

Intro to EHR Data

Reference Guide & Resources



The Intro to Electronic Health Record (EHR) Data reference guide helps researchers understand the EHR data within the *All of Us* dataset and its organization in the *All of Us* Researcher Workbench.

As you go through this reference guide, we hope you leave with a basic understanding of EHR data, Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), Athena, and more.

1. Electronic Health Records (EHR) data

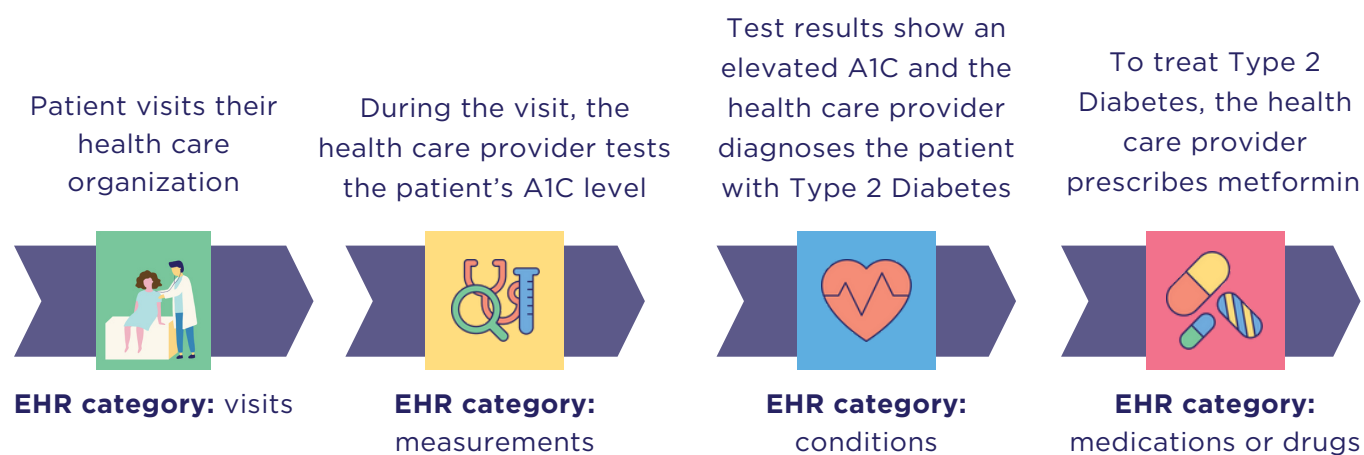
Electronic Health Record (EHR) data is generally organized into eight categories:

- Conditions (a diagnosis, a sign, or a symptom)
- Visits or encounters
- Procedures
- Measurements, including laboratory tests
- Drugs or medications
- Medical devices
- Other observations or clinical facts not included in the categories above

Each category represents clinical events that happened or were performed during a patient's visit to a health care organization or outside as part of their clinical care routine at home.

Let's consider the following scenario: a patient visits their health care organization (**EHR category:** visits) to test their A1C level (**EHR category:** measurements).

Test results show an elevated A1C and the provider diagnoses the patient with Type 2 diabetes (**EHR category:** conditions) and prescribes metformin (**EHR category:** medications or drugs).



All of the information collected during the patient's visit is then recorded into the health care organization's EHR system using codes related to clinical terminologies.

Clinical terminologies are the “standardized terms which record patient findings, circumstances, events, and interventions with details to support clinical care, decision support, outcomes research, and quality improvement; and can be efficiently mapped to broader classifications for administrative, regulatory, oversight and fiscal requirements.”

There are five standard clinical terminologies used in recording EHR data.

International Classification of Disease (ICD)	A classification coding system to represent conditions or diagnoses and it is used mainly for billing. There are multiple versions of ICD codes including ICD-9 and ICD-10.
Current Procedural Terminology (CPT)	A coding system that provides a coding scheme for surgical and diagnostic services and procedures.
Systematized Nomenclature of Medicine - Clinical Terminology (SNOMED-CT)	A multidisciplinary terminology that is considered the standard in representing diseases, clinical findings, procedures, and medications.
RxNorm	A system that provides normalized and standardized names for medications and drugs.
Logical Observation Identifiers Names and Codes (LOINC)	A collection of universal names and codes to represent laboratory tests and clinical observations.

Not all EHR databases are the same. Different health care organizations may use different clinical terminologies in their EHR data.

For example, the condition Type 2 diabetes may be recorded as SNOMED-CT code 44054006 at Health Care Organization A and ICD-10 code E11 at Health Care Organization B. This variability in EHR data coding impacts the feasibility of performing clinical research across different organizations.

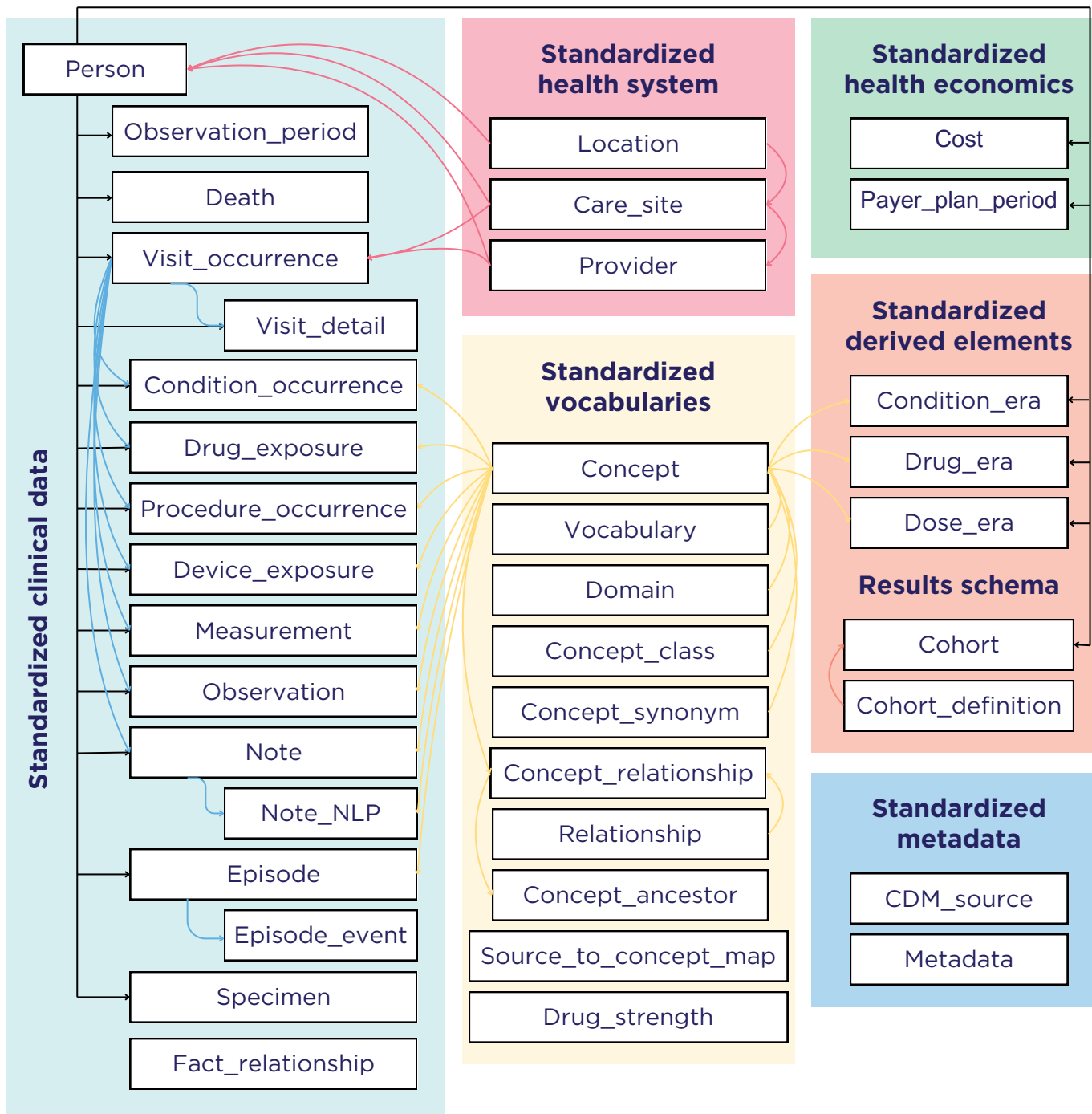
Common data models exist to standardize EHR data for clinical research. Examples of common data models are Informatics for Integrating Biology and the Bedside (i2b2) and Observational Medical Outcomes Partnership (OMOP).

2. Observational Medical Outcomes Partnership (OMOP)

The *All of Us* Research Program uses the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to store and standardize health and survey data collected from consented participants across the United States and its territories.

OMOP stores data in a relational database. Maintained by an international collaborative called the Observational Health Data Sciences and Informatics (OHDSI) program, OMOP contains **39 unique tables** (white boxes) within **six data categories** (color boxes) that relate to one another.

Think of the **39 unique tables** as spreadsheets that include patient demographics (person table), condition (condition_occurrence table), medication (drug_exposure table), procedure (procedure_occurrence table), and more. Each table contains a variety of fields (columns).



Some of these fields are unique to the table. For example, the drug_exposure table has a column that includes information about medication dose that does not exist in other tables.

Exercise 1: Let's apply the OMOP model with the earlier scenario about a patient visiting their health care organization to test their A1C level. The patient visited their health care organization where their A1C level was tested. During the visit, the provider diagnosed the patient with Type 2 diabetes and prescribed metformin.

Based on what we know about the OMOP model, what tables would we look at to find information about metformin, Type 2 diabetes, and A1C measurement?

Exercise 1 answer: drug_exposure (metformin); condition_occurrence (Type 2 diabetes); measurement (A1C measurement)

There are also identical fields across different tables that relate to one another called key variables. In the diagram, tables linked with an arrow indicate an identical field exists.

Exercise 2: Comparing the condition_occurrence and measurement tables, what identical field, or key variable, is present in both tables?

Condition_occurrence

1	condition_occurrence_id	person_id	condition_concept_id
2			
3			
4			
5			
6			
7			
8			
9			
10			

Measurement

1	measurement_id	person_id	measurement_concept_id
2			
3			
4			
5			
6			
7			
8			
9			
10			

Exercise 2 answer: person_id

Why do key variables matter? Key variables allow tables to be merged for research analysis. For example, we can merge, or match, condition_occurrence and measurement using the person_id to analyze the occurrence of Type 2 diabetes diagnosis with corresponding A1C measurements.

Condition_occurrence

CDM Field	User Guide	ETL Conventions	Datatype	Required	Primary Key	Foreign Key	FK Table	FK Domain
condition_occurrence_id	The unique key given to a condition record for a person. Refer to the ETL for how duplicate conditions during the same visit were handled.	Each instance of a condition present in the source data should be assigned this unique key. In some cases, a person can have multiple records of the same condition within the same visit. It is valid to keep these duplicates and assign them individual, unique, CONDITION_OCCURRENCE_IDs, though it is up to the ETL how they should be handled.	integer	Yes	Yes	No		
person_id	The PERSON_ID of the PERSON for whom the condition is recorded.		integer	Yes	No	Yes	PERSON	
condition_concept_id	The CONDITION_CONCEPT_ID field is recommended for primary use in analyses, and must be used for network studies. This is the standard concept mapped from the source value which represents a condition.	The CONCEPT_ID that the CONDITION_SOURCE_VALUE maps to. Only records whose source values map to concepts with a domain of "Condition" should go in this table. Accepted Concepts	integer	Yes	No	Yes	CONCEPT	Condition
condition_start_date	Use this date to determine the start date of the condition.	Most often data sources do not have the idea of a start date for a condition. Rather, if a source only has one date associated with a condition record it is acceptable to	date	Yes	No	No		

Measurement

CDM Field	User Guide	ETL Conventions	Datatype	Required	Primary Key	Foreign Key	FK Table	FK Domain
measurement_id	The unique key given to a Measurement record for a Person. Refer to the ETL for how duplicate Measurements during the same visit were handled.	Each instance of a measurement present in the source data should be assigned this unique key. In some cases, a person can have multiple records of the same measurement within the same visit. It is valid to keep these duplicates and assign them individual, unique, MEASUREMENT_IDs, though it is up to the ETL how they should be handled.	integer	Yes	Yes	No		
person_id	The PERSON_ID of the Person for whom the Measurement is recorded. This may be a system generated code.		integer	Yes	No	Yes	PERSON	
measurement_concept_id	The MEASUREMENT_CONCEPT_ID field is recommended for primary use in analyses, and must be used for network studies.	The CONCEPT_ID that the MEASUREMENT_SOURCE_CONCEPT_ID maps to. Only records whose SOURCE_CONCEPT_IDs map to Standard Concepts with a domain of "Measurement" should go in this table.	integer	Yes	No	Yes	CONCEPT	Measurement
measurement_date	Use this date to determine the date of the measurement.	If there are multiple dates in the source, date, order_date, draw_date, and result_date, choose the one that is closest to the date the sample was drawn from the patient.	date	Yes	No	No		
measurement_datetime	This is not required, though it is in v6. If a source does not specify the time to midnight (00:00:00).		datetime	No	No	No		
measurement_time	This is present for backwards compatibility and will be deprecated in an upcoming version.		varchar(10)	No	No	No		

Because the *All of Us* Research Program uses OMOP to store health and survey data, it's important to understand OMOP structure. All 39 tables and their corresponding fields are available online. [View the comprehensive list of OMOP tables and fields.](#)

OMOP standardizes medical terms and concepts. The *All of Us* Research Program collects EHR data from participants and health care organizations from across the United States and its territories.

Different health care organizations may use different clinical terminologies to code the same EHR concept. For example, the condition Type 2 diabetes may be recorded as SNOMED-CT code 44054006 at Health Care Organization A and ICD-10 code E11 at Health Care Organization B.

All of Us uses OMOP to map, or standardize, the different clinical terminologies, or source vocabulary, to a standard vocabulary. OMOP standardizes health data into six major OMOP EHR domains:

- Conditions
- Devices
- Drugs
- Measurements
- Observations
- Procedures

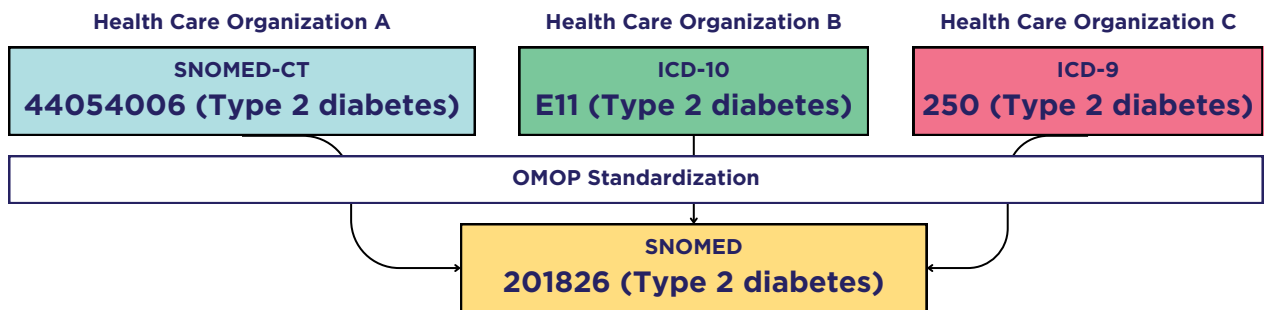
Each OMOP EHR domain has a set standard vocabulary that all source vocabulary is mapped to with OMOP standardization.

Domain	Standard Vocabulary
Conditions	SNOMED
Measurements	LOINC
Drugs	RxNorm
Procedures	SNOMED
Program Physical Measurements	SNOMED, LOINC, PPI*
Survey Questions and Answers	SNOMED, LOINC, PPI*

*Participant Provided Information (PPI)

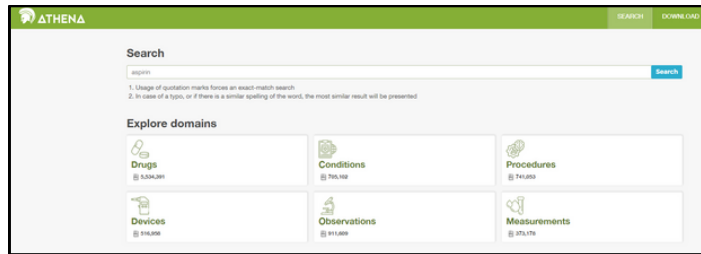
All source codes are mapped to OMOP’s standard vocabulary, which is SNOMED for the conditions domain. **Note:** all of the source codes are also retained in the OMOP database so that the data can still be searched using the source codes.

For example, the condition Type 2 diabetes may be recorded as SNOMED-CT code 44054006 at Health Care Organization A and ICD-10 code E11 at Health Care Organization B. During the standardization process, OMOP maps all the source vocabulary to a standard vocabulary defined by OMOP. For the conditions domain, sources codes are mapped to a standard SNOMED code.

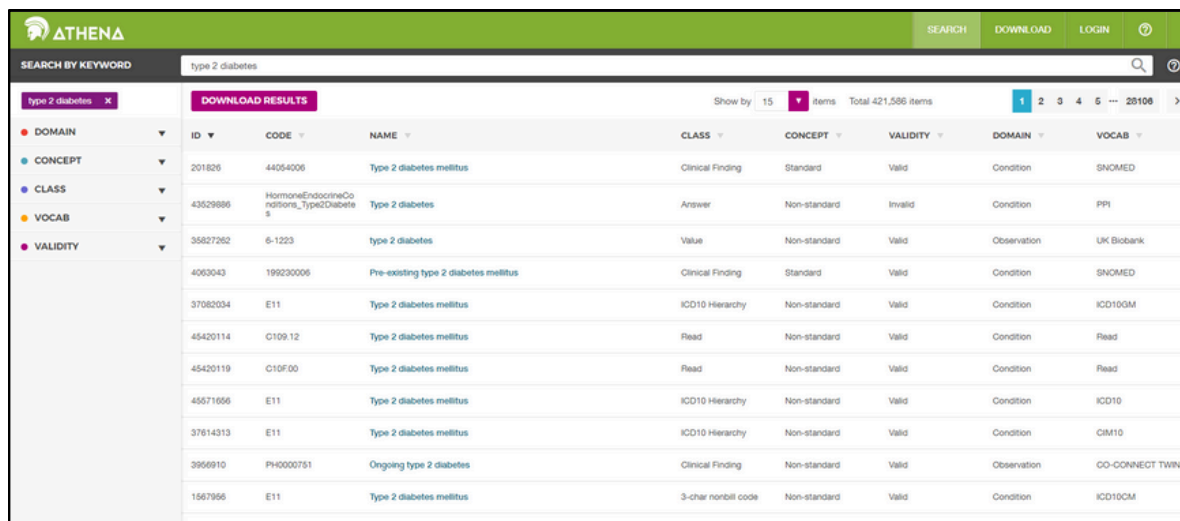


3. Athena

Athena is a publicly available website for browsing and exploring the standardized vocabularies used in OMOP. Using Athena, you can search by typing into the search bar or explore topics by the OMOP EHR domains.



Continuing our example scenario, let's search for Type 2 diabetes.



ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
201826	44054006	Type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED
43529886	HormoneEndocrineConditions_Type2Diabetes	Type 2 diabetes	Answer	Non-standard	Invalid	Condition	PPI
35827262	6-1223	type 2 diabetes	Value	Non-standard	Valid	Observation	UK Biobank
4063043	199230006	Pre-existing type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED
37082034	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICD10GM
45420114	C109.12	Type 2 diabetes mellitus	Read	Non-standard	Valid	Condition	Read
45420119	C10F00	Type 2 diabetes mellitus	Read	Non-standard	Valid	Condition	Read
45571656	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICD10
37614313	E11	Type 2 diabetes mellitus	ICD10 Hierarchy	Non-standard	Valid	Condition	ICM10
3956910	PH0000751	Ongoing type 2 diabetes	Clinical Finding	Non-standard	Valid	Observation	CO-CONNECT TWINS
1567956	E11	Type 2 diabetes mellitus	3-char nonbill code	Non-standard	Valid	Condition	ICD10CM

The search results display in a list format with multiple columns. The first and second columns are the concept ID and the concept code respectively followed by the name. The concept code is the source code retained, and the concept ID is a string of numbers given to the concept by OMOP for that specific concept code.

The list also includes if the concept is standard or not, which OMOP EHR domain the concept belongs to, and what the standard vocabulary is for the concept.

In the Type 2 diabetes search results, the first listing is marked as a standard concept while the second listing is marked as non-standard. When working with *All of Us* data, it's important to work with the standard concept.

Exercise 3

Using **Athena**, what are the **standard** concept ID and **standard** concept code for Type 2 diabetes?

Exercise 3 answer: 201826 (standard concept ID); 44054006 (standard concept code)



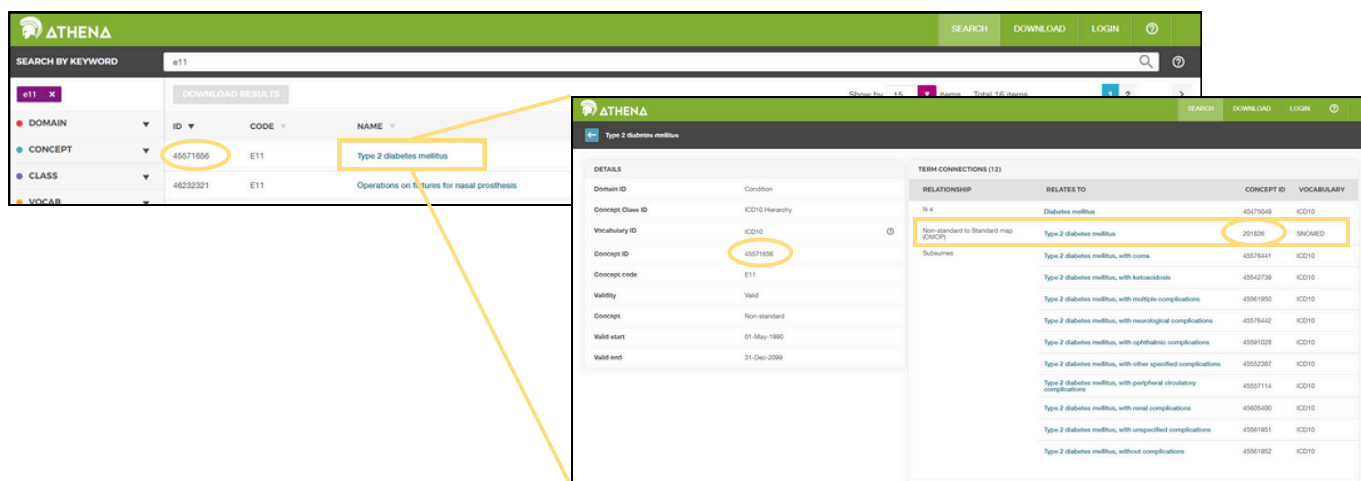
ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
201826	44054006	Type 2 diabetes mellitus	Clinical Finding	Standard	Valid	Condition	SNOMED
43529886	HormoneEndocrineConditions_Type2Diabetes	Type 2 diabetes	Answer	Non-standard	Invalid	Condition	PPI
35827262	6-1223	type 2 diabetes	Value	Non-standard	Valid	Observation	UK Biobank

Because OMOP retains the original source code when standardizing, searching is possible in multiple ways in Athena. You can type the name of a concept (i.e., Type 2 diabetes), the OMOP standard code for the concept (i.e., 201826), or the original source code for the concept (i.e., SNOMED-CT 44054006).

Exercise 4

Let's try searching with the source code instead of the name. Using Athena, what are the concept ID and **standard** concept ID for ICD-10 code E11?

Exercise 4 answer: 45571656 (concept ID); 201826 (standard concept ID)



Additional Resources

Below are additional, in-depth articles related to learning about the *All of Us* data, getting started in the Researcher Workbench, and analyzing data in the Workbench.

Learning the basics of the *All of Us* dataset

[Data curation process for the *All of Us* data](#)

[Participant privacy protections](#)

[Types of *All of Us* data](#)

[Understanding OMOP basics](#)

[Exploring concepts with OMOP and SQL](#)

Getting started in the Researcher Workbench

[Intro to the *All of Us* Researcher Workbench](#)

[About workspaces](#)

[Selecting participants using the Cohort Builder](#)

[Building a dataset with the Dataset Builder](#)

Analyzing data in the Researcher Workbench

[Overview of applications in the Researcher Workbench](#)

[Exporting and analyzing your data in the Researcher Workbench](#)

If you have questions or need assistance, please contact us at support@researchallofus.org.