Implementation Guide Appendix 1 - Minimum OMOP Dataset

Introduction

This guide outlines the minimum set of OMOP fields that are required in order to make your data observable on Cohort Discovery.

If you are not already familiar with OMOP then this link provides some useful background information on the Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM), including its role in health informatics and general design principles.

The rest of this section will focus on the application of OMOP in the context of the Cohort Discovery tool.

Background

OMOP is used globally by academic and healthcare organisations to support observational health research. Recently, the NHS Research Secure Data Environment (SDE) Network has agreed to adopt the OMOP CDM which aims to brings consistency to health data regardless of where it exists. It is patient centric, tabular, extendable, built for analytics and has a relational design making it simple to do analytics that are reproducible, interpretable and trustworthy.

> ℹ️ If you have already mapped your data to OMOP for general research purposes, this will likely already include the core OMOP fields required for Cohort Discovery.

Objective

The objectives are to:

- Create one standard OMOP map from each source dataset, using the 30 core minimum OMOP fields to be used for all Cohort Discovery. This will enable researchers to find cohorts of data relevant to their research using basic minimal information about a patient, their condition, procedure, drug exposure, observations or measurements.
- Once a researcher has identified a data source that contains data of interest in this core OMOP dataset using the Cohort Discovery tool, they would then submit a data access request to the custodian of that data in order to carry out their research.
- For subsequent analytics, i.e., once an appropriate cohort has been identified, and the researcher/analyst has gained access to the source data via data access requests, then additional purpose driven OMOP maps would need to be created to suit a particular type of analysis, or if OMOP is not suitable for a more complex study design then the researcher may need to request access directly to the underlying source data.

OMOP mapping for Cohort Discovery and Analytics is represented conceptually in figure O1 below. This shows how complex a data model is and the stage of a researchers journey that a model is most appropriate for. So for example, an OMOP data model for a Case Control study would be more complex than an OMOP data model for Cohort Discovery (because it would likely make use of a wider OMOP data space that utilises more than the 30 minimum OMOP fields), and it would be used further along the researcher journey.

Figure O1 - OMOP Mapping for Cohort Discovery and Analytics
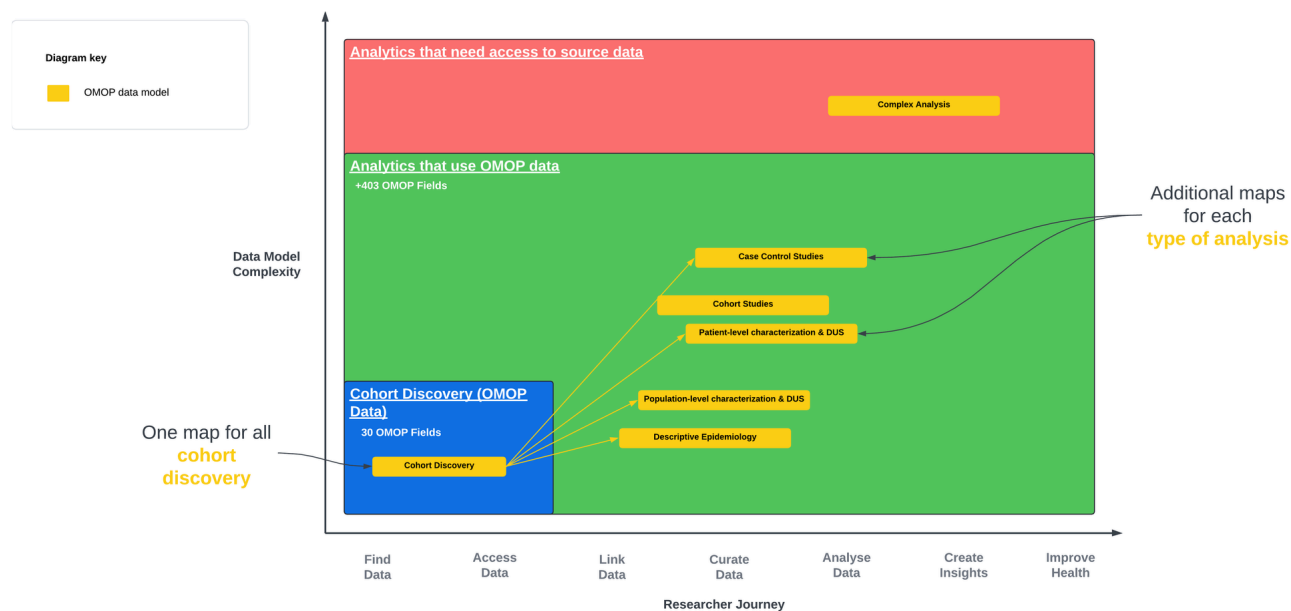
There are 30 core fields required for Cohort Discovery, and a further 403 (at the time of writing) OMOP fields available for additional analytical purposes.

As you develop mapping rules for additional fields, they can be added to your institutional "memory" so that they can be re-used in future studies. Mapping rules can be shared between source datasets and between institutions to support reproducible, interpretable and trustworthy analytics.

Cohort Discovery Minimum OMOP Dataset

The NHS Research Secure Data Environment (SDE) network and HDR have collaboratively agreed that source data should be mapped to a minimum set of OMOP fields to enable observability via the Cohort Discovery tool. This minimum dataset represents basic minimal information about a patient, their condition, procedure, drug exposure, observations and measurements that help researchers find and access cohorts for further analysis (see table O1 below).

Table O1 - Minimum viable OMOP fields for Cohort Discovery

OMOP Table	OMOP Field Names
Person	person_id, **gender_concept_id, year_of_birth, race_concept_id**
Condition_Occurrence	condition_occurrence_id, person_id, **condition_concept_id**, **condition_start_date**
Procedure_Occurrence	procedure_occurrence_id, person_id, **procedure_concept_id, procedure_date**
Drug_Exposure	drug_exposure_id, person_id, **drug_concept_id, drug_exposure_start_date**

Observation	observation_id, person_id, **observation_concept_id, observation_date**, value_as_number(*), value_as_concept_id(*)
Measurement	measurement_id, person_id, **measurement_concept_id**, **measurement_date**, **value_as_number, value_as_concept_id**
Death	person_id, **death_date, death_type_concept_id, cause_concept_id**
Visit_Occurrence	visit_occurrence_id(*), person_id(*), visit_concept_id(*), visit_start_date(*)

(*) - Optional, but useful

Fields in bold are the core content of interest for analysis. These are typically OMOP 'concepts' and any associated dates that the concept occurred. The remaining non-bold fields are fields required to link tables together.

A description of each field is in table 2 below and outlines the rationale behind its inclusion.

Table O2 - Minimum viable OMOP fields for Cohort Discovery with rationale

Include	Optional

cdmTableName	cdmFieldName	Rationale
CONDITION_OCCURRENCE	condition_occurrence_id	Unique identifier for each condition occurrence, facilitating tracking and analysis of individual diagnoses.
CONDITION_OCCURRENCE	person_id	Key
CONDITION_OCCURRENCE	condition_concept_id	Identifies the specific medical condition for cohort definition, disease progression analysis, and treatment research.
CONDITION_OCCURRENCE	condition_start_date	Tracks the date of condition onset for understanding disease trajectories and identifying cohorts based on diagnosis timing.
DEATH	person_id	Key
DEATH	death_date	Records the date of death for mortality-related cohorts and survival analysis.
DEATH	death_type_concept_id	Identifies the type of death (e.g., natural, accidental) for cohort definition and research on specific death causes.
DEATH	cause_concept_id	Identifies the underlying cause of death for mortality analysis and research on disease burden and risk factors.
DRUG_EXPOSURE	drug_exposure_id	Unique identifier for each drug exposure, allowing tracking and analysis of individual medication regimens.
DRUG_EXPOSURE	person_id	Key
DRUG_EXPOSURE	drug_concept_id	Identifies the specific drug for cohort definition, drug-drug interaction analysis, and medication effectiveness research.

DRUG_EXPOSURE	drug_exposure_start_date	Tracks the date of drug exposure initiation for understanding treatment patterns and defining exposure-based cohorts.
MEASUREMENT	measurement_id	Unique identifier for each measurement, facilitating tracking and analysis of individual measurements.
MEASUREMENT	person_id	Key
MEASUREMENT	measurement_concept_id	Identifies the specific type of measurement (e.g., blood pressure, lab test) for cohort definition and analysis of relevant clinical characteristics.
MEASUREMENT	measurement_date	Tracks the date of each measurement for understanding trends over time and defining cohorts based on measurement timing.
MEASUREMENT	value_as_number	Value of the measurement
MEASUREMENT	value_as_concept_id	Value of the measurement
OBSERVATION	observation_id	Unique identifier for each observation, enabling tracking and analysis of individual observations.
OBSERVATION	person_id	Key
OBSERVATION	observation_concept_id	Identifies the specific type of observation (e.g., physical exam finding, imaging result) for cohort definition and analysis of relevant clinical features.
OBSERVATION	observation_date	Tracks the date of each observation for understanding progress over time and defining cohorts based on observation timing.
OBSERVATION	value_as_number	Observation values possible of interest but not essential?
OBSERVATION	value_as_concept_id	Observation values possible of interest but not essential?
PERSON	person_id	Unique identifier for each individual, enabling linking across all tables.
PERSON	gender_concept_id	Identifies sex category for potential inclusion in sex-specific cohorts.
PERSON	year_of_birth	Determines age groups for cohort definitions and facilitates age-related analyses.
PERSON	race_concept_id	Potentially informative for understanding health disparities
PROCEDURE_OCCURRENCE	procedure_occurrence_id	Unique identifier for each procedure, enabling tracking and analysis of individual procedures.
PROCEDURE_OCCURRENCE	person_id	Key
PROCEDURE_OCCURRENCE	procedure_concept_id	Identifies the specific type of procedure for cohort definition, analysis of procedure utilisation and effectiveness research.
PROCEDURE_OCCURRENCE	procedure_date	Tracks the date of each procedure for understanding temporal trends and defining cohorts based on procedure timing.
VISIT_OCCURRENCE	visit_occurrence_id	Unique identifier for each visit, allowing differentiation and analysis of individual encounters.

VISIT_OCCURRENCE	person_id	Key
VISIT_OCCURRENCE	visit_concept_id	Identifies the type of visit (e.g., inpatient, outpatient) for cohort definition and resource utilisation analysis.
VISIT_OCCURRENCE	visit_start_date	Tracks the timing of each visit for temporal analysis and cohort inclusion based on visit timing.