Data Integration and Harmonisation

Maxim Moinat, Vaclav Papez and Spiros Denaxas

Abstract

Data harmonisation is an essential step for federated research, which often involves heterogeneous data sources. A standardised structure and terminology of the source allows application of standardised study protocol and analysis code. A Common Data Model (CDM) accompanied with standardised software supports standardised federated analytics. In this chapter we demonstrate the benefit of Common Data Models and the OMOP CDM in particular. We also introduce a general pipeline of an Extract Transform Load process to transform health data to the OMOP CDM and provide an overview of the supporting tooling that ensures a high-quality conversion. Finally, we discuss potential challenges of the harmonisation process and how to address them.

Keywords

Common data model · Electronic health records · Medical ontologies · OMOP · Mapping · Harmonisation

1 Introduction to Common Data Models

1.1 Introduction

The previous chapters, especially Chap. 3—*Data standards and terminology including Biomedical ontologies*, have introduced various standards used in healthcare and biomedical research. Each standard addresses a particular purpose and helps organising and interpreting data. Although many standards are global and used across domains, many data use local standards, like national drug coding or customised EHR systems built for a hospital.

For large-scale studies, fundamental for AI, it is essential to integrate data from various sources. For example to characterise treatment patterns at different healthcare settings [1] or predicting the risk of multiple outcomes after a

M. Moinat
The Hyve, Utrecht, Netherlands

Erasmus Medical Centre Rotterdam, Rotterdam, Netherlands

V. Papez (✉) · S. Denaxas
Institute of Health Informatics, University College London, London, UK
e-mail: v.papez@ucl.ac.uk

Health Data Research UK, London, UK

S. Denaxas
British Heart Foundation Data Science Centre, Health Data Research UK, London, UK

F. W. Asselbergs et al. (eds.), *Clinical Applications of Artificial Intelligence in Real-World Data*,
https://doi.org/10.1007/978-3-031-36678-9_4

COVID19 infection [2]. This enables interoperability and reusability of the collected information, which are two of the FAIR principles emphasising machine-actionability of data [3].

One way is to harmonise the data to a Common Data Model (CDM). Data harmonisation is not an easy task. Healthcare databases can consist of many tables from diverse systems, like inpatient, outpatient, lab, pharmacy. And the source model and the CDM might capture data at different granularities, leading either to loss of information or requiring to derive missing information. The choice of data model and terminology is important as is the support for the CDM of choice.

1.2 Common Data Models

An EMA workshop report from 2017 describes a CDM as: "a mechanism by which raw data are standardised to a common structure, format and terminology independently from any particular study in order to allow a combined analysis across several databases/datasets" [4]. In this report three CDMs (OMOP CDM, Sentinel, Pcornet) were compared for use for pan-European observational health studies to address regulatory questions in a timely manner. Specifically, to use a CDM for Post Authorisation Safety Studies, drug utilisation and drug effectiveness studies on a wide population.

This definition shows the main components of a CDM. The first is a common structure, where the elements of the model are defined. In traditional models this is the definition of the tables and fields, for graph databases these will be the attributes of nodes and edges. The second is a common format, the form in which the data is presented. This can be flat tables, preferably as a relational database, or nested documents, like JSON. The third is a common terminology, defining the semantics of the values in the model. For example, the target vocabulary used for diagnoses. Preferably the values are richly annotated with metadata about the terminology used.

All three elements are crucial for machines to process the data. Ideally the data is also richly annotated with interoperable metadata that describes the structure, format, and terminology of the data. This enables machines without any prior knowledge of the data to access it.

A CDM is not application specific. Therefore, in most cases the data is not stored natively in this model. Having data in a CDM requires extraction from the application specific system, applying transformations and loading it into the CDM.

1.3 Common Data Models in the Biomedical Domain

The notion of using a CDM for biomedical data is not new. For many years, data from different sources has been integrated at institutional, regional, and also global levels. Table 1 gives an overview of a selection of important open healthcare standards and their main purpose.

HL7 FHIR [5] and OpenEHR [6] are models that directly integrate with the systems of a clinical care site. Their aim is not so much on research, but on processing healthcare data for their primary purpose: patient care. These models are important, as they are important entry points for integrating with models aimed towards research.

The other models have their specific research purposes. The OMOP CDM, maintained by the global OHDSI open science collaborative [7], is the main topic of this section and will be addressed in detail later. The CDISC SDTM [8] is a well-established standard for submission of Clinical Trial data to regulatory bodies and is required by e.g., the FDA. The Sentinel CDM is at the basis of an FDA funded federated network of US claims data [4]. The i2b2 model is the only model that is aimed at translational medicine and can be used to combine real world data from healthcare and research data.

Table 1 Standards for biomedical data and their main purpose

Standard	Main purpose
HL7 FHIR	*Record Exchange*: Connecting digital resources like software and devices in order to improve healthcare delivery
OHDSI OMOP CDM	*Observational Research*: Representing clinical data to do reproducible large scale medical evidence generation
OpenEHR Archetypes	*Clinical Care*: Collecting and organising electronic health records (EHR) data at the source
CDISC SDTM	*Clinical Trial:* Submitting data from studies to regulatory bodies like the FDA
Sentinel CDM	*Regulatory Observational Analysis:* Studies on a FDA network of US claims data
i2b2 model	*Translational Medicine:* Integrating data from healthcare and research

HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resource, OHDSI: Observational Health Data Sciences and Informatics, OMOP CDM: Observational Medical Outcomes Partnership Common Data Model, CDISC SDTM: Clinical Data Interchange Standards Consortium Standard Data Tabulation Model, i2b2: Informatics for Integrating Biology and the Bedside

1.4 Benefits of Harmonisation to a CDM

One of the benefits of a CDM is to enable large scale evidence generation across a federated network of data sources [9]. We assume here that federation means that the analysis is run locally and only the study results are shared back with the central study coordinator. The analysis, or study code, consists of two main pieces: phenotype algorithms for the target, comparator and outcome cohorts, and a statistical program e.g., written in R, SAS, or SPSS.

Let us assume we want to execute a study protocol across a set of similar, but structurally and semantically different, datasets. The protocol can be as simple as characterising a population of interest or as complex as building and (externally) validating a predictive model. The study protocol describes in text all the definitions and analytical procedures needed to execute the study. This includes among other the inclusion/exclusion criteria, the medical codes used for each, statistical methods, and outcome measures.

Without a CDM, the protocol has to be translated into four separate pieces of study code (Fig. 1, top left). This can be implemented in any programming language or statistical framework.

The re-implementation of the study protocol is not only labour intensive but will also result in other issues. Different interpretations of the protocol can result in analysis code being implemented differently. If the analysis procedure is not identical across sources, it is difficult to determine if any differences observed are due to the data or due to the analysis. And variations in the output format of the study results make aggregation of the final results harder.

With a CDM, the protocol has to be translated to study code only once (Fig. 1, top-right) and the code is shared between sites. This ensures each site executes exactly the study definition and outputs results in the same format. However, there is a high upfront cost to harmonise each data source to a CDM. Regardless of the choice of CDM, this is a big amount of effort and also variations can occur between data sources on the conventions used to populate the CDM. A common data quality assessment is key to spot any issues early on, which we will elaborate in the section.

It might be clear that a CDM will make cross-institutional network studies more reliable. However, an observant reader might have noticed that with a CDM a total of five 'translations' are necessary (four CDM, one study

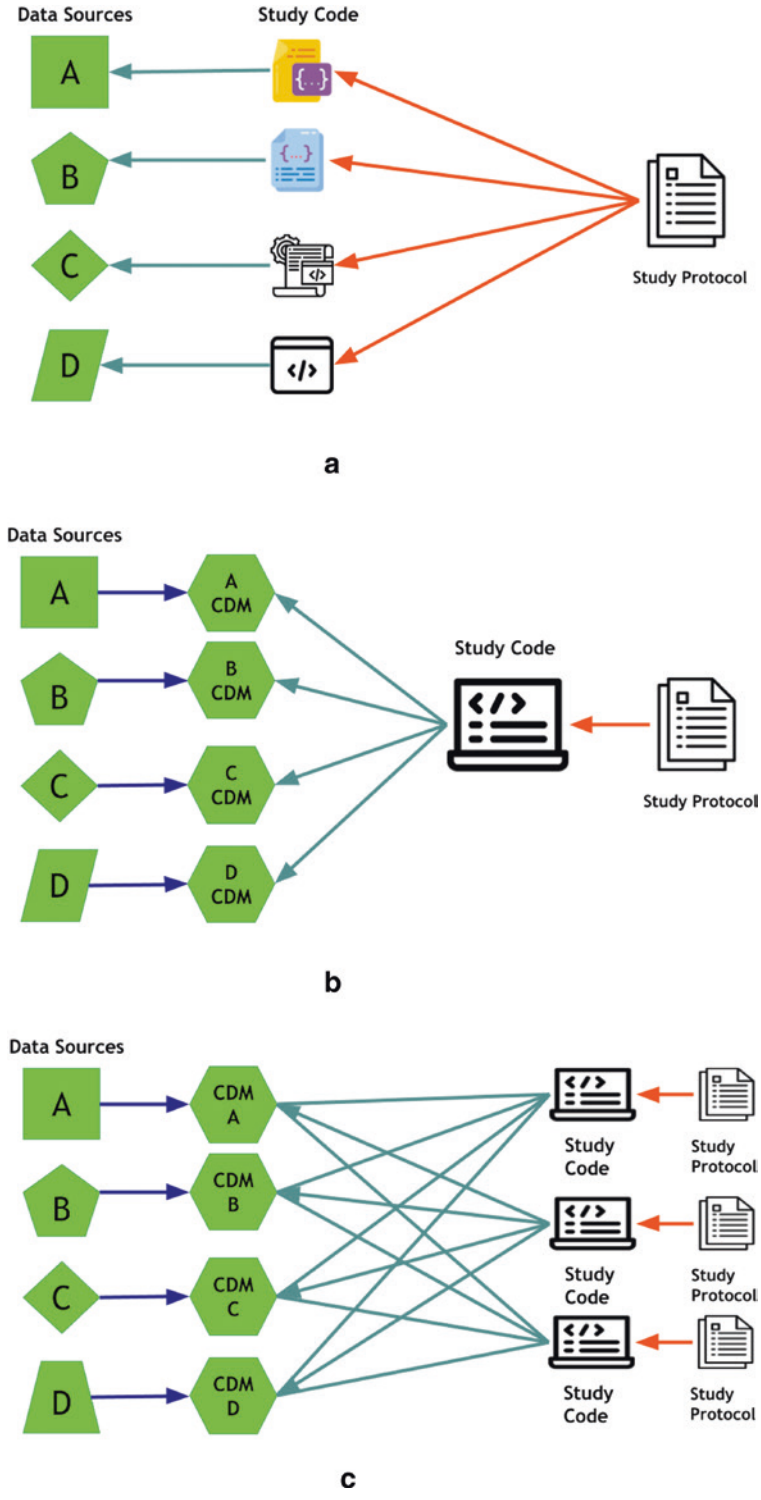

Fig. 1 Cross-institutional study of four structurally and semantically different databases (A, B, C, D). In the diagram on the top-left without a common data model. The protocol has to be 'translated' to study code for each of the data sources. In the diagram on the top-right each data source is harmonised to a CDM after which the protocol is 'translated' to one piece of study code that is executed against each CDM. In both scenarios the analysis is run locally and only study results are shared back with the central study coordinator. In the diagram on the bottom, performing multiple cross-institutional studies with a common data model is shown. After an initial harmonisation to a CDM, multiple studies are executed. Each requires translation to study code once

code) where without a CDM just four 'translations' are necessary (all study code). Also, harmonising a full data source to a CDM is often more work than creating a piece of study code focussed on a specific subset of variables. Thus, for one particular study using a CDM might not be worthwhile.

The real benefit of a CDM comes when executing a series of studies on the same network of data sources (Fig. 1, bottom). Without a CDM the number of code translations needed grows by multiplying the number of databases and studies. Executing one study across four databases requires 4 interfaces, executing ten studies across ten databases requires 100 interfaces. Instead of having to translate each protocol four times to code (resulting in twelve separate translations), this only has to be done three times in total (plus four CDM conversions). The number of databases is a constant for translations needed. And this scales of course when executing more studies across the network [10].

Furthermore, this goes beyond studies. A CDM enables the reuse of standard tooling for data quality assessment, visualisations, reporting and analysis. The OHDSI open science collaborative is a good example of a community that has produced a large library of standard tools and analytical methods around a CDM.

Standard research may be more costly for a single researcher compared with a bespoke study. But standardised research scales and benefits a community as a whole by enabling reuse. Akin 'Tragedy of Commons' where adding one cow to a field benefits a farmer, but degrades the field and negatively impacts the community as a whole [11].

Another benefit is that the conversion splits the path to evidence (i.e., study results) into two parts; the data harmonisation and the analysis execution. The harmonisation can be developed and evaluated separately from the analysis design.

2 The OMOP CDM

In this section we will dive deeper into one particular CDM, the OMOP CDM, which is used for research on real world healthcare data.

2.1 History

The OMOP CDM was born out of the Observational Medical Outcomes Partnership (OMOP), a public–private partnership chaired by the US FDA. This collaboration focussed on active medical product safety surveillance using observational healthcare data. In order to run studies across a heterogeneous set of databases, the OMOP Common Data Model was designed. This included standardised vocabularies for semantic interoperability. The OMOP studies showed successfully that it was possible to facilitate cross-institutional collaboration on safety studies [12].

After the lifetime of the OMOP project, the journey was continued as the currently well-known open science collaborative named OHDSI (Observational Health Data Sciences and Informatics, pronounced 'odyssey'). Under this collaboration, the use of the OMOP CDM was expanded to support a wide set of analytical use cases, like general comparative effectiveness of medical interventions, database characteristics and prediction models. All work is done collaboratively and published in the open domain. This includes data standards, ETL (Extract Transform Load) conventions, methodological research, and development of clinical applications.

2.2 The OMOP CDM

The OMOP CDM [13] is a relational database model consisting of 39 tables (Fig. 2), designed to store longitudinal health records

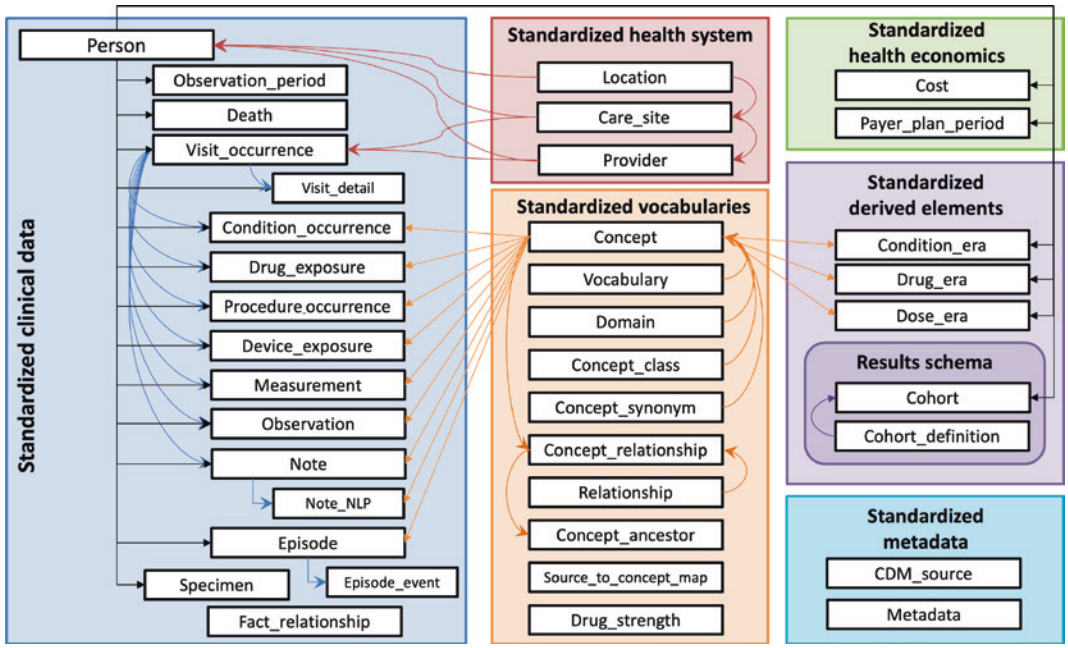

Fig. 2 The OMOP CDM overview of tables and relations between them [13]. The Person and Observation_period tables are the only ones required to be populated. The coloured boxes show the logical groupings of tables

collected from routine care. These are divided into seven logical groups. The tables from the 'Standardized clinical data' contain the main variables. Only the Person and Observation_period tables are required to be populated. The 'Standardized health system' tables provide additional context about who gave the care. The 'Standardized health economics' can contain associated costs of procedures and drugs and who pays these costs. Both the health system and economics data is often not made available by the source. The 'Standardized derived elements' are derived from the populated clinical data. The 'Standardized metadata' can provide information about the name of the data source, date of extraction and vocabulary version.

Every clinical event is captured in one of the eight domains, which each are stored in a separate table (Table 2). All clinical events, regardless of the domain, require at least a person_id (who), a fully specified date (when) and a concept_id (what). The concept_id has to refer to a standard concept from the OMOP Standardized vocabularies, explained in the next section.

Table 2 The eight domains of the OMOP CDM

Domain	Type of data
Condition occurrence	Diagnoses and symptoms
Drug exposure	Medications
Procedure occurrence	Diagnostic or surgical operations
Measurement	Lab results
Observation	Other clinical facts
Specimen	Sample, biopt
Device exposure	Medical equipment, Implantations, supplies
Note	Free text

Here we provide a short description of the most important tables in the OMOP CDM:

- Person contains demographic information. At least a year of birth and gender are required.
- Observation Period contains the periods of time for which we expect clinical events to be recorded for each person. This is important to determine 'healthy' time.

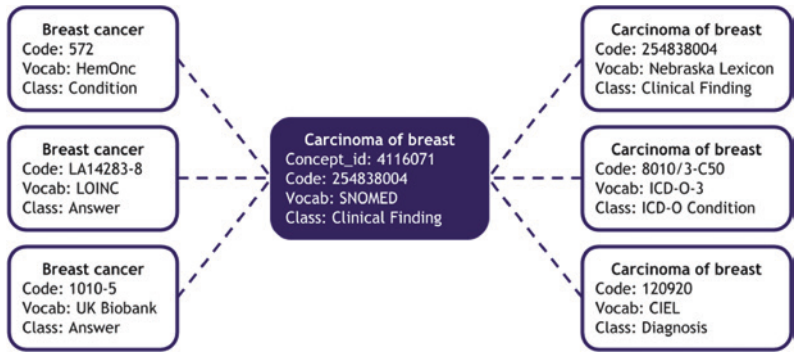

Fig. 3 The concept 'Carcinoma of breast' (SNOMED: 254838004) is the standard concept. Terms from other vocabularies with the same clinical meaning are mapped to this standard concept

- Death. At least the date is required, optionally the cause of death.
- Visit Occurrence contains the healthcare encounter, which can be anything between a short outpatient consult to a long hospitalisation.
- Drug Era is derived by combining single Drug Occurrences into longer periods of use of a particular ingredient.
- CDM Source. Contains the name of the dataset, date of extraction, link to ETL documentation, date of ETL process and vocabulary version.

2.3 The OMOP Standardised Vocabularies

"*The Standard Vocabulary is a foundational tool initially developed by some of us at OMOP that enables transparent and consistent content across disparate observational databases, and serves to support the OHDSI research community in conducting efficient and reproducible observational research.*" [14]

The OMOP Standardised Vocabularies provide semantic interoperability. It combines over 140 existing medical vocabularies, like ICD10, OPCS, SNOMED-CT, READ and RxNorm, into one vocabulary. See Chap. 3 for a more in-depth description of clinical terminologies. This is enriched with the mappings between the terms from these different vocabularies. Specifically, for each clinical idea (e.g. Type 2 Diabetes) one term is assigned as a **standard concept** and all similar terms are mapped to this standard concept (Fig. 3).

The latest release of the OMOP Standardised Vocabulary can be downloaded from Athena [15].

Not all medical ontologies are included in the OMOP Standardised Vocabularies. Especially local ontologies might be missing, for example a national medication vocabulary. In these cases for the mapping to the OMOP CDM, a manual conversion has to be created. This is explained in the sections below.

2.4 Use Cases from the OHDSI Community

The OHDSI community has created a wide range of tooling based on the OMOP CDM. We can roughly divide these tools into three categories: tools to help convert your data to the OMOP CDM, tools to design studies and tools to execute studies.

Using the study tooling, the OHDSI community has executed a quickly growing number of epidemiological studies. These studies can be separated into three pillars: characterization studies, comparative effectiveness/safety studies and prediction studies. Below we have selected three exemplary studies from the OHDSI community for each of these pillars. The focus is on

reproducible studies, each paper building new open-source standardised analytics or improving on existing analytics. All studies below are designed using Atlas [16]: a common analysis tool on a common data model.

3 General Pipeline of the Data Source Transformation to OMOP CDM Process

The ETL pipeline represents a series of steps which leads to a conversion of a source data model into a harmonised one. Whilst the desired goal is to automatize most steps in the pipeline, a manual intervention, mainly in source data preparation and terminology mappings, is often necessary.

A typical ETL pipeline consists of source preparation, environment setup, source data profiling, syntactic mapping, semantic mapping and finally validation and quality assessment of the target dataset. Some steps are usually realised iteratively, like going back to the syntactic mapping after quality assessment (Fig. 4, [17]).

Each of the ETL pipeline steps involve participation in one of more of the four typical roles. These groups are not necessarily disjunctive, and one person could fulfil multiple roles.

- Source data expert
- OMOP expert
- Technical ETL expert
- Clinical expert.

3.1 Source Preparation

By the source data we will assume a large dataset of structured (typically tabular) electronic health records (EHRs). This data needs to be analysed and prepared to be compatible with the ETL input interface. Patient level EHRs usually have restricted access and therefore a data governance process for corresponding roles is fundamental. For instance, source data experts and clinicians will typically have full access to the (pseudonymised) data, but OMOP or technical ETL experts might need only access to a subset or only a generated dataset.

Structured EHRs are usually stored in relational databases or plain text files like Comma Separated Values (CSV) files. In case of a plain text file, we need to know some basic file metadata: the coding set in which the files are saved, size of the files, container type if any (.zip archive,.tar.gz, etc.), presence of a table header row, separators between the table columns used (tabs, commas, semi-colons, etc.), quotation marks of character strings used (single or double quotation), end of the line characters used (linux based or windows based) and beginning of the line character used. In some cases this is well documented, in other cases this requires some investigation to get this information.

An upfront analysis of source data could help to estimate required computation power, storage, and free memory. Such information could help with setting up the environment to be supporting the ETL process.

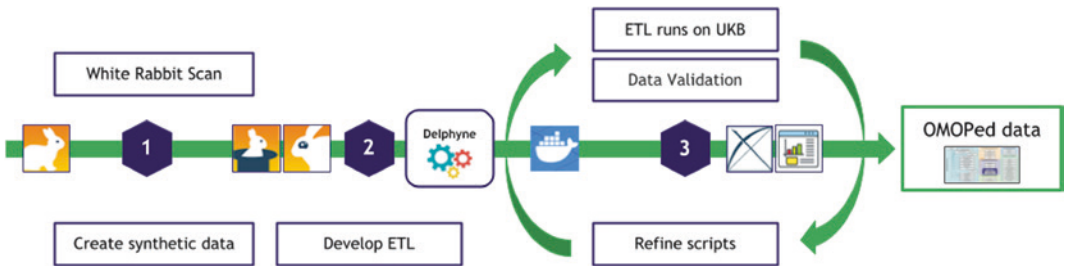

Fig. 4 ETL Pipeline—Transformation of UK Biobank into OMOP CDM use-case [17]

3.2 Environment Setup

The ETL environment consists of a database and an environment to run data transformation scripts. This runtime environment could be dedicated to the ETL or an existing shared environment could be used. Having a dedicated environment (both physical or virtual) means that all software requirements can be installed in isolation and hardware resources would not be shared with other processes. That decreases a risk of negative impact on a shared source data server as well as the ETL stability in case of a potential hardware overload or software incompatibility between the source server requirements and ETL requirements. A drawback of a fully dedicated environment could be a necessity of source data duplication. Also, a dedicated physical environment usually requires extra hardware, which adds overhead cost.

A specification of the ETL runtime environment requirements should contain hardware resources, hosting OS, required target DB system, required input form of source data, list of preinstalled tools, compilers, interpreters, and system and language specific libraries and packages. Main environmental dependency for OMOP CDM ETL is compatible DBMS. OMOP CDM v6 supports multiple DBMS including Oracle DB, PostgreSQL, and MS SQL Server. Other typical environmental requirements are Python 3 and R.

The minimal requirements for setting up an OMOP CDM and analysis environment are listed below:

- Server with about 3× the size of the source data (for raw source, OMOPed data, vocabulary data and Data Quality results)
- Relational database (Oracle DB, PostgreSQL, MS SQL Server)
- The OMOP vocabulary
- Java (White Rabbit, Usagi)
- R + OHDSI R packages for DQ (Achilles, DataQualityDashboard)
- Python (optional, being used as a workflow wrapper)
- OHDSI HADES R packages (analysis)
- OHDSI WebApi+Atlas (analysis)
- Bespoke mapping tools (optional, for example delphyne [17] or Perseus [18])

3.3 Data Profiling

A source data profile provides essential information required for ETL design, synthetic data generation (if necessary), data extraction code and validation test design. The data profile could be created using a dedicated tool like OHDSI WhiteRabbit [19] or by a direct query to all source data tables. Dedicated tools can be connected to the source data, and these will provide the report automatically. In both cases the analysis report ideally contains the following information for all tables:

- Table name with a description
- Field or attribute names
- Number of rows per table
- Number and/or percentage of values in each field—total, unique and empty
- Field data types
- List of most occurring values (e.g., diagnostic codes, measurement values, etc.) for each domain including their frequencies.

With a data profile, a data extraction and two types of transformation—syntactic and semantic—need to be performed. With syntactic mapping we describe a transformation of source attributes onto those of OMOP CDM tables and source values formatting. The semantic mapping covers a translation of source coding systems into systems supported by OMOP CDM.

3.4 Syntactic Mapping

In syntactic, or structural, mapping we define which source table fields/attributes map to which fields of the target model. This step could also include changes in source values structure, e.g., year taken from the date. An example of

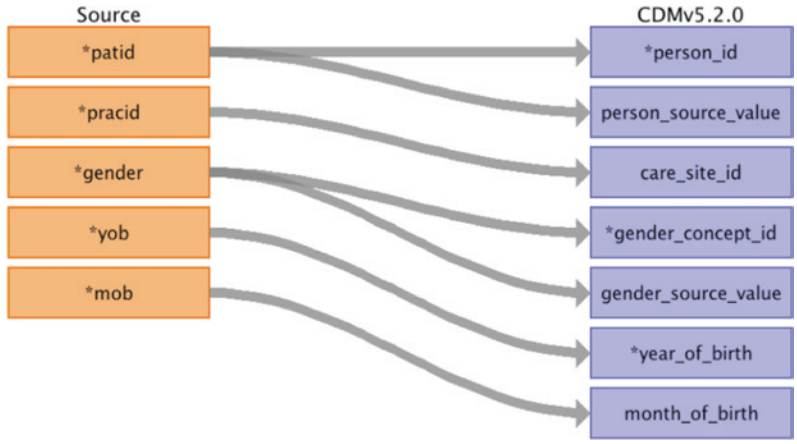

Fig. 5 A Syntax mapping between the CPRD patient table and Person table of OMOP CDM. Graphical representation was generated by the Rabbit-in-a-hat tool

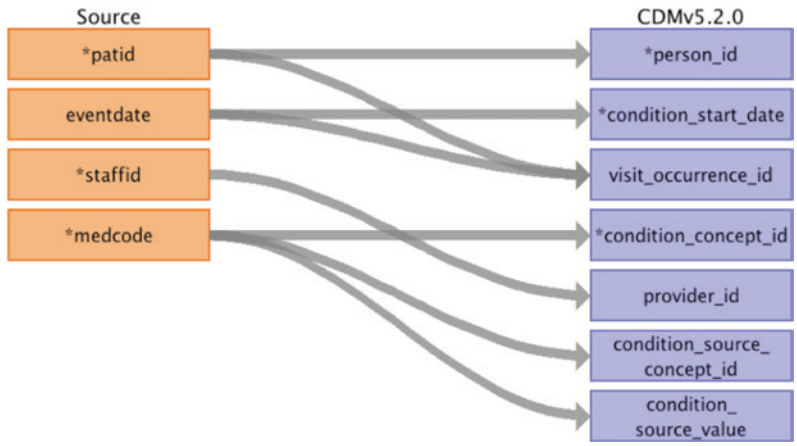

Fig. 6 A Syntax mapping between the CPRD clinical table and Condition Occurrence table of OMOP CDM. Graphical representation was generated by the Rabbit-in-a-hat tool

syntax mapping of a CPRD patient table onto OMOP CDM person table and CPRD clinical table onto OMOP CDM Condition Occurrence table can be seen in Figs. 5 and 6 respectively. The figures were generated by a Rabbit in a Hat tool [20]. Rabbit in a Hat is a syntax mapping assistant for OMOP CDM ETL development. Its graphical user interface (GUI) allows users to visualise syntax mapping between source data structure imported via WhiteRabbit scan report and target version of OMOP CDM. The tool helps with the manual mapping design via graphical representation and mapping document generation, however, the transformation code itself has to be implemented manually.

Two main issues for syntactic mapping could occur.

- the source data is missing for the required field in the target model
- source data elements do not have any equivalents in the target structure.

The first situation can be handled by a logic populating the missing and required target fields, e.g., a fixed value. The second situation

may represent a challenge. A main question in that case should be if the data without the equivalent fields in the target structure are necessary or if these could be omitted, e.g., administrative data may not be of interest for population research. If the data is necessary, then the solution depends on the flexibility and robustness of the target data model and potential workarounds. OMOP CDM provides categorised, yet generic, elements/fields suitable for most health-related data to minimise a potential data loss.

3.5 Semantic Mapping

The semantic mapping is often done in the first stages of the ETL development and applied at the same time as the syntactic mapping. i.e., when transforming a local source code field to a standard concept field, we apply the prepared semantic to translate one coding system into the other.

Electronic health data is captured using a variety of medical terminologies (see Chap. 3). These terminologies, or coding systems, allow us to structurally capture things like diagnosis codes, drug codes, measurement units, ethnicity, etc. Often data sites use a mix of local and global terminologies. For network research, we need to harmonise the local coding systems to an agreed upon global standard. For OMOP specifically, we need to map source codes to the standard OMOP vocabulary concepts (see Sect. 2 The OMOP CDM).

Whilst syntactic mapping is mainly manual work, semantic mapping could be effectively automated when a machine-readable validated dictionary lookup between source and target vocabularies exists. Within an OMOP vocabulary, such a lookup is called *concept mapping*. In general, a concept mapping between the source and target terminology could have four existential forms:

1. Direct concept mapping between the source and the target vocabulary exists
2. Direct concept mapping between the source and the target vocabulary does not exist, however an intermediate mapping exists and could be used
3. The concept mapping does not exist
4. Source and target use the same vocabulary (e.g., SNOMED CT).

In the first situation, the concept mapping could be implemented directly into the ETL scripts. A large repository of OMOP CDM compatible dictionaries could be found in the OHDSI Athena Repository [15]. An example of such a scenario could be a mapping between ICD10 terminology and SNOMED CT.

In the second situation, a chain of existing suitable concatenated mappings could substitute a missing direct trustworthy concept mapping. Such a solution is challenging and data loss risk increases with each additional mapping involved (see Sect. 4 Challenges). Figure 7 provides an example observed within a transformation of the CALIBER data source [21, 22].

Thirdly, when no direct or indirect mapping dictionary between the source and target vocabulary exists a new concept mapping needs to be created and reviewed by domain experts thoroughly. Tools designed to ease the new mapping development exist, like OHDSI Usagi [23]. Usagi provides a graphical user interface comparing the uploaded source terminology with selected standard terminologies supported by OMOP CDM. Within the comparison, Usagi calculates a match score based on a similarity between the source and target terms and automatically matches the most likely terms. Each suggestion has to be reviewed by a clinical expert to create a validated concept mapping. There can be thousands of codes that need to be reviewed, which is a considerable amount of work. We can prioritise this work by using the term frequency (Fig. 8).

Finally, when the source data uses a coding system that is already used as standard concepts in OMOP, only a simple lookup of the OMOP concept id is needed. For example, part of the UK data is coded at the source with SNOMED codes. This code is present in the OMOP vocabularies and can be retrieved with a simple SQL query. One consideration is to check whether the

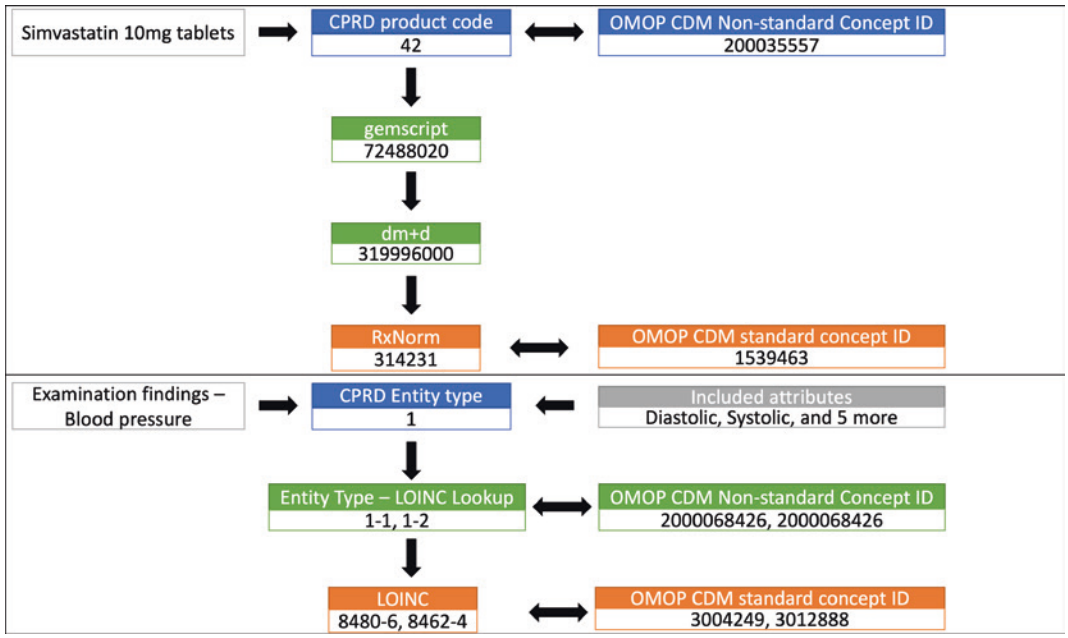

Fig. 7 Two examples where additional mappings were used. In the first case, CPRD product codes were translated into a gemscript terminology, then to dm+d terminology and finally to a target RxNorm terminology. In the second example, CPRD Entity types were firstly translated via a manual mapping file

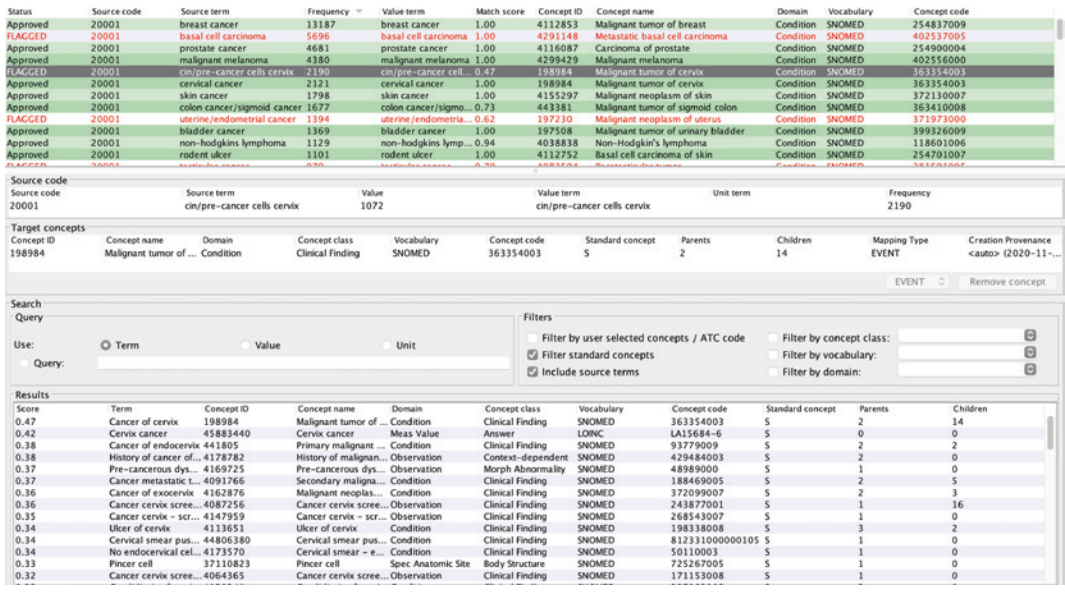

Fig. 8 OHDSI USAGI mapping tool comparing source participant self-reported cancer-illness UK Biobank vocabulary and target SNOMED vocabulary. Codes are ordered by the frequency of used terms

code in the OMOP vocabulary is still valid. If not, the OMOP vocabulary provides a mapping to the equivalent valid concept.

3.6 Validation

Validation starts during the ETL development by implementing a set of unit and/or end-to-end tests. Unit tests are for validating particular data manipulation functions and end-to-end tests allow validation of the whole pipeline by providing a known input and the expected output. The latter is especially important for validating the complete ETL pipeline. It makes it possible to detect any unwanted effects of code changes before running the ETL on actual data. We should note that it takes considerable effort to get a high coverage of tests, covering the most occurring scenarios.

Once the ETL is finished, a comprehensive validation of the target database including correctness of both semantic and syntactic mappings needs to be performed.

A first check on the ETL completeness is given by a comparison of general counts representing the dataset between the source and target databases. These counts typically include the number of patients, ratio between sex/ethnicity, average patient age, the number of events/prescriptions or median follow-up. Analytic tools like Achilles [24], Data Quality Dashboard (DQD) [25] and CDM Inspection [26] help to easily retrieve these overall counts from the OMOP CDM. The DQD provides a series of checks resulting in a data quality score. This score makes heterogeneous source datasets comparable on the same data quality metrics.

Another tool developed by OHDSI, the CDM Inspection report, contains a list of most used mapped and unmapped terms. Thus, the unmapped terms could be investigated individually based on their significance.

For use-case based (non-systematic) ETL evaluation miscellaneous codelists/cohort definitions to identify specific patient cohorts covering diverse fields of health care can be used. In previous research we have shown the validation process, comparing results on the source data and OMOP-transformed data for lifestyle data (smoking status, deprivation index), clinical measures (BMI, Blood pressure, haemoglobin concentration), clinical diagnosis (diabetes, cancer) or drug prescriptions (Beta blockers, loop diuretics) [22]. As the thorough test of all the used codes is time consuming, we should prioritise tests on the most frequently used codes and most needed codes according to the use case.

The OHDSI community has developed a tool, Cohort Diagnostics, that does something similar. Based on a set of phenotypes it will make suggestions on what other concepts are relevant and produce aggregate statistics to manually inspect [27].

4 Challenges of Harmonisation

Harmonisation of diverse data models into a common one in the health/bioinformatics domain is accompanied by several inevitable challenges.

4.1 Data and Information Loss

One of the most crucial challenges is to prevent the harmonisation from relevant data and/or information loss. Relevance of data depends on the purpose of the harmonised dataset, e.g., administrative details or internal hospital information would not be relevant for population-level studies and thus could be lost with no harm.

While data loss is mainly (not exclusively) caused by the structural mapping when part of the source data is not transformed into a target model, an information loss could be given also by the incorrect or imprecise interpretation and translation of the transformed data during the semantic mapping.

4.1.1 Data Loss

Data could get lost in the ETL process and/or due to issues/inconsistencies in the original datasets. Source data providers may use diverse

recording practices (table structures, used coding systems), documentation practices, management of missing data, technicalities of data distribution, data cleaning processes before the distribution, etc. Combination of these factors within the same source dataset could lead to scenarios predisposed to data losses, e.g.:

- A source record does not include a data field which is mandatory from the perspective of CDM. This can be handled in two ways—making an assumption for this field or removing the patient during the ETL, e.g., a registration date may be inferred from other fields, however, records belonging to patients with missing mandatory demographic details like gender or year of birth would be removed during the ETL. These patients are deemed to be of too low quality for population research.
- Unexpected value in a domain for a specific data field, e.g., values are expected to be positive only, however a negative value appears
- Diagnostic events happen outside the patient's observation period which starts with a patient's registration date at GP and ends with the last event or the patient's death.
- A broken follow up when a patient changes GP; the scenario could lead to a situation when one patient is being considered as two different ones.
- Same data field is using multiple different coding systems (e.g., ICD10 and SNOMED CT) and these are not explicitly distinguished.
- Source record contains a clinical code unrecognised in a mapping dictionary / target vocabulary used.
- Inconsistent records for unvarying data fields, e.g., a same patient would have a different sex during different visits.

Some of these scenarios can be fixed during the ETL (e.g., handling unexpected values), but others are inherent to incompatibilities between source and target data model. Therefore, a potential risk of data loss is inevitable.

4.1.2 Information Loss

Despite the correct and complete syntactic transformation, the information derived from the source records may not be fully reflected in the target CDM. Such information loss is often caused by an imprecise semantic translation from the source to the target coding system.

A source and target terminology could have a different level of granularity. This gives problems if the source terminology contains terms with more detail than the target terminology. Generalisation of the source term solves the problem at a cost of losing details. Incompleteness could be also found in a translation relation itself. Figure 9 shows the loss of information on a fragment of Chronic Obstructive Pulmonary Disease (COPD) phenotype. The loss of information of this type causes another secondary issue which is an incompatibility of source clinically approved phenotyping codelists with transformed CDM as these codelists cannot be precisely translated into the target terminology.

In the OMOP CDM, the granularity is preserved in the 'source concepts'. Locally, we can still define phenotypes based on the original codes. However, definitions based on source concepts instead of standard concepts are not executable at other data sites as these will not, very likely, share same local source concepts.

We can distinguish between several levels of equivalence of code translation (Table 3) [28]. The two top-levels without information loss are equal (exactly the same term) and equivalent (similar definition). Information loss occurs when a translation is wider (target term is more general), narrower (target term is more specific) or inexact (both source and target have meaning not covered in the other). The latter three levels still capture a part of the information but can lead to issues as described in Fig. 7. Most information loss occurs when a source code is unmatched in the target coding system (or 'unmapped'). Unfortunately, this is often unavoidable, and the percentage of unmapped codes is an important quality metric. In all cases this is

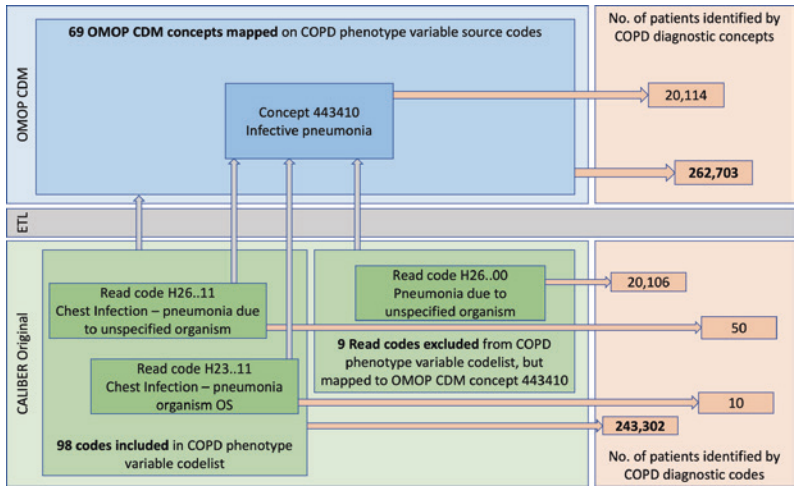

Fig. 9 Papez et al. [22] Example of inconsistency between original and converted records demonstrated with codelist from the Chronic Obstructive Pulmonary Disease (COPD) phenotype. Multiple source terminology terms codes (Read codes in green boxes) are mapped onto the same OMOP CDM target concept (blue box). The mapped concept however includes a broader set of clinical diagnoses which are not part of the original COPD phenotype. As a result, the number of patients retrieved (orange boxes) in the raw data using the original phenotype terms (243,302) is significantly lower than the number of patients retrieved using the OMOP CDM phenotype (262,703). Main result difference is caused by the Read code H26.0.00 Pneumonia due to unspecified organism used in more that 20,106 patients, which is excluded from COPD phenotype, but mapped to the same concept of Infective pneumonia as other Read codes from the phenotype

Table 3 Examples of equivalence levels

Equivalence level	Description	Example
Equivalent	Source and target contain the same information	Source—Depression Assessment Test Target—Assessment of depressed mood
Wider	The target is a more general concept than the source. In the mapping some information is lost, but the general information is captured	Source—Release of the median nerve at the carpal tunnel, by video surgery Target—Transposition of median nerve at carpal tunnel
Narrower	The target is a more specific concept than the source. In the mapping some information is added	Source—Corneal pachymetry Target—Ophthalmic ultrasound, diagnostic; corneal pachymetry, unilateral or bilateral (determination of corneal thickness)
Inexact	The target and source contain information that is not present in the other. In the mapping information is both lost and added	Source—Screening tests for deafness before the age of 3 years old Target—Ear disorder screening

a subject worthy of investigation whether it can be improved.

A key resource when fixing above mentioned issues is time. While the source and target terms with similar descriptions could be handled automatically, the others must be manually mapped or at least reviewed. Tools like Usagi provide a great help in sorting the terms by their frequency in the source dataset and calculation of text similarity weight between source and mostly probable target term. This speeds up the review process of mostly used terms rapidly. It is still good to be aware that even highly similar terms are not necessarily synonyms diverse in a

punctuation or case sensitivity but could differ in presented negation which changes their meaning; on the other hand, terms with almost 0% similarity could be synonyms, e.g., cancer and malignant neoplasm. However, as the similarity between terms together with their frequency decrease, time resource required per clinical record in the dataset grows massively and the rule of the vital few[1] is applied. A review of the controlled terminologies and mappings is a task for domain experts with a corresponding expertise. Such a review could increase demanded time resources to an unacceptable amount.

4.2 Data Privacy and Sensitivity

Working with personal-level health data is usually accompanied with a strict policy regarding data privacy and sensitivity. Usually, only a selected subset of people involved in the ETL development has an approval to access the health data that the ETL is being developed for. Also, a common practice is that the health data must not leave the datacenter the data is stored in, which in some cases differs from the centre where the ETL code is being developed, tested or even performed in case of the dedicated ETL environment. Therefore, the ETL development might have to be realised using synthetic data only. Despite the identical structure the synthetic data could have with the real data, unexpected differences in the value domains could appear (see Sect. 4.1.1 Data loss).

Usually, synthetic data are generated by bespoke tools designed for one specific purpose like the tool Tofu [29] for UK Biobank data or by a generic tool for synthetic EHRs like Synthea [30]. A usage of a generic synthetic data could lead to an additional challenge when the structure of the synthetic data needs to be transformed into a source data structure, i.e., additional syntactic ETL process.

Restricted patient-level data access is also related to derived reports. Data profiling reports should contain only those information which could be shared between all developers and testers who need it, e.g., data profiling report would contain aggregated information only.

[1] The Pareto Principle, also 80/20 principle; applied on the mapping problem the principle says that by covering 20% of most frequently used terms, an 80% of all records will be mapped correctly. In the opposite way, to cover/map the last 20% of source terms will take approx. 80% of time.

References

1. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A. 2016;113:7329–36. https://doi.org/10.1073/pnas.1510502113.
2. Williams RD, Markus AF, Yang C, et al. Seek COVER: development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. bioRxiv. 2020. https://doi.org/10.1101/2020.05.26.20112649.
3. FAIR principles. GO FAIR. 2017. https://www.go-fair.org/fair-principles/ (Accessed 29 Jun 2022).
4. EMA. A common data model in Europe? – Why? Which? How? European Medicines Agency. 2018. https://www.ema.europa.eu/events/common-data-model-europe-why-which-how (Accessed 29 Jun 2022).
5. FHIR v4.3.0. http://hl7.org/fhir/R4B (Accessed 29 Jun 2022).
6. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153–73. https://www.ncbi.nlm.nih.gov/pubmed/16160223.
7. OHDSI—observational health data sciences and informatics. http://ohdsi.org (Accessed 29 Jun 2022).
8. SDTM. https://www.cdisc.org/standards/foundational/sdtm (Accessed 29 Jun 2022).
9. Schuemie MJ, Ryan PB, Pratt N, et al. Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study. J Am Med Inform Assoc. 2020;27:1268–77. https://doi.org/10.1093/jamia/ocaa124.
10. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19:54–60. https://doi.org/10.1136/amiajnl-2011-000376.
11. Benson T, Grieve G. Principles of health interoperability: FHIR, HL7 and SNOMED CT. Springer Nature 2020. https://play.google.com/store/books/details?id=TiwEEAAAQBAJ.
12. Observational Health Data Sciences, Informatics. Chapter 1 the OHDSI community. 2021.https://ohdsi.github.io/TheBookOfOhdsi/OhdsiCommunity.html (Accessed 29 Jun 2022).

13. index.knit. https://ohdsi.github.io/CommonDataModel/index.html (Accessed 29 Jun 2022).
14. Data standardization. https://ohdsi.org/data-standardization/ (Accessed 29 Jun 2022).
15. Liu J, Li D, Gioiosa R, et al. Athena. In: Proceedings of the ACM international conference on supercomputing. New York, NY, USA: ACM 2021. https://doi.org/10.1145/3447818.3460355.
16. Kernighan BW, Plauger PJ. Software tools. SIGSOFT Softw Eng Notes. 1976;1:15–20. https://doi.org/10.1145/1010726.1010728.
17. Digital Natives. Mapping UK Biobank to the OMOP CDM using the flexible ETL framework Delphyne. the-hyve. https://www.thehyve.nl/cases/mapping-uk-biobank-to-omop-using-delphyne (Accessed 19 Jul 2022).
18. 'Perseus': Design and run your own ETL to CDM. https://ohdsi.org/2021-global-symposium-showcase-79/ (Accessed 19 Jul 2022).
19. *OHDSI WhiteRabbit tool*. Github https://github.com/OHDSI/WhiteRabbit (Accessed 25 May 2022).
20. Rabbit in a Hat. http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html (Accessed 29 Jun 2022).
21. Denaxas SC, George J, Herrett E, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41:1625–38. https://doi.org/10.1093/ije/dys188.
22. Papez V, Moinat M, Payralbe S, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open 2021;4:ooab001. https://doi.org/10.1093/jamiaopen/ooab001.
23. USAGI for vocabulary mapping. https://www.ohdsi.org/analytic-tools/usagi/ (Accessed 29 Jun 2022).
24. ACHILLES for data characterization. https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/ (Accessed 29 Jun 2022).
25. DataQualityDashboard: A tool to help improve data quality standards in observational data science. Github https://github.com/OHDSI/DataQualityDashboard (Accessed 29 Jun 2022).
26. CdmInspection: R Package to support quality control inspection of an OMOP-CDM instance. Github https://github.com/EHDEN/CdmInspection (Accessed 29 Jun 2022).
27. Diagnostics for OHDSI cohorts. https://ohdsi.github.io/CohortDiagnostics/ (Accessed 29 Jun 2022).
28. Valueset-concept-map-equivalence - FHIR v4.3.0. https://www.hl7.org/fhir/valueset-concept-map-equivalence.html (Accessed 29 Jun 2022).
29. Denaxas S. spiros/tofu: Updated release for DOI. 2020. https://doi.org/10.5281/zenodo.3634604.
30. syntheticHealth. Github https://github.com/syntheticHealth (Accessed 29 Jun 2022).