Introduction to AIR·MS

Health Data Literacy

Ashwin Sawant
Andrew Deonarine

Credits: Eugenia Alleva

Sept 15th, 2026

HPI·MS

Mount Sinai

Hasso Plattner Institute for Digital Health at Mount Sinai

Coding Systems & Concept Mapping

Mapping Overview

- We need to map information between Epic and AIRMS (OMOP)
- Epic stores medications, diagnoses, procedures, etc. In Epic ID codes, a proprietary commercial system they maintain
- These codes are then mapped to other coding systems like ICD-10, SNOMED, RxNorm, etc.
- Mount Sinai's OMOP has their own mapping tables (because we can't use the ones in Epic due to commercial license issues)
- Why can't we just map Epic codes to a standard set of OMOP codes in a big table? Due to codes being commercial properties, licensing, etc.

Exploring the OMOP Mapping

AIR·MS

- OMOP uses several standard vocabularies for mapping
- This table describes the relationships between the EHR coding system (ex. Epic, PowerChart, etc) and the OMOP Standard Vocabulary

EHR Source	Source Coding System	OMOP Standard Vocabulary
Diagnoses	ICD-10-CM, Epic Codes	SNOMED CT
Procedures	CPT, HCPCS	SNOMED CT / CPT4
Drugs	ATC, NDC	RxNorm
Labs	LOINC	LOINC
Vitals	LOINC	LOINC / Extensions
Devices	HCPCS	SNOMED CT
…		

https://athena.ohdsi.org/search-terms/start

MSDW has an “Extended” OMOP CDM

Column Type	OMOP CDM	MSDW Extension Tables
Standard OMOP columns	242	*n/a*
De-identification columns that mask standard OMOP PHI columns	104	0
Extension columns from Caboodle	297	54
Data lineage & ETL audit columns	479	45
TOTAL:	**1,122**	**99**

22 of 40 OMOP tables
(versions 5.3, 5.4, 6.0)

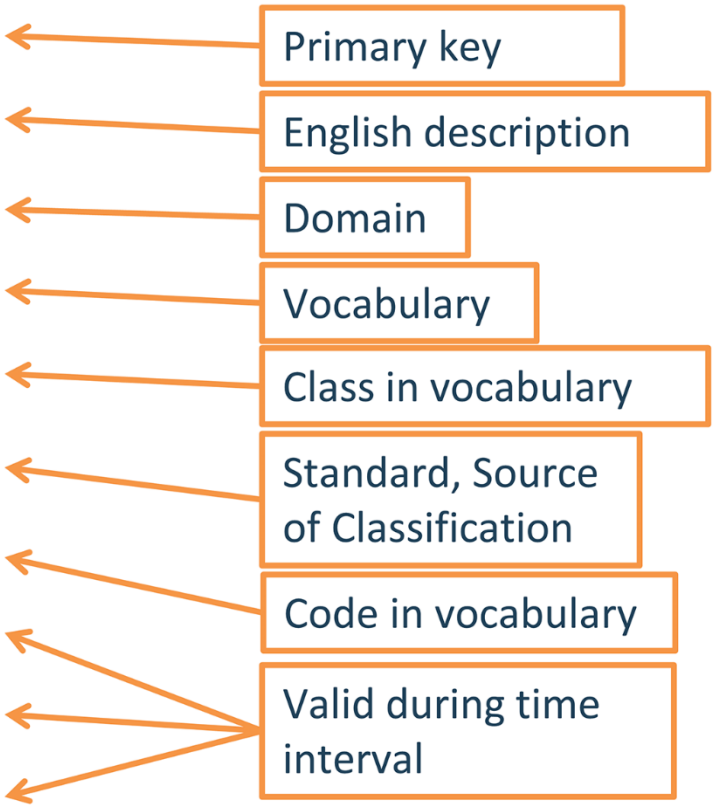
cdc_race_ethnicity_xtn
provider_attribute_xtn
date_xtn
time_xtn

Using the Data: Querying and Retrieving Data

AIR·MS

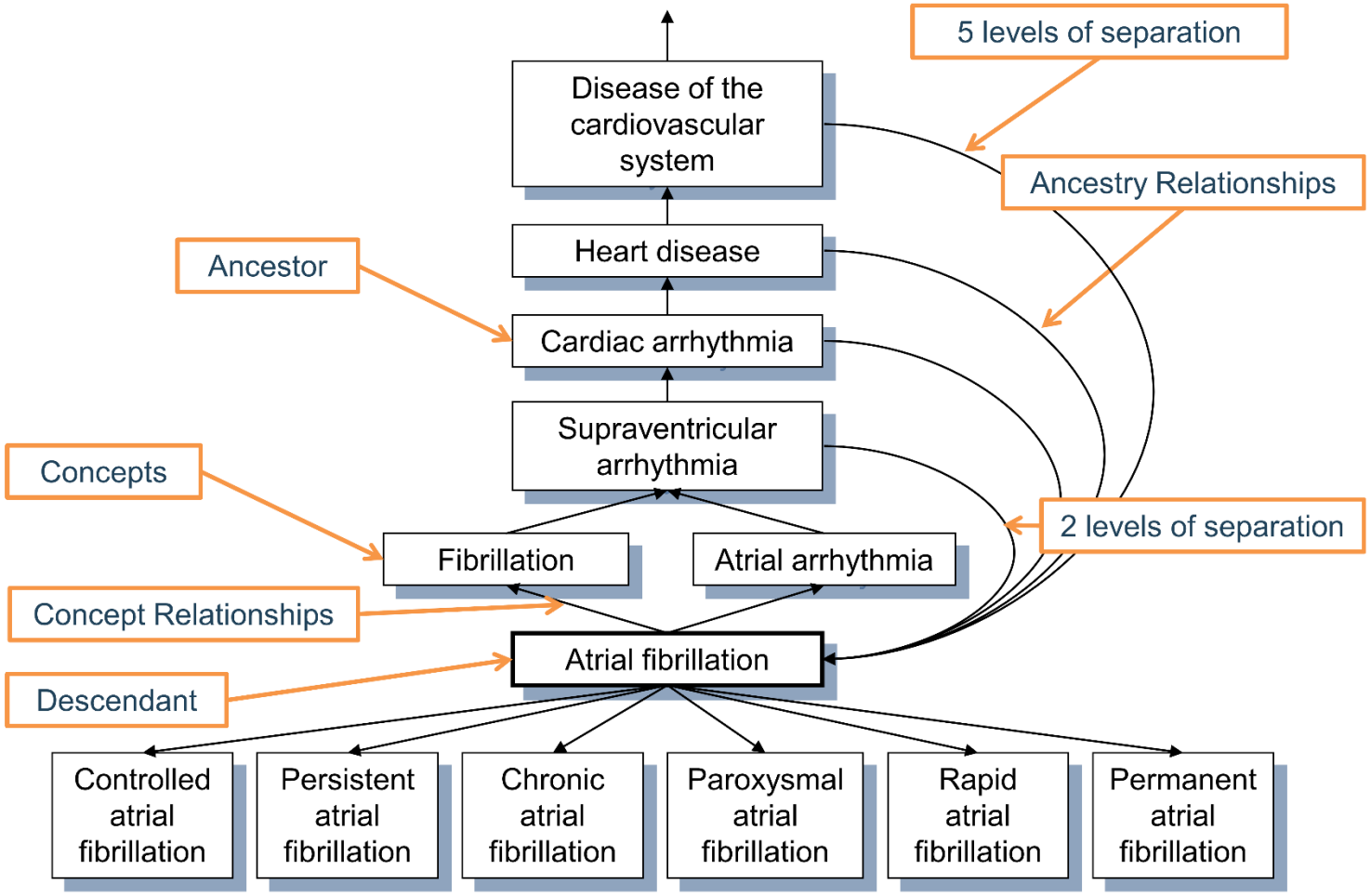
- Foundational to OMOP is a “Concept” (stored in the CONCEPT table)
- Concept domains: “Condition,” “Drug,” “Procedure,” “Visit,” “Device,” “Specimen,” etc.

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	



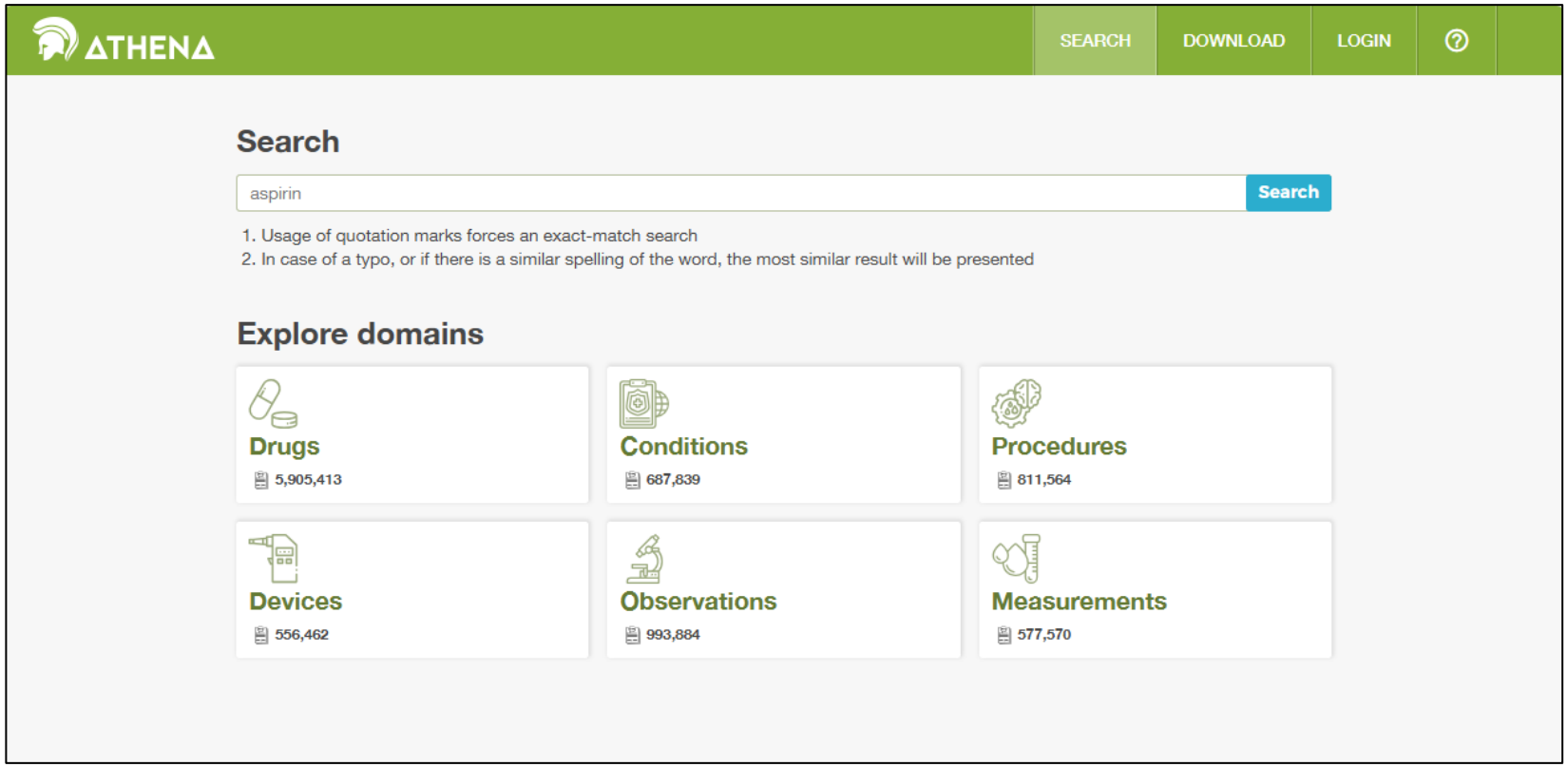
Using the Data: Querying and Retrieving Data (Cont'd)

- Different terminologies are mapped to the Systematized Nomenclature of Medicine (SNOMED), and there is an ontology which defines concept relationships



Exploring the OMOP Mapping

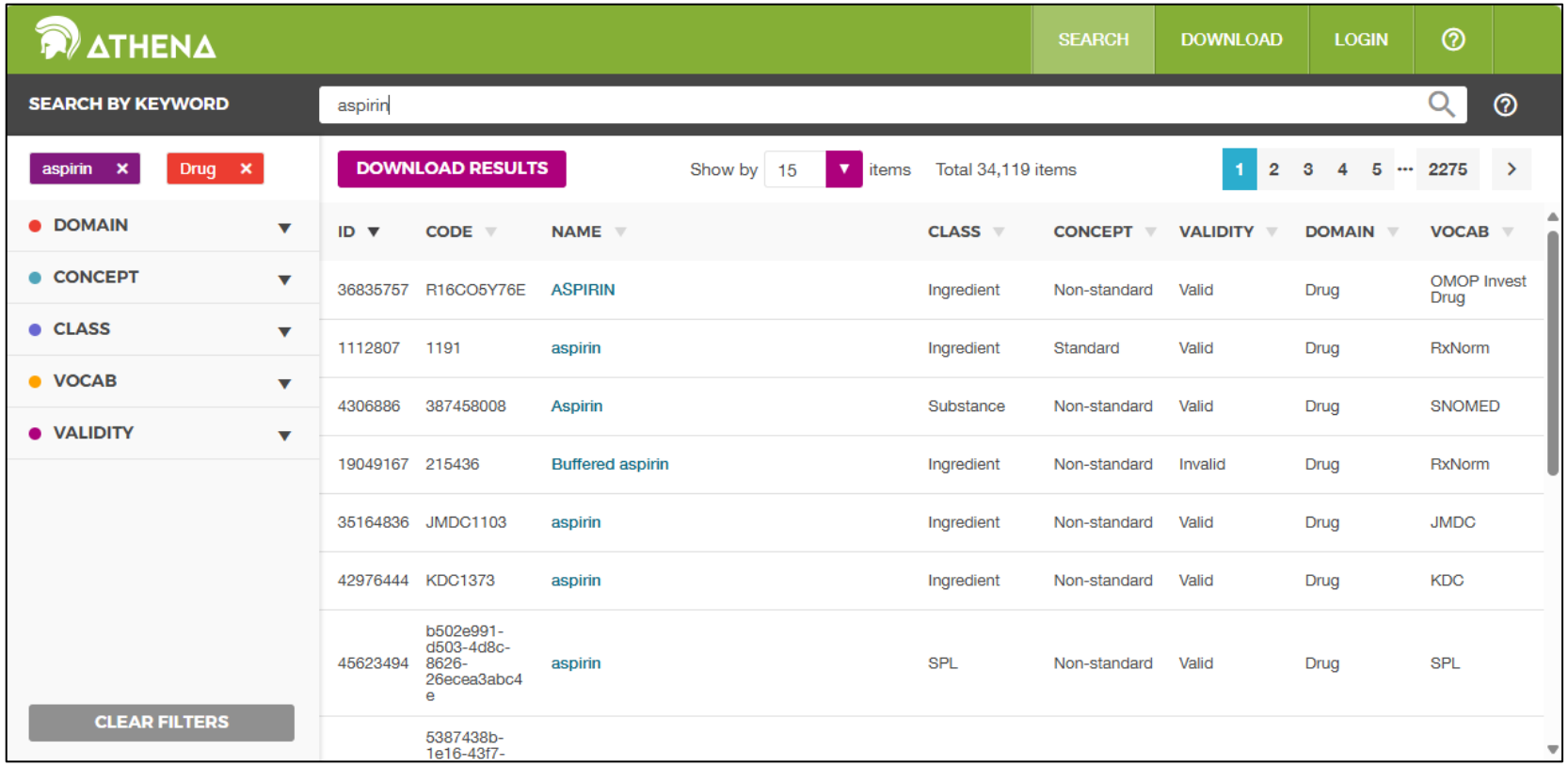
- There's a tool called Athena which contains concept mappings
- Contains standard mappings between several coding systems
- When you're building a list of codes to look for, you can start here to explore the concepts
- OMOP in Mount Sinai uses several of the Athena mappings



https://athena.ohdsi.org/search-terms/start

Exploring the OMOP Mapping

- If we do a search for aspirin, you can see the concepts that correspond (there are 34,119 items)
- Will find different vocabularies that are mapped to concepts – you can build a list of identifiers to search for
- Athena does not contain a mapping between Epic ID codes and other coding systems – due to commercial / license issues with Epic



https://athena.ohdsi.org/search-terms/start

Patients

observation		person	location		death	
Patient A	Race	Patient A	Patient A	Home Address A	Patient A	Death date
	Ethnicity					
	Language Preference					
	Sexual Orientation					
Patient B	Race	Patient B	Patient B	Home Address B		
	Marital Status					
	Gender Identity					
Patient C	Ethnicity	Patient C				
	Religious Affiliation					

Patient Demographics Variables

observation_concept_id	observation_concept_name	category_ count	patient_count	row_count
4148886	Confidential patient data held	7	14,267	14,432
4136468	Ethnic background	45	5,423,669	5,484,600
4271761	Ethnic group	42	81,419	81,419
4110772	Gender identity finding	9	1,047,144	1,047,144
432453	General clinical state finding	2	12,051,324	12,051,324
4181605	Language preference	134	5,922,499	5,922,499
4053609	Marital status	7	10,531,938	10,531,938
4013886	Race	96	7,235,748	7,383,325
3050381	Race or ethnicity	8	8,943,318	8,943,318
4052017	Religious affiliation	31	5,264,641	5,264,641
4283657	Sexual orientation	7	532,407	532,407
21494233	Tabulated ethnicity [CDC]	43	2,969,196	2,990,029
21494232	Tabulated race [CDC]	66	3,943,939	3,986,784

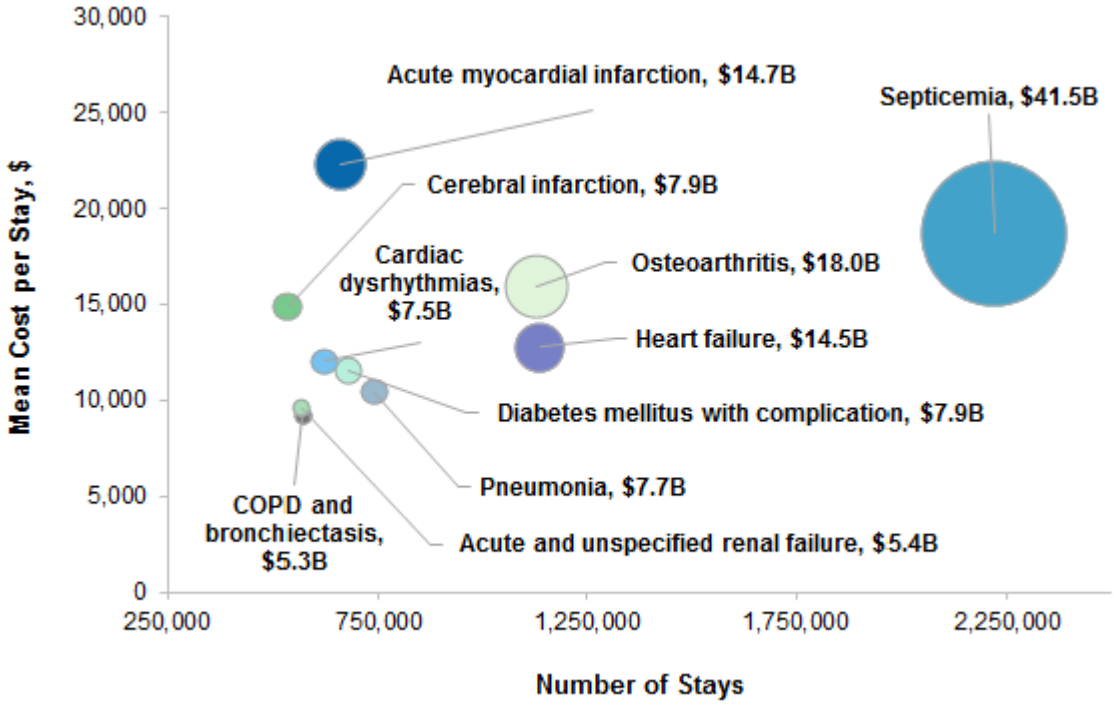
[5]:
```
# Quick smoke test: peek at PERSON table
sql = """
SELECT TOP 5 person_id, gender_concept_id, year_of_birth
FROM CDMDEID.PERSON
"""

airms.conn.sql(sql).collect()
```

Record counts as of April 21, 2025

Epic to OMOP Condition Mapping

Diagnosis Information

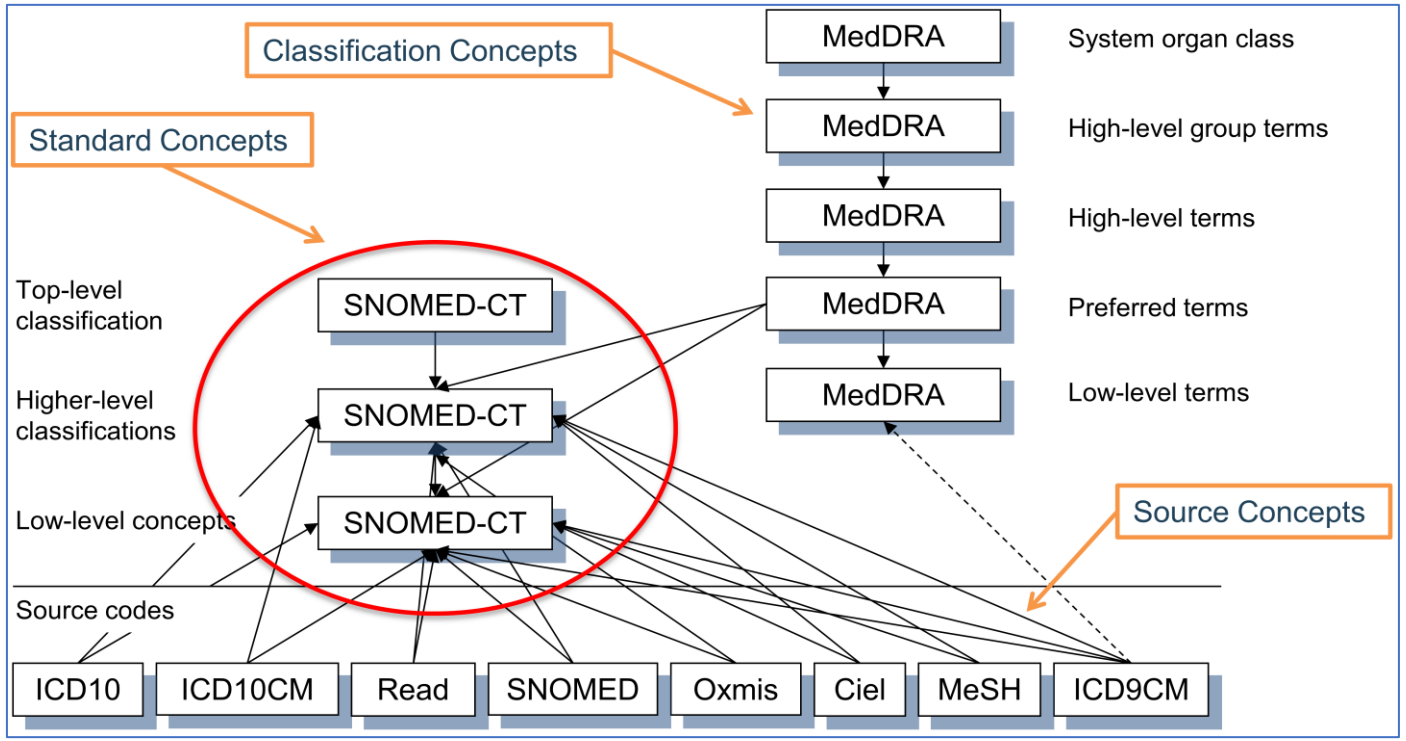
- During a visit (encounter) to the hospital, a patient may be diagnosed with a condition, such as diabetes, pneumonia, etc.
- Typically, diagnoses are stored using standard coding systems such as SNOMED and ICD-10
- These coding systems have been built over years by international consortia (for ICD-10) and are constantly being revised
- **When you look for diagnoses and pull up lists of conditions, you can consult with clinicians (physicians, specialists, etc) who will have a good idea of how frequent conditions are, how they rank, etc. – this is the process of clinically validating your query results**



https://hcup-us.ahrq.gov/reports/statbriefs/sb277-Top-Reasons-Hospital-Stays-2018.jsp

ICD-10 and SNOMED

- ICD-10: International Classification of Diseases, version 10 (developed internationally by the WHO)
- SNOMED: Systematized Nomenclature of Medicine (developed in the USA)
- There are 357,000 unique concepts in SNOMED, ICD-10-CM has about 70,000
- ICD codes are mapped to the low level and higher-level SNOMED-CT codes
- **AIRMS has a custom mapping produced by a 3rd party vendor between Epic IDs and SNOMED / ICD codes**



https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html

Examples – ICD-10

E11: Diabetes mellitus without complications

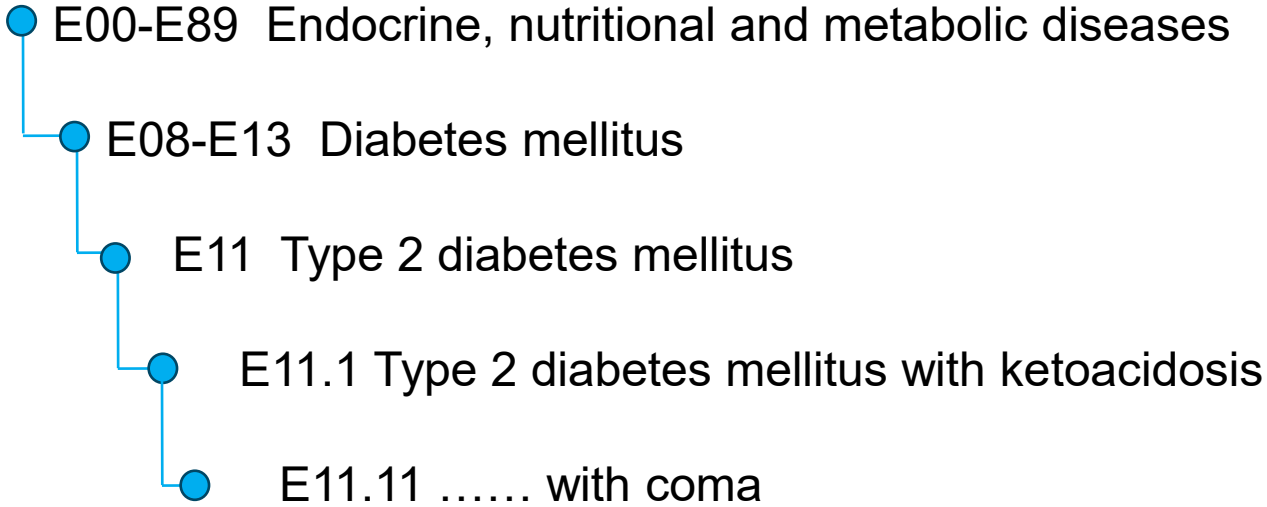
E11.0: Diabetes mellitus with hyperosmolarity

E11.1: Diabetes mellitus with ketoacidosis...

W61.62: Struck by duck

W61.62XA: Struck by duck, initial encounter

W61.62XD: Struck by duck, subsequent encounter...



Using the Data: Querying and Retrieving Data

AIR·MS

```
# Approach A
# Quick scan of atrial fibrillation by name

condition_name = '%atrial fibrillation%'

sql = f"""
SELECT TOP 20
    concept_id,
    concept_name,
    vocabulary_id,
    domain_id,
    standard_concept
FROM CDMDEID.CONCEPT
WHERE LOWER(concept_name) LIKE '{condition_name}'
ORDER BY standard_concept DESC, vocabulary_id, concept_name
"""
airms.conn.sql(sql).collect()
```

```
# Approach B
# ICD10CM I48* AF codes -> SNOMED standard concepts

code = 'I48%'
vocabulary = 'ICD10CM'
domain = 'Condition'
relationship = 'Maps to'

sql_maps_to = f"""
SELECT
    c1.concept_code    AS source_code,
    c1.concept_name    AS source_name,
    c1.vocabulary_id   AS source_vocab,
    c2.concept_id      AS standard_concept_id,
    c2.concept_name    AS standard_name,
    c2.vocabulary_id   AS standard_vocab
FROM CDMDEID.CONCEPT c1
JOIN CDMDEID.CONCEPT_RELATIONSHIP cr
  ON cr.concept_id_1 = c1.concept_id
JOIN CDMDEID.CONCEPT c2
  ON c2.concept_id = cr.concept_id_2
WHERE c1.vocabulary_id = '{vocabulary}'
  AND c1.concept_code LIKE '{code}'
  AND cr.relationship_id = '{relationship}'
  AND c2.standard_concept = 'S'
  AND c2.domain_id = '{domain}'
"""
mapped_af = airms.conn.sql(sql_maps_to).collect()
mapped_af
```

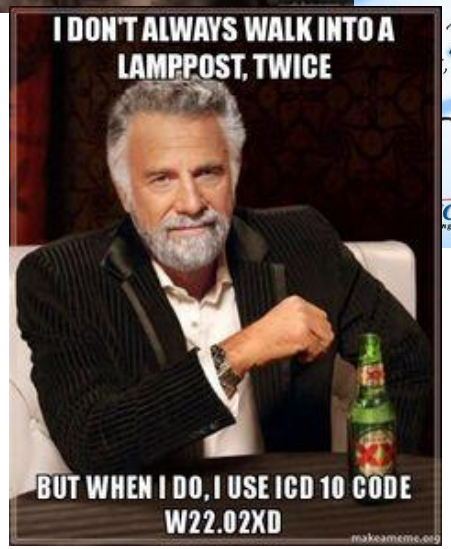
```
# Approach C
# Use mapped SNOMED parents -> expand to descendants
sql_af_conceptset = f"""
WITH mapped_snomed AS (
  {sql_maps_to}
)
SELECT DISTINCT
    ca.descendant_concept_id              AS concept_id,
    c.concept_name                        AS concept_name,
    c.vocabulary_id                       AS vocabulary_id,
    c.domain_id                           AS domain_id,
    c.standard_concept                    AS standard_flag
FROM CDMDEID.CONCEPT_ANCESTOR ca
JOIN CDMDEID.CONCEPT c
  ON c.concept_id = ca.descendant_concept_id
WHERE ca.ancestor_concept_id IN (
  SELECT standard_concept_id FROM mapped_snomed
)
ORDER BY c.vocabulary_id, c.concept_name
"""
af_concepts = airms.conn.sql(sql_af_conceptset).collect()
af_concepts
```

Using the Data: Querying and Retrieving Data

AIR·MS

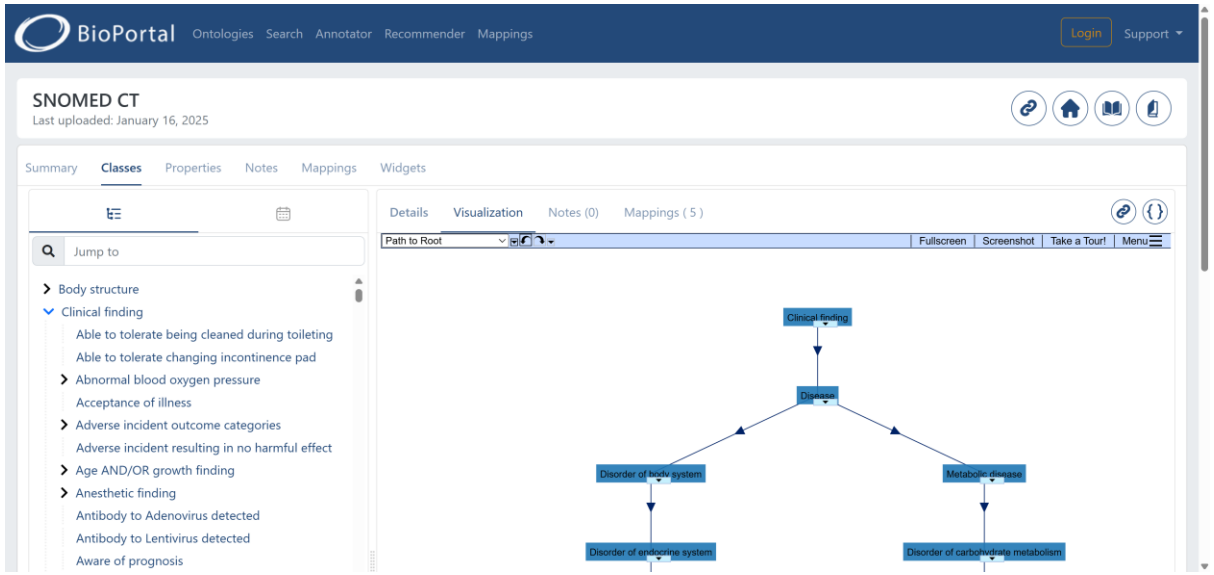
```
# QUALITY CONTROL STEP
# Identify which source concepts (ICD10CM etc.) actually contributed
# to the atrial fibrillation (AF) cohort in CONDITION_OCCURRENCE.
# This list should be reviewed by a clinical collaborator to confirm
# that all included codes are relevant, and that nothing important
# was missed.

sql_sources_for_review = f"""
WITH af_cs AS (
  {sql_af_conceptset}
),
af_hits AS (
  SELECT
      co.condition_source_concept_id
  FROM CDMDEID.CONDITION_OCCURRENCE co
  WHERE co.condition_concept_id IN (SELECT concept_id FROM af_cs)
    AND co.condition_source_concept_id IS NOT NULL
    AND co.condition_source_concept_id LIKE_REGEXPR '^[0-9]+$'
)
SELECT
    c.concept_id                AS source_concept_id,
    c.concept_code              AS source_code,
    c.concept_name              AS source_name,
    c.vocabulary_id             AS source_vocab,
    COUNT(*)                    AS n_occurrences
FROM af_hits h
JOIN CDMDEID.CONCEPT c
  ON c.concept_id = TO_INTEGER(h.condition_source_concept_id)
GROUP BY c.concept_id, c.concept_code, c.concept_name, c.vocabulary_id
ORDER BY n_occurrences DESC, source_vocab, source_code
"""

af_source_review = airms.conn.sql(sql_sources_for_review).collect()
af_source_review
```

Other ICD-10 Examples



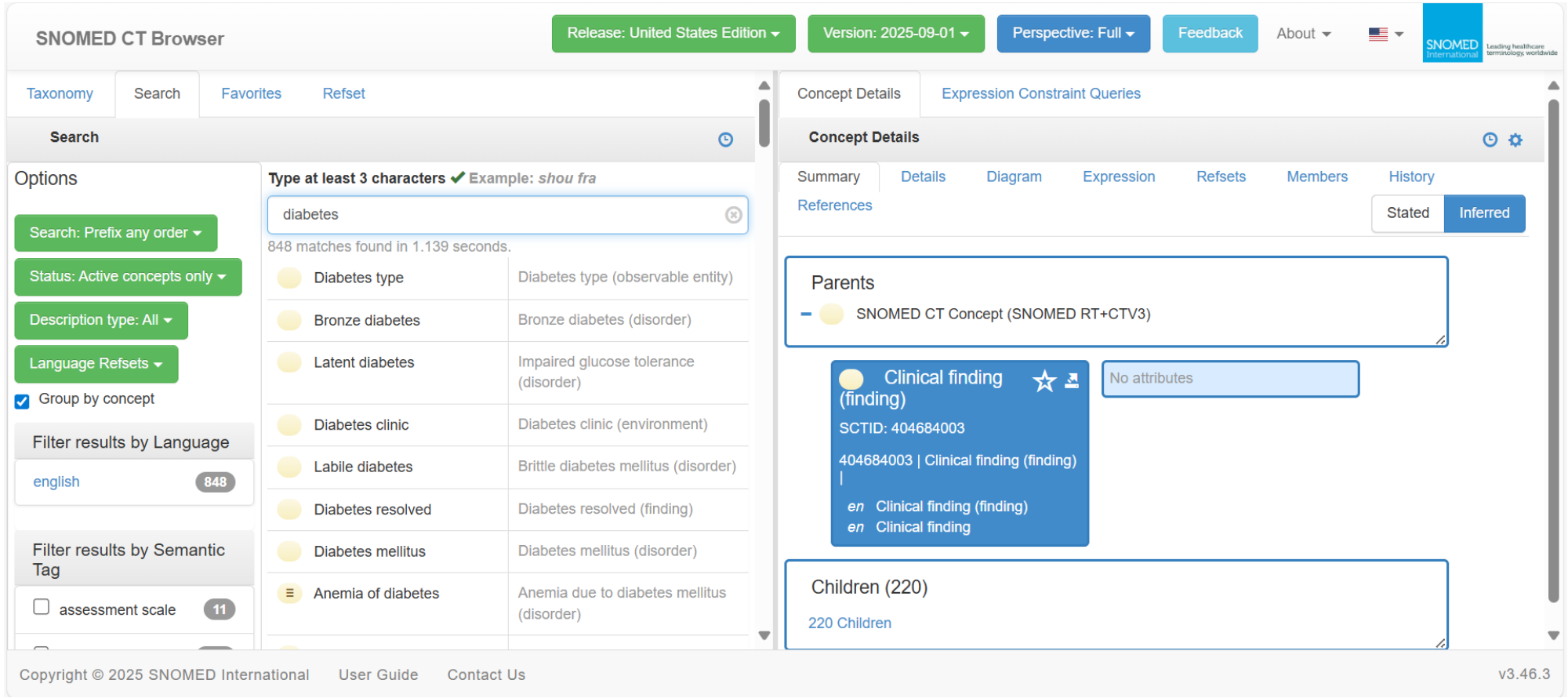
SNOMED CT

- SNOMED CT stands for Systematized Nomenclature of Medicine - Clinical Terms
- This is a comprehensive terminology for standardizing medical terms, including diagnoses, procedures, and other information
- SNOMED CT is organized into a hierarchical structure
- The main top-level categories include "Clinical Findings", "Procedures", "Organisms", etc.
- Relationships are also defined in SNOMED CT, the most common relationship being "is_a"
- Use Bioportal to browse these relationships: http://purl.bioontology.org/ontology/SNOMEDCT/313436004



https://bioportal.bioontology.org/

Examples – SNOMED



https://browser.ihtsdotools.org/

What about Epic IDs?

- Epic has its own internal identification system (Epic ID's) which map to diseases, procedures, medications, etc. **These are not standard codes – for research, we need standardized codes**
- It also has mapping between Epic IDs and SNOMED CT, ICD-10, etc.
- **However, these mapping are commercial, and cannot be used for research purposes, so external consultants and companies have to produce them.**
- This is why you typically won't find a mapping on the Internet of Epic ID to SNOMED, ICD-10, etc.
- AIRMS OMOP contains its own mapping between Epic IDs and SNOMED/ICD, which was produced with the help of a consulting company

Diagnoses

visit_occurrence		condition_occurrence			observation	
Patient A	Outpatient Visit 1				Patient A	Past Medical History F6
						Past Medical History K11
Patient B	Outpatient Visit 2	Patient B	Outpatient Visit 2	Encounter Diagnosis D4	Patient B	Past Medical History D4
				Encounter Diagnosis H8		Past Medical History L12
Patient C	Inpt Hospitalization 3	Patient C	Inpt Hospitalization 3	Encounter Diagnosis A1		
				Encounter Diagnosis B2		
				Hospital Problem C3		
				Hospital Problem D4		
				Problem List E5		
Patient D	Inpt Hospitalization 4	Patient D	Inpt Hospitalization 4	Encounter Diagnosis F6		
				Billing Diagnosis G7		
		Patient E		Problem List I9		
				Problem List J10		

Diagnosis Record Types

xtn_condition_type_source_concept_id	xtn_condition_type_source_concept_name	row_count
2000000108	Billing Diagnosis	73,040,692
2000000129	Encounter Diagnosis	120,323,291
2000000122	Hospital Problem	3,837,039
2000000120	Problem List	13,032,738

Record counts as of April 21, 2025

Using the Data: Querying and Retrieving Data

AIR·MS

```
# Build AF index dates: earliest AF diagnosis per person
sql_af_index = f"""
WITH af_concepts AS (
  {sql_af_conceptset}
)
SELECT
    co.person_id,
    MIN(co.condition_start_date) AS af_index_date
FROM CDMDEID.CONDITION_OCCURRENCE co
WHERE co.condition_concept_id IN (SELECT concept_id FROM af_concepts)
GROUP BY co.person_id
"""

af_index = airms.conn.sql(sql_af_index).collect()
af_index.head(10)
```

```
# How many patients have AF?
print("AF cohort size:", len(af_index))

# Make a year column for visualization
af_index["year"] = pd.to_datetime(af_index["AF_INDEX_DATE"]).dt.year

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 5))
ax = af_index["year"].value_counts().sort_index().plot(
    kind="bar",
    color="#1f77b4",
    edgecolor="black"
)
plt.title("Number of patients with first AF diagnosis per year", fontsize=14, pad=15)
plt.xlabel("Year of AF index", fontsize=12)
plt.ylabel("Patient count", fontsize=12)
plt.xticks(rotation=45, ha="right")
plt.grid(axis="y", linestyle="--", alpha=0.7)
plt.tight_layout()
plt.show()
```

How Do You Pick the Right Codes?

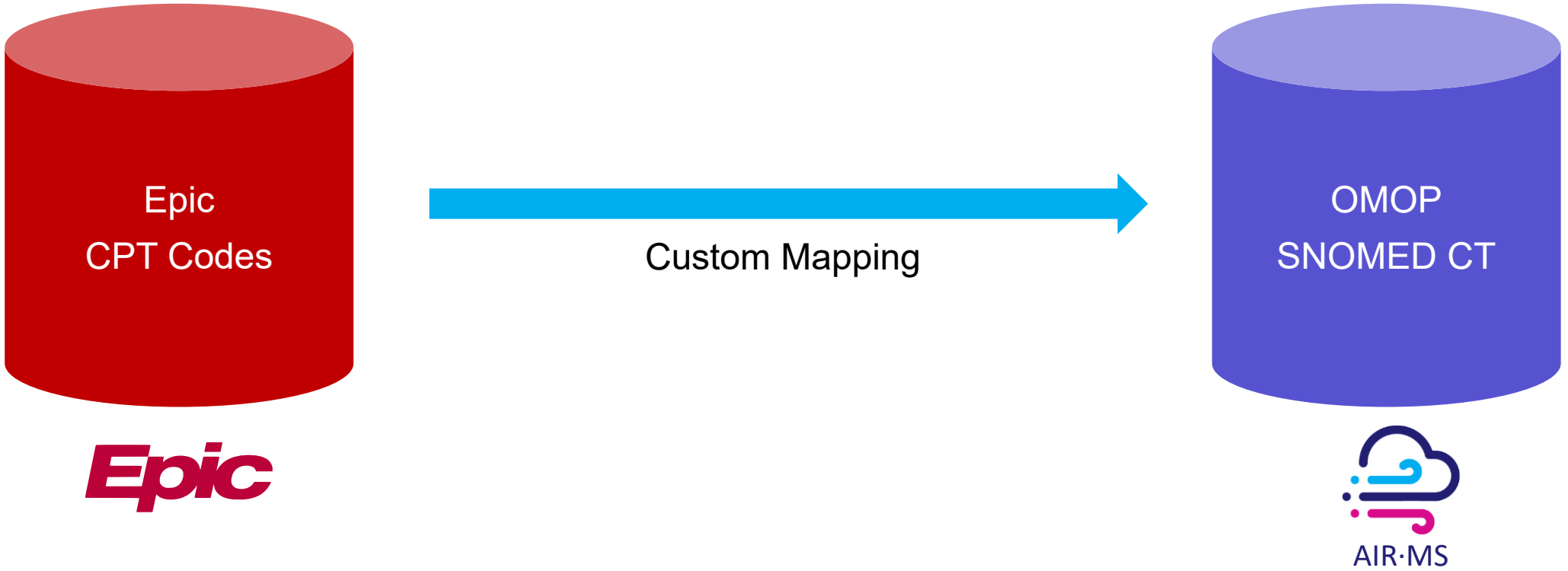
- It's tempting to do a keyword search for a condition, but this is not correct

- For example, if you query for "stroke" you'll miss "cerebral ischemia" (2025 ICD-10-CM Diagnosis Code I67.82)

- **As a result: look to existing published studies, or consult a clinician when building a list of conditions to look for**

- **Clinicians can also help you validate your query results: if you query for a few diseases among a specific demographic (ex. >65 year old males), certain diseases should come up as frequent**

Epic to OMOP Procedure Mapping



Procedure Data: CPT

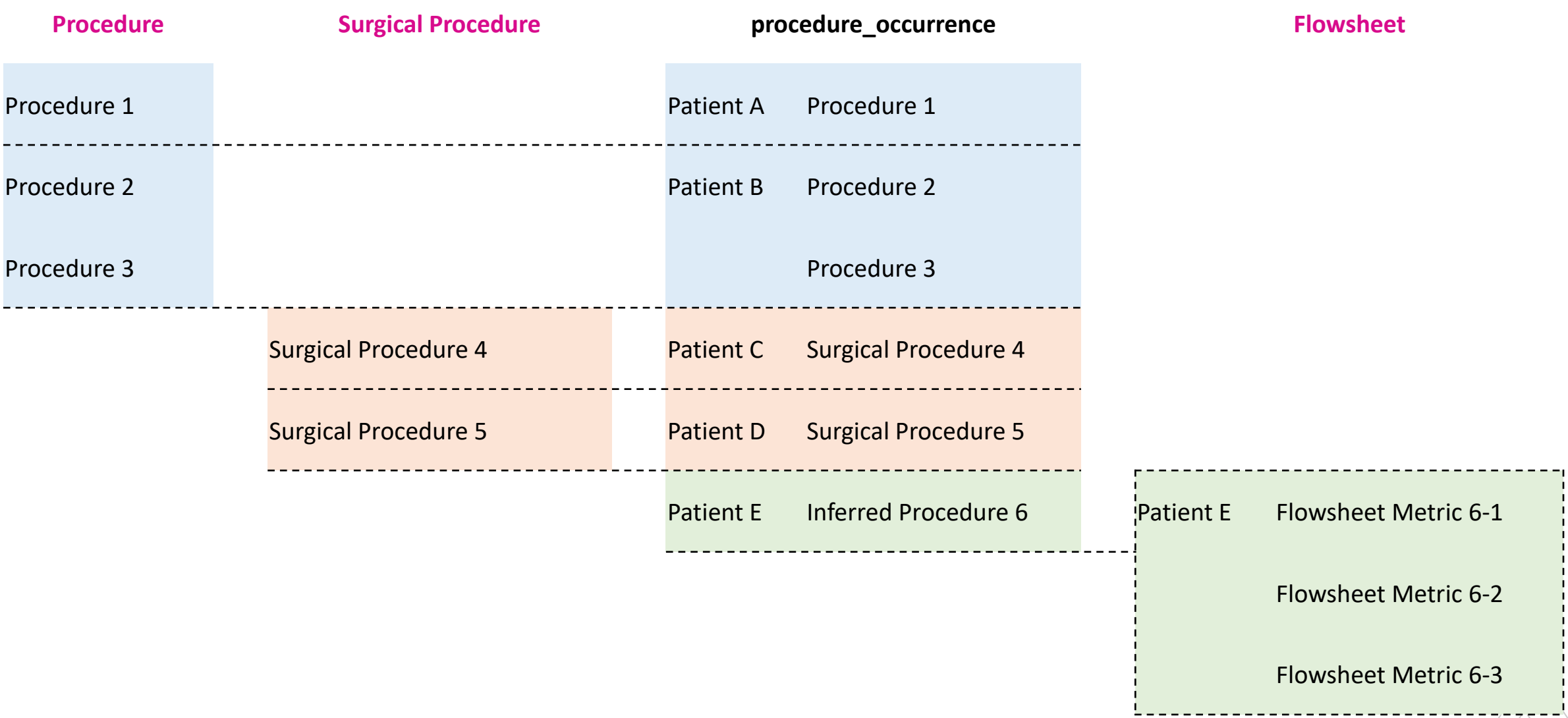
- CPT stands for Current Procedural Terminology
- A commercial code set of codes maintained by the American Medical Association (AMA), requires a license to use
- These codes have played an important role in medical billing, documentation, and reporting
- Currently there are > 11,000 CPT codes
- Commercial – will need subscription to tools like Codify
- Mount Sinai has a license, which the AIRMS data uses – limitation on distributing mapping

Code Range	Category
00100-01999	Anesthesia
95700-95811	Sleep Medicine Testing Procedures
10004-69990	Surgery
70010-79999	Radiology Procedures
80047-89398	Pathology and Laboratory Procedures
90281-99607	Medicine Services and Procedures
98000-99499	Evaluation and Management
0001F-9007F	Category II Codes
0002M-0020M	Multianalyte Assay
0042T-0987T	Category III Codes

Category II Codes – supplemental performance tracking codes (4 digits with "F" suffix)

Category III Codes – temporary codes for emerging technologies, etc. (4 digits with "T" suffix)

Procedures


Procedure
Surgical Procedure
procedure_occurrence
Flowsheet
Procedure 1
Procedure 2
Procedure 3
Surgical Procedure 4
Surgical Procedure 5
Patient A Procedure 1
Patient B Procedure 2
Procedure 3
Patient C Surgical Procedure 4
Patient D Surgical Procedure 5
Patient E Inferred Procedure 6
Patient E Flowsheet Metric 6-1
Flowsheet Metric 6-2
Flowsheet Metric 6-3

Procedure Record Types

xtn_procedure_type_source_concept_id	xtn_procedure_type_source_concept_name	row_count
2000000111	General Procedure	326,372,912
2002067235	Procedure Inferred from Flowsheet	281,426
2000000097	Surgical Procedure	962,017

Record counts as of April 21, 2025

```
# Let's say we are satisfied with all codes identified above. Let's see if we have additional sub-codes to extract
sql_pm_conceptset = f"""
WITH mapped_std AS (
  {sql_pm_recon}
)
SELECT DISTINCT
    ca.descendant_concept_id   AS concept_id,
    c.concept_name             AS concept_name,
    c.vocabulary_id            AS vocabulary_id,
    c.domain_id                AS domain_id,
    c.standard_concept         AS standard_flag
FROM CDMDEID.CONCEPT_ANCESTOR ca
JOIN CDMDEID.CONCEPT c
  ON c.concept_id = ca.descendant_concept_id
WHERE ca.ancestor_concept_id IN (SELECT concept_id FROM mapped_std)
ORDER BY c.vocabulary_id, c.concept_name
"""
pm_concepts = airms.conn.sql(sql_pm_conceptset).collect()
pm_concepts
```

```
# Approach A: text search
name = '%pacemaker%'
domain = 'Procedure'

sql_pm_recon = f"""
SELECT TOP 20
    concept_id,
    concept_name,
    vocabulary_id,
    domain_id,
    standard_concept
FROM CDMDEID.CONCEPT
WHERE domain_id = '{domain}'
  AND LOWER(concept_name) LIKE '{name}'
ORDER BY standard_concept DESC, vocabulary_id, concept_name
"""
airms.conn.sql(sql_pm_recon).collect()
```