Introduction to AIR·MS

Health Data Literacy

Eugenia Alleva
Andrew Deonarine

Sept 30th, 2025

HPI·MS
Mount Sinai
Hasso Plattner Institute for Digital Health at Mount Sinai

Overview

1. What is AIR•MS?
2. Mount Sinai Health System & Epic EHR
3. AIR•MS & OMOP
4. Coding Systems & Concept Mapping
5. Content Standardization via Concept Mapping
6. Data Contents of Mount Sinai's OMOP Research Data Repository
7. Putting It Together: Queries & Phenotypes

https://labs.icahn.mssm.edu/airms/

What is AIR·MS?

Artificial Intelligence-Ready Mount Sinai (AIR·MS) is a research platform that is composed of:

1) A (very fast) integrated database including Mount Sinai Data Warehouse (MSDW), Pathology and Radiology metadata; and the included data is growing

2) A Research Environment that allows interactions with the AIR·MS database from Python or R

3) An Application Tier to host a growing number of applications – including cohort building tools and annotation apps

Data Modalities in AIR·MS

AIR·MS

Currently Available Modalities:

- Mount Sinai Data Warehouse (MSDW), both containing protected health information (PHI) and DeID (deidentified) Observational Medical Outcomes Partnership (OMOP)-mapped electronic health record (EHR)
- Pathology Metadata
- Radiology Metadata
- BioMe/Sinai Million
- electrocardiogram (EKG)
- Echocardiography
- GI Research Database

Work in progress: Electroencephalogram (EEG), Endoscopy & Colonoscopy Reports

All modalities are stored in separate database schemas, and access is granted to each schema individually based on Institutional Review Board (IRB)

Request Access to AIR·MS in Sailpoint

1) Obtain Institutional Review Board (IRB) approval for your project if you want to access PHI data, including this indication "*we will use the AIR·MS platform (IRB # 20-01288) to access and store our data*"

2) Request Minerva/High Performance Computing (HPC) account

3) Request access to specific modalities (i.e. schemas) on SailPoint

4) After access approval you can get started with the data using our Getting Started Guides:

```
git clone https://github.mountsinai.org/AIRMS/airms-researcher-tutorials-minerva.git
```

Working with AIR·MS in Python

Connection: The airms_connect library enables easy connection to AIR·MS and is pre-installed on Minerva; install it separately when using conda/venv.

On Minerva vs Local: From a Minerva compute node, first establish an SSH tunnel via .on_minerva().

Interacting with HANA: The airms_connection() class exposes a conn attribute, a hana_ml.ConnectionContext object, enabling direct use of hana-ml.

- conn.sql() executes raw SQL queries.
- conn.table() accesses specific tables.

Both return a hana-ml DataFrame, which mimics pandas syntax and supports in-database operations. To pull data locally, use .collect() to return a pandas DataFrame.

```python
# import
from airms_connect.connection import airms_connection

# initialize connection
airms = airms_connection()

# establish tunnel on Minerva
login_host_name='li04e04'
airms.on_minerva(login_host_name=login_host_name)

# establish connection
airms.connect()

# query airms
print(airms.conn.table('PERSON', 'CDMDEID').head(10).collect())
# or
query = "SELECT TOP 10 * FROM CDMDEID.PERSON"
print(airms.conn.sql(query).collect())
```

Working with AIR·MS in R

Install Packages

To use AIR·MS with R, you need to install the following packages in your environment:

- install.packages("odbc")
- install.packages("RJDBC")

Download Java driver

You then need to point to the Java driver in your R script.

- On Minerva the driver can be found at /sc/arion/projects/airms/lib/ngdbc.jar

```
library(odbc)
library(RJDBC)
library(getPass)
driver_path <- path/to/driver
# Initialize the driver
jdbcDriver <- JDBC(driverClass="com.sap.db.jdbc.Driver", classPath=driver_path)
# Define connection context
hana_url <- "db.airms.mssm.edu:30041”
database_name <- "AIRMS”
username <- ”user-name”
connection_string <- sprintf( "jdbc:sap://%s/?databaseName=%s&user=%s
            &encrypt=TRUE&validateCertificate=FALSE&sslHostnameInCertificate=%s
            &connectTimeout=0&sslTrustStore=None",
            hana_url, database_name, username, hana_url)
# Connect
conn <- dbConnect(jdbcDriver, connection_string, username,
            password = getPass("Enter your password: "))
result <- dbGetQuery(conn, "SELECT TOP 10 * FROM CDMDEID.PERSON")
```

Working with AIR·MS in R vs R Studio on Minerva

AIR·MS

Working with R

Launch an interactive R session:

```
$ bsub -q interactive -P your_account -n 1 -W 1:00 -R rusage[mem=8000] -XF -Is /bin/bash
```

Load R module and start R

```
$ ml R/4.2.0
$ R
```

Download with R Studio

Launch R studio via `$ minerva-rstudio-web-airms.sh`

Rstudio is started on compute node lh06c28, port 8788

Access the RStudio Web using your web browser: http://10.95.46.94:53543 ...

Change the hana_url from `hana_url <- "db.airms.mssm.edu:30041”` to

```
hana_url <- paste("localhost",Sys.getenv("db_port"),sep=":")
```

Querying AIR·MS - SQL

We will give an introduction to SQL in Session II!

Most basic components:

- **Schema**: Is like a folder – each stores a collection of tables that can be connected to each other via keys (e.g. patient mrns)
- **Clause**: control the structure of a query

```
SELECT * FROM CDMDEID.PERSON
       WHERE YEAR_OF_BIRTH > 1988
```

- **Functions**: perform calculations (AVG, COUNT, ROUND, MIN, MAX, …)

```
SELECT AVG(2025-YEAR_OF_BIRTH) FROM CDMDEID.PERSON
       WHERE YEAR_OF_BIRTH > 1988

GROUP BY GENDER
```

-

Mount Sinai Health System & Epic Electronic Health Record (EHR)

The Health Data Ecosystem

AIR·MS

- The Mount Sinai Hospital was founded in 1852, one of the oldest and largest teaching hospitals in the United States
- The Mount Sinai Health System is one of the largest health systems in North America, and there are over 12 million patients in the MSHS information system

Mount Sinai

48,000 Employees

>12 Million Patients

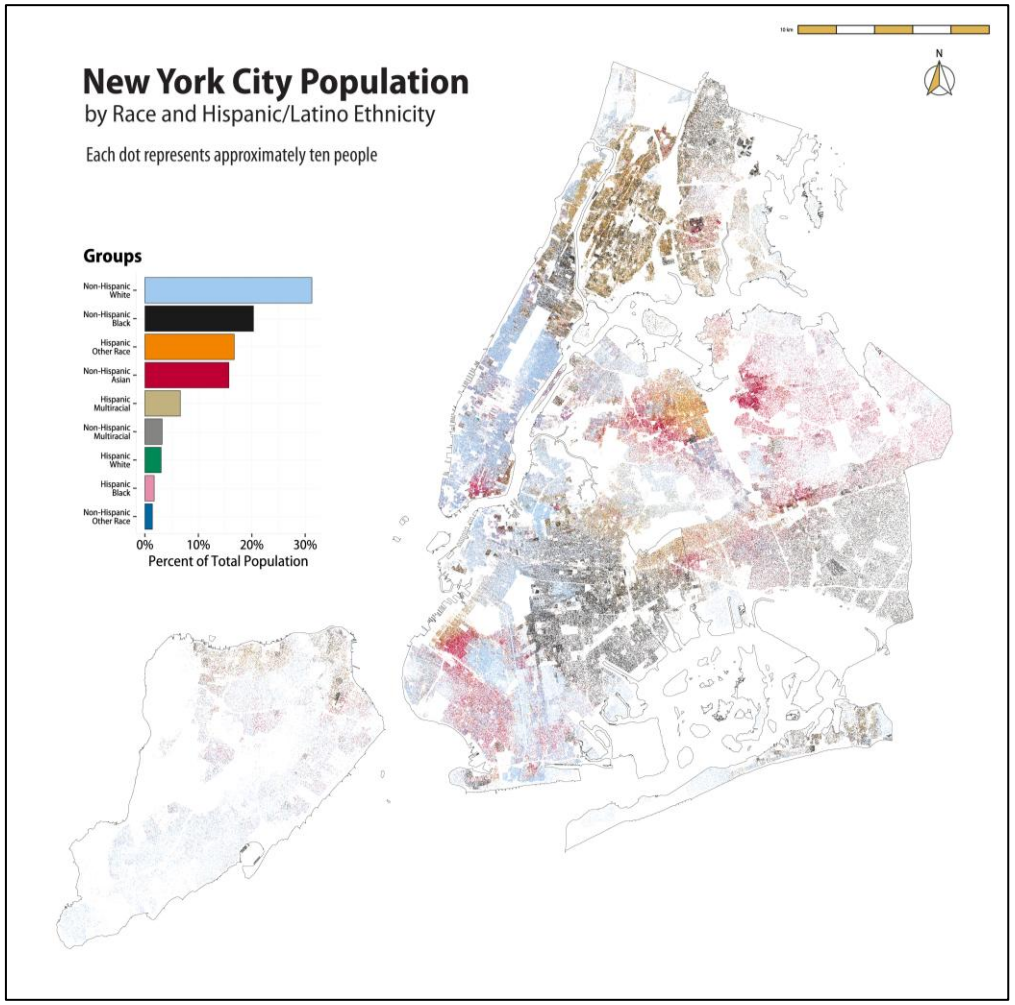
3,221 Beds

48 Institutes

7 Hospitals

New York City

- New York is one of the largest, most diverse cities in the world
- This is reflected in Mount Sinai's patient population as well
- Mount Sinai services the New York City area
- Models and research developed here can be applied to many other locations



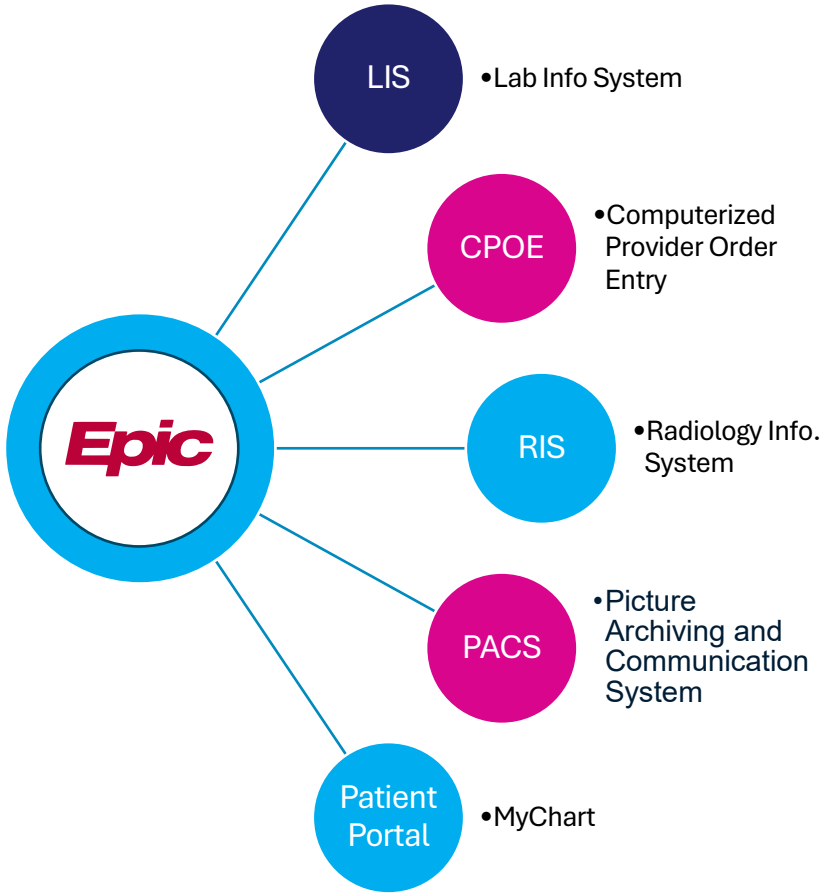
The Health Data Ecosystem: Epic

In MSHS, there are software systems we use:

- Electronic health record system (EHR) or electronic medical record system(EMR); biggest systems are Epic and PowerChart (Oracle Cerner)
- Picture Archiving and Communication System (PACS) – or storing imaging datalike CTs and MRIs
- Patient Portals (like MyChart)

- At the state level – there are HIEs (health information exchanges)

- Epic is used in the MSHS



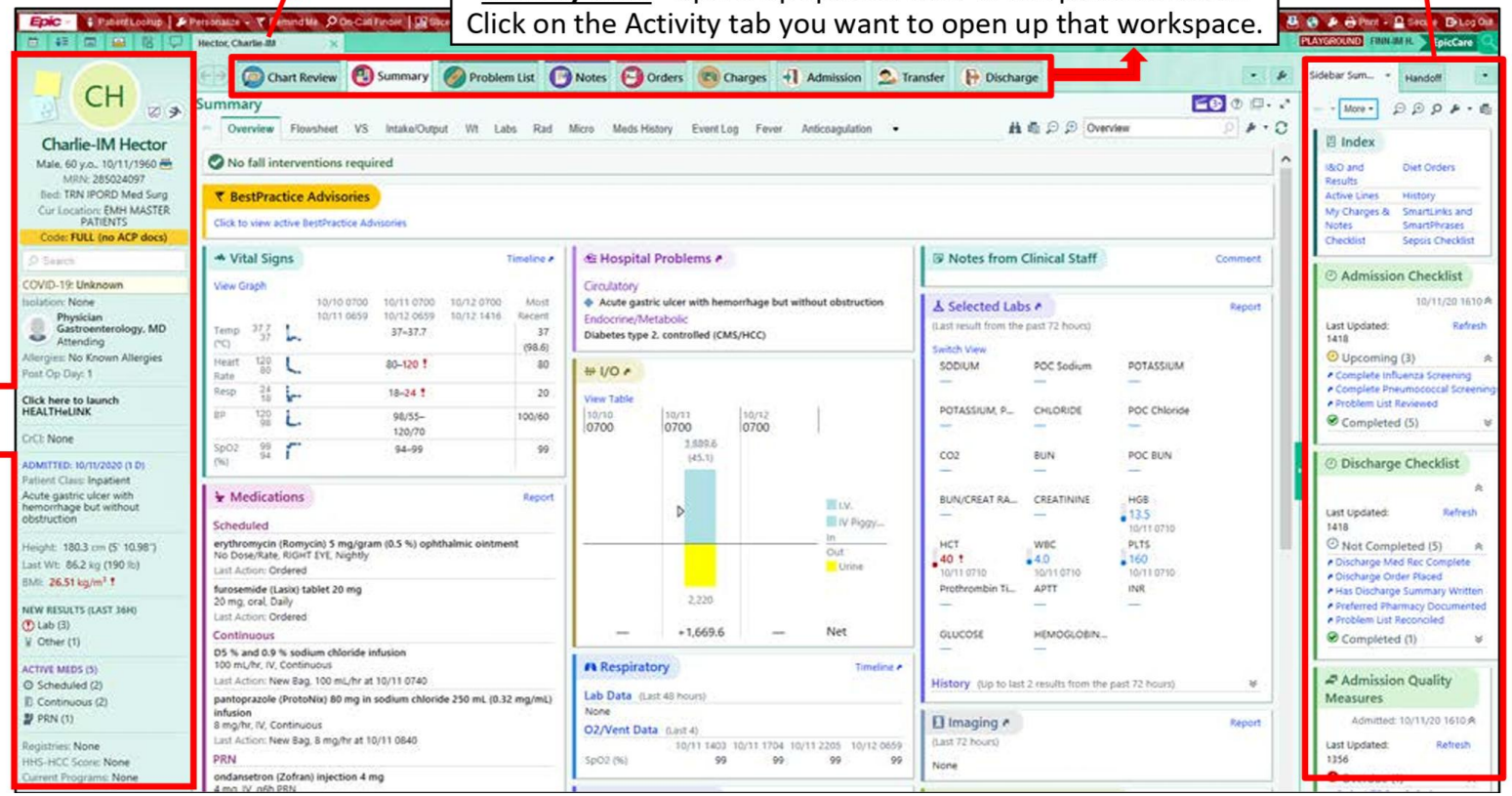
The Health Data Ecosystem: Epic

Patient Name

Sidebar

Activity Tabs - opens up specific view of the patients chart. Click on the Activity tab you want to open up that workspace.

Story Board
This displays useful patient information and hyperlinks that open up another part of the chart when you click on it. Hover over something and it opens up a window to discover more information.



A typical physician view in Epic, showing summary patient information

The Health Data Ecosystem: Epic

AIR·MS

Epic | Patient Lookup | ED Track Board | OR/Procedural Areas | Calculator | My Reports | On-Call Finder | Print | Log Out

Johnson, Ellie | Epicare SIDNEY A. | Search

Summary | Chart Review | Work List | MAR | Flowsheets | Notes | Education | Care Plan | Orders | Charges | Navigators

Ellie Johnson
Female, 11 y.o., 12/10/2007
MRN: 299038
Bed: 01
Cur Location: TRN Master Patient Pool Room
Code: FULL (no ACP docs)

Search

Isolation: None

ALLERGIES
No Known Allergies

Collection: Lab

ADMITTED: 9/12/2019 (1 D)
Appendicitis, acute
Martin Stitch, MD
Attending

Height: 144.8 cm
Last Wt: 44.5 kg
BMI: 21.21 kg/m²

ACKNOWLEDGE ORDERS (7)

NEW RESULTS, LAST 36H

ACTIVE MEDS (8)
Scheduled (4)
Continuous (1)
PRN (2)
One-Step (1)

MAR | Report | MAR Note | Messages | Legend | Show All Actions | Link Lines

ALL | Scheduled | PRN | Continuous | Respiratory | Due/Overdue Meds | Override Pulls | Chemo

Go to Now | or | Select Date: | Show All Details | Hide All Admins

Friday September 13, 2019

0800 | 0900 | 1000 | 1100 | 1200 | 1300 | 1400 | 1500

Ordered Admin Amount: 5 mL | Last Admin: Today 09/13/19 at 0130 (Given) | Dispense Location: Pharmacy to load in ADS

cetirizine (ZyrTEC) tablet 10 mg : Dose 10 mg : Oral : Daily
1230 Due
Ordered Admin Amount: 1 tablet (1 × 10 mg tablet) | Dispense Location: Pharmacy to load in ADS

dextrose 5 % and sodium chloride 0.45 % with KCl 20 mEq/L infusion : Dose 75 mL/hr : 75 mL/hr : Intravenous : Continuous
0931 Stopped
Last Admin: Today 09/13/19 at 0515 (Rate/Dose Verify) | Dispense Location: EMH Central Pharmacy

insulin glargine (Lantus) 100 UNIT/ML injection 15 Units : Dose 15 Units : Subcutaneous : 4 times daily before meals and nightly
1130 Given 15 Units
Ordered Admin Amount: 15 Units | Last Admin: Today 09/13/19 at 1130 (Given) | Dispense Location: EMH Central Pharmacy

insulin regular (HumuLIN R,NovoLIN R) 100 UNIT/ML injection 2-10 Units : Dose 2-10 Units : Subcutaneous : 4 times daily with meals and nightly
0800 Not Given | 1200 Due
Ordered Admin Amount: 2-10 Units | Last Admin: Yesterday 09/12/19 at 2100 (Given) | Dispense Location: EMH Central Pharmacy

Nurses see different views than physicians – MAR (medication administration record) https://www.youtube.com/watch?v=XTstvvjyrcE

Health Information Privacy Regulations

AIR·MS

State

NYHIPA

SHIELD

Public Health 27F

...

National

HIPAA

GINA

NIST 800-171

...

International

GDPR (EU/UK)

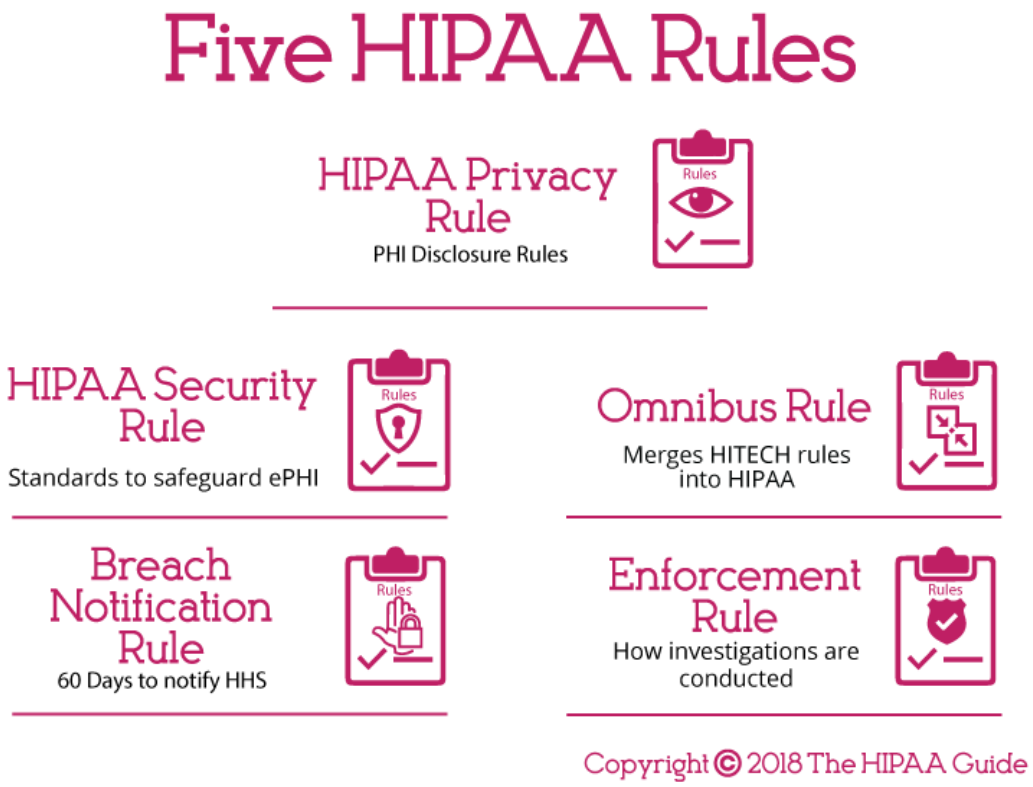
PIPEDA (Canada)

HIPAA (USA)

...

Health Regulations: HIPAA

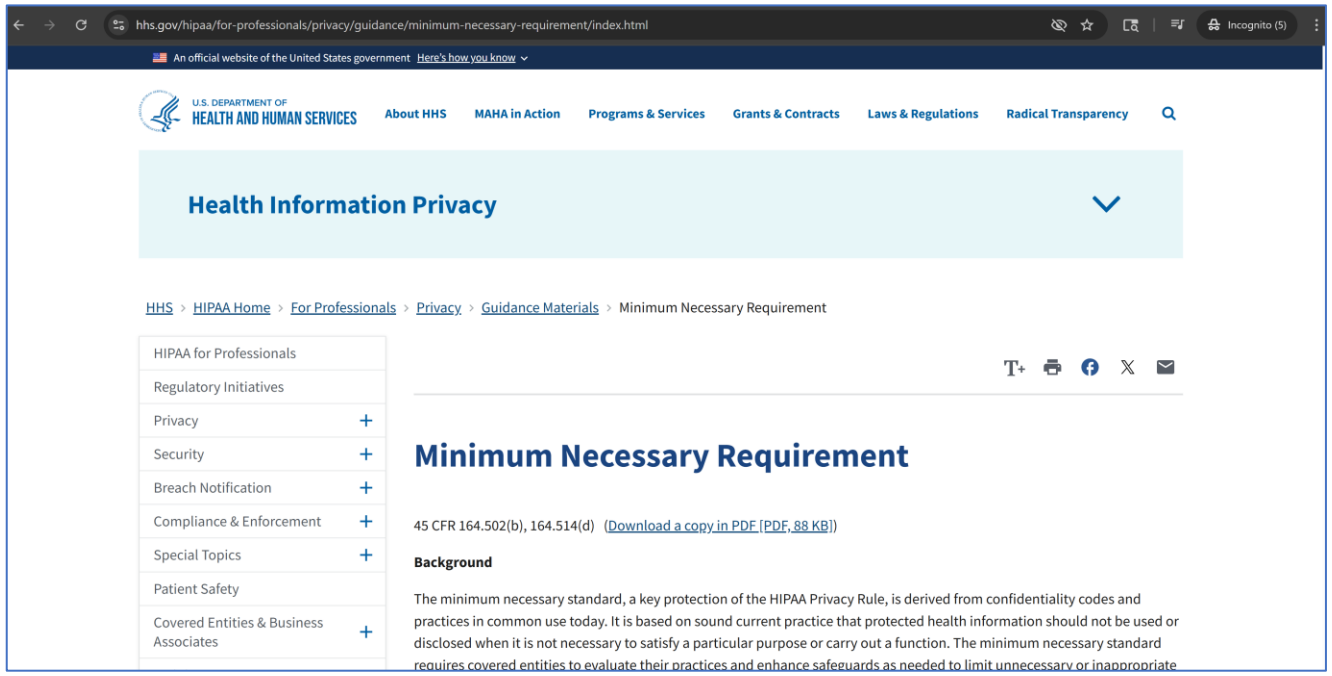
- Epic contains several modules and tens of thousands of data fields which capture data from doctors, nurses, and other health professionals
- Health information in the United States is governed by HIPAA regulations
- PHI = private health information
- Health information can be "de-identified" by review by an expert ("expert determination") or using "safe harbor" (removing 18 specific identifiers)



https://www.hipaaguide.net/hipaa-for-dummies/

Health Regulations: HIPAA

- The "Minimum Necessary Requirement" states that researchers should only use the minimal amount of data necessary to address a research question or use case
- This means making sure data is only provided for well-defined inclusion / exclusion criteria for a specific question
- Exploring broad datasets usually happens later in a study after initial limited dataset delivery
- If there's an incident or data is leaked – reduces the number of people impacted



https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html

The Patient Journey & Epic

- Let's take a look at a typical patient walkthrough with Epic
- Key workflows: admission, discharge, and transfer (ADT)

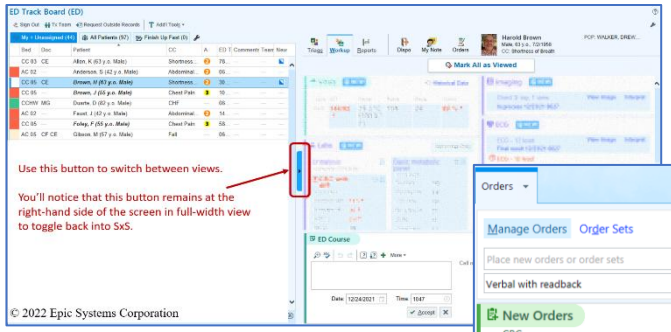
Admission	Orders	Document	Discharge	Pop. Health
Admit new patient, or register a new visit to hospital	*Physician can request medications, tests, etc.*	*SOAP Notes, Consults, Transfer Notes, etc.*	*Orders, discharge summaries, etc.*	*Slicer Dicer analysis*

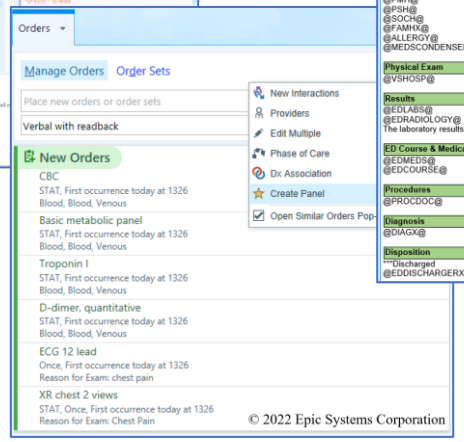
The Patient Journey & Epic

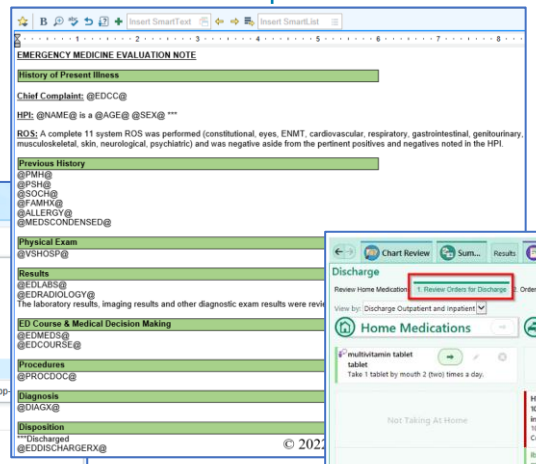
AIR·MS

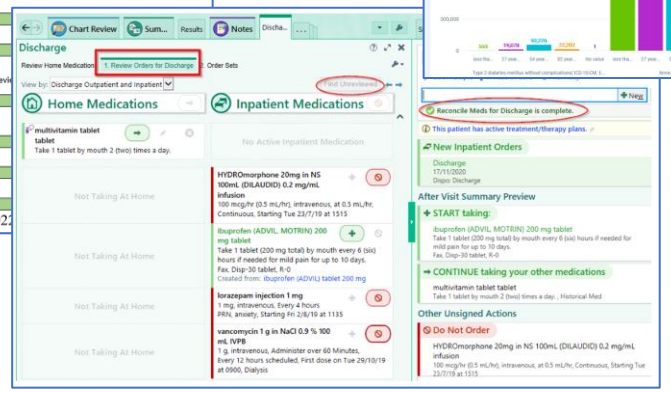
Epic

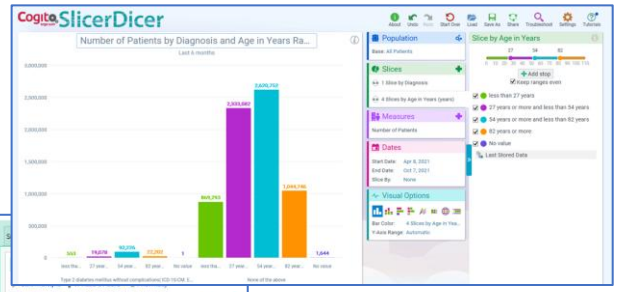
Admission → Orders → Document → Discharge → Pop. Health







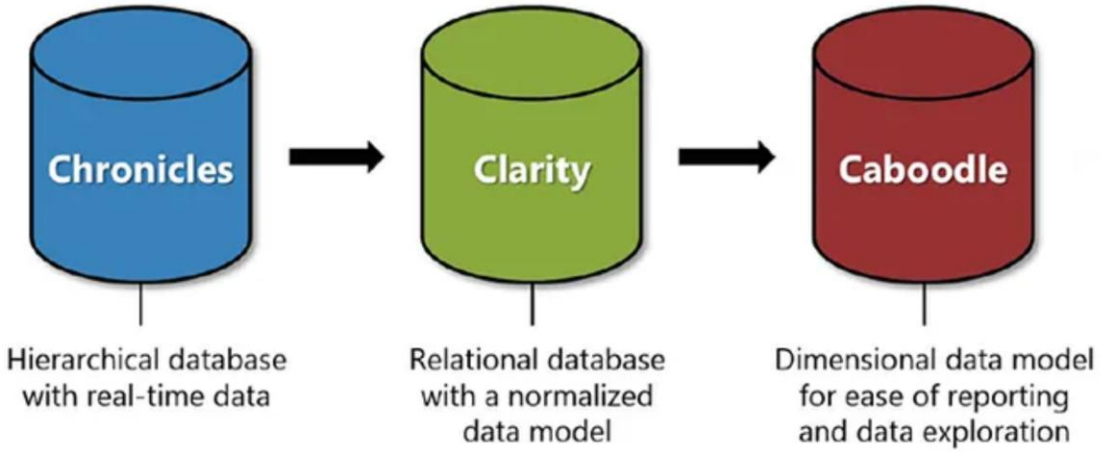




- Margaret Smith is a 75 year old female who has arrived at the emergency room today with dyspnea (shortness of breath) and angina (chest pain)
- Patient has a previous history of pneumonia, anxiety

The Health Data Ecosystem

- Once we have all this data, it then goes into a series of databases in Epic
- Epic has multiple databases that store information
- SQL = structured query language – Clarity, Caboodle
- NOSQL = other types of databases (ex. hierarchical, graph)- Chronicles



https://it.uclahealth.org/about/ohia/services/reports-dashboards/webi-dashboards

AIR•MS & OMOP

Using the Data: Introduction to OMOP

- Data is stored in “OMOP” format
- OMOP = Observational Medical Outcomes Partnership
- After the initial work completed in 2013, the OMOP group then became the Observational Health Data Sciences and Informatics (OHDSI) group (pronounced “Odessey”)
- The common data model (CDM) for OHDSI is OMOP
- A significant project is the OMOP vocabulary
- In the US – OMOP leadership is located at Columbia U.
- Visit https://ohdsi.org
- Mount Sinai's OMOP data is maintained by Dr. Timothy Quinn (Chief Data Architect)

Using the Data: Who Uses OMOP?

AIR·MS

- Several large collaborations use OMOP:

National COVID Cohort Collaborative

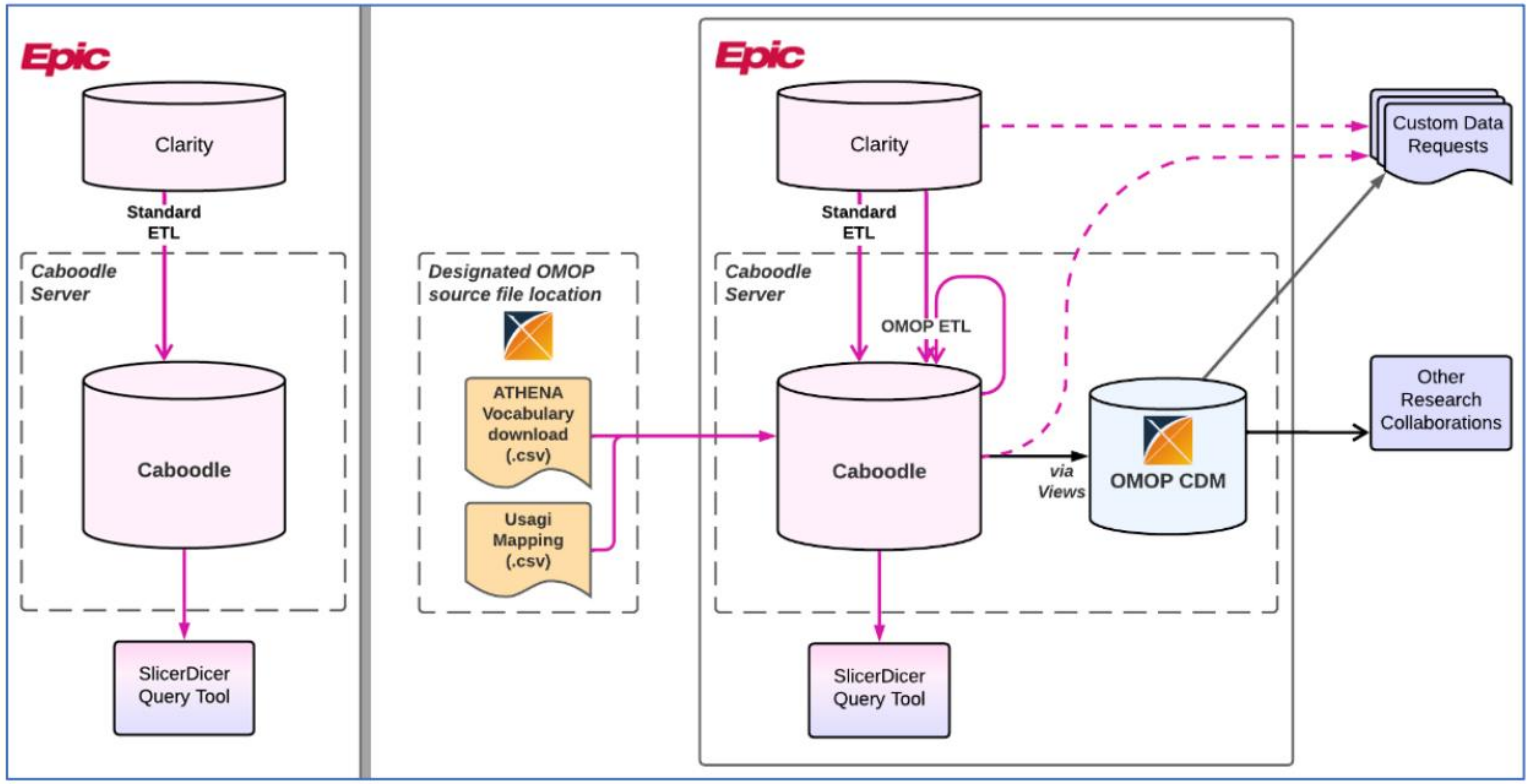
ARPA H

KPMP

All of Us RESEARCH PROGRAM

- Federal observational data research usually involves OMOP formatted data

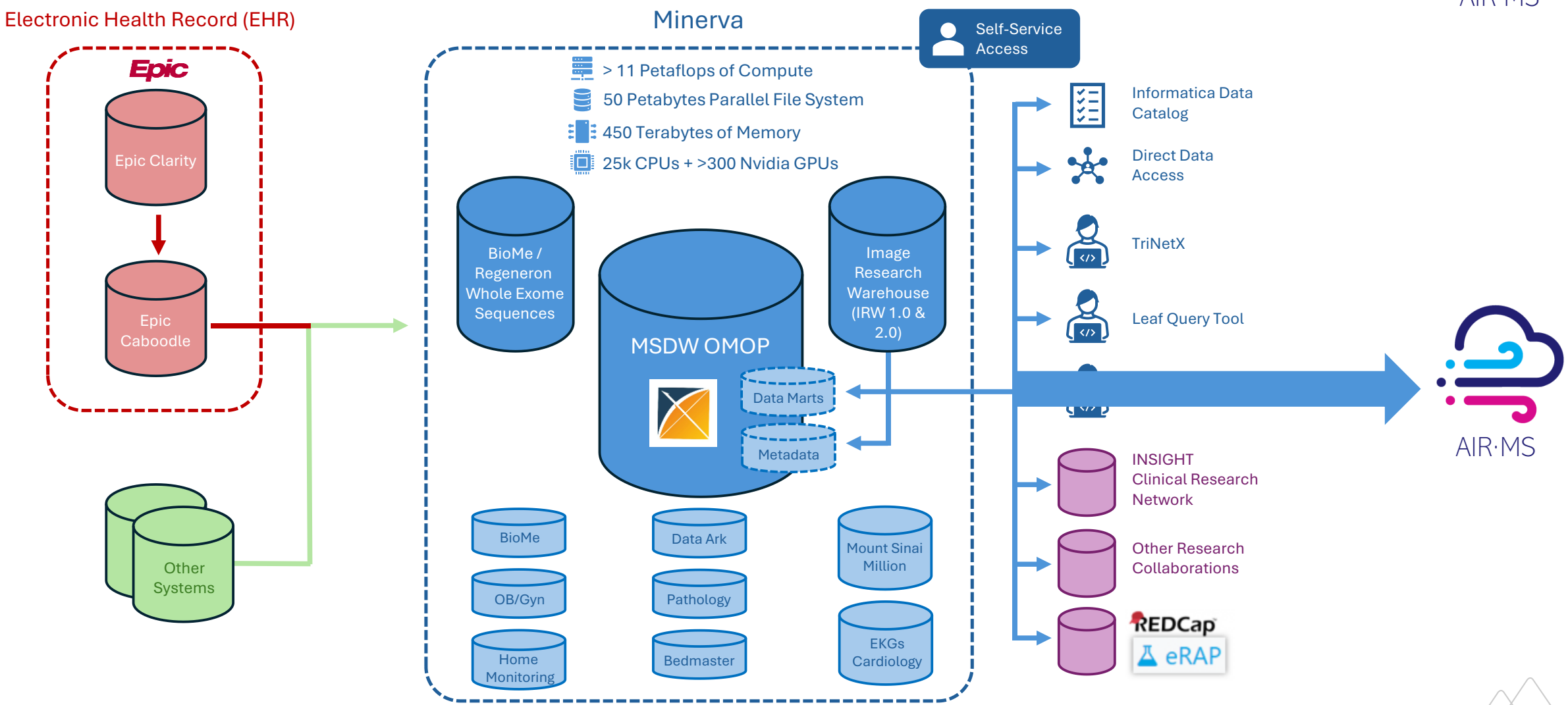
Using the Data: How Does OMOP Work?



https://www.ohdsi.org/wp-content/uploads/2023/10/10-Willett-BriefReport.pdf

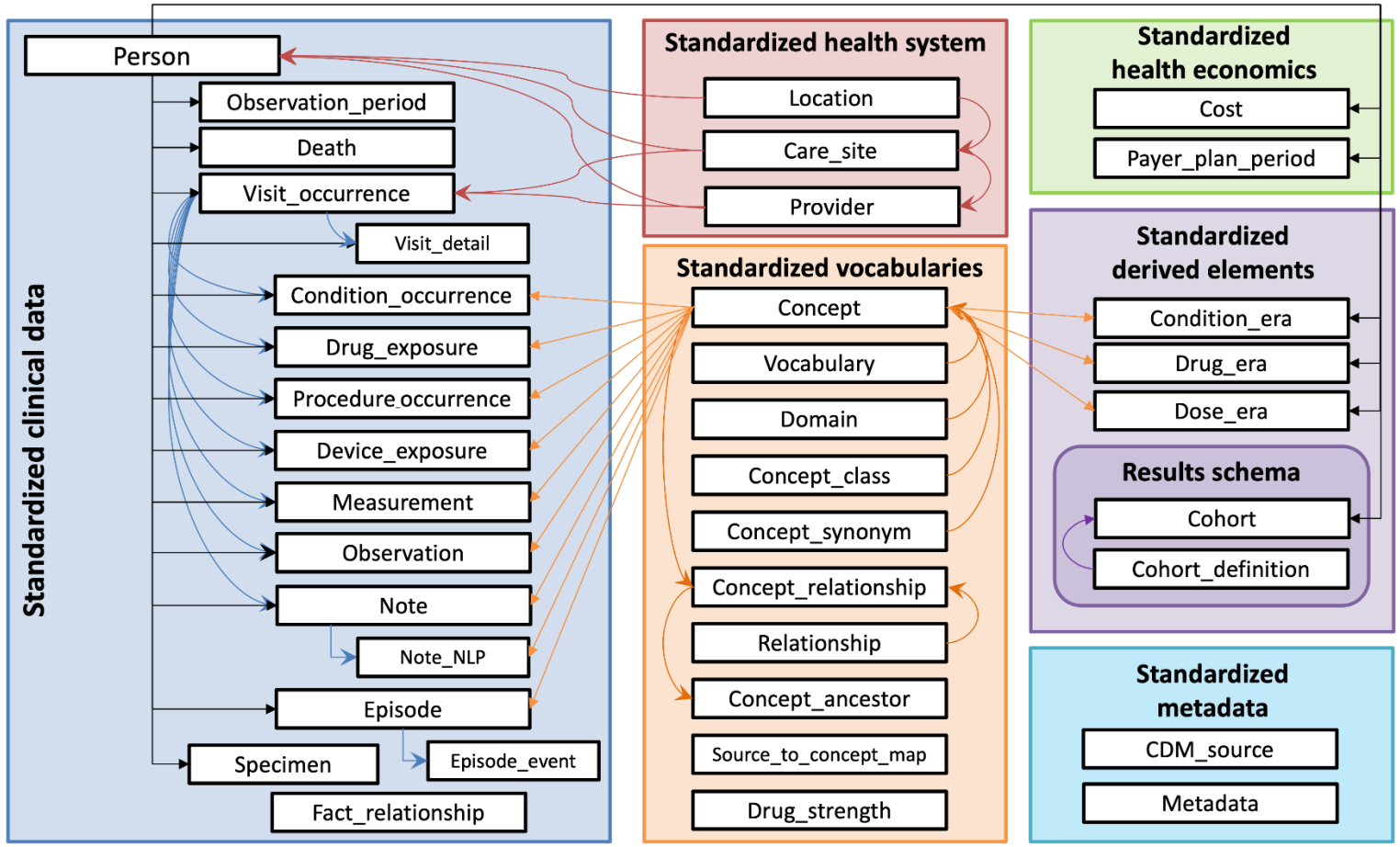
- Data is extracted from Caboodle into the OMOP Common Data Model (CDM)

Using the Data: How Does OMOP Work? (Cont'd)



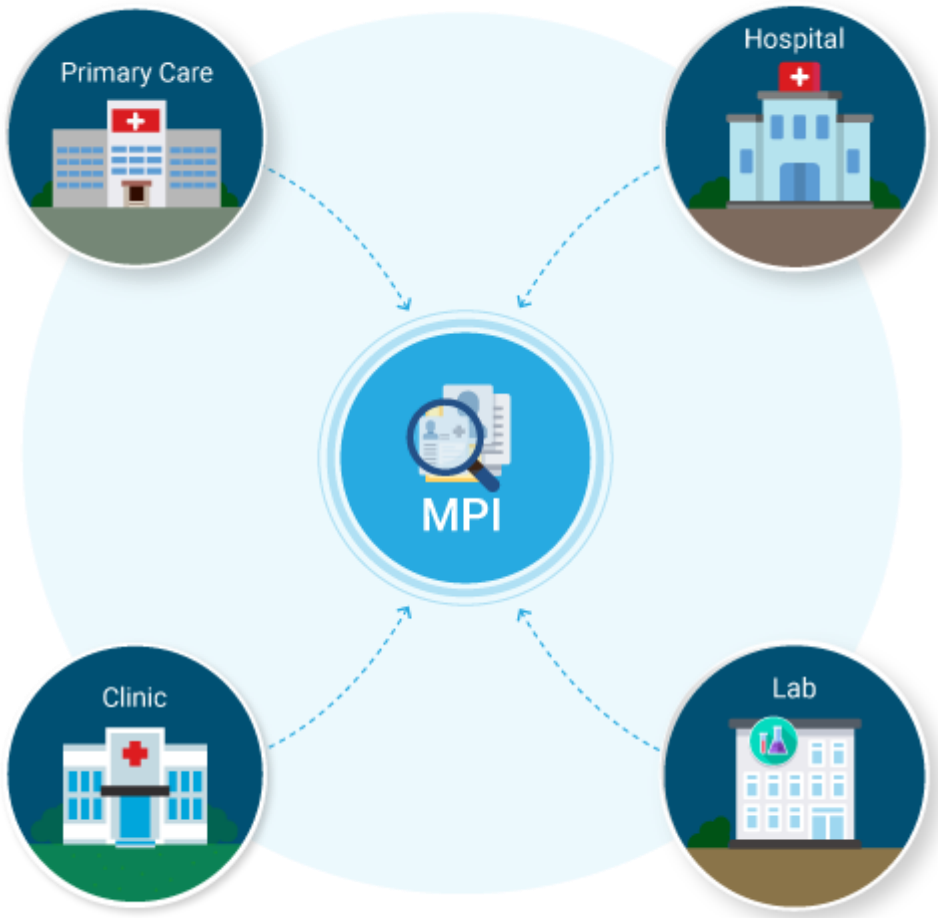
Using the Data: How Does OMOP Work? (Cont'd)

- The OMOP CDM consists of about 35 tables
- Very simplified compared to the ~20,000 tables in Clarity



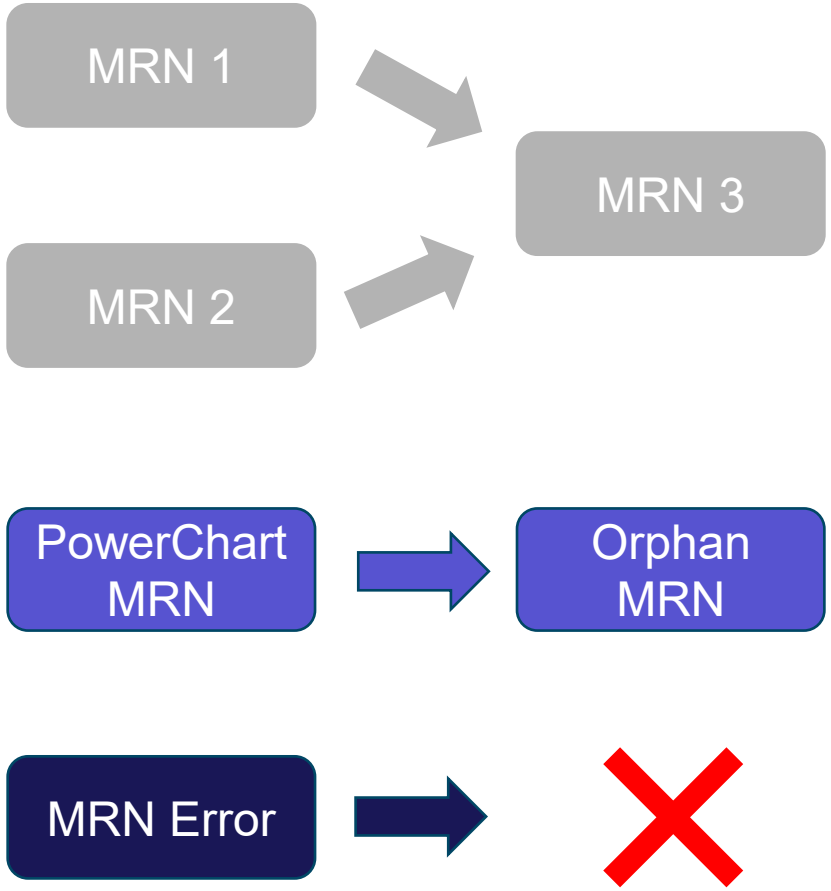
Medical Record Numbers (MRNs)

- Medical record numbers (MRNs) are the primary identifiers of patients
- They link information from across a health system
- Help with standardizing patient identifiers
- Identifiers can (1) merge, (2) be deleted, (3) be substituted/change
- These are tracked in the Master Patient Index (MPI)



Master Patient Index (MPI)

- In MSHS, several hospitals joined the system at different types, using different medical record systems (ex. PowerChart)
- Even though we have standardized over several years, there still may be old PowerChart MRNs in the system
- Additionally, some patient records have been merged across hospitals, so old identifiers may disappear



Two MRNs for same patient from separate hospitals

Old EHR System

Patient Accidentally Admitted

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

- 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

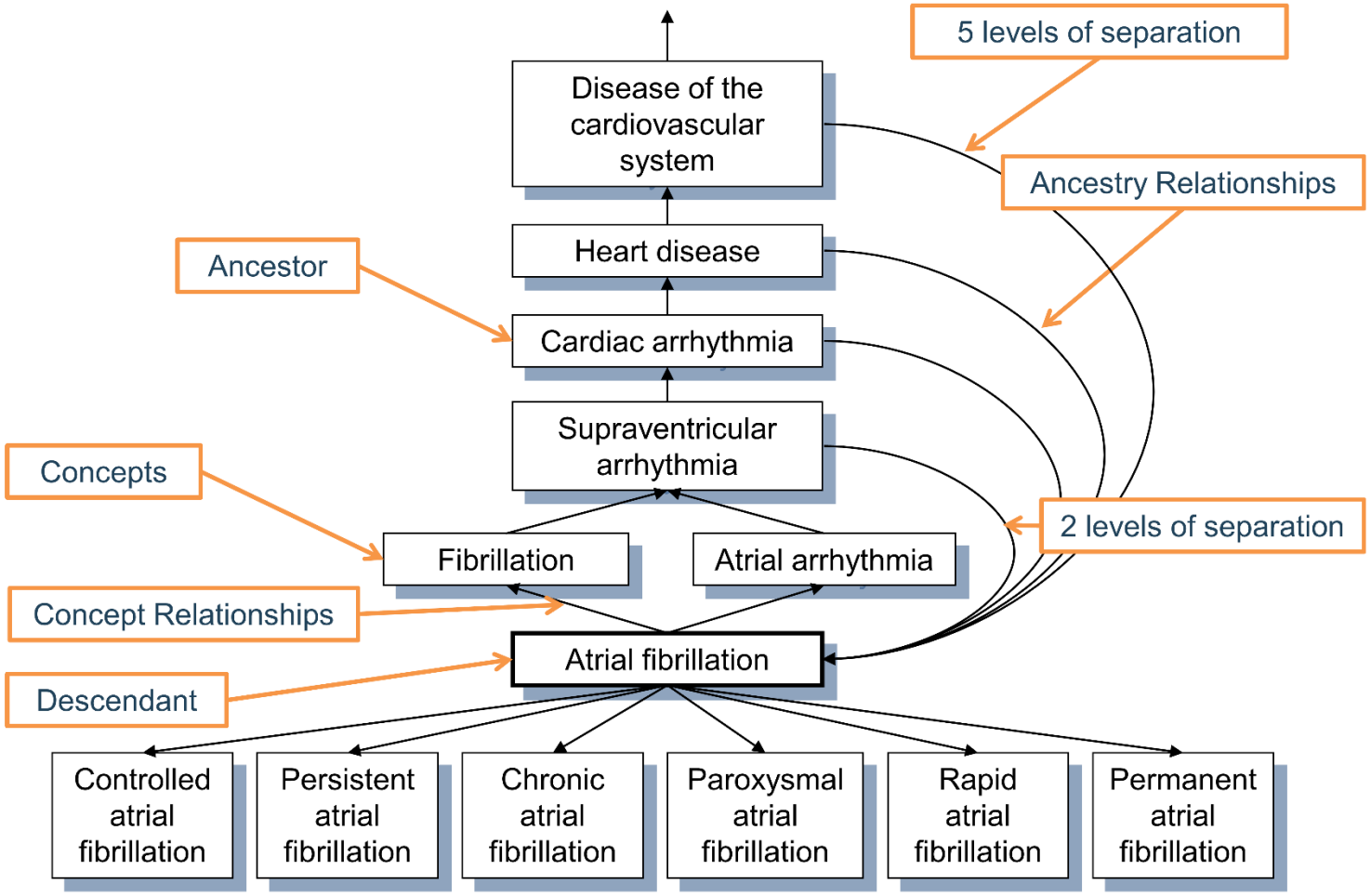
Using the Data: Querying and Retrieving Data

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

CONCEPT_ID	313217	Primary key
CONCEPT_NAME	Atrial fibrillation	English description
DOMAIN_ID	Condition	Domain
VOCABULARY_ID	SNOMED	Vocabulary
CONCEPT_CLASS_ID	Clinical Finding	Class in vocabulary
STANDARD_CONCEPT	S	Standard, Source of Classification
CONCEPT_CODE	49436004	Code in vocabulary
VALID_START_DATE	01-Jan-1970	Valid during time interval
VALID_END_DATE	31-Dec-2099	
INVALID_REASON		

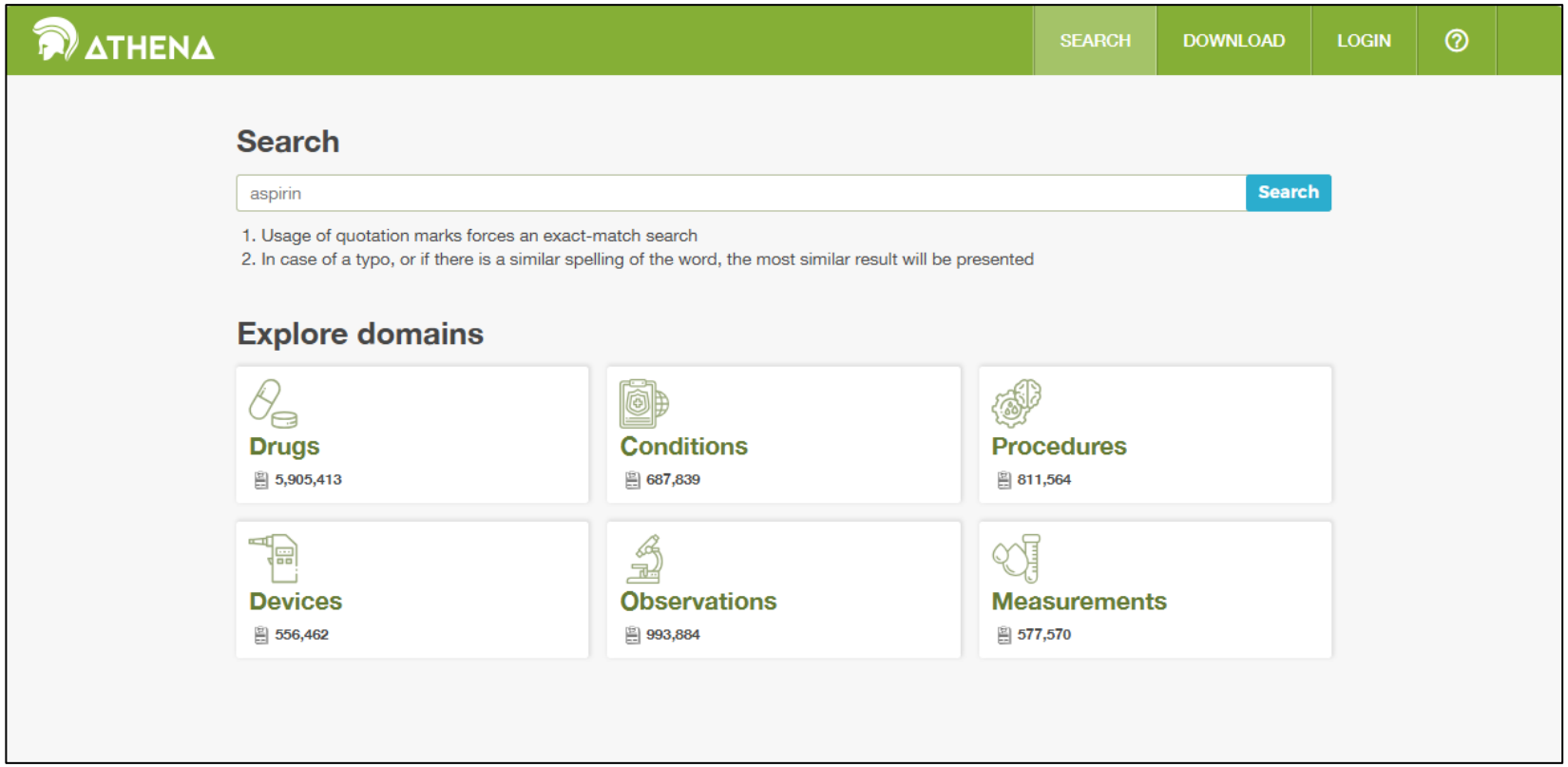
Using the Data: Querying and Retrieving Data (Cont'd)

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



Exploring the OMOP Mapping

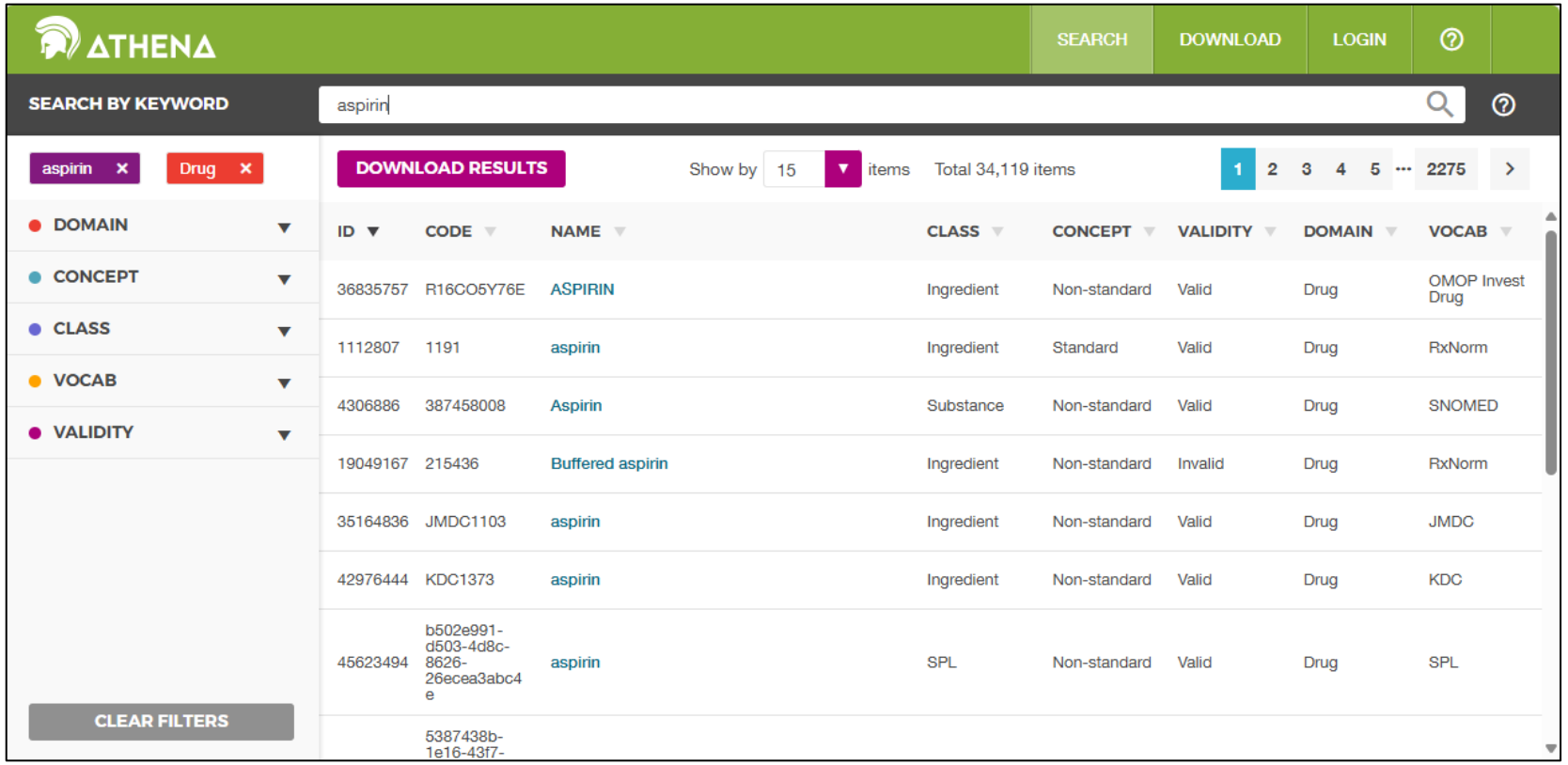
- There's a tool called Athena which contains concept mappings
- Contains standard mappings between several coding systems
- When you're building a list of codes to look for, you can start here to explore the concepts
- OMOP in Mount Sinai uses several of the Athena mappings



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



ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
36835757	R16CO5Y76E	ASPIRIN	Ingredient	Non-standard	Valid	Drug	OMOP Invest Drug
1112807	1191	aspirin	Ingredient	Standard	Valid	Drug	RxNorm
4306886	387458008	Aspirin	Substance	Non-standard	Valid	Drug	SNOMED
19049167	215436	Buffered aspirin	Ingredient	Non-standard	Invalid	Drug	RxNorm
35164836	JMDC1103	aspirin	Ingredient	Non-standard	Valid	Drug	JMDC
42976444	KDC1373	aspirin	Ingredient	Non-standard	Valid	Drug	KDC
45623494	b502e991-d503-4d8c-8626-26ecea3abc4e	aspirin	SPL	Non-standard	Valid	Drug	SPL
	5387438b-1e16-43f7-						

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