded, and the reliability of the data can be reduced.

3. Error types that deviated from the standard term values owing to input errors between concept_id and code data values were found. In addition, the problem of mapping values to nonstandard values was also derived. The importance of mapping international standard terms is mentioned often in existing healthcare studies, suggesting that it may be a problem in multicenter studies that use OMOP CDM.

4. A type of error regarding chronological relationships was also identified. This violates the precedence relationship between the patient’s date of birth and death and the observation period of each clinical information. This type of error draws attention to the implications of refining as errors that occur in the ETL process and errors that may occur in actual EHR systems.

5. A type of error that violates referential integrity was revealed. This is the type identified with most errors in this study. This error occurred because of the reference relationship between patient information and the treatment/diagnosis information table in the structure of the OMOP CDM. In other words, most of the data were loaded abnormally even though the data were updated, or the patient ID was present but could not be used for actual research because there was no examination history.

3.1.2. NDPR and WDPR

The data quality was examined based on five dimensions and the error rate was evaluated for each dimension. For each error type, the evaluation results were compared using the NDPR, which assessed only the number of errors, and the WDPR, which provided the weight of that type of error.

When checking the DQ4HEALTH dimension result, the quality level of the four dimensions was close to 99% or higher. Furthermore, the consistency dimension had the highest error rate, at 70.06% (1,338,817,961 records), of all the error data. As for the results of quality analysis, NDPR, which does not reflect the weight of consistency, was 90.66%, 76.52%, and 78.64% for institutions A, B, and C, respectively. When weights for each dimension evaluated by experts were provided, 98.22%, 94.74%, and 95.05% of institutions A, B, and C, respectively, showed results (Figure 2 and Table 5).

Depending on whether the experts’ weights were reflected, the difference in results was due to the following factors. Experts gave low weight to the consistency dimension in the case of tables that did not affect analysis and medical concepts that are hard to map using standard medical terms.

We adopted the chi-square analysis method to verify whether there is a level difference according to the quality results of all medical institutions and conducted a subsequent analysis. The result was $p < 0.001$, which confirmed that there was a difference in the quality of data from each hospital.

Additionally, we performed a chi-square analysis to determine whether there was any difference in quality for each OMOP CDM table at each medical institution. This allowed us to check the factors that affected the overall results. Of the 195 variables, all three medical institutions had no errors. Either that or two medical institutions excluded variables without errors from the analysis and we performed a chi-square analysis on 96 variables. The analysis confirmed that there was a difference in the quality of healthcare data between institutions.
Table 5. Multicenter OMOP CDM data quality assessment specific results.

| Center | DQ4HEALTH Dimension | Total Dimension Data Count | Total Dimension Error Count | Total Dimensions Error Rate | NDPR   | WDPR   |
|--------|----------------------|---------------------------|----------------------------|----------------------------|--------|--------|
| A      | Completeness         | 5,460,723,980             | 8,276                      | 0.01%                      | 99.99% | 99.99% |
|        | Validity             | 1,360,559,053             | 22,801,212                 | 1.67%                      | 98.33% | 99.70% |
|        | Accuracy             | 3,570,299,098             | 59,288,628                 | 1.66%                      | 98.34% | 99.69% |
|        | Uniqueness           | 840,625,891               | 239,985                    | 0.03%                      | 99.97% | 99.99% |
|        | Consistency          | 5,005,238,125             | 467,936,657                | 9.34%                      | 90.66% | 98.22% |
| B      | Completeness         | 2,619,120,230             | 1,399,297                  | 0.05%                      | 99.95% | 99.99% |
|        | Validity             | 644,669,318               | 11,173,281                 | 1.73%                      | 98.27% | 99.69% |
|        | Accuracy             | 1,847,001,586             | 333,479                    | 0.02%                      | 99.98% | 99.99% |
|        | Uniqueness           | 412,280,539               | 0                         | 0%                         | 100%   | 100%   |
|        | Consistency          | 2,835,935,266             | 816,059,524                | 28.77%                     | 71.23% | 94.74% |
| C      | Completeness         | 1,826,576,516             | 1,545,055                  | 0.08%                      | 99.92% | 99.98% |
|        | Validity             | 430,638,422               | 7,014,267                  | 1.62%                      | 98.38% | 99.71% |
|        | Accuracy             | 1,270,385,522             | 302,273                    | 0.00%                      | 99.99% | 99.99% |
|        | Uniqueness           | 291,598,022               | 0                         | 0%                         | 100%   | 100%   |
|        | Consistency          | 2,003,506,197             | 522,758,437                | 26.09%                     | 73.91% | 95.05% |

Figure 2. Comparison of NDPR and WDPR by consistency weights.

Consequently, it was confirmed that institution A had the highest quality data of the three medical institutions. Comparatively, institution B possessed low-quality data. Regarding institution B, the error derived from the consistency dimension was the highest of all three institutions. The consistency dimension was confirmed to be a factor with low quality.
4. Discussion

This study differs from previous studies on data quality because it developed an index that can evaluate the quality of multiple institutions using a large cohort.

Existing healthcare data quality studies suggest a conceptual model that can be applied to healthcare data through a literature review; however, few studies verify the proposed model using actual healthcare data [5,20,22,23,28,30]. The verified literature has the limitation of coming from a small cohort; therefore, the present study expanded itself to utilize a large-scale, cohort-based multicenter study [6,8,9,15,16,18,21,24,27].

In addition, an evaluation method was developed to compare the impact of errors on the healthcare quality results. The existing literature on data quality evaluation presents the net error rate and error distribution according to the quality dimension owing to the application of the data quality conceptual model. In this study, we propose a data quality evaluation method to review the causes of errors that affect healthcare data through multicenter quality comparisons according to the researcher’s quality study design by expanding the results of the net error. In other words, the quality evaluation method refers to four evaluation criteria (NPR, WPR, NDPR, and WDPR) for easy access to expert reviews in evaluating healthcare data.

Finally, when utilizing the opinions of experts, we can properly weight errors according to the degree of influence on the quality of medical institutions. Existing literature on data quality assessment emphasizes the importance of documentation and methods by which experts can review data quality results reports [8,11]. Therefore, in this study, weights were assigned based on expert evaluations so that expert opinions and reviews can be reflected. Therefore, this study complements the existing literature by addressing the existing limitations and intuitively suggesting effects on the quality of medical institutions according to expert reviews.

Our study has several limitations. Since the DQ4HEALTH model proposed in this study confirms and verifies the overall quality of OMOP CDM, more detailed and specific quality verification rules should be expanded when conducting research on specific diseases and medications. For example, Veronica Muthee conducted a healthcare data study centered on the HIV care data-based routine data quality assessment (RDQA) model [27]. This shows the detailed data quality point of view by verifying the missing values. In addition, continuous research on data quality tools that can intuitively express diagrams and visualization functions should be expanded by applying the DQ4HEALTH model. This was determined according to the multicenter automated quality evaluation function and quality evaluation results.

Despite these limitations, this study analyzes the types of errors by presenting a new model that can be applied to the OMOP CDM after considering and integrating healthcare data quality studies and applying it to multiple institutions. This can be utilized in future studies.

5. Conclusions

In this study, we developed a validation rule that can be applied to OMOP CDM by selecting frequent values through a review of previous studies on the existing information system quality and healthcare quality dimensions. Additionally, we proposed a new DQ4HEALTH model for OMOP CDM data quality management as a result of receiving expert advice based on the developed validation rule. The developed DQ4HEALTH model was applied to three institutions with more than two million CDM data to conduct an empirical healthcare data quality evaluation study.

As a result of analyzing the multicenter data quality error results with more than 2 million cohorts using the chi-square method, we confirmed that there is a difference in the quality of CDM data between hospitals. This means that even though the same OMOP CDM was applied, there was a difference in quality for each hospital. There was also a significant difference for each table. The types of errors presented in this study suggest
that the analysis results may be affected when conducting joint research using a common
data model.

In the future, it will be necessary to expand research to intuitively confirm the degree
of data quality improvement through comparison before and after cleansing the error data
derived from the data quality result. It is also necessary to expand the study on the effects
of analysis results before and after comparison [35–37]. Finally, this study contributes to
laying the foundation for the development of quality control tools using the developed
quality control rules and results analysis method [38].

Author Contributions: Conceptualization, K.-H.K. and I.-Y.C.; methodology, K.-H.K. and I.-Y.C.;
software, S.-H.C., K.-H.K. and S.-J.K.; validation, S.-J.K. and K.-H.K.; formal analysis, S.-H.C.; investiga-
tion, D.-J.K. and I.-Y.C.; resources, I.-Y.C. and D.-J.C.; data curation, I.-Y.C., D.-J.C. and Y.-W.C.;
writing—original draft preparation, W.C. and K.-H.K.; writing—review and editing, I.-Y.C., J.-K.K.,
and W.C.; visualization, W.C. and K.-H.K.; supervision, I.-Y.C.; project administration, D.-J.K. and
I.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Technology Innovation Program (20004927, Upgrade
of CDM-based Distributed Biohealth Data Platform and Development of Verification Technology)
funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Institutional Review Board Statement: The study was conducted in accordance with the guidelines
of the Declaration of Helsinki and approved by the Institutional Review Board of the Catholic Medical
Center (protocol code XC20RNDI0161 and 6 July 2021).

Informed Consent Statement: The requirement for written informed consent was waived by the Re-
search Ethics Committee of the Catholic Medical Centre, and this study was conducted in accordance
with relevant guidelines and regulations.

Data Availability Statement: Data sharing was not applicable to this study. Data supporting the
findings of this study are available from each hospital.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The Literature Review Result of Information System Dimension.

| DQ4HEALTH Dimensions | Definition | DQ Terminology | Authors |
|-----------------------|------------|----------------|---------|
| Completeness          | Evaluate missing data in the process of representing data in the real world as a system. | Completeness | [9,20,31–33] |
| null                  |            | Null Values    | [34]    |
|                      |            | Incompleteness | [30]    |
| Validity              | Evaluate whether it allows the scope of the data in the system. | Scope | [32] |
| Range                 |            | Value out of range etc. | [30] |
|                      |            | Format | [9,17,34] |
|                      |            | Correctness etc. | [20] |
|                      |            |                | [17,33,34] |
| Calculation           | Evaluate whether the calculation formula for items that are composed of multiple items is correct. | Accuracy | [9,32] |
|                      |            | Computation Conformance | [17] |
| Timeliness            | Evaluate time among data values expressed in the real world. | Timeliness | [9,31,32,34] |
|                      |            | Currency etc. | [20,33] |
|                      |            |                | [9,17] |
| Business Rule         | Evaluate whether business relevance (knowledge) among data values expressed in the real world is correctly expressed. | Accuracy | [9,31,32,34] |
|                      |            | (Atemporal) Plausibility etc. | [17,20] |
|                      |            |                | [30,33] |
| Uniqueness            | Evaluate whether duplicate values are allowed in the system. | Uniqueness (Plausibility) (Non)duplication | [17,34] |
|                       |            |                | [30,35] |
Table A1. Cont.

| DQ4HEALTH Dimensions | Definition | DQ Terminology | Authors |
|-----------------------|------------|----------------|---------|
| Consistency           | Standard   | Value Conformance, Incompatibility etc. | [17, 30, [9, 20] |
|                       | Relational | Relationship Conformance Etc. | [17, 20, 30, 32] |
| Relational            |            |                |         |

Appendix B

Table A2. DQ4HELTH (Data Quality for Healthcare) Model Development Result.

| Dimensions     | Definition                                                                 | OMOP CDM Rules Example                                                                 | Type | Rule count |
|----------------|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------------|------|-----------|
| Completeness   | -                                                                           | a. The patient number (person_id) column in the Person Table must not have a null value.  | E    | 85        |
|                |                                                                             | b. The Specimen Concept ID column in the Specimen table must not have a null value.    |      |           |
| Range          | This rule verifies that a data value is within a given range.                | a. The Measurement Result Value of measurement table should have a value greater than 0. | W    | 10        |
|                |                                                                             | b. The month of the patient’s date of birth must have a value between 1 and 12.      | E    |           |
|                |                                                                             | a. The year of birth in Person table should have a value in the format of a 4-digit number. | E    | 9         |
|                |                                                                             | b. The column of Measurement Time in the Measurement table should have a value in the format of 24H:MM:SS. |      |           |
| Validity       | Format                                                                      | a. The Measurement Result Value of measurement table should have a value greater than 0. | W    |           |
|                |                                                                             | b. The month of the patient’s date of birth must have a value between 1 and 12.      | E    |           |
|                |                                                                             | a. The year of birth in Person table should have a value in the format of a 4-digit number. | E    | 9         |
|                |                                                                             | b. The column of Measurement Time in the Measurement table should have a value in the format of 24H:MM:SS. |      |           |
| Accuracy       | Calculation                                                                 | a. Drug_exposure_end_date must be equal to Drug_exposure_start_date minus a value of −1. | E    | 1         |
|                |                                                                             | a. The value of the year of birth (YYYY) in the date of birth (Birth_Datetime) of the patient information and the value of the year of birth (year_of_birth) must have the same value. | W    | 58        |
|                |                                                                             | b. The Procedure_date in the Procedure table must occur after the date of birth and before the date of death. | E    |           |
|                |                                                                             | a. If one’s gender is female, they cannot have a diagnosis code for male disease.    | E    | 145       |
|                |                                                                             | b. The visit concept id should have a value of type of inpatient, outpatient, emergency, clinical trial, and medical examination. | W    |           |
| Uniqueness     | -                                                                           | a. The person id in the person table must have a unique value.                      | E    | 14        |
| Standard       | Consistency                                                                 | a. The Condition concept id of the Condition table must comply with the standard mapping of Domain = Condition, Standard concept = S, of Voca table | W    | 34        |
| Relationship   | If there is a referential relationship between tables, referential integrity is verified. | a. Location id of Person table should have the value of Location id of Location table. | E    | 44        |
33. English, L.P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*; John Wiley & Sons: Chicago, IL, USA, 1999. [CrossRef]

34. Loshin, D. *Enterprise Knowledge Management: The Data Quality Approach*; Morgan Kaufmann: Burlington, NJ, USA, 2001. [CrossRef]

35. Scannapieco, M. *Data Quality: Concepts, Methodologies and Techniques, Data-Centric Systems and Applications*; Springer: New York, NY, USA, 2006. [CrossRef]

36. Batini, C.; Cappiello, C. Methodologies for data quality assessment and improvement. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–52. [CrossRef]

37. Rahm, E. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* **2000**, *23*, 3–13. [CrossRef]

38. Bora, D.J. Big data analytics in healthcare: A critical analysis. In *Big Data Analytics for Intelligent Healthcare Management*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 43–57. [CrossRef]