OMOP Common Data Model Tutorial

With taking synthetic data as an example

Tutors: Honghan Wu, Minhong Wang, Debbie Wake

University of Edinburgh

April 2021

# Summary

In this tutorial, we will cover the streamlined pipeline of applying health data science to routine data for research. You will learn about the common data model for electronic health records (EHR) and multiple standardised terminologies. You will also learn about phenotype computation and data visualisation. You will have a hands-on experience of working with such type of data.

This tutorial includes multiple R files, as listed below. We recommend you to go through this document first then go through the R files in the order we provide here.

* HDR Tutorial – R basics.ipynb
* HDR Tutorial – Data model mapping.ipynb
* HDR Tutorial – Phenotype Computation.ipynb

# Background information

## OMOP

The Observational Health Data Sciences and Informatics maintains a standard, or Common Data Model (OHDSI CDM, often pronounced OMOP), to represent clinical data for research.

<https://www.ohdsi.org/data-standardization/the-common-data-model/>

### What is data standardisation and why is it important?

Data standardisation is the critical process of bringing data into a common format that allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies.

Healthcare data can vary greatly from one organisation to the next. Data are collected for different purposes, such as provider reimbursement, clinical research, and direct patient care. These data may be stored in different formats using different database systems and information models. And despite the growing use of standard terminologies in healthcare, the same concept (e.g., blood glucose) may be represented in a variety of ways from one setting to the next.

OHDSI provides resources to convert a wide variety of datasets into the CDM, as well as a plethora of tools to take advantage of data once it is in CDM format.

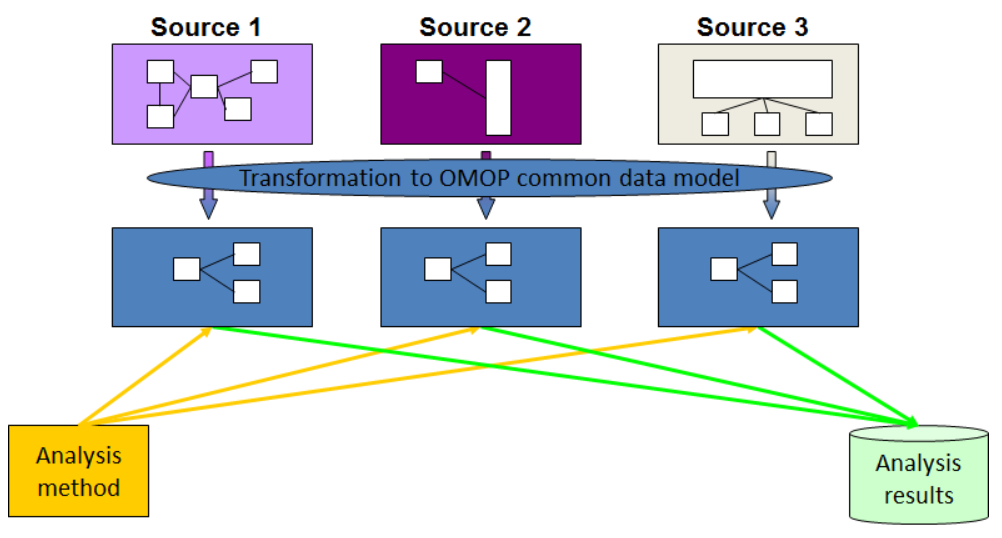


Figure 1: An illustrative plot of applying OMOP CDM on different sources.

### What is the OMOP CDM?

The OMOP CDM allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format.

More detailed information on data tables is provided here: <https://ohdsi.github.io/CommonDataModel/cdm60.html>

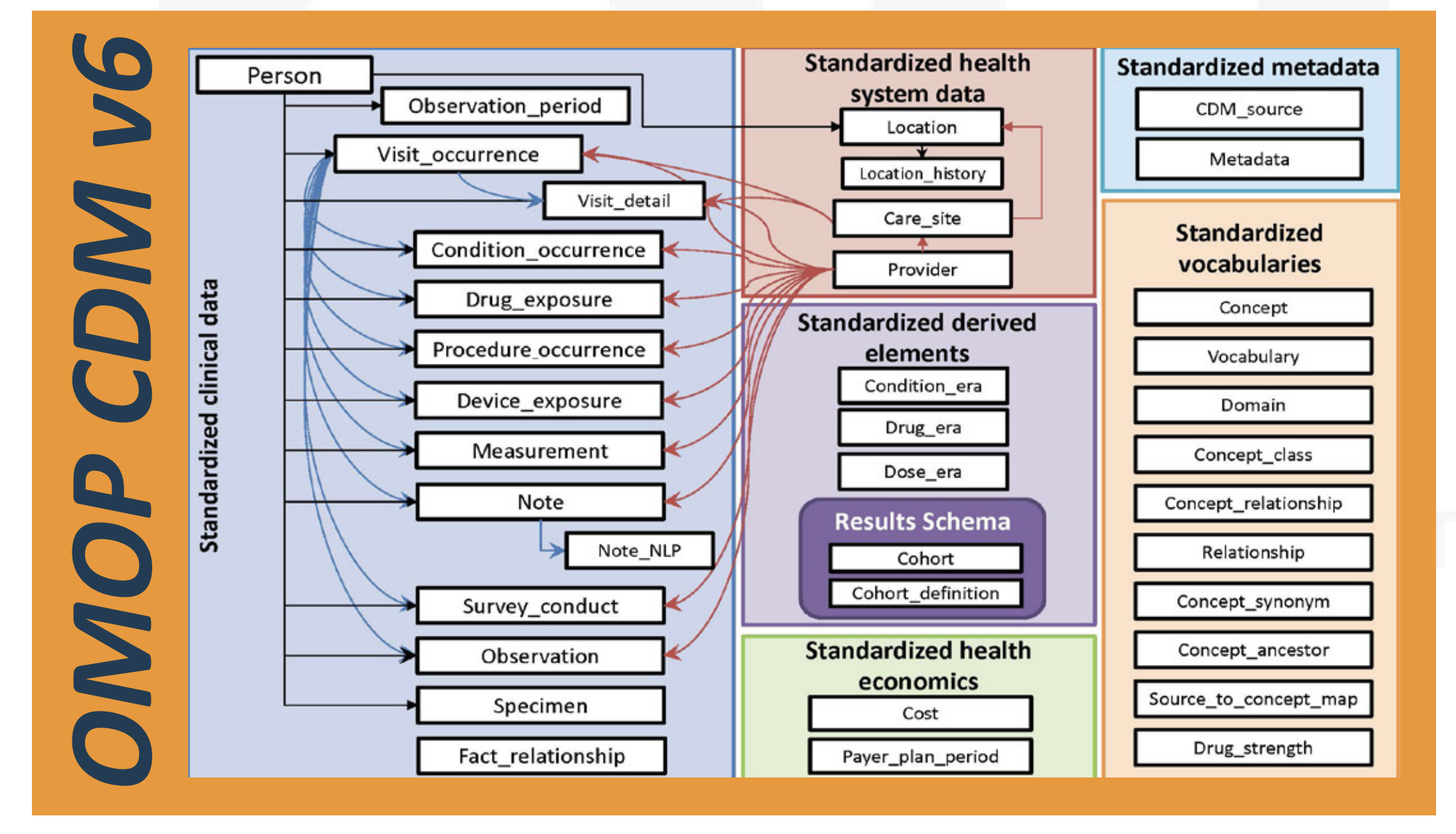


Figure 2: Overview of all tables in the CDM version 6.0. Note that not all relationships between tables are shown.

### Why do we need a CDM?

Observational databases differ in both purpose and design. Electronic Health Records (EHR) are aimed at supporting clinical practice at the point of care, while administrative claims data are built for the insurance reimbursement processes. Each has been collected for a different purpose, resulting in different logical organisations and physical formats, and the terminologies used to describe the medicinal products and clinical conditions vary from source to source.

The CDM can accommodate both administrative claims and EHR, allowing users to generate evidence from a wide variety of sources. It would also support collaborative research across data sources across the world, in addition to being manageable for data owners and useful for data users.

## Clinical Terminologies and Standards

These standards address a fundamental requirement for effective communication – the ability to represent concepts unambiguously between a sender and receiver of information. Most communication between health information systems relies on structured vocabularies, terminologies, code sets and classification systems to represent health concepts.

In this tutorial, we use ICD-10 for diagnosis, BNF for drug, Test codes for laboratory tests.

### ICD-10

ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organisation (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

<https://www.who.int/standards/classifications/classification-of-diseases>

### BNF

British National Formulary (BNF) is a reference book containing the standard list of medicines used in the UK prescribing.

# Outline

## Synthetic data generation

We use synthetic data in this tutorial. Synthetic medical data were generated by BadMedicine for this tutorial. Data generated by BadMedicine is based on (simple) models generated from live EHR datasets collected for over 30 years in Tayside and Fife (UK). This makes the data generated recognisable (codes used, frequency of codes etc.) from a clinical perspective and representative of the problems (ontology mapping etc.) that data analysts would encounter working with real medical data.

The following synthetic datasets were produced:

* Demography: address and patient details as might appear in the CHI register
* Biochemistry: Lab test codes as might appear in Sci Store lab system extracts
* Prescribing: Prescription data of prescribed drugs
* Hospital admissions: ICD9 and ICD10 codes for admission to hospital

We generated synthetic data for 1,000 patients with 10,000 records per dataset.

## Mapping data to the common data model (CDM)

Table mapping.ipynb provides the code we used for mapping the synthetic data to the common data model. Based on the synthetic data we have, we generated 6 tables for CDM:

* person: demographic data
* condition\_occurrence: diagnoses information
* drug\_exposure: drug use information
* measurement: lab tests results
* visit\_occurrence: hospital admission information
* concept: reference table

## Creating the database

The database was prepared for this tutorial following the instructions of OMOP. The tables we generated from the previous section were loaded already. The steps and scripts of constructing OMOP CDM are provided on <https://github.com/OHDSI/CommonDataModel/tree/master/PostgreSQL>

## Access the database for phenotype computing

HDR Tutorial - Phenotype Computation.ipynb illustrates the streamlined pipeline of applying health data science on routine data for research. In this section, we access the database from R and show some examples of retrieve patient data with specific conditions. We will retrieve type 2 diabetes patients based on diagnoses and lab test results. We also provide examples of achieving basic visualisation of retrieved patient data.