applied sciences

MDPI

Article

Using the Data Quality Dashboard to Improve the EHDEN Network

Clair Blacketer [1,2,3,*,†], Erica A. Voss [1,2,3,*,†], Frank DeFalco [1,3], Nigel Hughes [1,3], Martijn J. Schuemie [1,3], Maxim Moinat [2,3,4] and Peter R. Rijnbeek [2,3]

1 Janssen Pharmaceutical Research and Development LLC, Titusville, NJ 08560, USA; fdefalco@its.jnj.com (F.D.); nhughes@its.jnj.com (N.H.); mschuemi@its.jnj.com (M.J.S.)
2 Department of Medical Informatics, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands; maxim@thehyve.nl (M.M.); p.rijnbeek@erasmusmc.nl (P.R.R.)
3 OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York 10027, NY, USA
4 The Hyve, 3511 MJ Utrecht, The Netherlands
* Correspondence: mblacke@its.jnj.com (C.B.); evoss3@its.jnj.com (E.A.V.)
† Co-first authors.

Abstract: Federated networks of observational health databases have the potential to be a rich resource to inform clinical practice and regulatory decision making. However, the lack of standard data quality processes makes it difficult to know if these data are research ready. The EHDEN COVID-19 Rapid Collaboration Call presented the opportunity to assess how the newly developed open-source tool Data Quality Dashboard (DQD) informs the quality of data in a federated network. Fifteen Data Partners (DPs) from 10 different countries worked with the EHDEN taskforce to map their data to the OMOP CDM. Throughout the process at least two DQD results were collected and compared for each DP. All DPs showed an improvement in their data quality between the first and last run of the DQD. The DQD excelled at helping DPs identify and fix conformance issues but showed less of an impact on completeness and plausibility checks. This is the first study to apply the DQD on multiple, disparate databases across a network. While study-specific checks should still be run, we recommend that all data holders converting their data to the OMOP CDM use the DQD as it ensures conformance to the model specifications and that a database meets a baseline level of completeness and plausibility for use in research.

Keywords: data quality; OMOP CDM; EHDEN; healthcare data; real world data; RWD

Citation: Blacketer, C.; Voss, E.A.; DeFalco, F.; Hughes, N.; Schuemie, M.J.; Moinat, M.; Rijnbeek, P.R. Using the Data Quality Dashboard to Improve the EHDEN Network. *Appl. Sci.* **2021**, *11*, 11920. https://doi.org/10.3390/app112411920

Academic Editors: Elena Cardillo, Robert Vander Stichele and Dipak Kalra

Received: 31 October 2021
Accepted: 8 December 2021
Published: 15 December 2021

1. Introduction

Over the past 30 years, more observational health data have become available for use in research due to the digitization of health records and the comprehensive nature of administrative claims [1–3]. The potential of this data is well known; the non-invasive, passive method of data collection bypasses the ethical concerns of human subjects research and the speed with which it is collected foreshadows a future of near-real time evidence generation [4,5]. Professional organizations such as the American Thoracic Society have publicly announced their intention to use observational studies to inform clinical practice guidelines [6]. Similarly, both the United States Food and Drug Administration and the European Medicines Agency have begun to rely more heavily on real-world evidence (RWE) to support critical drug safety decisions [7,8].

To maximize the research capability of such large-scale observational health data, the Innovative Medicines Initiative made funding available to develop a network of health data sources to support outcomes-focused healthcare delivery in Europe [9]. This project is known as the European Health Data and Evidence Network (EHDEN) [10]. As of October 2021, there are 98 databases in the network representing a total of 23 countries and about

450 million patient records. With the rise of COVID-19, this need for a European-wide federated network of observational health data was thrown into sharp relief as health officials scrambled to understand the natural history of the disease [11]. To aid this effort, EHDEN held a Rapid Collaboration Call inviting institutions across Europe with COVID-19 data to participate [12]. This presented a challenge: how can we be sure the data are of high quality for research while quickly integrating them into the network?

EHDEN proposes to solve this problem of data integration by choosing to harmonize the network on the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [13]. European real-world clinical data is generated from diverse medical record systems, stored in different ways, captured in different languages, and controlled by differing policy restrictions. Conversion to the OMOP CDM allows for interoperability across data sources as it standardizes real world data to both a common structure (data model) and terminology (vocabulary). This then leaves the problem of ensuring these data are of high-quality to support evidence generation.

While there is little doubt in the research utility of observational health data, concerns are often voiced about the quality of the data since it is not primarily collected for research purposes. A study published in 2018 by Pacurariu et al. detailed a review of 34 observational health databases representing Northern, Central, and Western Europe [14]. The authors found that only half had published validation studies detailing the extent to which patients' records aligned with national registries or statistics reports [14]. Among databases studied, the approaches taken to assess data quality were observed to be very different. Clinical Practice Research Datalink (CPRD), for example, provides indicators of patient and practice acceptability to researchers [15]. In contrast, the Information System for the Development of Research in Primary Care (SIDIAP) from Catalonia, Spain created a scoring system to assess the completeness of the data provided by primary care practitioners [16].

To address this lack of standard data quality procedures, EHDEN, in conjunction with the Observational Health Data Sciences & Informatics (OHDSI) initiative [17,18], developed a novel process for assessing the quality of observational health data, known as the Data Quality Dashboard (DQD) (https://ohdsi.github.io/DataQualityDashboard/, accessed on 20 September 2021) [13,19,20]. While the DQD has proven effective when applied to a singular database, it has not yet been applied to a network of databases such as those within EHDEN. For this work, the COVID-19 Rapid Collaboration Call gave us the unique opportunity to study how the use of a standard data quality procedure improves the quality of data across a federated network.

With this work herein we present the process by which Data Partners (DPs) chosen for the COVID-19 Rapid Collaboration Call mapped their data to the OMOP CDM. Throughout the process, at least two DQD results were collected and compared for each DP to understand how the tool improves the quality of observational health databases at scale. This is the first study to apply the DQD to multiple, disparate databases representing different countries and data types, and is an important step forward towards transparent data provenance.

2. Materials and Methods

2.1. Data Conversion Process and DQD Collection

Twenty-five DPs were awarded a COVID-19 Rapid Collaboration Call grant and these DPs covered 11 different countries and represented over 1 million COVID-19 patients [12]. The data sources themselves ranged from small COVID-19 registries to large medical records systems that covered lives for nearly an entire country. Once a DP was notified that they would receive a grant, they were assigned an EHDEN COVID-19 Taskforce. This taskforce consisted of a small team of technical experts with the collective skills and experience to walk DPs through the Data Conversion to Analysis Pipeline shown in Figure 1.

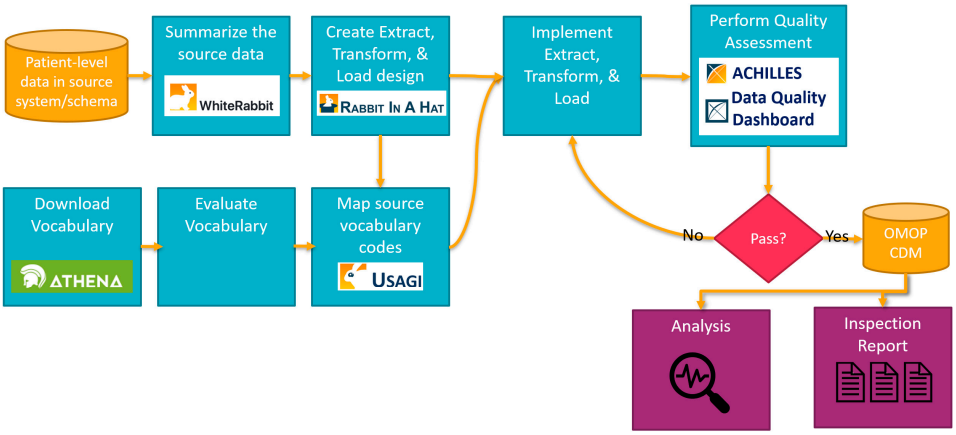

Figure 1. The Data Conversion to Analysis Pipeline to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) used by the European Health Data and Evidence Network (EHDEN).

This Data Conversion to Analysis Pipeline begins with WhiteRabbit and Rabbit-In-A-Hat [21], two tools maintained by EHDEN and used by the global OHDSI community. These tools facilitate the structural standardization of data by first revealing what tables, fields, and values are in a source data set (WhiteRabbit) and then allowing users to interactively map those source tables and fields to the CDM by way of a graphical user interface (Rabbit-In-A-Hat). Once the structural design specification is complete, standardization of the medical terminologies present in a source dataset can be done in two ways. First, one can leverage existing translations of common source terminologies (e.g., ICD10, Read, etc.) to standard concepts using the OMOP Vocabularies [22,23]. If the source data contains codes not found in the OMOP Vocabularies, a tool called Usagi [24] is used to suggest potential standard concept mappings, followed by manual review and correction. After designing the structural and semantic standardization, Rabbit-In-A-Hat produces a document containing the complete instructions to convert the source data to an OMOP CDM instance. DPs then implement these specifications in an ETL (Extract, Transform, & Load) program utilizing the hardware and technologies they have access to within their organization. Typically, this takes the form of SQL written in the database environment they already have.

Once data has been standardized, the next part of the Data Conversion to Analysis Pipeline is quality assessment, the cornerstone of which within EHDEN is the DQD [19,20]. DQD is an open-source R package supporting OMOP CDM versions 5.2.2 and 5.3.1 that uses a systematic approach to run data quality checks against an OMOP CDM instance. It works with multiple database management systems, including PostgreSQL, Microsoft SQL Server, and Amazon Redshift. This tool applies, to a given database, over 3300 data quality checks (Table A1) organized into categories first described by Kahn et al. [25]. These are *conformance* (requiring data to adhere to specified standards and formats), *completeness* (ensuring that data values are present), and *plausibility* (ensuring that the data values are believable). Using the MEASUREMENT table and the MEASUREMENT_CONCEPT_ID column as a motivating example, a *conformance* check ensures the column exists, a *completeness* check ensures it is non-zero above a given threshold, and a *plausibility* check for a given measurement (e.g., Calcium; total) and given unit (e.g., milligram per deciliter) ensures the value is biologically plausible.

The quality checks in the DQD tool can be described at a high-level by their check type. There are 20 types in total, the full list of which can be found in Appendix A. Each type represents a different data quality idea which is then automatically resolved and applied to the relevant tables and fields with data available as the DQD is run. These check types range in complexity from checking relational constraints to evaluating plausible values for measurement-unit pairs. For example, *cdmField*, a *conformance* quality check, is a check type

looking to see if all fields are present in the CDM as expected and for each field in this check type there is a result of either pass or fail based on if the field is correctly implemented.

Once the checks are applied and run, they each return the number of rows that failed the logical test and the total number of rows eligible to be inspected for the given quality metric. Dividing the number of failing rows by the total number of rows gives the proportion of failing rows for each data quality check. This proportion is then compared to a pre-specified threshold. If the proportion of failing rows exceeds the threshold, then the check is considered to have failed.

Once a DP completed the first iteration of their standardized database, they ran the DQD in their environment and produced the data quality check results. These results were then reviewed with the help of the EHDEN COVID-19 Taskforce using the DQD results viewer. An example of what the results viewer looks like can be found at https://data.ohdsi.org/DataQualityDashboardMDCD/ (accessed on 20 September 2021). Together, they determined whether the issues identified by the tool were most likely related to the source data or the ETL program. In the case of a source data problem the DPs would fix the issue, if possible, usually by instituting a rule in the ETL (e.g., dropping patients without a recorded year of birth). Occasionally issues were fixed at the source itself, such as the DP for whom everyone with a death date also had encounter records after death. This turned out to be a problem with their electronic health record system, which they went back and corrected.

In the case of a problem in the ETL program the EHDEN taskforce would work with the DP to pinpoint and repair the bugs in the code. If an issue could not be immediately fixed an explanation was written in the DQD tool clearly describing the reason for the check failure. After the issues in the source data and the ETL code were addressed, the DQD was run again. This process of running DQD and updating the ETL, source data, or DQD configuration continued until the taskforce felt confident that the database was of sufficient quality, as seen in Figure 2.

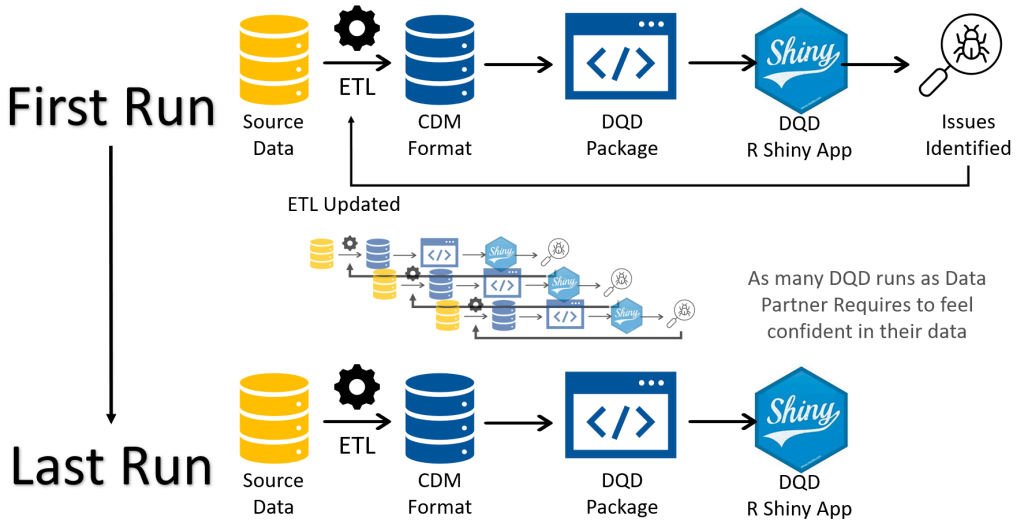

Figure 2. The Extract, Transform, & Load (ETL) to Common Data Model (CDM) to Data Quality Dashboard (DQD) Feedback Loop used in the European Health Data & Evidence Network (EHDEN) COVID-19 Rapid Collaboration Call.

2.2. Data Quality Dashboard Comparisons

At the end of the entire data conversion process, when a DP achieved an OMOP CDM-formatted database they and the taskforce were confident in, both the first and last DQD reports were collected. The data quality checks comprising these reports were then organized into four outcome categories: a *pass* indicated that the percentage of failing rows for a given quality check was below the threshold specified in the DQD tool, a *fail* indicated that the percentage of failing rows for a given quality check was above the threshold

specified, an *error* indicated that there was an error in the SQL and no result for the given quality check was returned, and *no data* indicated that there was no data in the database that satisfied the criteria for a given quality check.

$$\frac{\text{\# of passed checks}}{(\text{\# of passed checks} + \text{\# of failed checks} + \text{\# of check errors})} \times 100 \quad (1)$$

After organizing the data quality checks by outcome, the results from the first DQD report were compared with the results from the last DQD report. Initially, a simple check pass percentage (e.g., 95% of the DQD checks passed) was obtained by DP and time (first or last) using Equation (1) to determine if the percent of data quality checks passed in the last run of the DQD was higher than the percent of data quality checks passed in the first run of the DQD. It was then observed that almost all DPs ran a different number of quality checks at first versus at last. Upon inspection it was revealed that this phenomenon was due to a few factors: the iterative ETL process, the automatic nature of the tool, and the fact that DPs removed checks with no data to support them from the calculation. For example, many DPs entered the Data Conversion to Analysis Pipeline beginning with the information they believed was easiest to standardize, typically patient demographics and patient diagnoses. Once they achieved the first standardized database containing this information, many ran the DQD with only a subset of the total checks turned on. These tended to be checks that evaluated conformance to the model and completeness of standard vocabulary mapping. As they iterated through the process, the DPs continued to turn on more checks, usually those evaluating plausibility. To account for this, and to enable more appropriate comparisons, we conducted a second analysis limiting to only those data quality checks that were run both during the first DQD and the last DQD.

As mentioned above, the DQD determines a pass or a fail by comparing the percentage of rows that fail a check to a pre-specified threshold. While this gives an immediate idea of how well a quality check performed in a database, it only provides part of the story. The binary nature of assigning a pass or fail label overlooks the notion of improvement. Since we were interested in how well the DPs improved the quality of their database, we quantified, by DP, the number of checks that decreased the percent failing rows from first to last DQD run in addition to the passes and fails.

To gain insight into which types of checks failed most often in the beginning compared to the end of the data conversion process we also stratified the percent of passing checks by check type and time (first or last).

3. Results

At the time of this writing, 15 of the 25 DPs completed their DQD review, 8 had not completed, and 2 were not considered, as they did not run DQD for various reasons. Table 1 lists, in alphabetical order, the 15 DPs that contributed data for this work. These 15 DPs represented 10 countries, primarily were electronic medical record (EHR) data, and ranged in size from 766 subjects to 2.8 M subjects. The types of subjects captured within each DP were diverse in type (e.g., general practitioner data versus hospital data), depth (e.g., data only about COVID-19 care to data on all clinical care) and complexity (e.g., 1 table to 100s of tables). These DPs also represented diverse technology platforms that include several types of database management systems (the most popular being PostgreSQL, Microsoft SQL Server, and Amazon Redshift), differing infrastructure set up (e.g., on premise versus in the cloud), and differing number of processors (1 to 56) and memory sizes (8 GB to 1.5 TB) at their disposal. Additional information about each DP can be found at https://www.ehden.eu/datapartners/ (accessed on 1 October 2021). For the rest of this paper, these DPs will be anonymized.

Table 1. Data Partners (DPs) that contributed data (alphabetical order).

Data Partner	Data Type	Country	Total No of Subjects
Amsterdam University Medical Centers—COVID-19 ICU Decision Support (Amsterdam UMC) *	EHR	The Netherlands	1.8 K
Clinical Practice Research Datalink (CPRD) AURUM	EHR	United Kingdom	409 K **
Danish Colorectal Cancer Group (DCCG) with Center for Surgical Science	Patient Registry	Denmark	76.8 K
Fondazione IRCCS Istituto Neurologico Carlo Besta COVID-19 Database (FINCB) *	Patient Registry	Italy	766
Health Data Warehouse of Assistance Publique—Hopitaux de Marseille (AP-HM)	EHR	France	2.5 M
Health Informatics Centre (HIC)	EHR	United Kingdom	1.2 M
Heliant—University Clinical Center of Serbia (CCSERBIA)	EHR	Serbia	860 K
Information System for Research in Primary Care—Hospitalization Linked Data (SIDIAP)	EHR	Spain	7.8 M
Istanbul University—Faculty of Medicine (IU) *	EHR	Turkey	59.6 K
LynxCare	EHR	Belgium	10 K
Medaman Hospital Data (MHD)	EHR	Belgium	104 K
Parc de Salut Mar Barcelona and Hospital del Mar Medical Research Institute (IMIM)	EHR	Spain	956 K
Servicio Cántabro de Salud and IDIVAL (IDIVAL)	EHR	Spain	580 K
UK Biobank via University College London Institute of Health Informatics	Biobank	United Kingdom	500 K
Unidade Local de Saúde de Matosinhos (ULSM) COVID-19 *	EHR	Portugal	9.7 K

EHR = electronic health record; * = Common Data Model only contain COVID-19 subjects, ** = small subset of data for conversion purposes.

For each of the 15 DPs, the minimum number of DQD runs was 2 times, maximum 33, with a median of 4 times. The total execution time on the first run of the DQD ranged from 0.02 h to 10.5 h with a median of 0.7 h. The total execution time on the last run of the DQD ranged from 0.02 h to 25.0 h with a median of 0.7 h. These ranges were heavily dependent on the size of the database, the size of the server the DP used, and the amount of compute available. The duration of time over which the first and last DQD were run ranged from 10 days to 190 days, with a median of 90 days. The number of DQD runs and time to completion were impacted by many things such as the number of issues identified by DQD that needed to be fixed, desire to fix all issues identified by DQD, the amount of time a specific DP could afford to spend working on their ETL conversion, and experience with ETL conversions and the ETL technology required to do so (thus higher number of subjects did not necessarily equate to more DQD issues as there are many factors at play).

After limiting the comparison to the checks that were run both at first and at last, all DPs had a higher percent passing in the last run of the DQD compared to the first (Table 2), with a median of 883 total checks evaluated. We also quantified the number of checks that saw a reduction in the percent of failing rows even though they still failed the DQD check. Most DPs only had 1 or 2 checks that were labeled as failures though they showed improvement, except for DP8 and DP9, which had 12 and 8 respectively.

Table 2. The proportion of data quality checks that either passed or reduced failing rows between the first and last run of Data Quality Dashboard (DQD), by Data Partner (DP), for checks in common between the two.

Data Partner	First DQD Run			Last DQD Run			
	Passing Checks	Number of Checks	% Passing	Passing Checks	Checks with Failing Rows	Number of Checks	% Passing or Decrease Failing Rows
1	981	988	99.3%	984	0	988	99.6%
2	677	696	97.3%	678	1	696	97.6%
3	712	723	98.5%	721	0	723	99.7%
4	738	963	76.6%	911	2	963	94.8%
5	799	925	86.4%	916	0	925	99.0%
6	1015	1055	96.2%	1046	2	1055	99.3%
7	872	877	99.4%	875	0	877	99.8%
8	744	773	96.2%	751	12	773	98.7%
9	922	959	96.1%	942	8	959	99.1%
10	697	803	86.8%	802	0	803	99.9%
11	778	883	88.1%	826	2	883	93.8%
12	668	696	96.0%	687	1	696	98.9%
13	858	952	90.1%	937	4	952	98.8%
14	949	1044	90.9%	1028	2	1044	98.7%
15	822	871	94.4%	861	0	871	98.9%

When expanded to all checks run (not just those in common between the first and last) and stratified by check type, we found that more total checks were evaluated in the last run for all check types except cdmField and measurePersonCompleteness (Table 3). With this increase in total checks, we also saw an increase in the percent of passing checks for all check types from first to last.

Table 3. The proportion of passing data quality checks between the first and last run of Data Quality Dashboard (DQD), for all checks run, by high-level check type and Summarized across all Data Partners (DPs).

Kahn Category	Check Type	First DQD Run			Last DQD Run		
		Passing Checks	Number of Checks	% Passing	Passing Checks	Number of Checks	% Passing
Conformance	FKCLASS	8	12	66.7%	14	14	100.0%
Conformance	CDMDATATYPE	1143	1147	99.7%	1268	1268	100.0%
Conformance	ISREQUIRED	740	768	96.4%	853	854	99.9%
Conformance	CDMFIELD	4215	4341	97.1%	4011	4017	99.9%
Conformance	ISPRIMARYKEY	159	162	98.1%	183	185	98.9%
Conformance	ISSTANDARDVALIDCONCEPT	344	407	84.5%	415	423	98.1%
Conformance	ISFOREIGNKEY	862	1116	77.2%	1097	1119	98.0%
Conformance	FKDOMAIN	235	332	70.8%	328	340	96.5%
Completeness	MEASUREVALUECOMPLETENESS	2244	2313	97.0%	2499	2501	99.9%
Completeness	SOURCECONCEPTRECORD COMPLETENESS	128	141	90.8%	142	146	97.3%
Completeness	SOURCEVALUECOMPLETENESS	237	254	93.3%	271	281	96.4%
Completeness	MEASUREPERSONCOMPLETENESS	176	216	81.5%	193	201	96.0%
Completeness	STANDARDCONCEPTRECORD COMPLETENESS	327	381	85.8%	377	416	90.6%
Plausibility	PLAUSIBLEGENDER	1017	1228	82.8%	1410	1481	95.2%
Plausibility	PLAUSIBLEVALUELOW	383	435	88.0%	584	618	94.5%
Plausibility	PLAUSIBLETEMPORALAFTER	265	334	79.3%	355	377	94.2%
Plausibility	PLAUSIBLEDURINGLIFE	146	207	70.5%	242	263	92.0%
Plausibility	PLAUSIBLEVALUEHIGH	372	406	91.6%	514	594	86.5%

We found that the data quality checks that failed most often in the first run of the DQD were related to checks ensuring that the database conforms to the specifications of the OMOP CDM. The worst offenders in the first run were the checks in the fkclass and fkdomain check types, which are conformance checks reviewing if the classes and domains of mapped terminology are appropriate. Initially only 66.7% passed for fkclass and 70.8%

passing for fkdomain. While overall the conformance checks failed most often in the first run of the DQD, they also showed the most improvement from first to last, achieving between 100% and 96.5% passing at last run.

The percent of passing completeness and plausibility checks also increased from first to last run of the DQD, but not as drastically as the conformance checks. The most improved completeness checks belonged to the measurePersonCompleteness check type, going from 81.5% passing at first to 96% passing at last. Similarly, the most improved plausibility checks belonged to the plausibleDuringLife check type, going from 70.5% passing at first to 92% passing at last.

4. Discussion

Throughout the COVID-19 Rapid Collaboration Call, selected DPs mapped their data to the OMOP CDM and applied the DQD at least twice along the way. These participants were from different countries with different types of data, all working toward the goal of developing a research-ready standardized database. Multiple DQD results per DP allowed for comparisons to understand how standard data quality procedures improve the quality of a federated network. When limited to checks that were in common between the first and last run of the DQD, we saw an increase in the percent of passing data quality checks for all DPs. In other words, the DQD helped all DPs improve their data quality.

Moreover, when we expanded our scope to look at all checks that were run, stratifying on check type, we found that a higher number of checks were run at last then at first. The only two check types that did not follow this pattern were cdmField and measurePersonCompleteness because some DPs turned off checks for tables that did not have data. Equation (1) removes almost all checks without data to support them from the percent passing calculation; however, based on how the cdmField and measurePersonCompleteness check types function, they would still report a result even if there was no data in the table. Turning off the checks in the tool for these tables with no data helped the DPs to remove failures that were uninformative for their data.

As the DPs continued to iterate through the process and run the DQD, the increase in the total number of checks run tells us that the tool verified the past issues and then continued to check new items. These new items were a result of either the DPs turning on checks they intentionally turned off in the first run, they were a result of continued vocabulary mapping, or additional data being added to the OMOP CDM. The more proprietary source codes were mapped to standard concepts, the more checks the DQD could leverage against those records. Specifically, these are checks evaluating measurement plausibility as they are dependent on individual concepts representing individual source codes. Ultimately, the DQD became more and more comprehensive the better the DPs got at mapping their data to the OMOP CDM.

We also found that the conformance checks improved the most from first to last as compared with the completeness and plausibility checks. The DQD seemed the most effective at iterating on and helping to improve those checks that assess the *conformance* to the model. While not perfect at the last run, all *conformance* checks achieved a percent passing of 96.5% or higher, with two achieving 100%. The handful of primary key and foreign key checks that did not pass during the last DQD run were found to be related to the VISIT_OCCURRENCE_IDs for two DPs. These are required to be unique in the VISIT_OCCURRENCE table. One DP was perfect on the first run but showed a handful of duplicates in the last run. Most likely this was due to the increase in records from first to last; the ETL performed well on a small subset but introduced duplicates when expanded to the full database. The other DP showed great improvement on this issue between first and last, going from 1.3 M to 4 K offending records. Since it was not 100% perfect in the last run it was still labelled as a failure though the DP clearly worked to address the problem.

Viewed from the network level this focus on adherence to model specifications is extremely important. A federated network functions with each data holder maintaining their OMOP CDM compliant database behind a firewall. They participate in studies by

running standardized code against their OMOP CDM instance and reporting the results. Therefore, if the database does not conform to the model, they would not be able to run the study code, let alone report results.

This was especially evident as DPs easily ran a network-based research study after the quality assessment was completed. When the first handful of DPs achieved a standardized database both they and the taskforce were confident in, they were invited to participate in an ongoing research study on adverse events of special interest (AESI) for the COVID-19 vaccines [26,27]. Thanks to the data quality review all DPs were able to successfully run the package without any major *conformance* issues hindering their involvement.

While the DQD functioned well to help DPs address issues of *conformance*, we saw less improvement in the *completeness* and *plausibility* checks. There was still an increase in the percent of passing checks from first to last run, but it was less dramatic than that of the *conformance* checks. This was partly due to the amount of work needed to improve each one. *Completeness*, in terms of the OMOP CDM, can mean data completeness but also completeness of standard vocabulary mapping. If a DP has a large source vocabulary that needs to be mapped to standard terminologies, this work can take time to address, and it is only once hundreds or thousands of codes are mapped to standard concepts that the data quality check moves from a fail to a pass.

In addition to the amount of work required, we often found it difficult to assess the *completeness* and *plausibility* checks when considering the entire database. The number of records that needed to be addressed was daunting and there was no way to know which ones would potentially impact an analysis. Therefore, our recommendation is to institute not only database-level data quality checks but study-specific checks. It is important to know the database is conformant enough to run a study, but once divided into cohorts, the field of focus for data quality issues is narrowed and may reveal items that were masked at the higher level. For example, when running the AESI COVID-19 vaccine study, one DP found that most of the patients in the defined study cohort did not have visit information. However, another discovered that many of the lab values of interest for the analysis did not have units of measure mapped to standard concepts. Both of these items were identified during execution of the study, addressed, and then the study was re-run.

5. Conclusions

This is the first study to date that demonstrates the DQD, as an analytic tool, can be applied to a diverse collection of databases representing different types of data across many countries. The DQD improved the quality of all participating databases, rendering them ready for analysis. We recommend this practice to all data holders either working to convert their data to the OMOP CDM or who already have an OMOP CDM instance. While study-specific quality checks should still be performed, the DQD ensures *conformance* to the model specifications and that a database meets a baseline level of *completeness* and *plausibility* for use in research.

Author Contributions: Conceptualization, C.B., E.A.V., M.J.S. and P.R.R.; Data curation, C.B. and E.A.V.; Formal analysis, C.B. and E.A.V.; Visualization, E.A.V. and F.D.; Writing—original draft, C.B. and E.A.V.; Writing—review & editing, F.D., N.H., M.J.S., M.M. and P.R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 research and innovation programme and EFPIA, grant number 806968. The APC was funded by Janssen Research & Development, LLC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors want to thank the data partners who participated in this research and the Innovative Medicines Initiative.

Conflicts of Interest: C.B., E.A.V., F.D., N.H. and M.J.S. are employees of Janssen Research & Development, LLC and Johnson & Johnson shareholders. M.M. and P.R.R. has no conflicts to declare.

Appendix A. Data Quality Check Types (Ideas) by Context and Category

The information within this table can additionally be found here: https://ohdsi.github.io/DataQualityDashboard/articles/CheckTypeDescriptions.html (accessed on 1 October 2021).

Table A1. Kahn category.

Check Name	Check Description	Kahn Category [1]
cdmField	A yes or no value indicating if all fields are present in the @cdmTableName table as expected based on the specification.	Conformance
isRequired	The number and percent of records with a NULL value in the @cdmFieldName of the @cdmTableName that is considered not nullable.	Conformance
cdmDatatype	A yes or no value indicating if the @cdmFieldName in the @cdmTableName is the expected data type based on the specification.	Conformance
isPrimaryKey	The number and percent of records that have a duplicate value in the @cdmFieldName field of the @cdmTableName.	Conformance
isForeignKey	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that does not exist in the @fkTableName table.	Conformance
fkDomain	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that do not conform to the @fkDomain domain.	Conformance
fkClass	The number and percent of records that have a value in the @cdmFieldName field in the @cdmTableName table that do not conform to the @fkClass class.	Conformance
isStandardValid Concept	The number and percent of records that do not have a standard, valid concept in the @cdmFieldName field in the @cdmTableName table.	Conformance
measureValue Completeness	The number and percent of records with a NULL value in the @cdmFieldName of the @cdmTableName.	Completeness
measurePerson Completeness	The number and percent of persons in the CDM that do not have at least one record in the @cdmTableName table	Completeness
standardConcept RecordCompleteness	The number and percent of records with a value of 0 in the standard concept field @cdmFieldName in the @cdmTableName table.	Completeness
sourceConcept RecordCompleteness	The number and percent of records with a value of 0 in the source concept field @cdmFieldName in the @cdmTableName table.	Completeness
sourceValue Completeness	The number and percent of distinct source values in the @cdmFieldName field of the @cdmTableName table mapped to 0.	Completeness
plausibleValueLow	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName table less than @plausibleValueLow.	Plausibility

Table A1. *Cont.*

Check Name	Check Description	Kahn Category [1]
plausibleValueHigh	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName table greater than @plausibleValueHigh.	Plausibility
plausible TemporalAfter	The number and percent of records with a value in the @cdmFieldName field of the @cdmTableName that occurs prior to the date in the @plausibleTemporalAfterFieldName field of the @plausibleTemporalAfterTableName table.	Plausibility
plausible DuringLife	If yes, the number and percent of records with a date value in the @cdmFieldName field of the @cdmTableName table that occurs after death.	Plausibility
plausibleValueLow	For the combination of CONCEPT_ID @conceptId (@conceptName) and UNIT_CONCEPT_ID @unitConceptId (@unitConceptName), the number and percent of records that have a value less than @plausibleValueLow.	Plausibility
plausibleValueHigh	For the combination of CONCEPT_ID @conceptId (@conceptName) and UNIT_CONCEPT_ID @unitConceptId (@unitConceptName), the number and percent of records that have a value higher than @plausibleValueHigh.	Plausibility
plausibleGender	For a CONCEPT_ID @conceptId (@conceptName), the number and percent of records associated with patients with an implausible gender (correct gender = @plausibleGender).	Plausibility

[1] Kahn, M.G.; Callahan, T.J.; Barnard, J.; Bauck, A.E.; Brown, J.; Davidson, B.N.; Estiri, H.; Goerg, C.; Holve, E.; Johnson, S.G.; et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* **2016**, *4*, 1244, doi:10.13063/2327-9214.1244.

References

1. Mitchell, J.B.; Bubolz, T.; Paul, J.E.; Pashos, C.L.; Escarce, J.J.; Muhlbaier, L.H.; Wiesman, J.M.; Young, W.W.; Epstein, R.S.; Javitt, J.C. Using Medicare claims for outcomes research. *Med. Care* **1994**, *32*, 38–51. [CrossRef]
2. Lewis, N.J.W.; Patwell, J.T.; Briesacher, B.A. The Role of Insurance Claims Databases in Drug Therapy Outcomes Research. *PharmacoEconomics* **1993**, *4*, 323–330. [CrossRef] [PubMed]
3. Adler-Milstein, J.; DesRoches, C.M.; Kralovec, P.; Foster, G.; Worzala, C.; Charles, D.; Searcy, T.; Jha, A.K. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. *Health Aff.* **2015**, *34*, 2174–2180. [CrossRef] [PubMed]
4. Scherer, D.G.; Annett, R.D.; Brody, J.L. Ethical Issues in Adolescent and Parent Informed Consent for Pediatric Asthma Research Participation. *J. Asthma* **2007**, *44*, 489–496. [CrossRef] [PubMed]
5. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
6. Gershon, A.S.; Lindenauer, P.K.; Wilson, K.C.; Rose, L.; Walkey, A.J.; Sadatsafavi, M.; Anstrom, K.J.; Au, D.H.; Bender, B.G.; Brookhart, M.A.; et al. Informing Healthcare Decisions with Observational Research Assessing Causal Effect. An Official American Thoracic Society Research Statement. *Am. J. Respir. Crit. Care Med.* **2021**, *203*, 14–23. [CrossRef] [PubMed]
7. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. Available online: https://www.fda.gov/media/120060/download (accessed on 12 October 2021).
8. Heads of Medicines Agency. European Medicines Agency HMA-EMA Joint Big Data Taskforce Summary Report. Available online: https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf (accessed on 12 October 2021).
9. IMI2–Call 12. Available online: http://www.imi.europa.eu/apply-funding/closed-calls/imi2-call-12 (accessed on 15 October 2021).
10. European Health Data Evidence Network. Available online: https://www.ehden.eu/ (accessed on 12 March 2021).
11. Hrabovszki, G. COVID-19: Latest Updates. Available online: https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/covid-19-latest-updates (accessed on 15 October 2021).
12. 04/2020—COVID19 Rapid Collaboration Call. Available online: https://www.ehden.eu/open-calls/04-2020-covid19-data-partner-call/ (accessed on 20 September 2021).
13. OMOP Common Data Model. Available online: http://ohdsi.github.io/CommonDataModel/ (accessed on 14 January 2021).

14. Pacurariu, A.; Plueschke, K.; Mcgettigan, P.; Morales, D.R.; Slattery, J.; Vogl, D.; Goedecke, T.; Kurz, X.; Cave, A. Electronic healthcare databases in Europe: Descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open* **2018**, *8*, e023090. [CrossRef] [PubMed]
15. Herrett, E.; Gallagher, A.M.; Bhaskaran, K.; Forbes, H.; Mathur, R.; van Staa, T.; Smeeth, L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* **2015**, *44*, 827–836. [CrossRef] [PubMed]
16. Del Mar García-Gil, M.; Hermosilla, E.; Prieto-Alhambra, D.; Fina, F.; Rosell, M.; Ramos, R.; Rodriguez, J.; Williams, T.; van Staa, T.; Bolibar, B. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). *Inform. Prim. Care* **2011**, *19*, 135–145. [CrossRef] [PubMed]
17. Observational Health Data Sciences & Informatics (OHDSI). Available online: www.ohdsi.org (accessed on 20 September 2021).
18. Hripcsak, G.; Duke, J.D.; Shah, N.H.; Reich, C.G.; Huser, V.; Schuemie, M.J.; Suchard, M.A.; Park, R.W.; Wong, I.C.K.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Re-searchers. *Study Health Technol. Inform.* **2015**, *216*, 574–578.
19. Blacketer, C.; Defalco, F.J.; Ryan, P.B.; Rijnbeek, P.R. Increasing trust in real-world evidence through evaluation of observational data quality. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2251–2257. [CrossRef] [PubMed]
20. DataQualityDashboard (DQD). Available online: https://ohdsi.github.io/DataQualityDashboard/ (accessed on 10 October 2021).
21. WhiteRabbit. Available online: https://github.com/OHDSI/WhiteRabbit (accessed on 20 September 2021).
22. Vocabulary. Available online: https://github.com/OHDSI/Vocabulary-v5.0 (accessed on 12 October 2021).
23. The Book of OHDSI. Available online: https://ohdsi.github.io/TheBookOfOhdsi/ (accessed on 22 September 2021).
24. Usagi. Available online: https://github.com/OHDSI/Usagi (accessed on 22 September 2021).
25. Kahn, M.G.; Callahan, T.J.; Barnard, J.; Bauck, A.E.; Brown, J.; Davidson, B.N.; Estiri, H.; Goerg, C.; Holve, E.; Johnson, S.G.; et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Gener. Évid. Methods Improv. Patient Outcomes)* **2016**, *4*, 18–1244. [CrossRef] [PubMed]
26. Releases OHDSI/CommonDataModel. Available online: https://github.com/OHDSI/CommonDataModel/releases (accessed on 10 October 2021).
27. Calculating the Background Rates of Adverse Events of Special Interest (AESI) for the COVID Vaccines. Available online: https://github.com/ohdsi-studies/Covid19VaccineAesiIncidenceRate (accessed on 18 October 2021).