HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2020 March 01.

Published in final edited form as:

J Biomed Inform. 2019 March ; 91: 103119. doi:10.1016/j.jbi.2019.103119.

ADEpedia-on-OHDSI: A Next Generation Pharmacovigilance Signal Detection Platform Using the OHDSI Common Data Model

Yue Yu, PhD[1], **Kathryn J. Ruddy, MD, MPH**[2], **Na Hong, PhD**[1], **Shintaro Tsuji, PhD**[1], **Andrew Wen, MS**[1], **Nilay D. Shah, PhD**[1,3], and **Guoqian Jiang, MD, PhD**[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[2]Department of Oncology, Mayo Clinic, Rochester, MN

[3]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

Abstract

Objective—Supplementing the Spontaneous Reporting System (SRS) with Electronic Health Record (EHR) data for adverse drug reaction detection could augment sample size, increase population heterogeneity and cross-validate results for pharmacovigilance research. The difference in the underlying data structures and terminologies between SRS and EHR data presents challenges when attempting to integrate the two into a single data base. The Observational Health Data Sciences and Informatics (OHDSI) collaboration provides a Common Data Model (CDM) for organizing and standardizing EHR data to support large-scale observational studies. The objective of the study is to develop and evaluate an informatics platform known as ADEpedia-on-OHDSI, where spontaneous reporting data from FDA's Adverse Event Reporting System (FAERS) is converted into the OHDSI CDM format towards building a next generation pharmacovigilance signal detection platform.

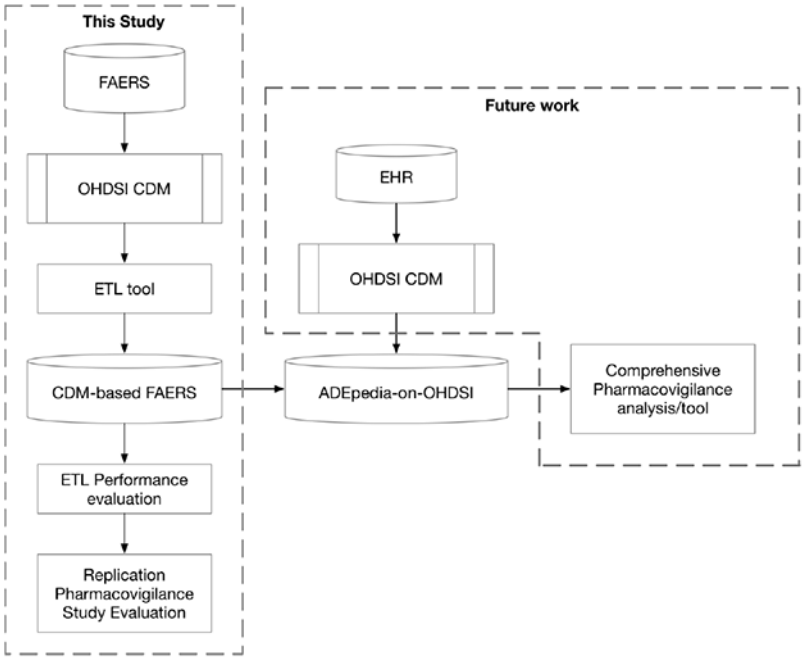
Methods—An extraction, transformation and loading (ETL) tool was designed, developed, and implemented to convert FAERS data into the OHDSI CDM format. A comprehensive evaluation, including overall ETL evaluation, mapping quality evaluation of drug names to RxNorm, and an evaluation of transformation and imputation quality, was then performed to assess the mapping accuracy and information loss using the FAERS data collected between 2012 and 2017. Previously published findings related to vascular safety profile of triptans were validated using ADEpedia-on-OHDSI in pharmacovigilance research. For the triptan-related vascular event detection, signals were detected by Reporting Odds Ratio (ROR) in high-level group terms (HLGT) level, high-level terms (HLT) level and preferred term (PT) level using the original FAERS data and CDM-based FAERS respectively. In addition, six standardized MedDRA queries (SMQs) related to vascular events were applied.

Results—A total of 4,619,362 adverse event cases were loaded into 8 tables in the OHDSI CDM. For drug name mapping, 93.9% records and 47.0% unique names were matched with RxNorm

codes. Mapping accuracy of drug names was 96% based on a manual verification of randomly sampled 500 unique mappings. Information loss evaluation showed that more than 93% of the data is loaded into the OHDSI CDM for most fields, with the exception of drug route data (66%). The replication study detected 5, 18, 47 and 6, 18, 50 triptan-related vascular event signals in MedDRA HLGT level, HLT level, and PT level for the original FAERS data and CDM-based FAERS respectively. The signal detection scores of six standardized MedDRA queries (SMQs) of vascular events in the raw data study were found to be lower than those scores in the CDM study.

Conclusion—The outcome of this work would facilitate seamless integration and combined analyses of both SRS and EHR data for pharmacovigilance in ADEpedia-on-OHDSI, our platform for next generation pharmacovigilance.

Graphical Abstract



Keywords

OHDSI Common Data Model; FAERS; Pharmacovigilance; ETL tool; Data Standardization

1 Introduction

Pharmacovigilance is a science which focuses on the collection, detection, assessment, and prevention of adverse effects associated with usage of pharmaceutical products. The US Food and Drug Administration (FDA)'s Adverse Event Reporting System (FAERS), a spontaneous reporting system (SRS), is a valuable resource for pharmacovigilance [1]. A key challenge with the use of FAERS is data quality: adverse events are reported by manufacturers, physicians, pharmacists, nurses and consumers, which leads to data duplication and a lack of standardization. Further, because not all adverse drug reactions (ADRs) are captured in FAERS, longitudinal observational databases like Electronic Health

Records (EHRs) and transactional claims can be used as additional data sources for pharmacovigilance to address gaps in coverage[2]. Using both SRS and EHR data simultaneously for pharmacovigilance could augment sample size, increase population heterogeneity and cross-validate results [3]. The difference of the underlying data models and terminologies between SRS and EHRs, however, presents challenges. It is thus beneficial to standardize different data sources, both in terms of how the data is modeled and the terminologies used to express concepts. A common data model (CDM) is a logical and semantic data model that can be used to standardize multiple data sources into a common format. Additionally, as part of the CDM definition itself, standard terminologies are adopted for normalizing the data semantics. Large-scale analyses can then be conducted using the same query to each of the individual data sources when a CDM is adopted [4]. Normalizing data into a CDM has been extensively employed to integrate and standardize heterogeneous data sources for pharmacovigilance research [5, 6] and other types of clinical research [7, 8].

One of the popular CDMs for pharmacovigilance is the CDM developed by the Observational Health Data Sciences and Informatics (OHDSI) organization, a multi-stakeholder, interdisciplinary collaboration that creates open-source solutions to facilitate the use of observational health data for pharmacovigilance and clinical research [9]. This CDM defines both a data model and a standardized vocabulary for standardizing records. The OHDSI CDM could be used to facilitate research on identifying and assessing associations between medical interventions and health-related outcomes, which could be especially useful in pharmacovigilance studies.

The OHDSI CDM has been adopted for pharmacovigilance and pharmacoepidemiologic research in multiple studies. Zhou et al. [3] transformed the UK's Health Improvement Network (THIN) database into the OHDSI CDM format. Three ADR analysis methods were conducted to assess the practical value of the CDM for pharmacovigilance by validating results of several published studies [10]. The study demonstrates that despite information loss as a result of incomplete mapping between medical concepts and the OHDSI standard vocabulary, the CDM-based THIN database outperformed the original THIN database in both analysis and runtime performance. Following this success, the adoption of the OHDSI CDM was considered to be a viable method for pharmacovigilance. Overhage et al. [11] further supported this consideration after transforming 10 disparate US observational healthcare databases into the OHDSI CDM format and evaluating the suitability of the model and its associated standardized vocabulary for active pharmacovigilance studies. In order to address the structure and coding problems in epidemiologic analysis and comparisons with other databases, Matcho et al. [12] converted the Clinical Practice Research Datalink (CPRD) database into the OHDSI CDM format. After an extraction, transformation and load (ETL) process, 99.9% of the conditions and 89.7% of the medications were mapped correctly, demonstrating high utility of the OHDSI CDM. The study also replicated a published case-control study [13] through the use of some queries on nonsteroidal anti-inflammatory drugs (NSAIDs) and the risk of first-time acute myocardial infarction (AMI) and obtained comparable results between the raw CPRD data and the CDM-based CPRD with less programmatic work.

FAERS collects suspected adverse event reports from health care professionals, patients and pharmaceutical manufacturers in the USA and other countries and includes information about patient demographics, drug, adverse event and patient outcome information. FAERS is a prominent SRS that has been employed widely for pharmacovigilance research [1]. As previously noted, integrating data from SRSs and longitudinal observational databases is a recent trend in active pharmacovigilance research. While the OHDSI CDM has previously been adopted for tandardizing the data from a variety of EHR databases, few studies have examined the coverage and nformation loss associated with converting FAERS data into the OHDSI CDM. Transforming the FAERS data into the OHDSI CDM format has the potential to improve the quality of adverse event reporting data, support seamless data integration between FAERS and EHRs, and enable the standardization and reproducibility of pharmacovigilance analyses using common vocabularies.

In the ADEpedia project [14, 15]. we created a scalable and standardized knowledge base of adverse drug events (ADEs) for pharmacovigilance. In addition, as a part of the ADEpedia project, Wang et al, built a standardized FAERS dataset in a previous study [16]. Similarly, Banda et al., [17] developed AEOLUS (Adverse Event Open Learning through Universal Standardization) to build a standard process for FAERS data deduplication and tooling for mapping drug names to RxNorm concepts and outcomes to SNOMED CT concepts. Although their research increases the level of standardization in FAERS and provides a tool for drug name and outcome standardization which can be used to facilitate FAERS conversion into the OHDSI CDM, FAERS has not yet been completely transformed into the OHDSI CDM. It is important to design an ETL tool that covers all of the following aspects: 1) database structure mapping; 2) concept mapping for several fields such as patient demographic data and unit data that lack of a system for normalization into standard concepts; and 3) data imputation for some required fields in OHDSI CDM such as year of birth that could not be mapped directly from FAERS. That being said, further investigation on the information loss during the ETL process is required to ensure the feasibility and accuracy of the transformation process. We therefore propose to extend the ADEpedia project leveraging some of the tooling developed in AEOLUS with the objective of developing a next generation signal detection platform known as ADEpedia-on-OHDSI that would be capable of fully converting the FAERS data into the OHDSI CDM format. We also assessed the accuracy of the conversion and appraised the ppropriateness of the OHDSI CDM for the FAERS data. We aimed to standardize the FAERS data and share the platform with the community to promote the integrative data analyses of FAERS and EHRs for improving signal detection.

2 Methods

2.1. Converting the FAERS data into the OHDSI CDM format

2.1.1. FAERS Source Data—Updated quarterly, the FAERS database can be downloaded from the FDA website [18]. The database has two versions with different formats, one version is referred as the Legacy version, which covers data from January 2004 to August 27, 2012, and the other is named as FAERS, which covers data after September 2012. In this study, we used the FAERS version and downloaded the FAERS database with

reports collected from September 2012 through March 2017, distributed quarterly. Table 1 provides the description of the tables in the FAERS database. Note that adverse events in the FAERS database are captured using the MedDRA Preferred Terms (PTs), where MedDRA is a standardized medical terminology, aiming to facilitate standardization of regulatory activities terms for human medical products. MedDRA concepts are organized into 5 hierarchical levels - System Organ Classes (SOCs), High Level Group Terms (HLGTs), High Level Terms (HLTs), Preferred Terms (PTs) and Lowest Level Terms (LLTs). However, due to data duplication and the lack of standardization in drug names, data cleaning and drug name normalization must be conducted prior to the CDM conversion process.

2.1.2. The OHDSI CDM—There are 39 data tables in the OHDSI CDM version 5.3.0 (Feb 18, 2018) for standardizing vocabularies, meta-data, clinical data, health system data, health economics, and derived data elements. Another critical feature of the OHDSI CDM is the utilization of standard concepts to describe data in its tables. Standard concepts in the OHDSI CDM are collected from various medical terminologies or ontologies such as SNOMED CT, RxNorm, and LOINC, and stored in the vocabulary tables. In this study, drug names in FAERS are mapped to RxNorm, a preferred standard terminology in the OHDSI CDM for drugs. Meanwhile, adverse vents/indications/outcomes are mapped to SNOMED CT, a preferred standard terminology in the OHDSI CDM for clinical concepts.

2.1.3 Data Cleaning and Drug/Outcome Mapping—The following procedures from AEOLUS were implemented to conduct data cleaning and Drug/Outcome Mapping in this study:

Impute missing values: AEOLUS defines that a fully populated case record contains at least four ‘key’ demographic fields - event date, age, sex and reporter country. Records were aggregated by case ID, and missing fields were populated with data from other records in the same aggregation, if present.

De-duplicate case records: A two-round case de-duplication was performed to eliminate the data redundancy problem caused by multiple case versions. First, case versions were aggregated by case ID: if the data in all of the other 9 chosen fields (case id, case initial/follow-up code, event date, age, sex, reporter country, drug names, and reaction/outcomes preferred terms) of interest were the same, only the latest version of the case record was retained. Secondly, in order to solve the problem that a duplicate case version was not linked to the original case id by the FDA, for those case records with the same value in the four ‘key’ demographic fields, only the latest record was retained.

Map drug names to RxNorm: In this step, drug names in FAERS were mapped to RxNorm using the dictionary lookup approach. Furthermore, New Drug Application (NDA) drug names were mapped to the FDA Orange book of NDA ingredients first and then mapped to an OHDSI concept.

Map adverse events/indications/outcomes to SNOMED CT: As SNOMED CT is the preferred standard vocabulary for clinical concepts in the OHDSI CDM, AEOLUS

implements a process to map MedDRA preferred terms (PTs) to SNOMED CT concepts. However, not all MedDRA PTs are mappable to SNOMED CT. Only 64% of MedDRA PTs in indications and 80% of PTs in reactions in FAERS can be mapped to SNOMED CT concepts[17]. In our implementation, we map MedDRA PTs to SNOMED CT concepts if applicable and keep the original MedDRA PTs for those failed to be mapped.

The implementation of AEOLUS yields a de-duplicated and standardized view of the FAERS database.

2.1.4 Transforming FAERS Tables into the OHDSI CDM—

We developed a conversion tool to transform the FAERS tables into OHDSI CDM tables following OHDSI's recommended ETL process. The following details the conversion process.

Define and execute structure mapping (extraction): We defined structure mapping by choosing the appropriate tables/fields from the OHDSI CDM for tables/fields in the FAERS database. It was achieved manually through multiple rounds of discussions between two experts with medical informatics background. A total of 8 out of 39 OHDSI tables were chosen to map with 6 FAERS tables. Table 2 describes the content of the 8 OHDSI tables. And the details of the table level mappings are shown in Figure 1. For field level mappings, fields in FAERS tables and OHDSI tables were connected if they had the same interpretation. In addition, the case id in FAERS was mapped as a person_id in OHDSI PERSON table to be a record identifier that can be used as a foreign key in other tables. The details of the field level mapping are provided in Supplementary File 1. Note that the drug-event or drug-indication pairs in FAERS were mapped to the FACT_RELATIONSHIP table in the OHDSI CDM. To simplify the use of the data for pharmacovigilance research, each pair was mapped to two FACT_RELATIONSHIP records. For example, if the medication with drug_exposure_id as 1 causes the adverse event with observationjd as 1, then there are two records in the FACT_RELATIONSHIP table: (Drug, 1, Observation, 1, causes) and (Observation, 1, Drug,1, caused by).

Transform data: In this step, we conducted data conversion and data imputation to transform the FAERS data into the OHDSI CDM format.

First, after the structure mapping between FAERS tables and OHDSI CDM tables was created, additional data conversion is needed prior to the loading of the data into the OHDSI CDM format. For example, values in the age, drug dose and date fields in the FAERS tables need to be converted into the corresponding data type in the OHDSI CDM. We also mapped values in some other fields such as sex in the DEMO table, unit and route in the DRUG table to standard concepts in the OHDSI vocabulary manually.

Second, some of the fields in those mapped OHDSI tables are required fields, an imputation process was conducted as follows: 1) For the year_of_birth field in the PERSON table, the date of the adverse event occurred minus age was calculated to determine the year of birth of the patient; 2) For the condition_start_date field in the CONDITION_OCCURRENCE table, due to the indication date is not provided by the FAERS database, so we chose the therapy date to populate the CONDITION_OCCURRENCE table; 3) For the death_date field in the

DEATH table, if one case was reported death, the latest therapy end date would be extracted as the death date; 4) For the days_supply field in the DRUG_EXPOSURE table, if the FAERS therapy duration data was valid, we converted the duration time into days. If duration data was null, we used therapy end time minus start time to compute the number of days of supply for the drug in question. In addition, if supply time was 0 days for a record after computation, to the supply time was set to 1 day.

Load data: After data transformation, 8 OHDSI tables were created to store the normalized data in a PostgreSQL database. Users can download a virtual machine containing a full OHDSI SQL database and related tools at OHDSI's website (http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:ohdsi-in-a-box). Note that although we had normalized and imputed data for the required fields in OHDSI table, many required fields of our 8 tables contained null values due to failed concept mapping or there being no equivalent value in the FAERS tables to use as a source. In these instances, we used the default value or "Not Available" (concept_id = 0) recommended by OHDSI. For example, as the gender field is required in the OHDSI PERSON table, we used the concept id "0" as the value for the gender field for those records with a missing value. After data loading was completed, we also added a description of the source FAERS database into the OHDSI "CDM_SOURCE" table.

2.2. Evaluation Experiments

We conducted an experiment to evaluate the appropriateness of the OHDSI CDM for representing the FAERS data and validate the data mapping and conversion process. In addition, we also evaluated the utility of the CDM-based FAERS through a replication study.

2.2.1. Evaluation of the ETL Process—We assessed the ETL process by computing the overall statistics before and after the ETL process and investigating field level mapping quality. We also assessed the completeness of the mapping of drug names into RxNorm concepts by providing descriptive statistics to describe the overall mapping results and summarizing the characteristics of the unmatched drug names. To assess the accuracy of the drug name mappings, we randomly sampled 500 unique mappings and two annotators manually verified the correctness of the mapping. Kappa coefficient [19] was calculated to assess the inter-annotator agreement. We also calculated the data transformation rate between the original FAERS and the CDM-based FAERS to assess the information loss during the ETL process.

2.2.2. Utility Evaluation through a Replication Study—We replicated the study of Roberto et al.[20] using our CDM-based FAERS to assess its utility. In Roberto's study, legacy FAERS data from 2004 to 2010 was used to determine the vascular safety profile of triptans. We replicated this analysis using the FAERS data collected after 2012 and compared the results obtained using the original FAERS and the CDM-based FAERS. We used both the triptans' brand name/unique concept ID and ingredient name/unique concept ID to search the triptan-related event reports.

Specifically, for ADR signal detection, a disproportional algorithm was used to detect the triptan related signals. Reporting odds ratio (ROR) and relevant 95% confidence intervals (95% Cl) were calculated to measure the ratio of the odds of case/non-case [21]. Figure 2 shows the calculation method of ROR. In the contingency table, one case is defined as a triptan-related adverse report, which may have more than one vascular event, whereas non-cases were all reports of vascular events without the use of triptans. Seven triptan drugs and 1,689 PTs belonging to the MedDRA SOCs 'Cardiac disorder' or 'Vascular disorders' were applied to retrieve triptan related vascular event reports. When the case number > 3 and the lower limit of 95% Cl of ROR > 1, a triptan-related vascular event was considered to be a positive ADR signal.

In order to get a comprehensive signal detection result, the detection process was performed in three different hierarchical levels. For the primary analysis, signals were detected using high-level group terms (HLGT) and high-level terms (HLT) from MedDRA to increase the sensitivity of detection. The secondary analysis, which is the PT level adverse drug reaction signal detection aiming to enhance the specificity of analysis [22], was then performed in PTs belonging to those HLTs which were discovered as positive signals in the primary analysis. Finally, six standardized MedDRA queries (SMQs) related to vascular events were applied. Those SMQs are a group of MedDRA PTs describing the same vascular medical condition, which are validated and pre-determined by MedDRA.

3 Results

3.1. ETL Performance

3.1.1. Overall ETL Evaluation—After data extraction, transformation and loading, data in FAERS was converted into 8 OHDSI CDM tables. A total of 4,619,362 adverse event cases were transformed into the OHDSI CDM format. The average age of patients was 55.2 years. Of all the patients, 2,577,989 (55.8%) were female, 1,603,982 (34.7%) were male and 437,391(9.5%) were Unknown/Not Specified.

Table 3 presents a comparison of the generated statistics for records from the original FAERS dataset and the CDM-based FAERS database. Among 8 CDM-based FAERS tables, the PERSON table, DRUG_EXPOSURE table and CONDITION_OCCURRENCE table have the same amount of data as the original tables (DEMO, DRUG, INDI) in FAERS. In the OHDSI CDM version of FAERS, adverse event information and patient outcomes were transformed into the OBSERVATION table. Thus, records in the OBSERVATION table can be considered to be equivalent to the sum of the REAC table and the OUTC table in FAERS. In the FACT_RELATIONSHIP table, it should be noted that all relationships are directional, and as such each relationship record will be represented twice symmetrically. In regards to the four relationship types in FACT_RELATIONSHIP table, 61,739,303 drug-adverse event combinationsv were stored twice for relationship type "Causes" and "Caused by" symmetrically, 9,593,169 drug-indication combinations were also saved twice for relationship type "May treat" and "May be treated by" respectively.

Figure 3 shows an overview of the mapping quality of FAERS into the OHDSI CDM format at the field level. For most of OHDSI CDM tables, the primary key is automatically

generated by the system, except for the PERSON table. For the other fields, we perform the mapping, calculation and transformation process to ensure that the data in the field is completely imported. The remaining fields in the OHDSI CDM were left empty if the FAERS data set did not provide any associated information. Note that there are two required fields, race_concept_id and ethnicity_concept_id in the PERSON table which were not available to be mapped from FAERS directly. So, in order to comply with OHDSI specifications, the recommended blank value with the concept id “0” was used.

3.1.2. Mapping Quality of Drug Names to RxNorm—

Drug name mapping results were compared across the original FAERS data and CDM-based FAERS to evaluate the drug name mapping quality. Of the 15,438,807 medication records in the OHDSI DRUG_EXPOSURE table, 14,502,476 (93.94%) records were mapped to RxNorm codes. Of the 340,249 unique drug names in the FAERS source data, 159,995(47.02%) unique names are matched into 10,897 RxNorm codes.

We investigated the results of drug name mappings. Table 4 shows the comparison of the top10 drug names in the original FAERS data and CDM-based FAERS data. Amongst the top 10 drug names before and after the mapping normalization was done, 6 drug names (Humira, Enbrel, Aspirin, Xarelto, Revlimid and Xyrem) appeared in both the top 10 drug names before mapping and the top 10 drug names after mapping. We also found an increase in the number of records for those 6 drugs. These results indicate that the mapping leads to better concept-level integration of records. The integration may be associated with the standardization of uppercase and lowercase letters (for example, source drug name “ENBREL” and “Enbrel” are both mapped to concept “Enbrel (RxCUl = 216891)” in RxNorm), normalization of special characters (for example, source drug name “ASPIRIN.” Is mapped to the concept “Aspirin (RxCUl = 1191)” in RxNorm) and the concept-level integration of differing input formats for the same drugs (for example, source drug name “HUMIRA” and “HUMIRA 40 MG” are both matched to concept “Humira (RxCUl = 353484)” in RxNorm).

For the manual verification of 500 unique mappings, the mapping accuracy of the two annotators was determined to be 96.8% and 96.6% respectively with the observed kappa value of 0.91, indicating high mapping performance and “almost perfect agreement” [23]. The following summarizes the mapping errors : 1) For multi-ingredient medicines, all the active ingredients may not have been completely identified, e.g., “Lisinopril HCTZ” mapped to only “Lisinopril”; 2) Active ingredients of medicines were not exactly mapped, e.g. “MAGNESIUM SULPHATE” matched to “Magnesium”; 3) Non-specific drug names in original FAERS database, e.g., “ANTIDEPRESSANTS” matched to “Clonazepam”; 4) Errors caused by some medical devices or medications not collected by RxNorm, e.g. “PARAGARD T 380A” matched to “Copper”.

A total of 936,331 (6.06%) of records and 180,254 (52.98%) unique drug name terms in the original FAERS data were not mapped with RxNorm codes, of which 164,171(91.08%) unique drug names have an occurrence of less than 5 records. The top 10 unmapped drug names are shown in Table 5. The following reasons may result in a mapping failure: 1) The drug name was too complicated to match, such as “DIANEAL LOW CALCIUM

PERITONEAL DIALYSIS SOLUTION WITH DEXTROSE"; 2) Some drug names were not included in RxNorm due to it is not being a US drug trade name, such as "DUODOPA" and "LOXONIN"; 3) Drug names that were not specific enough caused mapping failures, such as "ANTIHYPERTENSIVES".

3.1.3. Transforming and Imputing Quality of other Additional Health Data—13 fields in the CDM-based FAERS were involved in data type conversion or concept mapping for the ETL process. The information loss of those fields is shown in Figure 4. Among all the 13 fields, we observed a total of 10 fields with very low information loss. Among them, gender_concept_id, condition_concept_id, lot_number, observation_concept_id and measurement_date were completely populated with the data from FAERS and for the other five fields, drug_exposure_start_date, drug_exposure_end_date, observation_date, value_as_number and unit_concept_id, more than 99% of the fields were transformed and loaded into the CDM-based FAERS. The following highlights the underlying reasons why those 10 fields have such a low information loss: 1) For condition concept and observation concept, all the terms of ADE or indication in the original FAERS is recorded as MedDRA PTs. And as we mentioned in the method section, those concepts were loaded into the OHDSI CDM either by SNOMED CT concepts or MedDRA PTs, so there is no information loss during the ETL process. 2) Only a few date values are not in a standard date format, resulting in a high loading rate of data for those 4 date fields. 3) For the gender and patient's weight unit, standard terms were used as data input. 4) For lot number and weight value, only data types were changed during the data transformation.

For each of the remaining 3 fields, drug_concept_id, locationjd, and route_concept_id, the information loss is greater than 6%. The information loss of the drug_concept_id field has been discussed in the section above. For the route_concept_id field, the information loss rate is about 34%, the highest among all the fields. That is primarily caused by the lack of standardization for route in FAERS. The information loss in location_id is 6.75%, primarily due to the error input in the original FAERS data since the country name was standardized using the ISO 3166-1 standard in the ETL process.

We also analyzed the information loss (shown as table 6) for those 4 fields loaded with the imputation data: 1) For the year_of_birth field, only when the data of age, age unit or event date field are completely available in the same raw patient's record, the birth year can be calculated. Of the 2,751,210 records with age information in the original FAERS database, only 1,990,826 (72.36%) was transformed into the CDM-based FAERS. 2) For condition_start_date, we imputed 99.63% of the values based on the therapy date from the original FAERS database. 3) For death_date, we imputed 100% of the death cases using the last therapy end date. 4) For days_supply, we imputed 2,287,681 records. Note that only 23,977 records in the original FAERS have the supply information.

3.2 Result of the Replication Study

We investigated the triptan-related vascular events in the original FAERS and CDM-based FAERS, respectively. We retrieved 1,101,856 reports containing at least one PT from the 'Cardiac disorder' or 'Vascular disorders' SOCs. Table 7 lists the retrieval results of triptan-

related reports. The number of triptan-related reports is 24,251 and 24,499 using the original FAERS and the CDM-based FAERS, respectively. We noticed after the ETL process, more triptan-related reports were retrieved. Of all the specific triptan drugs, sumatriptan had the most adverse event reports in our studies (the original FAERS: 16,521; CDM-based FAERS: 16,767).

We compared the numbers of positive signals detected by ROR to three studies in HLGT, HLT, PT levels, and their overlap between CDM-based FAERS and original FAERS study. Interestingly, we found slightly more ADRs in CDM-based FAERS than those in the original FAERS study in both HLGT level (6 vs 5) and PT level (50 vs 47) with the same number of ADRs at the HLT level (18). This finding indicated that the CDM-based platform does have an impact on the signal detection results.

Figure 5 shows the signal detection results of SMQ analysis. 2 queries detected a positive ADR in both our original FAERS data study (Central nervous system hemorrhages and cerebral conditions: 1.77, 1.00-3.13; Cerebrovascular disorders: 9.44, 7.41-12.02) and CDM-based FAERS study (Central nervous system hemorrhages and cerebral conditions: 1.92, 1.11-3.32; Cerebrovascular disorders: 9.58, 7.53-12.18). The signal detection scores of all the six queries in the original FAERS data study were a bit lower than those scores in the CDM-based FAERS study because the CDM data study could collect more triptan-related event cases.

4 Discussion

Our evaluation of the ETL process demonstrated the feasibility of converting the FAERS database to the OHDSI CDM format. Even with some information loss during the ETL process, the process was evaluated to be reasonably accurate and meaningful. The study provides a way to facilitate the effective integration of the spontaneous reporting data with the EHR data through the adoption of the OHDSI CDM. There are significant advantages when adopting the OHDSI CDM for FAERS: most significantly, it will improve the precision of ADR signal detection through standardization. Furthermore, the ability to seamlessly integrate EHR and other forms of longitudinal data make it possible to further discover and understand additional knowledge about adverse events such as causes, confounders and possible corrective actions. The adoption also allows a series of open-source applications to assist data analysis such as ATLAS [24]. In addition, we also released the ETL process open source to the broad community so other teams can adopt the ETL process to save time and facilitate the usage of CDM.

For drug mapping, we used the matching method as developed by Banda, et al [17], and we achieved similar information loss to drug exposure in our research (6.06%) as in Banda's research (7%). This is regarded as an acceptable result when also compared with 7% in Zhou's research [3] and 10.3% in Matcho's research [12]. In order to identify in further detail the reason why those drug names could not be matched or mismatched with the OHDSI standard concepts, we conducted a manual review for the drug names which were not matched or mismatched and concluded the reasons as the following: 1) Some drug names, such as those with a brand name registered outside the US, medical devices, or new

drugs that were not collected by RxNorm, could not be mapped to OHDSI concepts correctly, although Banda's study linked the New Drug Application (NDA) drug names to the FDA orange book of NDA ingredients; 2) For some drug combinations or complicated drug names, active ingredients may not have been accurately identified due to the match rules used in Banda's study; 3) Some drug names were not clearly recorded in FAERS.

In order to improve the accuracy of drug name mappings, we plan to investigate the following areas in future work. First, there are a number of standardized vocabularies (e.g., WHO Anatomical Therapeutic Chemical Classification, ATC) which collect drug concepts such as non-US drug brand names. These vocabularies are also loaded into the OHDSI CDM. Although these vocabularies are not preferred for the OHDSI CDM and not recommended to annotate drug exposure data, drug name mapping results may be improved by other vocabularies as a supplement to RxNorm. The OHDSI CDM provides semantic relationships between concepts from different vocabularies, which could be used to create these RxNorm extensions. Secondly, compared with the OHDSI CDM version we used, the version for OHDSI will update and comes with the newest RxNorm version, which covers more drug names, especially for new drugs. We thus consider updating the CDM version used by the database in the future, which may improve the drug name mapping accuracy. Finally, we can conduct a manual mapping using the open source OHDSI Usagi application [25] for those drug names that remain unmapped.

Another problem regarding drug name mapping is version updates of RxNorm. RxNorm releases monthly updates adding new drugs and retiring inactive or deprecated drugs. Thus, if we do not update the drug name mapping with the latest version of the RxNorm in a timely manner, some bias for pharmacovigilance studies which involve detecting adverse events in drug class level may occur, especially in the case of inactive drug involvement. So, in future work, we will develop tools for our ADEpedia-on-OHDSI platform to accelerate the drug name mapping update process.

Due to our focus on developing an ETL tool for the current FAERS database, we did not use the same collection period as in Banda's research to evaluate the drug mapping results. Although our information loss rate of drug exposure was comparable to Banda's, some reasons behind mismatched drug names in the legacy AERS database may remain undiscovered. In our future research, we will extend our ETL tool to meet the transformation request for mapping the legacy AERS database to the OHDSI CDM format.

Another issue for mapping is observation and condition concepts. While the preferred vocabulary for observation/condition concept representation is SNOMED CT in the OHDSI CDM, all the indication names and ADE names are recorded by MedDRA PTs in FAERS, which could not be mapped to SNOMED CT concepts completely. In order to reduce information loss, we conducted a compromise process in which loaded the MedDRA code directly for those terms that could not match with SNOMED CT concepts. In a future study, we will extend the existing mapping between MedDRA and SNOMED CT in UMLS (according to Bodenreider's study, about 58% of the PTs in MedDRA could be mapped into SNOMED CT concepts through the UMLS [26].) to improve the match rate. A comprehensive semantic analysis of non-mapped MedDRA PTs will be conducted to

facilitate the mapping. Post-coordination and natural language processing methods may also be used to improve the concept mapping between MedDRA and SNOMED CT.

The information loss of other CDM-based FAERS tables was also assessed in our study. We believe that information loss does not hugely affect the results of pharmacovigilance studies because most information loss is caused by the error or ambiguous input in the original FAERS database itself. On the other hand, for some required fields, we performed a series of data calculations and imputations. Although some of those imputed data may not accurate, this process ensures the integrity of the data and improves the quality of data. We will design a more rigorous validation to evaluate the impact of computation algorithms in future research.

Our comparison of the original FAERS and the CDM-based FAERS in the vascular safety profiling of triptans shows more ADRs detected using the CDM-based FAERS, illustrating the utility of the CDM-based FAERS for ADR detection. We examined the difference of the number of triptan-related reports and the number of ADRs detected. The number of triptan-related reports captured in the CDM-based FAERS was 1.02% higher than that found in the original FAERS. Through an investigation of the drug source names of retrieval results, we found that the retrieval results of CDM-based FAERS were more comprehensive and accurate because of the use of concept mapping and normalization. This validates the value of drug name standardization. In addition, there are some other advantages of drug name standardization: 1) a more accurate data retrieval result could improve the signal detection precision. 2) we can choose the search concepts directly using the RxNorm codes, greatly reducing the design time and complexity of retrieval query, and increasing the portability of the query. 3) executing SQL-based queries in CDM-based FAERS will consume far less time than running a regular expression query in the original FAERS. For instance, it took less than a minute to execute a query in CDM-based FAERS as opposed to about 7 minutes in the original FAERS for the same sumatriptan-related event report retrieving task. Furthermore, there is only a slight difference for the ROR and 95% CI value of the signal detection between the original FAERS and the CDM-based FAERS. This indicates that although the drug mapping algorithm could be further improved, the transformation does not excessively impact our signal detection result. In addition, the ROR of signals detected by the CDM-based FAERS was greater than the ROR of signals detected by the original FAERS, which suggests that the sensitivity of CDM-based FAERS detection was better than that of the original FAERS detection.

We also compared the signal detection results of the original FAERS study vs. CDM-based FAERS study to further validate whether CDM-based FAERS was a good approximation of the original FAERS data. More ADR signals and higher ROR value were detected in the CDM-based FAERS study than that in the original FAERS study at different adverse event levels, which indicates the CDM-based approach is more powerful in the ADR signal detection study. In future research, we will leverage advanced technologies, such as text mining EHRs, to verify all the detected signals and further validate the reliability of the CDM-based FAERS.

As previously mentioned, in this study, we aimed to build a platform which contains both standardized SRS data and EHR data to facilitate next generation pharmacovigilance signal detection. As such, we are actively working on identifying use cases (e.g., signal detection of the immune-related adverse events) [27] that integrate both EHR data and FAERS data for improved signal detection. In future work, we will develop new methods to conduct comprehensive pharmacovigilance signal detection utilizing the ADEpedia-on-OHDSI platform.

Conclusion

In this study, we extracted, transformed and loaded the FAERS spontaneous reporting data into an integrated data repository based on the OHDSI CDM in our ADEpedia-on-OHDSI platform to support the needs for the next generation signal detection. The outcome of the work would facilitate seamless integration and combined analyses of multiple datasets in our platform, particularly through leveraging EHR data, so as to improve signal detection. The open-source conversion tool is available at https://github.com/adepedia/adepedia-on-ohdsi.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was funded in part by FDA HHSF223201710167C and NIH U01 HG009450. The authors would like to thank Hongfang Liu and Liwei Wang for their previous work in FAERS standardization and their support of revising this paper.

References

1. Harpaz R, et al., Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. Clinical Pharmacology & Therapeutics, 2012 91(6): p. 1010–1021. [PubMed: 22549283]
2. Harpaz R, DuMochel W, and Shah NH, Big Data and Adverse Drug Reaction Detection. Clin Pharmacol Ther, 2016 99(3): p. 268–70. [PubMed: 26575203]
3. Zhou X, et al., An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf, 2013 36(2): p. 119–34. [PubMed: 23329543]
4. Garza M, et al., Evaluating common data models for use with a longitudinal community registry. Journal of Biomedical Informatics, 2016 64: p. 333–341. [PubMed: 27989817]
5. Voss EA, et al., Accuracy of an automated knowledge base for identifying drug adverse reactions. Journal of Biomedical Informatics, 2017 66: p. 72–81. [PubMed: 27993747]
6. Li Y, et al., A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions. Drug Safety, 2015 38(10): p. 895–908. [PubMed: 26153397]
7. Boyce RD, et al., Preparing Nursing Home Data from Multiple Sites for Clinical Research - A Case Study Using Observational Health Data Sciences and Informatics. EGEMS (Wash DC), 2016 4(1): p. 1252. [PubMed: 27891528]
8. Rosenbloom ST, et al., Representing Knowledge Consistently Across Health Systems. Yearb Med Inform, 2017 26(1): p. 139–147. [PubMed: 29063555]
9. Hripcsak G, et al., Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform, 2015 216: p. 574–8. [PubMed: 26262116]
10. Hardoon SL, et al., Trends in longer-term survival following an acute myocardial infarction and prescribing of evidenced-based medications in primary care in the UK from 1991: a longitudinal

population-based study. J Epidemiol Community Health, 2011 65(9): p. 770–4. [PubMed: 20515898]

11. Overhage JM, et al., Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc, 2012 19(1): p. 54–60. [PubMed: 22037893]
12. Matcho A, et al., Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf, 2014 37(11): p. 945–59. [PubMed: 25187016]
13. Schlienger RG, Jick H, and Meier CR, Use of nonsteroidal anti-inflammatory drugs and the risk of first-time acute myocardial infarction. British Journal of Clinical Pharmacology, 2002 54(3): p. 327–332. [PubMed: 12236854]
14. Jiang G, Solbrig HR, and Chute CG, ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annu Symp Proc, 2011 2011: p. 607–16. [PubMed: 22195116]
15. Jiang G, et al., ADEpedia 2.0: Integration of Normalized Adverse Drug Events (ADEs) Knowledge from the UMLS. AMIA Jt Summits Transl Sci Proc, 2013 2013: p. 100–4. [PubMed: 24303245]
16. Wang L, et al., Standardizing adverse drug event reporting data. J Biomed Semantics, 2014 5: p. 36. [PubMed: 25157320]
17. Banda JM, et al., A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data, 2016 3: p. 160026. [PubMed: 27193236]
18. FDA. FAERS. Available from: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm.
19. Cohen J, A Coefficient of Agreement for Nominal Scales. Educational & Psychological Measurement, 1960 20(1): p. 37–46.
20. Roberto G, et al., Triptans and serious adverse vascular events: data mining of the FDA Adverse Event Reporting System database. Cephalalgia, 2014 34(1): p. 5–13. [PubMed: 23921799]
21. Bate A, et al., A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol, 1998 54(4): p. 315–21. [PubMed: 9696956]
22. Pearson RK, et al., Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform, 2009 78(12): p. e97–e103. [PubMed: 19230751]
23. Landis JR and Koch GG, The measurement of observer agreement for categorical data. Biometrics, 1977 33(1): p. 159–74. [PubMed: 843571]
24. OHDSI. ATLAS. Available from: www.ohdsi.org/web/atlas/.
25. OHDSI. Usagi. Available from: https://github.com/OHDSI/Usagi.
26. Bodenreider O, Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. AMIA Annu Symp Proc, 2009 2009: p. 45–9. [PubMed: 20351820]
27. Yu Y, et al., Developing A Standards-based Signal Detection and Validation Framework of Immune-related Adverse Events Using the OHDSI Common Data Model, in AMIA Annu Symp Proc. 2018 (In press)

Highlight

- A next generation pharmacovigilance platform using the OHDSI CDM is designed
- An ETL tool for converting FAERS to the OHDSI CDM is developed
- A comprehensive ETL performance evaluation for CDM-based FAERS is conducted
- A replication study is conducted to validate the utility of CDM-based FAERS

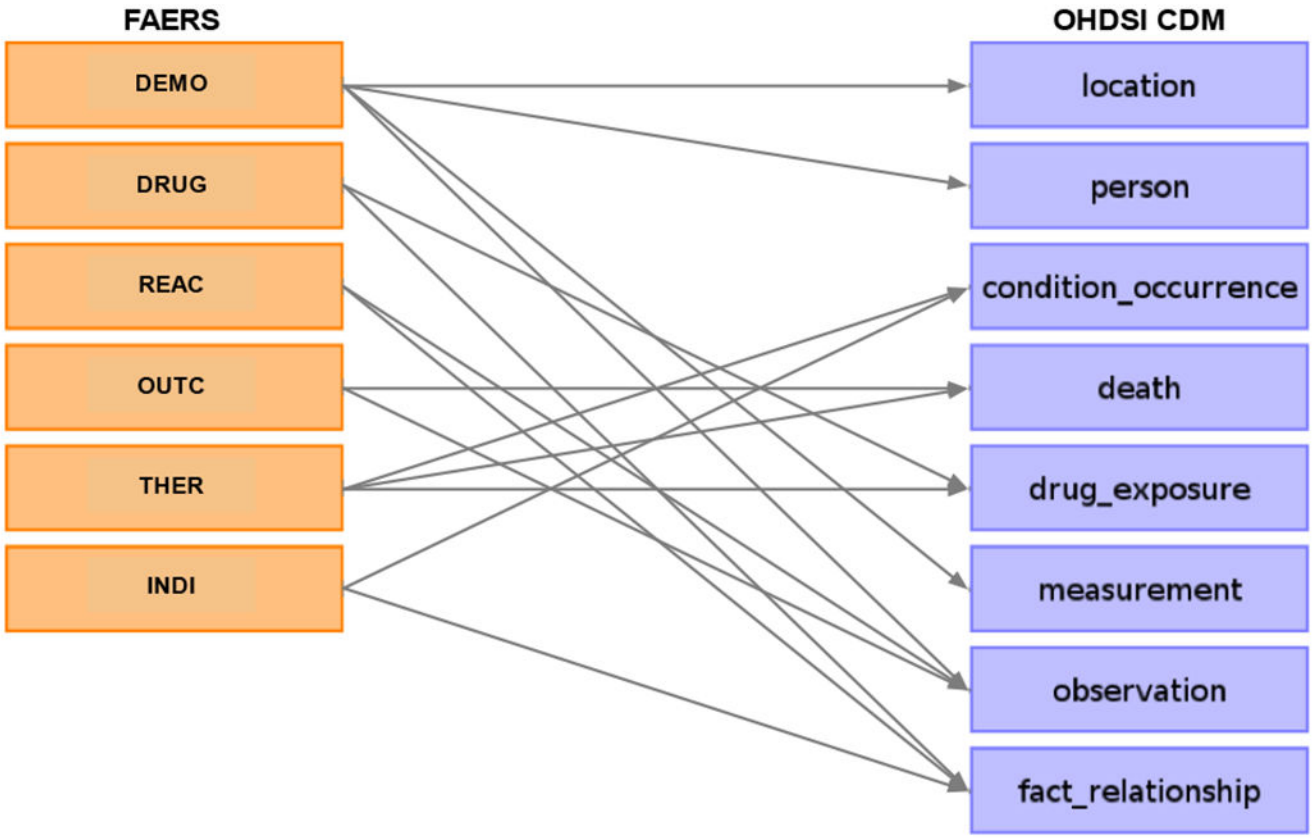


Figure 1.
Table level mapping between FAERS and the OHDSI CDM

	Reports with target event	Reports without target event
Reports with triptans	A	B
Reports without triptans	C	D

$$\mathrm{ROR} = \frac{a/b}{c/d}$$

Figure 2.
Calculation method of ROR.

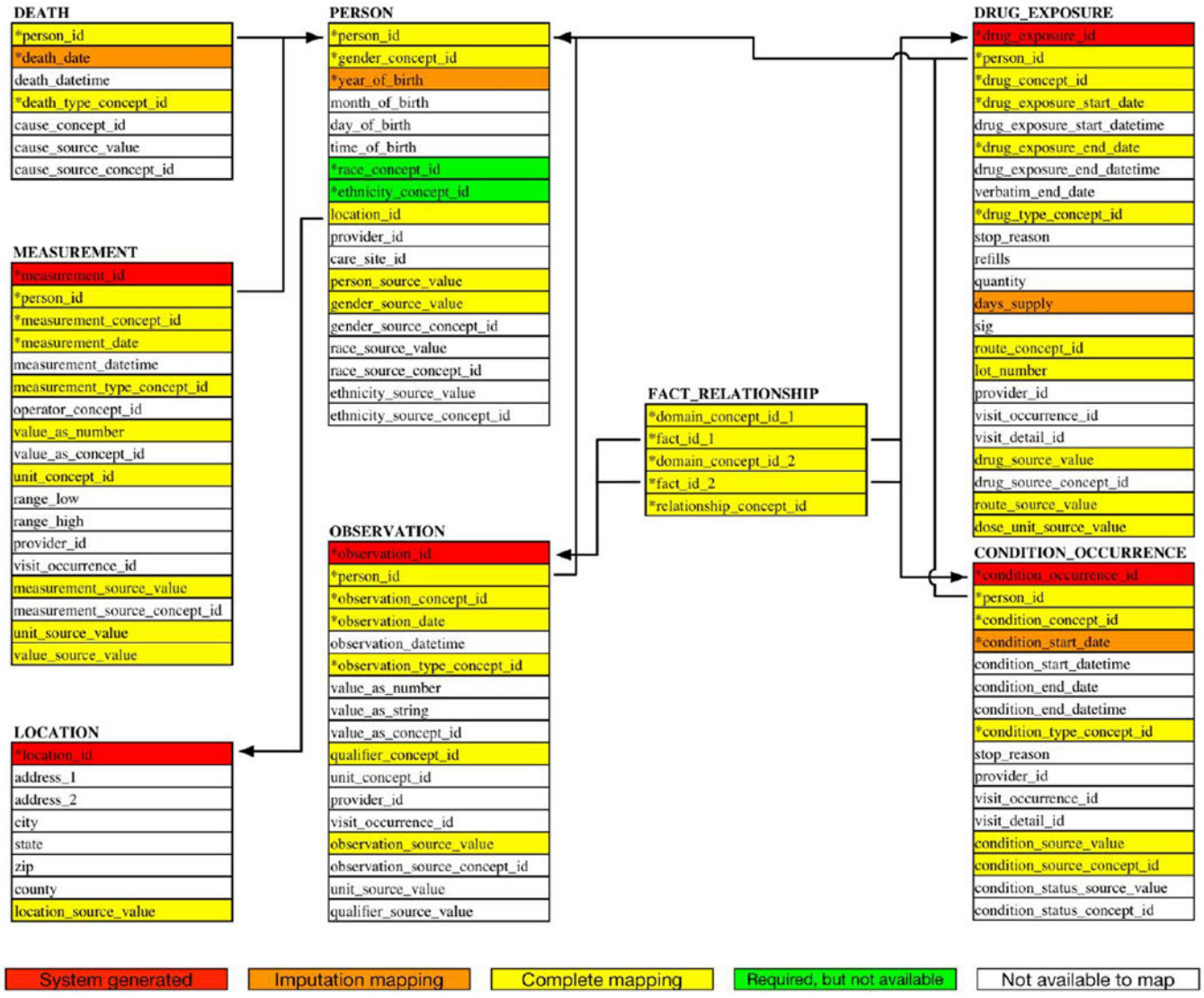


Figure 3.
Database heat map of mapping quality. * represents the required fields.

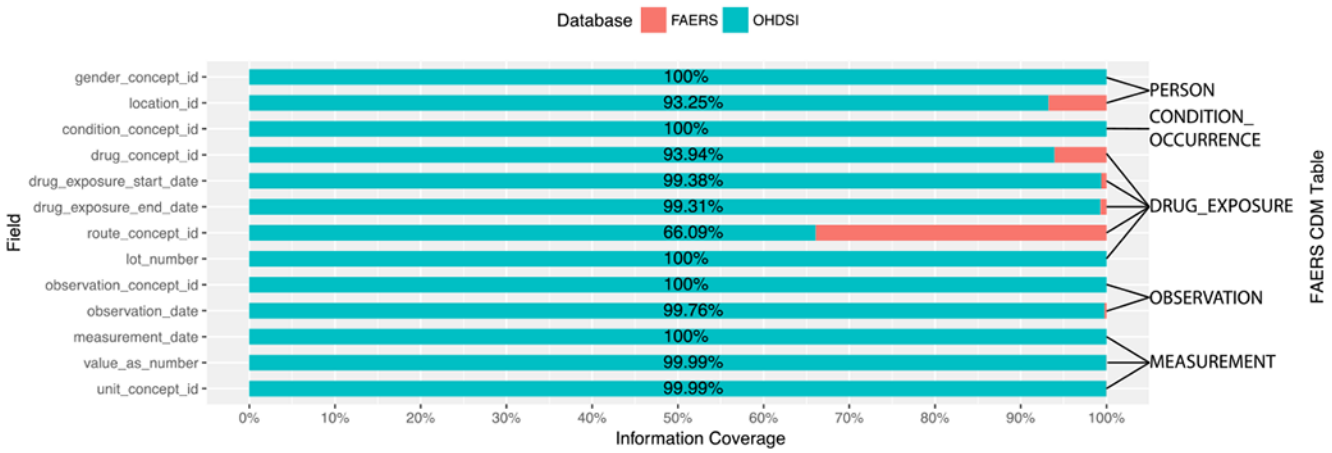


Figure 4.
Information loss of ADEpedia-on-OHDSI ETL process.

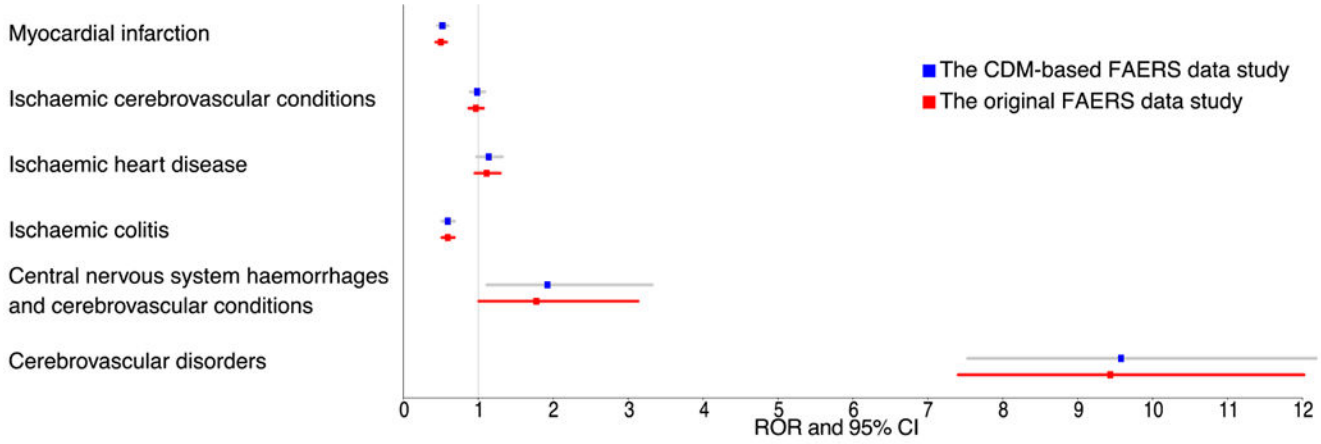


Figure 5.

The associations between triptans and 6 vascular related SMQs in the analysis of two studies. Abbreviations: ROR: Reporting Odds Ratio; 95% CI: 95% Confidence interval.

Table 1.

FAERS source table description

Table name	Description
DEMO	Includes patient demographic and administrative information
DRUG	Includes drug or biologic information
REAC	Includes adverse event coded by MedDRA terms
OUTC	Includes patient outcomes
RPSR	Includes report sources
THER	Includes drug therapy start and end dates
INDI	Includes indication for drugs or diagnosis coded by MedDRA terms

Abbreviations: MedDRA, Medical Dictionary for Regulatory Activities

Table 2.

OHDSI CDM table description

OHDSI CDM Table Name	OHDSI CDM Table Description
LOCATION	captures location or address information of patients and health institutions
PERSON	contains patient demographic information
CONDITION_OCCURRENCE	records disease or medical conditions of patients
DEATH	contains cause of death and time of death for relevant patients
DRUG_EXPOSURE	captures records regarding patient drug exposures
OBSERVATION	captures clinical facts about a patient that cannot be represented by other tables
FACT_RELATIONSHIP	contains relationships between facts from different OHDSI CDM tables
MEASUREMENT	contains the information of examination or testing

Table 3.

Basic Statistics of OHDSI CDM Tables after ETL Process.

FAERS Table name	Records	OMOP CDM Table name	Records
DEMO	4,619,362	PERSON	4,619,362
DRUG	15,438,807	DRUG_EXPOSURE	15,438,807
INDI	9,593,169	CONDITION_OCCURRENCE	9,593,169
REAC	13,142,011	OBSERVATION	16,303,802
OUTC	3,161,791	FACT_RELATIONSHIP	142,664,944
THER	5,863,850	DEATH	440,562
RPSR	372,284	MEASUREMENT	977,450
		LOCATION	210

Table 4

Top 10 drug names before and after mapping

No.	Top 10 unique drug names before mapping	Records	No.	Top 10 RxNorm concepts after mapping (RxCUl)	Records
1	HUMIRA	297,057	1	Humira (353484)	297,990
2	ENBREL	174,208	2	Enbrel (216891)	270,576
3	XARELTO	160,440	3	Aspirin (1191)	210,856
4	REVLIMID	157,543	4	Xarelto (1114199)	161,414
5	XYREM	100,265	5	Revlimid (337535)	157,754
6	Enbrel	96,118	6	Prednisone (8640)	121,733
7	DIANEAL LOW CALCIUM PERITONEAL DIALYSIS SOLUTION WITH DEXTROSE	93,291	7	Metformin (6809)	110,579
8	AVONEX	91,681	8	Xyrem (353098)	100,270
9	TYSABRI	87,050	9	Methotrexate (6851)	99,311
10	ASPIRIN.	87,007	10	Lyrica (593441)	98,368

Abbreviations: RxCUI, RxNorm Concept Unique Identifier

Table 5

Top 10 unmatched drug name terms in FAERS

No.	Drug names in FAERS	Records
1	DIANEAL LOW CALCIUM PERITONEAL DIALYSIS SOLUTION WITH DEXTROSE	93,291
2	HUMIRA 40 MG/0.8 ML PEN	19,210
3	DUODOPA	6,900
4	SERETIDE	6,577
5	HUMIRA 40 MG/ 0.8 ML PRE-FILLED SYRINGE	5,654
6	NOVORAPID	5,488
7	LOXONIN	3,788
8	ANORO ELLIPTA	3,525
9	DIANEAL PD2	3,290
10	ADDERALL XR	3,021

Table 6

Information loss of imputation fields

Field name	Data No. before ETL	Data No. after ETL	Transformation Rate
year_of_birth	2,751,210	1,990,826	72.36%
condition_start_date	4,036,046	4,021,248	99.63%
death_date	136,538	136,538	100%
days_supply	23,977	2,287,681	NA

Table 7.

Total triptan-related reports and number of cases with vascular events based on three studies.

Active ingredient	The original FAERS		CDM-based FAERS	
	Total reports	Cases	Total reports	Cases
Sumatriptan	16,521	3,869	16,767	3,981
Eletriptan	2,549	835	2,552	838
Zolmitriptan	1,596	643	1,615	649
Rizatriptan	3,696	1,227	3,645	1,203
Naratriptan	557	190	560	193
Frovatriptan	398	136	400	138
Almotriptan	303	104	346	122
Total	24,251	6,516	24,499	6,622

Cases: number of reports of triptan-related vascular events