Paper RE04

A Customized Algorithm to Evaluate the Quality and Reliability of Real-World Lab Data

Lihai Song, Merck & Co., Inc., Rahway, NJ, USA
Xing He, University of Florida, Gainesville, FL, USA
Ashwin Rai, Evidera, Bethesda, MD, USA
Xinyue Liu, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

The FDA has recently released guidance on the use of real-world data (RWD), emphasizing the importance of ensuring the relevance and reliability of data through governance and established processes. In response to this guidance, we have developed a customized algorithm to assess the quality of lab data in RWD. This algorithm, based on the OMOP Common Data Model (CDM), employs predefined criteria to evaluate conformance, completeness, and plausibility. The resulting assessment report, available as an R Shiny dashboard and a detailed algorithmic spreadsheet, enables researchers to determine the reliability and feasibility of lab data within RWD databases. Our innovative approach, with its well-documented process and user-friendly implementation, provides a practical means of evaluating the usability of RWD. Additionally, it serves as a valuable example of a complaint process for the effective use of RWD in regulatory submissions.

INTRODUCTION

Real-world data (RWD) in healthcare research is defined as data collected outside of clinical trials. Common RWD examples are electronic health records (EHRs), insurance claims and patient registries. The quality and reliability of RWD are essential for successful RWD studies and need to be systematically assessed. Laboratory data (lab data), generated from laboratory tests (lab tests) of a variety of patients' samples (e.g., tissue, blood, urine, other body fluid etc.), has become an integral part of many databases while facing quality and reliability challenges. To better follow FDA's guidance on the use of RWD and fill the knowledge gap, we developed a customized algorithm to evaluate the quality and reliability of lab data in RWD and applied algorithm to multiple RWD sources.

Emerging from the FDA's Real-world Evidence (RWE) program, real-world data (RWD)—particularly electronic health records (EHRs), administrative claims, and billing information—are becoming increasingly vital in supporting a broad spectrum of healthcare and regulatory decisions. The FDA has established a framework for its RWE program, issued guidance on using EHR data in clinical investigations, and provided instructions on submitting documents that employ RWD and RWE for drug and biologic evaluations[4]. Consistently, the FDA points out the importance of data quality (DQ) and reliability (i.e., "the strength of RWE submitted in support of a regulatory decision depends on the clinical study methodology and the reliability [data accrual and DQ control (data assurance)] and relevance of the underlying data.").

However, the quality of RWD raises concerns due to widely reported DQ issues such as incompleteness, inconsistency, and inaccuracy, which can undermine its reliability and applicability. These challenges are particularly significant given the growing reliance on RWD for decision-making in clinical research, regulatory assessments, and healthcare practices. To maximize the utility of RWD, it is crucial to systematically assess and understand DQ. A recent study has comprehensively reviewed existing DQ assessment frameworks and methods, revealing a few gaps in the current approaches to address data quality issues, thereby highlighting the need for more rigorous and standardized solutions. The quality of RWD continues to be a significant concern, as deficiencies in data quality can profoundly impact the reliability of subsequent analyses and the validity of study outcomes. Recognizing this issue, the FDA has stressed the importance of ensuring that RWD is fit-for-purpose when used to generate real-world evidence for decision-making in healthcare and research.

Laboratory tests are critical in healthcare for diagnosing diseases, predicting outcomes, and monitoring treatment responses. However, within EHRs, laboratory test results collected during routine care are often not uniformly coded or documented in a standardized manner across different organizations or even within the same organization over time. This inconsistency poses challenges for data integration and analysis. To address these challenges, assessing the DQ of laboratory data—focusing on completeness, consistency, and trends—is essential. Furthermore, adopting standardized coding systems, such as the Logical Observation Identifiers, Names, and Codes (LOINC) system, can significantly enhance the usability and comparability of laboratory data, ultimately improving its utility for clinical and research purposes.

METHODOLOGY

To formulate laboratory data quality (DQ) checks, we conducted a systematic evaluation of existing DQ methodologies for electronic health record (EHR) lab data. This process involved a comprehensive review of the literature and online resources related to lab DQ, during which we identified critical DQ dimensions and established a robust framework for assessing lab DQ. A scoping review targeting DQ issues related to EHR lab data was executed in adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. We developed search terms such as "(electronic health records OR real-world data) AND (lab data) AND (quality assessment)" to query literature databases like PubMed. Detailed inclusion and exclusion criteria were established to effectively filter relevant articles. Two independent reviewers screened titles and abstracts for inclusion, resolving any disagreements through discussions and consensus with an expert team. The same reviewers conducted full-text screenings for articles that met eligibility criteria after the title and abstract review. See Table 1 for examples of literature review criteria.

Table 1. Literature review criteria

Database	("RWD1"[Title/Abstract] OR " RWD2"[Title/Abstract] OR " RWD3"[Title/Abstract] OR " RWD4"[Title/Abstract] OR RWD5"[Title/Abstract
Disease	AND ("lung cancer"[Title/Abstract] OR "cervical cancer"[Title/Abstract] OR "cervix cancer"[Title/Abstract] OR "diabetes"[Title/Abstract] OR "hepatitis C"[Title/Abstract] OR "HCV"[Title/Abstract] OR "metabolic syndrome"[Title/Abstract])
Lab test	AND ("lab"[Title/Abstract] OR "laboratory"[Title/Abstract] OR "Cancer antigen 125"[Title/Abstract] OR "CA125"[Title/Abstract] OR "Lactate dehydrogenase"[Title/Abstract] OR "LDH"[Title/Abstract] OR "Epidermal growth factor receptor"[Title/Abstract] OR "EGFR gene mutation"[Title/Abstract] OR "Estrogen receptor"[Title/Abstract] OR "ER "[Title/Abstract] OR "Human Papillomavirus molecular test"[Title/Abstract] OR "HPV molecular test"[Title/Abstract] OR "Human Papillomavirus test"[Title/Abstract] OR "HPV test"[Title/Abstract] OR "Cervical smear"[Title/Abstract] OR "HbA1c"[Title/Abstract] OR "Hemoglobin A1C"[Title/Abstract] OR "Alanine transaminase"[Title/Abstract] OR "ALT"[Title/Abstract] OR "Aspartate transaminase"[Title/Abstract] OR "AST"[Title/Abstract] OR "Alkaline phosphatase"[Title/Abstract] OR "ALP"[Title/Abstract] OR "Gamma-glutamyl transferase"[Title/Abstract] OR "GGT"[Title/Abstract] OR "Bilirubin"[Title/Abstract] OR "White blood cells"[Title/Abstract] OR "White blood cell"[Title/Abstract] OR "Lymphocytes"[Title/Abstract] OR "Neutrophils"[Title/Abstract] OR "Platelet counts"[Title/Abstract] OR "Platelet count"[Title/Abstract] OR "Red blood cell"[Title/Abstract] OR "Red blood cells"[Title/Abstract] OR "Hemoglobin"[Title/Abstract] OR "High-density lipoprotein"[Title/Abstract] OR "HDL"[Title/Abstract] OR "low-density lipoprotein"[Title/Abstract] OR "LDL"[Title/Abstract] OR "Triglycerides"[Title/Abstract])
	limit to 3 year (after 2020) for diabetes, and 5 year (after 2018) for other diseases: returned 103 literatures
If restricted to data quality related	AND ("data quality"[Title/Abstract] OR "data accuracy"[Title/Abstract] OR "data reliability"[Title/Abstract] OR "data validity"[Title/Abstract] OR "data consistency"[Title/Abstract] OR "data completeness"[Title/Abstract] OR "data error"[Title/Abstract] OR "performance measure"[Title/Abstract] OR "quality measure"[Title/Abstract] OR "quality measurement"[Title/Abstract] OR "quality assurance"[Title/Abstract] OR "missing data"[Title/Abstract] OR "data use"[Title/Abstract] OR "data evaluation"[Title/Abstract] OR "data verification"[Title/Abstract])

Additionally, we thoroughly reviewed two online resources related to lab DQ assessment: 1) the National Patient-Centered Clinical Research Network (PCORnet) data check and 2) the Observational Health Data Sciences and Informatics (OHDSI) data quality dashboard (DQD). Both resources are crucial for evaluating the quality of RWD, including lab data, within established clinical research networks that maintain extensive RWD collections.

Lab tests and their corresponding codes were selected based on relevance and criteria for a sufficient sample size, ensuring a robust dataset for accurate analysis. Specific lab DQ checks were implemented for the labs of particular interest. See Table 2 for the list of Lab of interest and target population.

Table 2: Lab of interest and target population.

Lab of interest	Target population
Oncology-related labs	Lung cancer
Cancer antigen 125 (CA125)	
Lactate dehydrogenase (LDH)	
Epidermal growth factor receptor (EGFR) gene mutation	
Estrogen receptor (ER)	
Human Papillomavirus (HPV) molecular test	Cervical cancer
Cervical smear	
HbA1c	Diabetes
Liver panel/disease	Hepatitis C
Alanine transaminase (ALT)	
Aspartate transaminase (AST)	
Alkaline phosphatase (ALP)	
Gamma-glutamyl transferase (GGT)	
Bilirubin	
Complete blood count (CBC)	General population
White blood cells	
Lymphocytes	
Neutrophils	
Platelet counts	
Red blood cell	
Hemoglobin	
Lipid panel	Metabolic syndrome
High-density lipoprotein (HDL)	
low-density lipoprotein (LDL)	
Triglycerides	

The checks list and implementation of SQL scripts were continuously refined and expanded. This iterative process included developing checks for labs of interest, adding additional standard codes for lab tests, and the execution of algorithms.

For each check—conformance, completeness, and plausibility—along with their corresponding subcategories (such as atemporal, temporal, relational, uniqueness, and value), assessments are conducted at the field, table, and concept levels. The failure rate is calculated and compared against predefined thresholds to determine whether each check passes or fails. Abnormal lab test results are investigated to ensure that failures are attributable to data quality issues rather than algorithm errors. A final report is compiled across all Real-World Data (RWD) and published to a Shiny app, allowing researchers to evaluate lab quality results.

This customized algorithm is based on Observational Health Data Sciences and Informatics (OHDSI) network, the Data Quality Dashboard[2,3](DQD). This user-driven approach enables tailored assessments that meet the unique needs of different datasets and research objectives. The use cases were derived from lab-specific studies, integrating these additional lab DQ checks into the algorithm to expand the check beyond general data quality assessments.

Our method follows these steps:

1. Identify the target disease.
2. Conduct a literature review to identify specific lab tests and results associated with the disease.
3. Utilize the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) for lab test codes.
4. Establish thresholds for testing failures.
5. Write and test SQL queries against in-house RWD.
6. Integrate the tested SQL queries into the existing data quality check R package.
7. Execute this package across multiple RWD databases.
8. Review logs and results, identify algorithm errors, and refine the algorithm.
9. Repeat steps 5 to 8 until the algorithm meets the established criteria (no errors, no unexplained outliers).
10. Compile the quality check results and dissemination.

The customization of the OHDSI Data Quality Dashboard (DQD) R package was to focus on laboratory data in the MEASUREMENT table of the OMOP CDM. The MEASUREMENT table contains records of measurements, i.e., structured values (numerical or categorical) obtained through systematic and standardized examination or testing of a Person or Person's sample. The MEASUREMENT table contains both orders and results of such measurements as laboratory tests, vital signs, quantitative findings from pathology reports, etc. Measurements are stored as attribute value pairs, with the attribute as the Measurement Concept and the value representing the result. The value can be a Concept (stored in VALUE_AS_CONCEPT), or a numerical value (VALUE_AS_NUMBER) with a Unit (UNIT_CONCEPT_ID).

The DQD's flexible framework, which utilizes CSV files and SQL scripts for data quality (DQ) checks, allowed us to efficiently align and format our draft set of labs specific DQ checks to match the DQD's structure. Specifically, new SQL queries based on DQD's template queries were written and integrated into the system. Key modifications included adjustments to the R code were conducted to cover complex data checks.

A list of SQL files containing customized queries is saved in the corresponding SQL subfolder within the inst directory of the R package. By doing this, we can customize a series of SQL queries to check specific lab tests. Figure 1a provides examples of entries in the check_description.csv file, including the purpose of the check (checkDescription), the context and category of each check, and the corresponding SQL query file name.

Figure 1a. Example of check_description.csv file in the customized R package.

checkDescription	kahnContext	kahnCategory	kahnSubcat	sqlFile
Among the patients who have a positive COVID test, the number and percent of patients who do not have COVID diagnosis.	Validation	Completeness		2.sql
Among the patients who have ARF billing code, the number and percent of patients who do not have serum creatinine >=1.2mg/dL.	Validation	Completeness		3.sql
Among the patients who have COVID diagnosis, the number and percent of patients who do not have a positive COVID test.	Validation	Completeness		4.sql
Among the general population, the number and percent of patients who do not have any corresponding target lab test.	Validation	Completeness		6.1.sql
The number and percent of obese patients who do not have any target lab tests.	Validation	Completeness		7.sql
The number and percent of overweight patients who do not have any target lab tests.	Validation	Completeness		8.sql
The number and percent of patients who have no one target lab test	Validation	Completeness		9.sql
For a specific target lab test @conceptId (@conceptName), the number and percent of results whose value are outliers (>= mean + 3	Validation	Plausibility	Atemporal	12.sql
The number and percent of lab test results that are mapped to LAB_LOINC, and do not have value in either value_as_number or valu	Validation	Completeness		13.sql
The number and percent of lab tests that do not have laboratory result dates.	Verification	Completeness		14.sql
The number and percent of inpatient/emergency room visits that do not have any lab tests in the first 72 hours after admission.	Validation	Completeness		15.sql

The following SQL script in Figure 1b is for calculating the violation rate per predefined criteria.

Figure 1b. SQL script for the number and percent of inpatient/emergency room visits that do not have any lab tests in the first 72 hours after admission.

```
SELECT
    denominator.num_rows - num_not_violated_rows as num_violated_rows,
    CASE WHEN denominator.num_rows = 0 THEN 0
    ELSE 1.0 * (denominator.num_rows - num_not_violated_rows) / denominator.num_rows END AS pct_violated_rows,
    denominator.num_rows as num_denominator_rows
FROM
    (
        SELECT
            COUNT(DISTINCT vo.visit_occurrence_id) AS num_not_violated_rows
        FROM
            @cdmDatabaseSchema.VISIT_OCCURRENCE as vo,
            @cdmDatabaseSchema.MEASUREMENT as m
        where
            visit_concept_id in (262, 8717, 8870, 9201, 9203)
            and m.person_id = vo.person_id
            and vo.visit_start_date <= m.measurement_date
            and DATEADD(day, 3, vo.visit_start_date) >= m.measurement_date
    ) violated_row_count,
    (
        SELECT
            COUNT(DISTINCT visit_occurrence_id) AS num_rows
        from
            @cdmDatabaseSchema.VISIT_OCCURRENCE
        where
            visit_concept_id in (262, 8717, 8870, 9201, 9203)
    ) denominator
```

This R package can be installed internally and executed on five real-world databases that adhere to the OMOP Common Data Model (CDM). Once the initial configuration is complete, including setting the necessary parameters such as database schema, vocabulary, and output folder for results, no manual intervention is required during the execution of this algorithm.

Figure 2 below is the R code for running the customized algorithm, which performs data quality checks across the five Real-World Data (RWD) databases. The schema names for the five RWD databases are saved in omopdata, while sourcename is used to store the alias names for these datasets, making them easily identifiable later. The JSON files generated by the algorithm are named in outputFile and can be retrieved from the outputfolder specified in the code. These files can be processed to write assessment reports and develop a Shiny app.

This R code is in the R folder along with other components such as inst, tests, docs, etc., as defined in a typical R package.

The check results for each database are saved as separate JSON files, named after the respective database, in the output folder specified in the running call function. Additionally, log files and error files are generated in this folder during the execution of the algorithm. These files provide essential evidence and documentation for further troubleshooting and dissemination of the results.

RESULTS

Utilizing this customized algorithm, we performed 104 unique checks from the field, concept, and table levels in terms of validation and verification. There are total of 6255 checks on 3 categories (Completeness, Conformance, and Plausibility) and 6 subcategories(Atemporal, Relational ,Temporal, Uniqueness, and Value). The passed, failed and non-application distribution for table, field, and concept levels can be seen in Figure 3.

Figure 2. R code that performs data quality checks across the five Real-World Data (RWD).

```r
omopdata <- c("RWD1","RWD2","RWD3","RWD4","RWD5")
sourcename <- c("Source1","Source2"," Source3"," Source4"," Source5")
for (i in 5){
  cdmDatabaseSchema <- omopdata[i]
  vocabDatabaseSchema <- omopdata[i]
  resultsDatabaseSchema <- "results"
  resultsDatabaseSchema <- str_glue("work_",Sys.getenv('UID'))
  cdmSourceName <- sourcename[i]
  numThreads <- 20
  sqlOnly <- FALSE
  outputFolder <- "DQD_output"
  outputFile <- glue("{omopdata[i]}.json")
  verboseMode <- TRUE
  writeToTable <- FALSE
  checkLevels <- c("TABLE", "FIELD", "CONCEPT")
  checkNames <- c()
  DataQualityDashboard::executeDqChecks(connectionDetails = connectionDetails,
                                        cdmDatabaseSchema = cdmDatabaseSchema,
                                        resultsDatabaseSchema = resultsDatabaseSchema,
                                        vocabDatabaseSchema = vocabDatabaseSchema,
                                        cdmSourceName = cdmSourceName,
                                        numThreads = numThreads,
                                        sqlOnly = sqlOnly,
                                        outputFolder = outputFolder,
                                        outputFile = outputFile,
                                        verboseMode = verboseMode,
                                        writeToTable = writeToTable,
                                        checkLevels = checkLevels,
                                        checkNames = checkNames,)
}
```

After the successful implementation of this algorithm, five JSON files were compiled, and two types of reports were generated.

1. **R Shiny Dashboard**: An interactive dashboard summarizes the data quality assessment. This internally developed R Shiny app includes an overview of the quality assessment, results, and metadata. Users can choose which database to explore and filter the results by "Failed," "Not Applicable," or "All." The results displayed in the app can be downloaded in various file formats for documentation purposes. Figure 4a shows the interface of the internally developed Shiny dashboard, while Figure 4b demonstrates the assessment results for one of the databases selected through the Shiny app menu with the applied filter.

2. **Detailed Algorithmic Spreadsheet**: A comprehensive spreadsheet containing metadata, an overall quality summary, SQL query scripts, query file names, query purposes, thresholds, pass/fail flags, and failure rates for each check. This standalone spreadsheet serves as the exact output generated by the algorithm.

Figure 3: Distribution of DQD check levels and check count

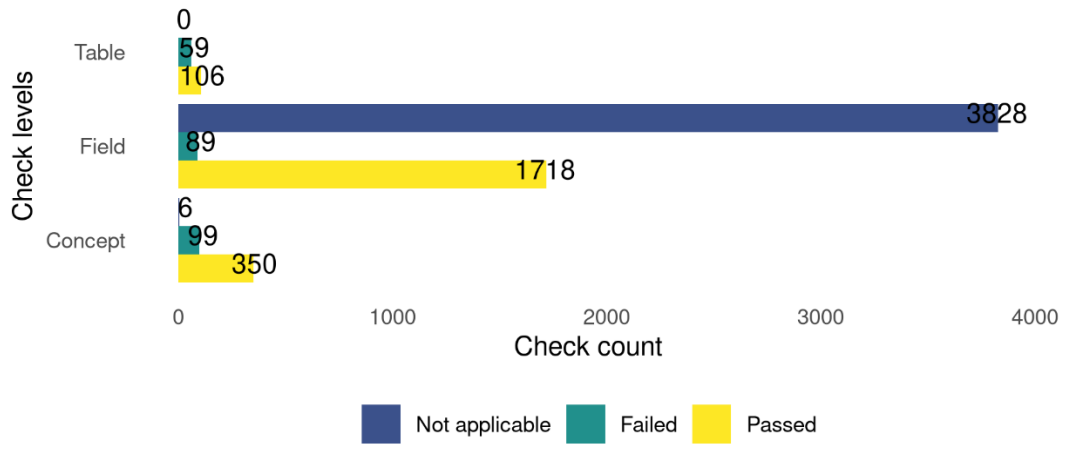


Figure 4a: Shiny dashboard interface

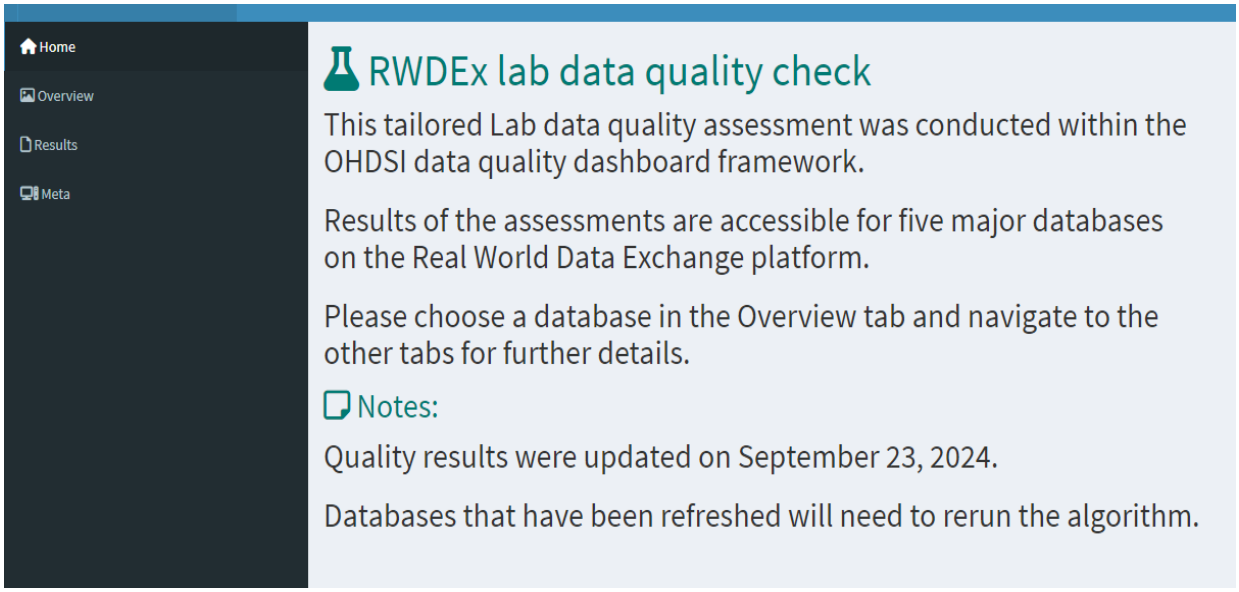
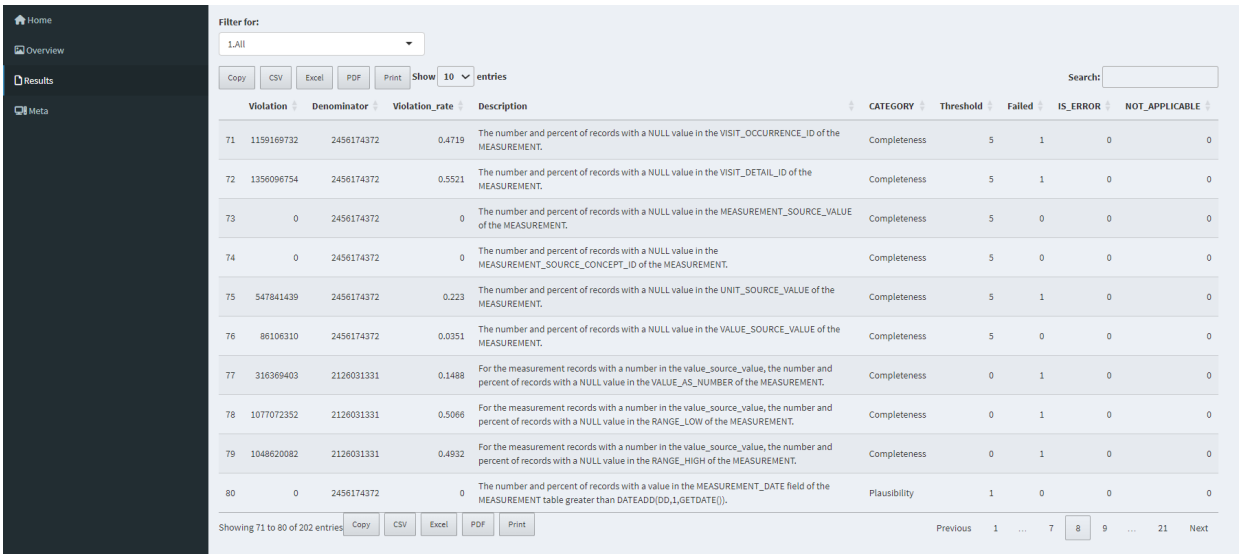


Figure 4b: Assessment results for one of the five selected databases.



Both the Shiny app and the quality check report spreadsheets offer different methods for presenting data quality assessment results, serving various purposes.

This customized Data Quality Dashboard (DQD) performs 104 unique checks. It revealed various data quality issues, including a high percentage of NULL values in specific columns of the MEASUREMENT table, discrepancies in unit standardization, and potential misclassification of categorical values. While the DQD system efficiently flagged these potential data quality issues, we will not fix the Real-World Data (RWD) issues identified in this evaluation, instead detailed check results, SQL query scripts, thresholds, and violation rates was provided to the Extraction, Transformation, and Loading (ETL) team. This information will aid the ETL team in conducting root cause analysis for future corrections. In the meantime, an R Shiny dashboard was made available for researchers to evaluate and plan their data sources for real-world data (RWD) studies.

DISCUSSION

This study reviewed existing literature and online resources, specifically focusing on data quality (DQ) issues related to EHR lab data. We customized the OHDSI Data Quality Dashboard (DQD) framework to include the DQ checks identified during the review and applied the customized DQD to internal OMOP databases.

Customization of the OHDSI DQD

The customization of the OHDSI DQD to incorporate new laboratory DQ checks highlighted several implementation challenges. First, analytical Barriers we face with returned results includes SQL errors and OMOP CDM non-compliance. The Initial SQL implementations often failed due to syntax conversions from SQL Server to Redshift, performed by SQLRender. The target databases sometimes deviated from the standard OMOP CDM by omitting certain columns or tables, which could only be resolved through collaboration between developers and data analysts.

Second, failure to meet thresholds. Failed DQ checks often result from flawed SQL implementation, ETL process issues, or inherent problems with the data sources. A thorough resolution required collaborative analysis with independently verify and adjust the data queries.

Recommendations for Future Projects

The initial framework design catered primarily to simple and moderately complex SQL queries, for more complex projects necessitated additional customizations. Enhancements to the R package were essential to implement specific DQ checks. Current DQ checks only identified the number of rows violating specified rules. A more sophisticated DQ framework is needed to advance beyond this basic indicator.

DQ issues identified using the customized DQD

Our methodology focused on using the DQ results as initial indicators of potential data problems. This process involved a detailed examination of failed DQ checks, with a primary focus on the OMOP database and its corresponding source data. Teamwork was essential in identifying the underlying causes of data quality (DQ) issues. A more in-depth collaboration with the ETL team is essential to comprehensively understand and resolve these DQ issues.

Key areas for improvement include: 1). Investigating ETL processes: It's crucial to examine the ETL procedures to identify and rectify issues that may arise during data transformation. Addressing these problems is vital for enhancing the overall DQ. 2) Source data verification: It is important to document and make these known publicly for issues stemming directly from the initial

data collection. This transparency will alert data users to potential inaccuracies or anomalies in the data, guiding more informed analysis.

In summary, while the current process effectively identifies problematic data through OMOP, a more thorough investigation involving original data sources and ETL processes is necessary for a deeper understanding and resolution of these quality concerns. This approach will clarify the origins of data discrepancies and enhance the integrity and reliability of the data analysis conducted by users.

LIMITATION

While this customized algorithm is flexible and user-friendly, there are some limitations to consider. Firstly, some complex or sophisticated quality check assessments may exceed the capabilities of SQL queries, necessitating the use of other open-source languages such as Python and R. Secondly, SQL queries may require syntax conversion when transitioning between different platforms, such as from SQL Server to Amazon Redshift. Lastly, given the extensive variety of unique laboratory tests available (e.g., approximately 55,000 unique LOINC[6] codes), creating a comprehensive list of data quality checks for all potential tests was deemed impractical within the constraints of this study. Despite these limitations, the study maintained a focus on two primary objectives:

1) Establishing a normalized process: The aim is to develop a standardized methodology for creating new lab DQ checks that are tailored to specific areas of interest. This process involves identifying essential DQ dimensions, formulating checks that address these dimensions, and continuously refining these checks based on implementation feedback.
2) Developing a technical platform: We are constructing a robust technical platform based on the OHDSI DQD. This platform will facilitate the implementation and operation of the new DQ checks, ensuring their integration into existing data systems is both effective and efficient.

The current analysis primarily inspected the transformed OMOP data without delving into the original data sources or examining the Extract, Transform, Load (ETL) processes. This approach potentially overlooked issues originating earlier in the data handling pipeline.

CONCLUSION

This paper discusses the effectiveness of the customized algorithm in evaluating RWD quality highlighting its fit-for-purpose assessment approach. Our whole process from literature review, algorithm tweaking, lab test codes refining and expanding, to implementation of R package are all documented, which increases the reproducibility, traceability, and transparence. The Shiny dashboard provides a convenient way to display check results across all databases without the need to launch multiple apps.

This customized algorithm leveraging existing open-source tools with a well-documented process and easily reproducible codes, exemplifies compliance with FDA guidance.

Although tailored to lab test data, this process can be applied to other data elements, such as medical diagnosis, procedure, and pharmacy data. Its successful implementation demonstrates a cost-efficient approach for assessing RWD usability, reliability, and a compliant process for regulatory submissions involving RWD.

REFERENCE:

1. Office of the Commissioner. Real-World Evidence. U.S. Food and Drug Administration. April 25, 2023. Accessed April 23, 2024. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
2. U.S. Food & Drug Administration. FRAMEWORK FOR FDA'S REAL-WORLD EVIDENCE PROGRAM. https://www.fda.gov/media/123160/download
3. Center for Drug Evaluation, Research. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry. U.S. Food and Drug Administration. May 8, 2020. Accessed June 20, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry
4. Center for Drug Evaluation, Research. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drug and Biological Products. U.S. Food and Drug Administration. March 15, 2023. Accessed June 20, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drug-and-biological-products
5. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit Transl Bioinform*. 2010;2010:1-5.
6. Bae CJ, Griffith S, Fan Y, et al. The Challenges of Data Quality Evaluation in a Joint Data Warehouse. *EGEMS (Wash DC)*. 2015;3(1):1125.

7. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc*. 2020;27(12):1999-2010.
8. Logical Observation Identifiers, Names and Codes. LOINC. Accessed June 20, 2024. https://loinc.org/
9. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473.
10. PCORnet Data Checks v13. https://pcornet.org/wp-content/uploads/2022/12/PCORnet-Data-Checks-v13.pdf
11. *DataQualityDashboard: A Tool to Help Improve Data Quality Standards in Observational Data Science*. Github Accessed June 20, 2024. https://github.com/OHDSI/DataQualityDashboard
12. OMOP Common Data Model. Accessed January 7, 2025. https://ohdsi.github.io/CommonDataModel/

ACKNOWLEDGMENTS

We would like to thank Dr Jian Bian and Tianchen Lyu from Health Outcomes & Biomedical informatics, University of Florida for their support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Xinyue Liu
Company: Merck & Co., Inc.
Address: 126 E Lincoln Ave, Rahway, NJ 07065 USA
Work Phone: +1 (267) 3055965
Email: xinyue.liu1@merck.com

Author Name: Lihai Song
Company: Merck & Co., Inc.
Address: 126 E Lincoln Ave, Rahway, NJ 07065 USA
Work Phone: +1 (267) 3054827
Email: lihai.song@merck.com