Journal of the American Medical Informatics Association, 28(10), 2021, 2251–2257
doi: 10.1093/jamia/ocab132
Advance Access Publication Date: 27 July 2021
Research and Applications

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD

Research and Applications

Increasing trust in real-world evidence through evaluation of observational data quality

Clair Blacketer [1,2], Frank J. Defalco,[1] Patrick B. Ryan,[1,3] and Peter R. Rijnbeek[2]

[1]Observational Health Data Analytics, Janssen Research and Development, LLC, Titusville, New Jersey, USA, [2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands, and [3]Department of Biomedical Informatics, Columbia University, New York, New York, USA

Corresponding Author: Clair Blacketer, MPH, Observational Health Data Analytics, Janssen Research and Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560, USA (mblacke@its.jnj.com)

Received 25 March 2021; Revised 25 May 2021; Editorial Decision 13 June 2021; Accepted 15 June 2021

ABSTRACT

Objective: Advances in standardization of observational healthcare data have enabled methodological breakthroughs, rapid global collaboration, and generation of real-world evidence to improve patient outcomes. Standardizations in data structure, such as use of common data models, need to be coupled with standardized approaches for data quality assessment. To ensure confidence in real-world evidence generated from the analysis of real-world data, one must first have confidence in the data itself.

Materials and Methods: We describe the implementation of check types across a data quality framework of conformance, completeness, plausibility, with both verification and validation. We illustrate how data quality checks, paired with decision thresholds, can be configured to customize data quality reporting across a range of observational health data sources. We discuss how data quality reporting can become part of the overall real-world evidence generation and dissemination process to promote transparency and build confidence in the resulting output.

Results: The Data Quality Dashboard is an open-source R package that reports potential quality issues in an OMOP CDM instance through the systematic execution and summarization of over 3300 configurable data quality checks.

Discussion: Transparently communicating how well common data model-standardized databases adhere to a set of quality measures adds a crucial piece that is currently missing from observational research.

Conclusion: Assessing and improving the quality of our data will inherently improve the quality of the evidence we generate.

Key words: Data Quality, Data Standardization, Common Data Model, Real World Evidence

INTRODUCTION

As the amount of observational health data available to researchers continues to grow, regulatory agencies like the US Food and Drug Administration (FDA)[1] and the European Medicines Agency (EMA)[2] have seen the value of using real-world evidence, but there is still some concern and lack of trust in real-world data. Use of standardized data structures like the Observed Medical Outcomes Partnership Common Data Model (OMOP CDM)[3] have allayed some of these fears, but standardization alone is not enough. Rigorous data quality assessments are needed to evaluate the quality of data with which evidence is generated.

The use of the OMOP CDM has lain the groundwork for impressive methodological and clinical breakthroughs in the areas of population-level effect estimation, patient-level prediction, and clini-

© The Author(s) 2021. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

cal characterization.[4–6] The recent *Large-Scale Evidence Generation and Evaluation across a Network of Databases* (LEGEND) study published in the Lancet is an exemplar of these ideas as the authors delivered not only relevant information about first-line antihypertensive drugs but also a novel approach to generating evidence using a systematic framework.[7] However, in order for regulatory agencies and clinicians to make decisions using such evidence, there needs to be trust not only in the methodologies employed but the underlying data itself.

The pitfalls of the secondary use of observational data to support research are well documented.[8–13] Typically these data are collected either for billing or diagnostic purposes and not with research endpoints in mind. Von Lucadou, et al notes that information in the electronic health record may not be as granular as data captured during the course of a clinical trial and time stamps of clinical events should be examined prior to inferring temporal relationships.[9] Additionally, it is often the case that clinical ideas are captured in free-text fields, as described by Varela, et al.[11] It can be easy to overlook the multiple ways a white blood cell count is recorded, for example. One physician might use "WBC" while another uses "White Blood Cell" and yet another "White BC." These types of inconsistencies can lead to misclassification and measurement error.

Such issues are concerning but not unknown to clinical research networks (CRNs). The Sentinel Initiative,[14,15] the National Patient-Centered Clinical Research Network (PCORnet),[16,17] and the Pediatric Learning Health System (PEDSnet),[18,19] among others, have all built processes and tools meant to identify data quality problems well in advance of any analytics that might be performed using the data. Historically, the Observational Health Data Sciences and Informatics (OHDSI) initiative promoted a tool known as the Automated Characterization of Health Information at Large-scale Longitudinal Evidence System (ACHILLES) to assess the quality of data in the OMOP CDM format.[20] ACHILLES is primarily used for database characterization. It computes a set of aggregate summary statistics, such as gender and age stratifications, of persons included, average follow-up time, and distribution of diagnosis codes, among others. These statistics are then assessed for quality by running the ACHILLES Heel rules against the aggregated summaries rather than on the database itself. These rules include looking for patients with an age less than zero and prescription dispensing records with implausible drug quantities.

A comparison of the data quality assessment programs across 6 different CRNs in 2017 revealed that OHDSI had the fewest data quality checks in place (172) while the other networks ranged from 875 up to 3434 checks.[21] The reason for this difference is that OHDSI, as an open collaborative, has traditionally left data quality assessment to the individual data owners. Lead protocol investigators are responsible for making a "fitness-for-use" decision and independently determine if a dataset is suitable to answer a clinical question. As OHDSI continues to move in the direction of large-scale network research,[22–24] a more robust data quality tool is needed to ensure that a participant's data comply with community defined standards.

It is our goal to address the need for better data quality processes with the development of the Data Quality Dashboard (DQD). In this article, we describe the methods used to design the tool through community engagement. Next, we describe the inner workings of the DQD and run it against a US claims database as a proof-of-concept to show utility in practice. Finally, we discuss future enhancements and the potential for this tool to change the way observational data is utilized within OHDSI and beyond.

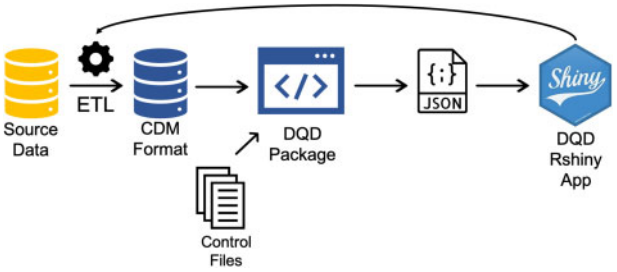

Figure 1. The standard quality control framework.

OBJECTIVE

To take advantage of the tools and methodologies[4,25] available to the OHDSI community, collaborators must first convert their data to the OMOP CDM. Data owners, clinicians, and OMOP experts all come together to standardize a source database by putting it in the structure of the model and applying agreed upon conventions through a process known as extract, transform, load (ETL). This is the expected function of any standard data model, but where OMOP differs is that not only the structure but also the content of the data is standardized—harmonizing on the SNOMED vocabulary[26] for conditions and RxNorm vocabulary[27] for drugs, for example. It is this semantic standardization that facilitates international adoption and fosters rapid collaboration.

Once a database is converted to the CDM, ideally, it should be assessed for quality prior to using it for generating evidence. Some collaborators, like PEDSNet, designed their own data quality tools to keep track of metrics of interest to them.[19] For many collaborators, the ACHILLES Heel report served this function up until now. The tool runs a set of checks against a CDM instance and reports them back to the user as an available option in the Achilles characterization package (https://github.com/ohdsi/achilles). The Heel report gives an overview of potential data quality issues, but it does not allow the user to change how a pass or fail for a given check was determined, does not share the Structured Query Language (SQL) query that was run to produce the result, nor give the option to capture any metadata about why a perceived quality issue might be occurring. There is also no process for extending the tool either by adding features or new data quality checks.

We therefore developed the DQD as a stand-alone R package to improve upon prior work and to fill in the gaps left by the Achilles Heel report. By focusing our attention on assessing data after the conversion to the OMOP CDM, the standardized structure and content directly enabled the creation of a standardized quality control framework with the DQD at the center (Figure 1).

MATERIALS AND METHODS

To lead this initiative, a few interested OHDSI collaborators formed a committee to identify high-level data quality issues important to the community. We aligned with the framework described by Kahn et al[28] as a way to organize our approach. Kahn and his colleagues identified the categories of conformance, completeness, and plausibility into which most, if not all, data quality checks can be grouped.

Conformance checks measure how well a database conforms to specified formats and relational constraints. The committee applied this idea to the OMOP CDM, describing issues such as whether a CDM instance contains all required fields, if fields that are defined

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

as primary keys contain unique values, and if a foreign key value is present in its corresponding primary key field.

Completeness checks look at the frequency of values in a given dataset without examination of the values themselves. In terms of the CDM, the idea of completeness can likewise be used to understand the quality of vocabulary mapping. The committee identified this as an area of importance for an OMOP data quality solution, including checks to evaluate the proportion of source values (ICD10CM, CPT4, Read, etc) that were not mapped to standard concepts.

Plausibility checks are meant to gauge the believability of values in a dataset. This can take many forms, like making sure a person's healthcare encounters all occur on or after their birth date or looking to see that no one has a weight of zero kilograms recorded.

After deciding on the data quality checks to implement in a new data quality tool, the committee turned their sights to the features they would like included. These were chosen based on discussions with key stakeholders from both academia and industry about their needs in this space. Considering the myriad of infrastructure constraints present in the community, it should be a stand-alone application that any user can run out-of-the box. It should be easily scalable to allow inclusion of additional checks over time. It should also be flexible to allow adjustment of the checks and failure thresholds based on apriori knowledge of a database. The results of the data quality assessment should be easily shareable in some form or fashion that is not a burden to the data owner.

With these requirements in mind, we devised a framework that would allow us to use the structure of the data model to our advantage. Since we already know the schema of every database that will run this tool, we didn't have to focus on how to write and execute individual data quality checks; instead, we were able to abstract a layer and define data quality *ideas*. For example, what if it is important to assess the number of persons in the Person table that don't have a record in the Visit_Occurrence table? That check can be assessed using a simple SQL statement against an OMOP CDM instance:

```
SELECT COUNT (DISTINCT P.person_id)
FROM PERSON as P
LEFT JOIN VISIT_OCCURRENCE as V
ON P.person_id = V.person_id
WHERE V.person_id IS NULL
```

As a data quality idea, we are simply evaluating the degree to which persons in the Person table are represented in a fact table (Visit_Occurrence, in this example). In the OMOP CDM, the field Person.person_id is a primary key with corresponding foreign keys in all clinical fact tables. Using that constraint as our guide, the abstraction of this data quality check to a data quality idea results in a SQL statement like this:

```
SELECT COUNT (DISTINCT P.person_id)
FROM PERSON as P
LEFT JOIN @cdmTable
ON P.person_id = @cdmTable.person_id
WHERE @cdmTable.person_id IS NULL
```

Where the object @cdmTable represents the universe of clinical fact tables that have a foreign key to the Person table. Each of them can be rotated into the SQL statement to take the place of the parameter. If there are 15 of such tables then the above SQL would automatically generate 15 data quality checks from 1 data quality idea.

Using the process described above as our foundation, we developed the Data Quality Dashboard (DQD) R package that systematically runs and evaluates data quality checks based on the structure of the CDM and prespecified failure thresholds.

To test it, the DQD was run on the IBM® Marketscan® Multi-State Medicaid (MDCD) database. This database contains adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses and procedures, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. The major data elements contained within this database are outpatient pharmacy dispensing claims, inpatient, and outpatient medical claims. The data does not contain laboratory results.

The use of IBM® MarketScan® Multi-State Medicaid Database was reviewed by the New England Institutional Review Board and determined to be exempt from broad Institutional Review Board approval, as this research project did not involve human patient research.

RESULTS

In total, 20 high-level data quality ideas were identified: 5 evaluating completeness, 8 evaluating conformance, and 7 evaluating plausibility, the complete list of which is included in Supplementary Appendix 1. The example given above to assess the extent to which persons in the Person table are represented in the clinical fact tables became the data quality idea, or check type, *measurePersonCompleteness*. In terms of OMOP, completeness means not only missingness but vocabulary mapping completeness. The results of 2 check types, *standardConceptRecordCompleteness* (SCRC) and *sourceValueCompleteness* (SVC), work together to show how well the diagnostic, procedural, and drug codes, etc (source values) in a database have been mapped to the standard terminology as defined by OMOP. SCRC counts the number of records for a given table that have been mapped to a Concept_Id of zero where a Standard Concept is expected. If applied to the Condition_Occurrence table, for example, it would count the number of records with a Condition_Concept_Id of zero. In a situation where a source value in a database cannot be mapped to a Standard Concept, a zero is used to denote "No matching concept." Therefore, if a large number of records in a table have a Standard Concept of zero, those records cannot be used in an analysis, as standardized analytics rely on standardized vocabularies.

The SVC check type, on the other hand, quantifies the number of distinct source values in a database that have been mapped to zero. Using the example from earlier, the values "WBC," "White Blood Cell," and "White BC" are all ways in which the clinical idea of a white blood cell count measurement might be represented. These values are all free text and as such do not have an automatic mapping to a Standard Concept. Thus, records with these values would be given a Standard Concept Id of zero during the ETL conversion process. If database A has 5000 records in the resulting Measurement table with these source values, then the SCRC check type applied to the Measurement_Concept_Id field in the Measurement table would return 5000 records mapped to zero while the SVC check type applied to the same field would return 3 distinct values mapped to zero. This can be interpreted to mean there are 5000 records representing 3 distinct values that are missing a mapping to a Standard Concept in the Measurement_Concept_Id field of database A. The relationship between these 2 checks will signal different vocabulary or ETL changes to be made. A high SCRC and high SVC

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

DATA QUALITY ASSESSMENT

IBM® MARKETSCAN® MULTI-STATE MEDICAID DATABASE

DataQualityDashboard Version: 1.0.0
Results generated at 2020-08-24 15:44:34 in 3 hours

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1849	6	1855	100%	281	6	287	98%	2130	12	2142	99%
Conformance	550	13	563	98%	80	0	80	100%	630	13	643	98%
Completeness	322	5	327	98%	12	0	12	100%	334	5	339	99%
Total	2721	24	2745	99%	373	6	379	98%	3094	30	3124	**99%**

Figure 2. The overview tab of the DQD showing data quality check pass percentages by category in MDCD.

might indicate that the database has a proprietary coding system not represented in the OMOP vocabulary. In such a case, either the coding system is added to the vocabulary or the source codes are manually mapped to standard concepts. A high SCRC and low SVC is often due to a small number of catch-all values like "unknown" or "UNK." These do not have any real meaningful analytic use, so the records are either ignored (by increasing the failure threshold) or they are removed, in which case they would no longer be visible to the DQD. A low SCRC and high SVC usually means that the source values representing the highest number of records were mapped to standard concepts and the rest were mapped to zero. This is an accepted ETL practice, as the records mapped to zero are retained for later use if necessary. A low SCRC and low SVC means that a large number of source codes representing a large number of records were mapped to standard concepts. This is the ideal scenario for an ETL, though it is important to monitor these values over time in the event the source coding practices change.

The *plausibleTemporalAfter* check type evaluates the temporal relationships between date values. Applied to the Visit_Occurrence table this check quantifies the number of visits with a visit end date prior to the visit start date. If such visits are found, the ETL must make a choice how to handle them. The most common practice is to choose either the start or end date as listed in the source data and use that value for both fields (Visit_Start_Date, Visit_End_Date) in the standardized dataset. This allows for the information from the visit to be retained while eliminating the temporal inconsistency.

When the *plausibleTemporalAfter* check type is applied to the Condition_Occurrence table, it becomes a way to measure the number of condition records that occur prior to a person's birth. There can be many reasons for records to be written with incorrect dates, but usually these are patient history records where no date was given by the patient and, instead, a default date is assigned by the electronic health record. As it is impossible to discern the correct date for these records the, typical recourse is to remove them.

These check types were all written as parameterized SQL statements. To resolve and then run these SQL statements. the DQD reads a set of included control files that detail each table and field in the CDM, their constraints, and their relationships to one another. These files also indicate which data quality check types should be run on which fields and what thresholds should be applied to the results to determine a pass or fail, all of which can be edited by the user. The tool then takes these files and swaps out the parameters in the SQL statements for the values indicated in the control files using the SqlRender R package.[29]

With this approach the 20 checks types are resolved to over 3300 individual quality checks: 396 evaluating completeness, 779 evaluating conformance, and 2126 evaluating plausibility. The DQD then compiles all results into a text- based file as the default output, though the user can also specify the option to write the results back to a table in the database. The resulting file (or table) contains all information produced by the tool and is read by the package to render an interactive Rshiny[30] user interface (Figure 1).

The first screen that greets the user is the overview tab, shown in Figure 2 displaying the result of running the DQD on the MDCD database. The tables' Note, Note_Nlp, and Specimen are not populated for MDCD, resulting in a lower total number of checks at 3124 instead of 3300. The total run time of the DQD was 7 hours for this database, sized around 700GB in a cloud-based parallel environment.

The overview tab gives a high-level summary of the total number of passes and failures by Kahn category. In Figure 2, the database examined has 13 conformance failures, or instances where it does not conform to the specifications of the OMOP CDM. It has 5 completeness failures related to potentially missing data and 12 plausibility failures which could indicate a myriad of issues including incorrect dates or implausible measurement values. To explore each of these failures, the results tab (Figure 3) shows 1 line per check run. Across the top is the option to filter by pass/fail status, CDM table, and Kahn context. The plus sign on each line expands to show the exact SQL query that was run to achieve that result, allowing the user to pinpoint the identified failing records in their dataset. A publicly available instance of these data quality results can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/, and the text file with the full result set is available in the supplementary materials.

One of the major findings from this exercise showed a high SCRC (19.20%) and high SVC (23.34%) for the Procedure_Concept_Id field of the Procedure_Occurrence table.

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

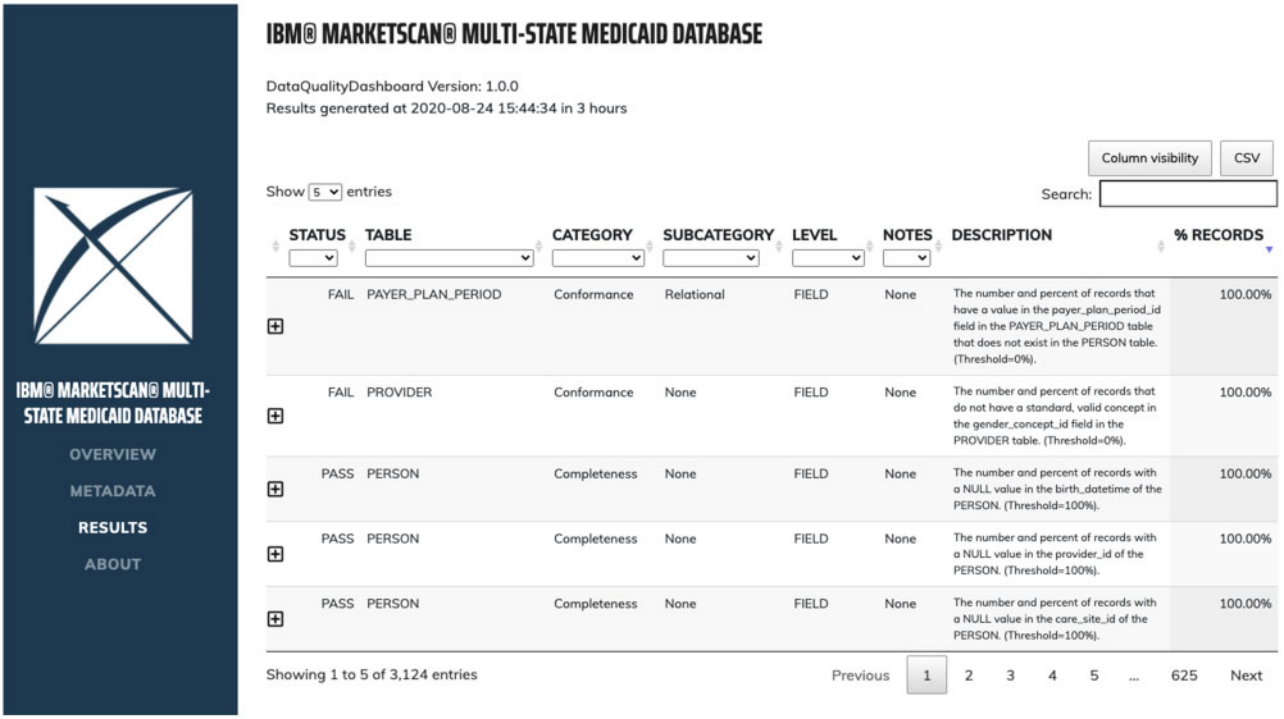

Figure 3. DQD results page with all data quality checks run and their outcomes.

Further investigation found the problem was due to a previously unaccounted-for source vocabulary. Unlike most commercial health insurance plans, state Medicaid coverage usually includes routine dental care. These procedures are coded using Current Dental Terminology (CDT). This had not been documented in the ETL, so all dental records were being mapped to a Procedure_Concept_Id of zero. To fix this issue, the CDT vocabulary was added to the ETL and the ETL was rerun.

In this case, a full batch job was completed in which the entirety of the ETL and DQD were rerun and examined. In absence of a fix, ETLs are most often run as the source data are updated. Increasing numbers of data owners have moved to an incremental model where only the newly acquired data are run through the ETL process, as this method is both time- and cost-effective. While research is ongoing in this area the current best practice is to break the DQD into 2 steps. As data is initially mapped to the OMOP CDM and the ETL is set up, only conformance and completeness checks are run to ensure the newly minted CDM database conforms to the standards set by the OHDSI community. Once those checks are satisfied the plausibility checks are run initially and each time an incremental load is processed. The underlying assumption is that once conformance and completeness is satisfied, the structure of the database should not change as long as the version of the CDM and vocabulary does not change. Plausibility should be assessed with each load to confirm that the new data complies with the gender, temporal, and measurement value quality measures.

The DQD is available on GitHub[31] as a fully executable R[32] package, supporting OMOP CDM versions 5.2.2 and 5.3.1. It works with multiple database management systems, including PostgreSQL, Microsoft SQL Server, and Redshift. All documentation can be found at https://ohdsi.github.io/DataQualityDashboard/.

DISCUSSION

As use of observational health data continues to increase, 2 of the largest regulatory agencies in the world have established plans for how these data can be used to support data-driven regulatory decision-making. The US Food and Drug Administration (FDA) published the framework for FDA's Real-World Evidence Program in 2018, which details how and in what capacity real-world evidence generated from real-world data might be evaluated for its potential use to support approvals of new drug indications or to satisfy post-marketing safety studies.[1] Similarly, the Heads of Medicines Agency (HMA) and the European Medicines Agency (EMA) in 2017 initiated a joint Big Data Task Force. A summary report from 2019 describes a strategy to understand observational data such that they might be ready to make use of it in a regulatory capacity.[2] These reports are robust, especially when discussing how real-world evidence supporting drug safety and effectiveness research should be conducted. Both groups agree that data quality is important and should be considered when determining whether a set of data is suitable to answer specific questions. This is especially timely given the COVID-19 pandemic and rush to publish any information that may further understanding of the natural history and clinical treatment of the disease.[33]

The DQD is well-poised to answer the call for higher quality data. It is at the center of the standard quality control framework (Figure 1), providing critical insights to data owners that then feed back into the ETL process, ultimately resulting in research-ready data. There is evidence of broad adoption of the DQD at the network level as the European Health Data and Evidence Network (EHDEN)[34] and the National COVID Cohort Collaborative (N3C)[35] both leverage the framework to ensure that participating databases pass critical quality control measures. As these networks

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

conduct research and learn from their data, any quality issues identified during analysis will be incorporated back into the tool. Continuous iteration on the set of data quality checks in coordination with research creates a living system that improves as we advance our understanding. The DQD will also continue to add features and expand the check types it covers. Initial roadmaps include enabling it to be run on a cohort rather than the entire database, extending the check types to include the OMOP vocabulary, incorporating prevalence measures, and evaluating temporal stability.

Other CRNs have built tools and processes for evaluating and managing the quality of data at the network level. The PEDSNet DQ Workflow[36] utilizes a similar software architecture in that it relies heavily on R, is designed to be applied to databases on the OMOP CDM standard (modified slightly for pediatric use), and can be run on multiple different database management systems (PostgreSQL, Oracle, MySQL, SQLite, or SQL Server). PEDSnet is supported by the Patient Centered Outcomes Research Institute (PCORI) and brings together data from various hospitals into 1 large, pediatric database. This structure is reflected in the DQ workflow, as the coordinating center has more control over how data quality issues are reported and tracked across the network. Sites run the quality assessment, the results of which are then sent to the coordinating center. Results are reviewed by a data scientist who discusses any issues and corrections needed with the site. The Sentinel Initiative[14,15] has a similar system where data partners run a data quality study package, results are sent to the Sentinel Operations Center, and the operations center reviews and recommends changes to the ETL.[37] Sentinel data quality checks are divided into different levels and each data partner must pass all level 1 checks before moving on to level 2, etc. PCORNet has a 5-step data curation process that also involves review of the data quality results by a coordinating center. This review is done in cycles whereby network partners are expected to correct any model conformance issues before running completeness checks, for example.

OHDSI, in contrast, is a distributed data network with no central coordination of data. Each member is responsible for their own data and ETL processes. This is a decided choice that allows for faster, broader reaching collaboration. The OMOP CDM could of course be used in a distributed network but sharing protocols and analytic code rather than data eliminates the need to take data beyond the firewall of the data holder. Such a system allows far more sites around the world to contribute as there is no longer an issue of data governance. Therefore, to ensure the quality of the OHDSI network, the DQD has to both assess and communicate the quality of a database while taking differences at the source into account. This led to a tool built using a dynamic framework that expands 20 check types into over 3300 individual checks. Data owners have full control over which checks are run and how they are assessed for failure while the resulting report details every single check, pass or fail, in an easy-to-share way. It is our vision that this becomes the way data quality is shared among networks, with reviewers, and with regulators. The most recent articles published using the PEDSnet, PCORNet, and Sentinel networks[38–40] either do not mention how the data used were assessed for quality or they touch on it very briefly. Instead, with a tool like the DQD, a file can be shared as additional material to a publication, detailing all data quality checks that were run on the database in which the study was executed. This level of transparency is unprecedented in current literature but is necessary in an era where we are asked to trust evidence generated by the scientific community with little to no insight into the data used.

CONCLUSION

The DQD has established a new way to garner trust in real-world data. Transparently communicating how well CDM standardized databases adhere to a set of quality measures adds a crucial piece that is currently missing from observational research. Assessing and improving the quality of our data will inherently improve the quality of the evidence we generate.

FUNDING

This study was sponsored by Janssen Research & Development, LLC. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

AUTHOR CONTRIBUTIONS

Conception and design—all authors; collection and assembly of data—all authors, but primarily CB; data analysis and interpretation—all authors, but primarily CB; manuscript writing—all authors; and final approval of manuscript—all authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at the *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY

The Data Quality Dashboard R package is available at https://github.com/OHDSI/DataQualityDashboard/. A publicly available instance of the data quality results for IBM MarketScan® Multi-State Medicaid can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/.

CONFLICT OF INTEREST STATEMENT

CB, FJD, and PBR are Janssen employees and own stock or stock options. PRR receives an unconditional grant from Janssen Research & Development and funding through the Innovative Medicines Initiative.

REFERENCES

1. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018. https://www.fda.gov/media/120060/download. Accessed January 10, 2020.
2. Heads of Medicines Agency, European Medicines Agency. HMA-EMA Joint Big Data Taskforce Summary Report. 2019. https://www.hma.eu/fileadmin/dateien/HMA_joint/00-_About_HMA/03-Working_Groups/Big_Data/2019_02_HMAEMA_Joint_Big_DataTaskforce_summary_report.pdf. Accessed January 10, 2020.
3. OMOP Common Data Model. http://ohdsi.github.io/CommonDataModel/. Accessed January 14, 2021.
4. Reps JM, Schuemie MJ, Suchard MA, *et al*. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018; 25 (8): 969–75.
5. Ryan PB, Schuemie MJ, Gruber S, *et al*. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Saf* 2013; 36 (Suppl 1): S59–72. doi:10.1007/s40264-013-0099-6

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021

6. Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA* 2016; 113 (27): 7329–36.
7. Suchard MA, Schuemie MJ, Krumholz HM, *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019; 394 (10211): 1816–26.
8. Botsis T, Hartvigsen G, Chen F, *et al.* Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl Bioinform* 2010; 2010: 1–5.
9. von Lucadou M, Ganslandt T, Prokosch H-U, *et al.* Feasibility analysis of conducting observational studies with the electronic health record. *BMC Med Inform Decis Mak* 2019; 19 (1): 202.
10. Prokosch HU, Ganslandt T. Perspectives for medical informatics. reusing the electronic medical record for clinical research. *Methods Inf Med* 2009; 48 (1): 38–44.
11. Otero Varela L, Wiebe N, Niven DJ, *et al.* Evaluation of interventions to improve electronic health record documentation within the inpatient setting: a protocol for a systematic review. *Syst Rev* 2019; 8 (1): 54.
12. Alla F, Rosilio M, Funck-Brentano C, *et al.* Participants of round table N° 2 of Giens Workshops XXVIII (th). How can the quality of medical data in pharmacovigilance, pharmacoepidemiology and clinical studies be guaranteed? *Therapie* 2013; 68 (4): 209–23.
13. Callahan A, Shah NH, Chen JH. Research and reporting considerations for observational studies using electronic health record data. *Ann Intern Med* 2020; 172 (Suppl 11): S79–84.
14. Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System–a national resource for evidence development. *N Engl J Med* 2011; 364 (6): 498–9.
15. Ball R, Robb M, Anderson SA, *et al.* The FDA's sentinel initiative–a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016; 99 (3): 265–8.
16. Collins FS, Hudson KL, Briggs JP, *et al.* PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014; 21 (4): 576–7.
17. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating foundational data quality in the national patient-centered clinical research network (PCORnet®). *EGEMS (Wash DC)* 2018; 6 (1): 3.
18. Forrest CB, Margolis PA, Bailey LC, *et al.* PEDSnet: a national pediatric learning health system. *J Am Med Inform Assoc* 2014; 21 (4): 602–6.
19. Utidjian LH. Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network. American Academy of Pediatrics, 2015. https://aap.confex.com/aap/2015/webprogrampress/Paper30131.html. Accessed January 13, 2021.
20. Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite evaluation of a data quality tool for patient-level clinical data sets. *EGEMS (Wash DC)* 2016; 4 (1): 1239.
21. Callahan TJ, Bauck AE, Bertoch D, *et al.* A comparison of data quality assessment checks in six data sharing networks. *EGEMS (Wash DC)* 2017; 5 (1): 8.
22. Morales DR, Conover MM, You SC, *et al.* Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis. *Lancet Digit Health* 2020: 17.
23. Schuemie MJ, Ryan PB, Pratt N, *et al.* Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study. *J Am Med Inform Assoc* 2020; 27 (8): 1268–77.
24. Golozar A, Lai LY, Sena AG, *et al.* Baseline phenotype and 30-day outcomes of people tested for COVID-19: an international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain, and the United States. *MedRxiv Prepr Serv Health Sci* 2020: 27. doi:10.1101/2020.10.25.20218875
25. Schuemie M, Suchard M, Ryan P. *CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models*. 2020. https://ohdsi.github.io/CohortMethod/index.html. Accessed January 21, 2021.
26. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html. Accessed March 4, 2021.
27. UMLS Metathesaurus - RXNORM (RXNORM) - Synopsis. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RXNORM/index.html. Accessed March 4, 2021.
28. Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 1244.
29. Schuemie M, Suchard M. *SqlRender: Rendering Parameterized SQL and Translation to Dialects*. 2020. https://ohdsi.github.io/SqlRender/. Accessed January 21, 2021.
30. Chang W, Cheng J, Allaire JJ, *et al. Shiny: Web Application Framework for R*. 2020. https://CRAN.R-project.org/package=shiny. Accessed January 10, 2021.
31. Blacketer C, DeFalco F, Londhe A, *et al.* DataQualityDashboard. https://ohdsi.github.io/DataQualityDashboard/. Accessed January 21, 2021.
32. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org/
33. Brainard J. Scientists are drowning in COVID-19 articles. Can new tools keep them afloat? *Sci AAAS* 2020. https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat. Accessed January 21, 2021.
34. European Health Data Evidence Network. ehden.eu. https://www.ehden.eu/. Accessed March 12, 2021.
35. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
36. Khare R, Utidjian LH, Razzaghi H, *et al.* Design and refinement of a data quality assessment workflow for a large pediatric research network. *EGEMS (Wash DC)* 2019; 7 (1): 36.
37. How Sentinel Gets Its Data. Sentin. Initiat. 2019. https://www.sentinelinitiative.org/about/how-sentinel-gets-its-data. Accessed February 9, 2021.
38. Bailey LC, Razzaghi H, Burrows EK, *et al.* Assessment of 135 794 pediatric patients tested for severe acute respiratory syndrome coronavirus 2 across the United States. *JAMA Pediatr* 2021; 175 (2): 176–84.
39. Bachmann KN, Roumie CL, Wiese AD, *et al.* Diabetes medication regimens and patient clinical characteristics in the national patient-centered clinical research network, PCORnet. *Pharmacol Res Perspect* 2020; 8 (5): e00637.
40. Eworuke E, Haug N, Bradley M, *et al.* Risk of non-melanoma skin cancer in association with use of hydrochlorothiazide-containing products in the United States. *JNCI Cancer Spectr* 2021; 5 (2). doi:10.1093/jncics/pkab009.

Downloaded from https://academic.oup.com/jamia/article/28/10/2251/6328963 by guest on 29 October 2021