Introduction to AIR·MS

Health Data Literacy

Ashwin Sawant
Andrew Deonarine

Credits: Eugenia Alleva

Sept 15th, 2026

HPI·MS

Mount Sinai

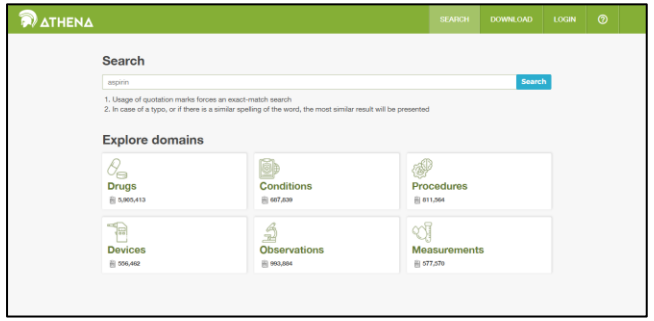
Hasso Plattner Institute for Digital Health at Mount Sinai

Epic to OMOP Medication Mapping

AIR·MS

Epic
ATC Codes

Epic

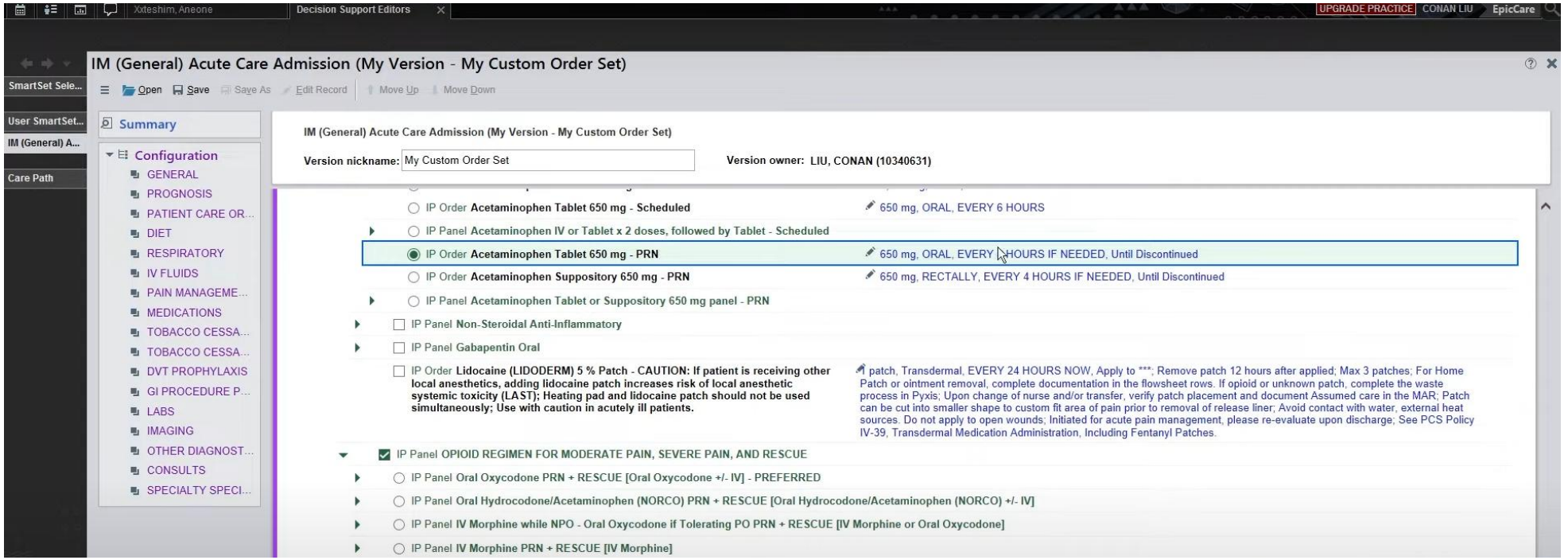


https://athena.ohdsi.org/

OMOP
RxNorm

AIR·MS

Medication Data



- Medications are stored in AIRMS, and contain a wealth of data
- This information is mapped from ATC codes (in Epic) to RxNorm (in OMOP)

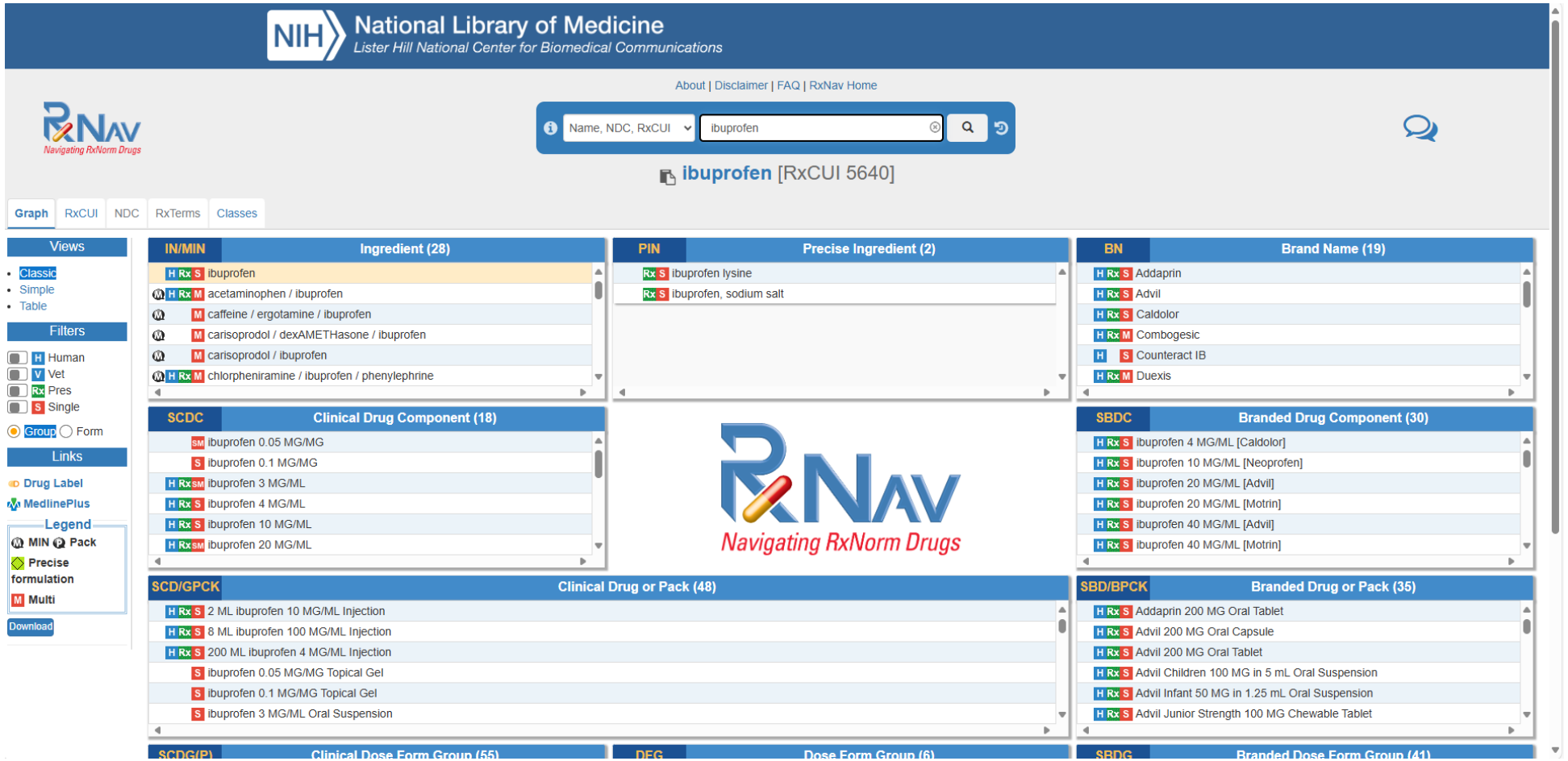
Medication Data: RxNorm

- RxNorm – this coding standard is used in the OMOP CDM
  - It is developed by the National Library of Medicine (NLM)
  - A comprehensive system that presents a standardized way of representing medications
  - They use unique identifiers, namely a Concept Identifier (RXCUI), that is linked to each medication concept.
  - Identifiers are consistent across different sources and versions

Ingredient (IN)
└── Precise Ingredient (PIN)
└── Clinical Drug Form (CDF)
└── Clinical Drug (SCD)
└── Branded Drug Component (SBDC)
└── Branded Drug (SBD)
└── Branded Pack (BPCK)

Multiple Ingredients (MIN)
└── Clinical Drug (SCD)
└── Branded Drug (SBD)
└── Pack (BPCK)

Medication Data: RxNorm



https://mor.nlm.nih.gov/RxNav/search

RxNorm Examples

1. Ingredient (IN)

The basic chemical component of a drug.

Example: Ibuprofen

2. Precise Ingredient (PIN)

A more specific form of an ingredient, like a salt or ester.

Example: Ibuprofen sodium

3. Multiple Ingredients (MIN)

Represents a combination of ingredients.

Example: Ibuprofen / Hydrocodone

4. Clinical Drug Form (CDF) [Optional level in hierarchy]

Combines an ingredient with a dosage form (without strength).

Example: Ibuprofen Oral Tablet

5. Clinical Drug (SCD = Semantic Clinical Drug)

Includes the ingredient, strength, and dose form, but no brand.

Example: Ibuprofen 200 MG Oral Tablet

6. Branded Drug Component (SBDC = Semantic Branded Drug Component)

A branded version of the SCD with one ingredient.

Example: Advil 200 MG Oral Tablet [Ibuprofen]

7. Branded Drug (SBD = Semantic Branded Drug)

A branded medication with all details (strength, dose form, etc.).

Example: Advil 200 MG Oral Tablet

8. Branded Pack (BPCK) / Clinical Pack (GPCK)

Represents packaging of one or more SBDs or SCDs.

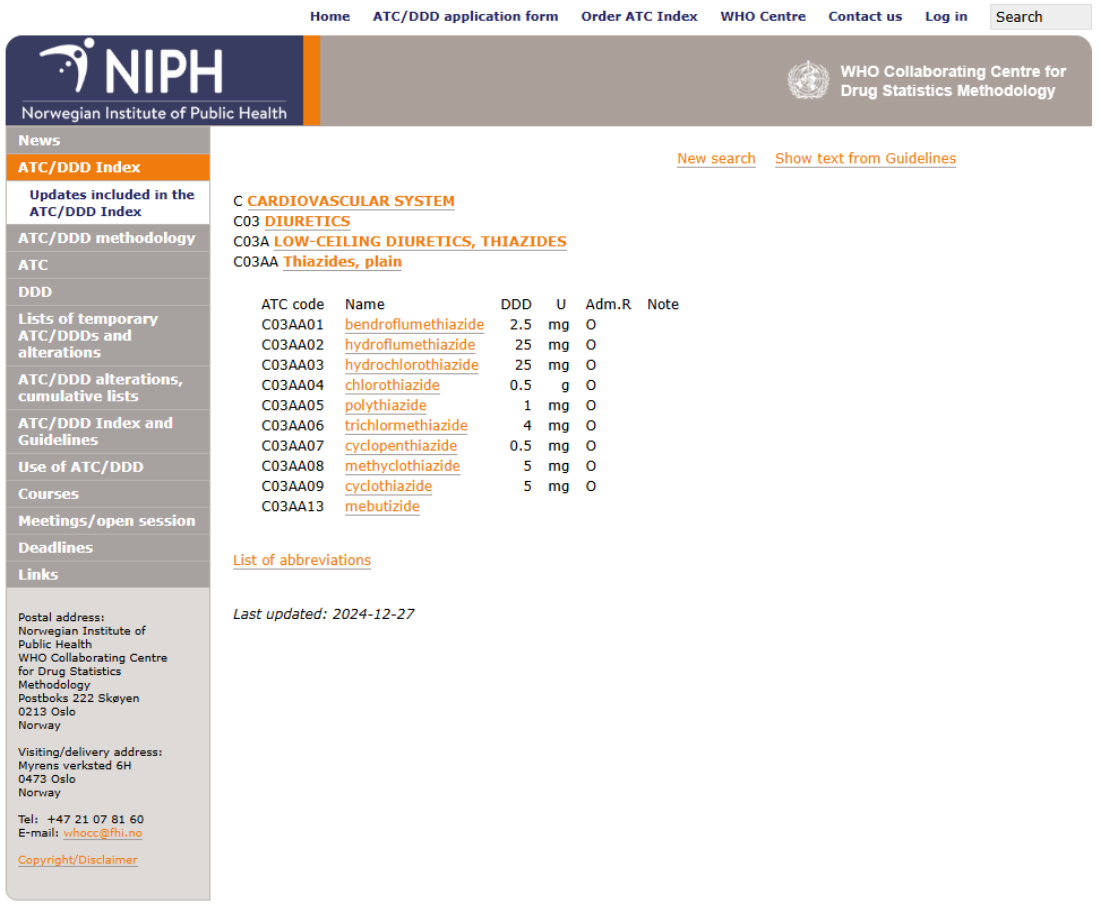
Example: Advil 200 MG Oral Tablet Pack

Medication Data: ATC Codes

- Anatomical Therapeutic Chemical (ATC) Classification System - classifies drugs based on the organ system they act on, and their properties (chemical, pharmacological, or therapeutic)
- Produced by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOCC)
- First published in 1976
- The lowest level of the ATC classification contains 5067 codes
- Simple drug browser available from https://atcddd.fhi.no (Norwegian Institute of Public Health)

A	Alimentary tract and metabolism (1st level, anatomical main group)
A10	Drugs used in diabetes (2nd level, therapeutic subgroup)
A10B	Blood glucose lowering drugs, excl. insulins (3rd level, pharmacological subgroup)
A10BA	Biguanides (4th level, chemical subgroup)
A10BA02	metformin (5th level, chemical substance)

Medication Data: ATC Codes



Home | ATC/DDD application form | Order ATC Index | WHO Centre | Contact us | Log in | Search

NIPH Norwegian Institute of Public Health

WHO Collaborating Centre for Drug Statistics Methodology

- News
- ATC/DDD Index
- Updates included in the ATC/DDD Index
- ATC/DDD methodology
- ATC
- DDD
- Lists of temporary ATC/DDDs and alterations
- ATC/DDD alterations, cumulative lists
- ATC/DDD Index and Guidelines
- Use of ATC/DDD
- Courses
- Meetings/open session
- Deadlines
- Links

Postal address:
Norwegian Institute of Public Health
WHO Collaborating Centre for Drug Statistics Methodology
Postboks 222 Skøyen
0213 Oslo
Norway

Visiting/delivery address:
Myrens verksted 6H
0473 Oslo
Norway

Tel: +47 21 07 81 60
E-mail: whocc@fhi.no

Copyright/Disclaimer

New search | Show text from Guidelines

C CARDIOVASCULAR SYSTEM
C03 DIURETICS
C03A LOW-CEILING DIURETICS, THIAZIDES
C03AA Thiazides, plain

ATC code	Name	DDD	U	Adm.R	Note
C03AA01	bendroflumethiazide	2.5	mg	O	
C03AA02	hydroflumethiazide	25	mg	O	
C03AA03	hydrochlorothiazide	25	mg	O	
C03AA04	chlorothiazide	0.5	g	O	
C03AA05	polythiazide	1	mg	O	
C03AA06	trichlormethiazide	4	mg	O	
C03AA07	cyclopenthiazide	0.5	mg	O	
C03AA08	methyclothiazide	5	mg	O	
C03AA09	cyclothiazide	5	mg	O	
C03AA13	mebutizide				

List of abbreviations

Last updated: 2024-12-27

https://atcddd.fhi.no/atc_ddd_index/?code=B02BC&showdescription=no

Medications

Medication Order		Medication Dispense		Medication Administration		Immunization Event		drug_exposure		
						Patient A	Vaccine Z	Patient A	Immuniz Z1	Vaccine Z
Patient B	Inpt Med Order 2					Inpt Med Order 2	Vaccine Y	Patient B	Inpt Med Order 2	Vaccine Y
Patient C	Inpt Med Order 3			Inpt Med Order 3	Med Admin 3-1			Patient C	Med Admin 3-1	Drug X
					Med Admin 3-2				Med Admin 3-2	Drug X
Patient D	Inpt Med Order 4	Inpt Med Order 4	Dispense 4-1	Inpt Med Order 4	Med Admin 4-1			Patient D	Med Admin 4-1	Drug W
			Dispense 4-2		Med Admin 4-2				Med Admin 4-2	Drug W
Patient E	Inpt Med Order 5	Inpt Med Order 5	Dispense 5-1					Patient E	Dispense 5-1	Drug V
Patient F	Outpt Med Order 6	Outpt Med Order 6	Dispense 6-1					Patient F	Dispense 6-1	Drug U
Patient G	Outpt Med Order 7	Outpt Med Order 7	Dispense 7-1					Patient G	Dispense 7-1	Drug T
			Dispense 7-2						Dispense 7-1	Drug T
Patient H	Outpt Med Order 8							Patient H	Outpt Med Order 8	Drug S

Medication Record Types

xtn_drug_type_source_concept_id	xtn_drug_type_source_concept_name	row_count
2002056202	Immunization Administration	7,835,276
2002056203	Immunization from Medication Order	289,032
2002056204	Immunization from Procedure Order	3,488,313
2000000121	Medication Order	13,670,853
2000000110	Medication Order with Administration	138,120,557
2000000098	Medication Order with Dispense	16,351,836
2000000096	Outpatient Medication Order	48,588,821
2000000109	Outpatient Medication Order with Dispense	151,086

Record counts as of April 21, 2025

```
# Confirm the ancestor concept (ATC class)

antiplatelet_concept = 35807468

sql_ap_atc_info = f"""
SELECT concept_id, concept_name, vocabulary_id, concept_class_id, domain_id
FROM CDMDEID.CONCEPT
WHERE concept_id = {antiplatelet_concept}
"""
airms.conn.sql(sql_ap_atc_info).collect()
```

```
# Pull RxNorm Ingredients that descend from the ATC Antiplatelet class (concept_id = 35101523)
vocabulary = 'RxNorm'
concept_class = 'Ingredient'

sql_ap_ingredients_from_atc = f"""
SELECT
    ca.max_levels_of_separation,
    c.concept_id          AS ingredient_id,
    c.concept_name        AS ingredient_name,
    c.vocabulary_id,
    c.concept_class_id,
    c.domain_id,
    c.standard_concept
FROM CDMDEID.CONCEPT_ANCESTOR ca
JOIN CDMDEID.CONCEPT c
  ON c.concept_id = ca.descendant_concept_id
WHERE ca.ancestor_concept_id = {antiplatelet_concept}
   AND c.vocabulary_id = '{vocabulary}'
   AND c.concept_class_id = '{concept_class}'
   AND c.invalid_reason IS NULL
ORDER BY ca.max_levels_of_separation, c.concept_name
"""
ap_ingredients = airms.conn.sql(sql_ap_ingredients_from_atc).collect()
ap_ingredients
```

Lab Results & Vital Signs (measurements)

Lab Order

Patient A	Lab Order 1
Patient B	Lab Order 2

Lab Component Result

Lab Order 1	WBC Count
	RBC Count
	Hct Result
	Hgb Result
Lab Order 2	Glucose Result
	Calcium Result
	Sodium Result
	BUN Result

measurement

Patient A	WBC Count
	RBC Count
	Hct Result
	Hgb Result
Patient B	Glucose Result
	Calcium Result
	Sodium Result
	BUN Result
Patient C	Height
	Weight
	Temp
Patient D	Metric 3
Patient E	Metric 4

Flowsheet

Patient C	Height
	Weight
	Temp
Patient D	Metric 3
Patient E	Metric 4

Measurement Record Types

xtn_measurement_type_source_concept_id	xtn_measurement_type_source_concept_name	row_count
2002067233	Flowsheet Measurement	254,696,901
2000000100	Lab Component Result	1,107,913,246
2000000123	Vital Signs	692,312,136

Record counts as of April 21, 2025

```
## BMI Measurement near AF Index

In OMOP, anthropometric values like BMI live in the **MEASUREMENT** table.

- Each row represents a quantitative measurement (`value_as_number`) recorded at a date (`measurement_date`).
- Concepts come from vocabularies like **LOINC** or **SNOMED**.

For BMI, we use the standard concept **"Body mass index"** and its descendants.

**Approach:**
1. Identify BMI concepts (LOINC + SNOMED descendants).
2. Pull all BMI rows for AF patients.
3. For each patient, find the **BMI value closest in time to AF index**.
```

```sql
# Find BMI concepts in OMOP (LOINC + SNOMED)

concept_name = 'body mass index'
concept_class = 'Clinical Observation'
domain = 'Measurement'

sql_bmi_conceptset = f"""
SELECT concept_id, concept_name, vocabulary_id, domain_id, concept_class_id
FROM CDMDEID.CONCEPT
WHERE domain_id = '{domain}'
  AND standard_concept = 'S'
  AND concept_class_id = '{concept_class}'
  AND LOWER(concept_name) LIKE '%{concept_name}%'
  AND invalid_reason IS NULL
```

Content Standardization via Concept Mapping

OMOP Data Example (*synthetic*)

condition_occurrence	Value	Definition
condition_type_concept_id	32827	OMOP's standard record type "EHR encounter record"
xtn_condition_type_source_concept_id	2000000129	MSDW's source record type "Encounter Diagnosis"
visit_occurrence_id	999888777	Unique identifier for the encounter
person_id	987654321	Unique identifier for the patient
provider_id	123456789	Unique identifier for the provider recording the diagnosis
condition_concept_id	4193704	OMOP's identifier for **SNOMED code 313436004**
condition_source_concept_id	2000602205	MSDW's identifier for **Epic diagnosis code 521601**
condition_start_date	1/1/2022	The date of condition onset or documentation per provider
condition_end_date	NULL	The date on which the condition resolved (if any)

concept_relationship	Value	Definition
concept_1	2000602205	MSDW's identifier for **Epic diagnosis code 521601**
relationship_id	Maps to non-standard	Text string denoting the type of **mapping relationship** between concept_id_1 & concept_id_2
concept_2	35206882	OMOP's identifier for **ICD-10-CM code E11.9**

OMOP Data Example (*synthetic*)

condition_occurrence	Value
condition_type_concept_id	32827
xtn_condition_type_source_concept_id	2000000129
visit_occurrence_id	999888777
person_id	987654321
provider_id	123456789
condition_concept_id	4193704
condition_source_concept_id	2000602205
condition_start_date	1/1/2022
condition_end_date	NULL

concept_relationship	Value
concept_1	2000602205
relationship_id	Maps to non-standard
concept_2	35206882

concept

concept_id	vocabulary_id	concept_code	concept_name
32827	Type Concept	OMOP4976900	EHR encounter record
2000000129	MSDW Src Rec Type	Encounter Diagnosis	Encounter Diagnosis
4193704	SNOMED	313436004	Type 2 diabetes mellitus without complication
2000602205	EPIC EDG .1	521601	Type 2 diabetes mellitus without complications
2000602205	EPIC EDG .1	521601	Type 2 diabetes mellitus without complications
35206882	ICD10CM	E11.9	Type 2 diabetes mellitus without complications

Visualizing and Analyzing Data in Jupyter

Using the Data: Querying and Retrieving Data

Age and Sex from PERSON

We now extend the cohort with **demographics**:

- **Sex** is stored in `PERSON.gender_concept_id`.
  We join to the CONCEPT table for a readable label.
- **Age at AF index** is calculated from `PERSON.birth_date` relative to `AF_INDEX_DATE`.

This information is critical for describing the cohort and adjusting analyses.

```
# Querying age and sex
sql_pm_af_ap_age_sex = f"""
WITH cohort_base AS (
  {sql_pm_af_ap_bmi}
)
SELECT
    cb.*,
    p.birth_datetime,
    g.concept_name AS sex,
    FLOOR(MONTHS_BETWEEN(p.birth_datetime, cb.af_index_date) / 12) AS age_at_af
FROM cohort_base cb
JOIN CDMDEID.PERSON p
  ON p.person_id = cb.person_id
JOIN CDMDEID.CONCEPT g
  ON g.concept_id = p.gender_concept_id
ORDER BY cb.af_index_date
"""

pm_af_ap_age_sex = airms.conn.sql(sql_pm_af_ap_age_sex).collect()
pm_af_ap_age_sex.head()
```

Using the Data: Querying and Retrieving Data

AIR·MS

Exploratory Data Analyses

We summarize cohort size, demographics, therapies near diagnosis, device procedures, and BMI capture.
Figures are de-identified and derived from the **de-identified OMOP** dataset.

Cohort anchor: first AF diagnosis (`AF_INDEX_DATE`).
Therapy window: antiplatelet exposure overlapping `[AF_INDEX_DATE, +6 months]` .
Pacemaker outcome: first pacemaker procedure on/after `AF_INDEX_DATE` .
BMI: nearest measurement within ±180 days of `AF_INDEX_DATE` .

Preparing the data ¶

```
# Build final dataset (called cohort)
cohort  = pm_af_ap_age_sex
```

```
# Build by_month & cumulative counts to extract important overall metrics
af_series = pd.to_datetime(cohort['AF_INDEX_DATE'], errors='coerce').dropna()
by_month = (
    af_series.dt.to_period('M')
    .value_counts()
    .sort_index()
    .to_timestamp()
)
cum_by_month = by_month.cumsum()

# Extract number of unique patients in the cohort
N = cohort['PERSON_ID'].nunique()

# Extract age statistics
age_vals = cohort['AGE_AT_AF'].dropna().astype(float)
age_med  = np.nanmedian(age_vals) if len(age_vals) else np.nan
```

Using the Data: Querying and Retrieving Data

AIR·MS

```
# Find BMI concepts in OMOP (LOINC + SNOMED)

concept_name = 'body mass index'
concept_class = 'Clinical Observation'
domain = 'Measurement'

sql_bmi_conceptset = f"""
SELECT concept_id, concept_name, vocabulary_id, domain_id, concept_class_id
FROM CDMDEID.CONCEPT
WHERE domain_id = '{domain}'
  AND standard_concept = 'S'
  AND concept_class_id = '{concept_class}'
  AND LOWER(concept_name) LIKE '%{concept_name}%'
  AND invalid_reason IS NULL
"""
bmi_concepts = airms.conn.sql(sql_bmi_conceptset).collect()
bmi_concepts
```

```
# We want the actual ration that was observed (i.e. 3038553)
sql_bmi_conceptset = 'SELECT * FROM CDMDEID.CONCEPT WHERE CONCEPT_ID=3038553'
```

```
# Build on your existing cohort (sql_pm_after_af) and append BMI nearest to AF index
# Assumes sql_pm_after_af returns: person_id, af_index_date, first_pm_date, days_to_pacemaker,
#                                  on_antiplatelet_within_6mo, first_ap_start_in_window, last_ap_end_in_window

sql_pm_af_ap_bmi = f"""
WITH cohort_base AS (
  {sql_pm_af_ap}
),

bmi_cs AS (
  {sql_bmi_conceptset}
```

Observations

Patient Demographics		Surgical History		Social History		Family History		Patient Allergy		observation	
Patient A	Race									Patient A	Race
	Ethnicity										Ethnicity
	Language Preference										Language Preference
	Sexual Orientation										Sexual Orientation
Patient B	Race									Patient B	Race
	Marital Status										Marital Status
	Gender Identity										Gender Identity
Patient C	Ethnicity									Patient C	Ethnicity
	Religious Affiliation										Religious Affiliation
		Patient A	Procedure Z							Patient A	Procedure Z
			Procedure Y								Procedure Y
		Patient B	Procedure X							Patient B	Procedure X
			Procedure W								Procedure W
				Patient C	Soc Hx Item 1-1					Patient C	Soc Hx Item 1-1
					Soc Hx Item 1-2						Soc Hx Item 1-2
					Soc Hx Item 1-3						Soc Hx Item 1-3
						Patient D	Fam Hx 2-1			Patient D	Fam Hx 2-1
							Fam Hx 2-2				Fam Hx 2-2
								Patient E	Allergy 3-1	Patient E	Allergy 3-1

Observation Record Types

xtn_observation_type_source_concept_id	xtn_observation_type_source_concept_name	row_count
2002056205	Allergy	4,277,961
2000000128	Family History	65,724,728
2002067234	Flowsheet Observation	114,620,841
2000000124	Past Medical History	8,610,801
2000000116	Patient Demographics	64,233,860
2000000118	Social History	145,273,185
2000000126	Surgical History Procedure	3,681,599

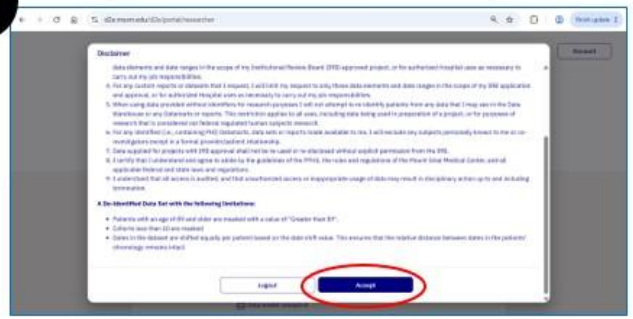
Record counts as of April 21, 2025

```
SELECT
  t.xtn_observation_type_source_concept_id
, t.xtn_observation_type_source_concept_name
, FORMAT(COUNT_BIG(*), 'N0') AS row_count
FROM omop.cdm_phi.observation t
WHERE t.observation_id <> 0
GROUP BY
  t.xtn_observation_type_source_concept_id
, t.xtn_observation_type_source_concept_name
ORDER BY
  t.xtn_observation_type_source_concept_name
```

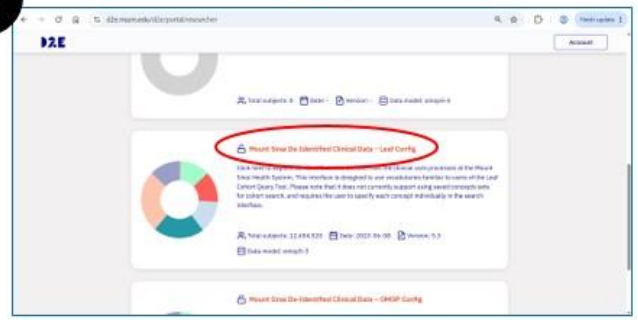
Putting it Together: Queries & Phenotypes

Get Started with Data2Evidence and AI Agent

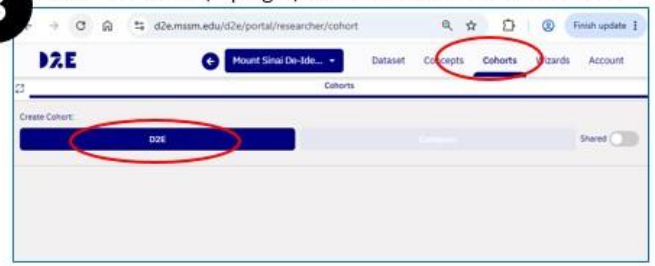
1. Visit https://d2e.airms.mssm.edu , Click “Accept”



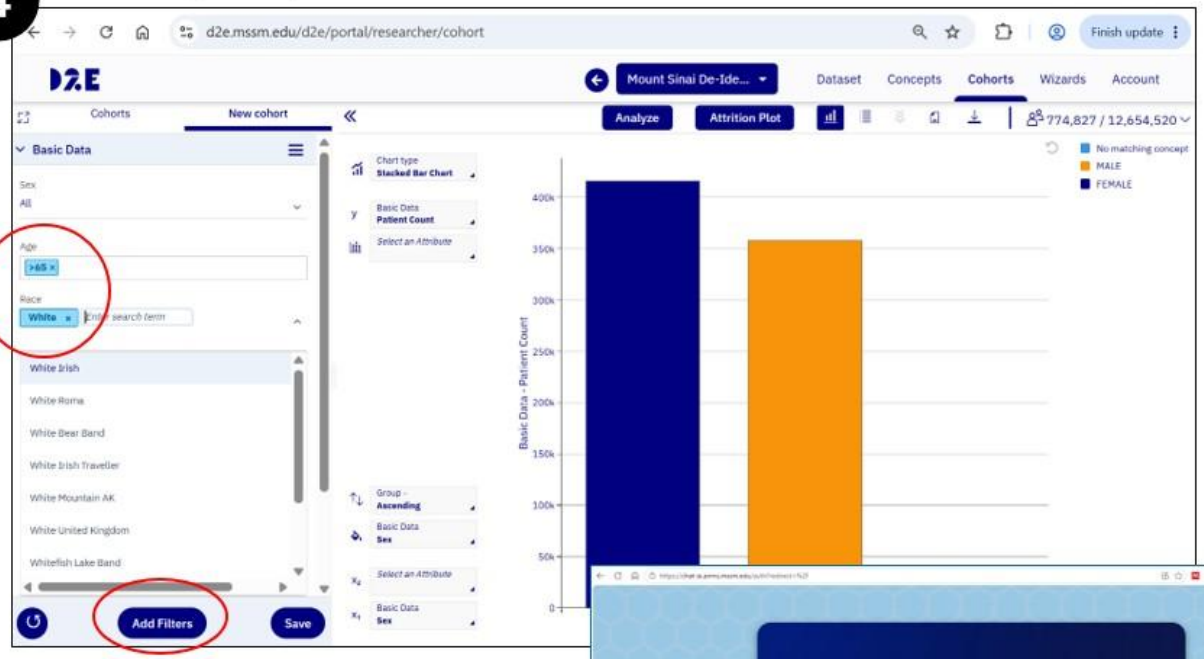
2. Select De-Identified Data (Leaf)



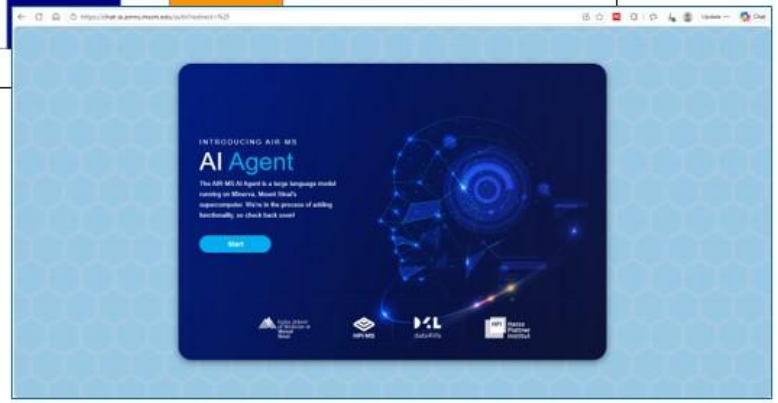
3. Click “Cohorts” (top right) and then click the “D2E” button



4. Start searching: enter Age: “>65” and Race: “WHITE”, click “Add Filters” to add diseases



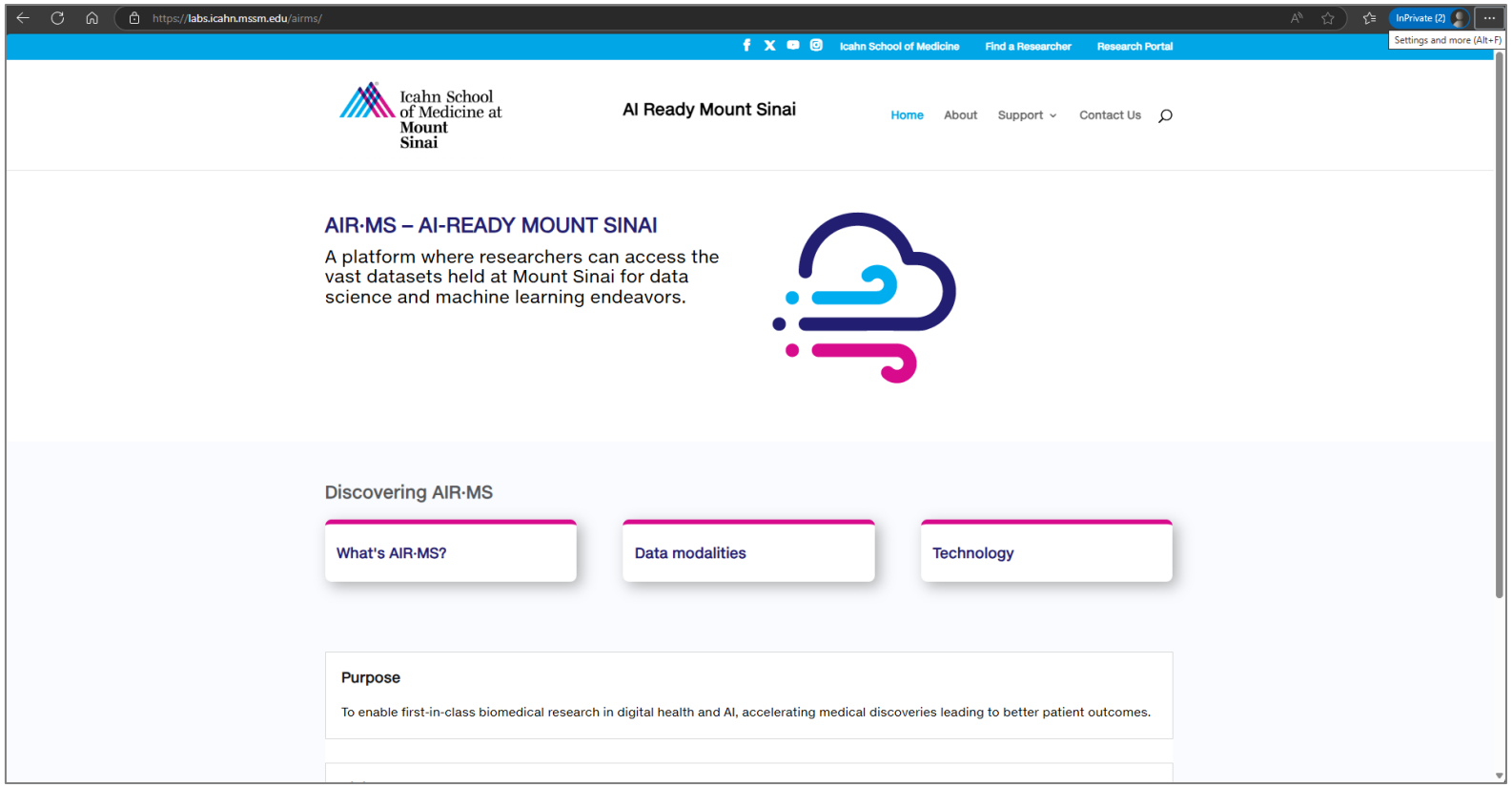
Also visit our new AI Agent:
https://chat-ai.airms.mssm.edu/
(under development)



https://d2e.airms.mssm.edu

The AIR•MS Chatbot

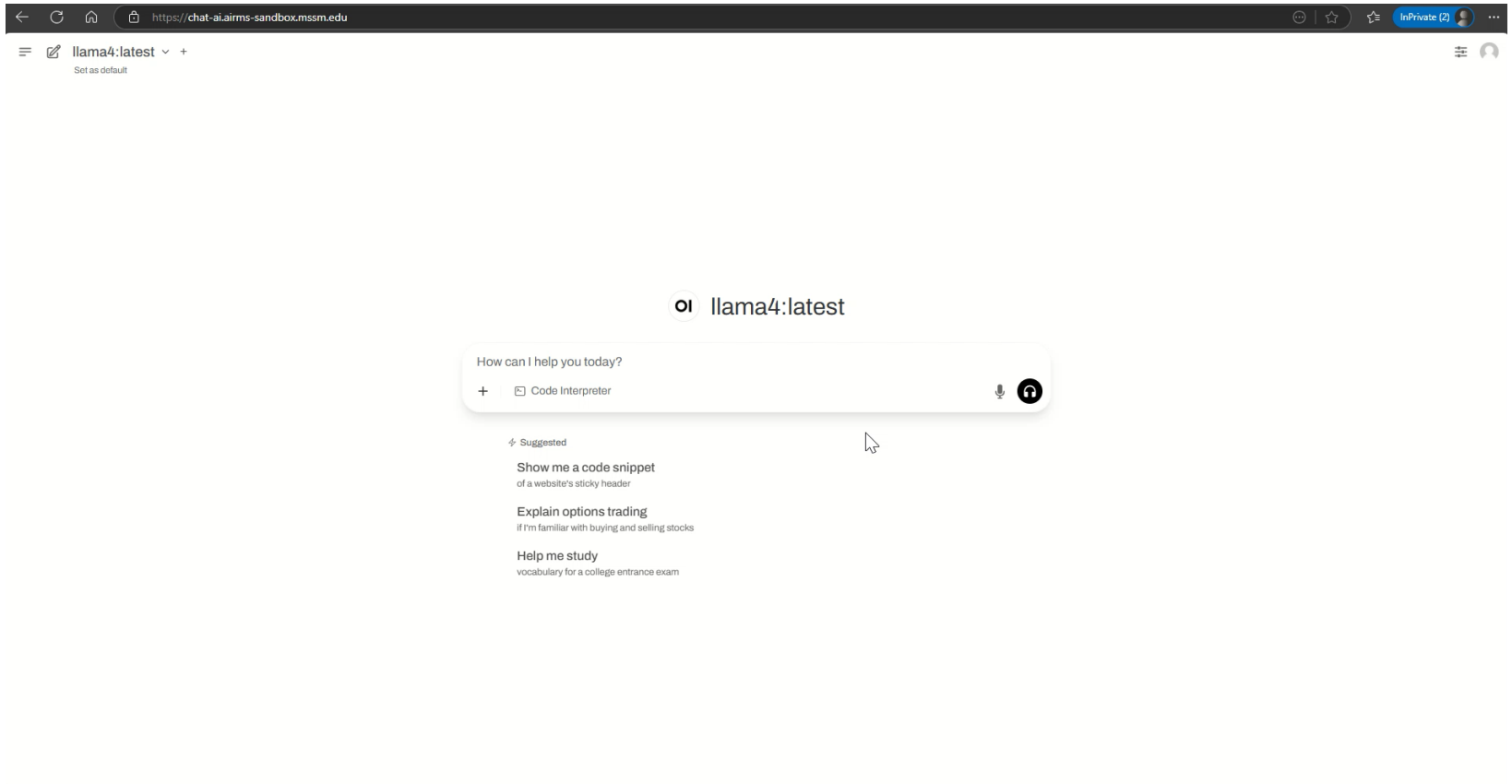
- The chatbot is part of AIR•MS (AI Ready Mount Sinai) which is a centralized platform that enables AI throughout the Mount Sinai System
- Eventually the chatbot will help users answer questions about Mount Sinai clinical data
- An online presence for AIR•MS is currently under construction and will be frequently updated
- Built as part of a collaboration between **Data4Life** (D4L), **Scientific Computing and Data** (SCD), **AI and Human Health** (AIHH), and the **Hasso Plattner Institute** (HPI)



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

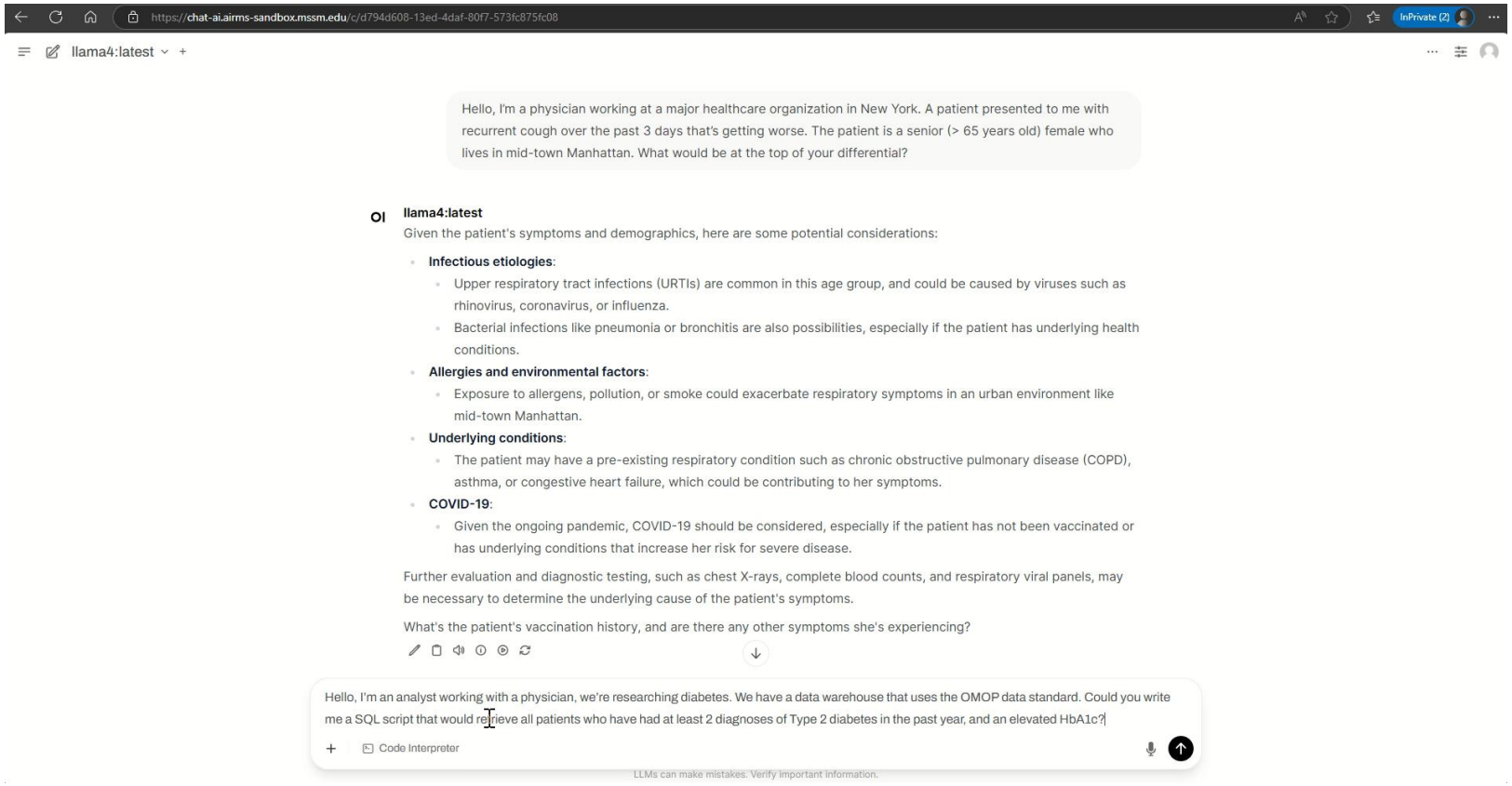
A Simple Prompt

- “Hello, I’m a physician working at a major healthcare organization in New York. A patient presented to me with recurrent cough over the past 3 days that’s getting worse, and the patient has a history of breast cancer. The patient is a senior (> 65 years old) female who lives in New York City. What would be at the top of your differential?”



A More Complex Prompt

- “Hello, I’m an analyst working with a physician, we’re researching diabetes. We have a data warehouse that uses the OMOP data standard. Could you write me a SQL script that would retrieve all patients who have had at least 2 diagnoses of Type 2 diabetes in the past year, and an elevated HbA1c?”



Querying the Data & Phenotypes

- When you're researching a disease (ex. Lyme disease) – just looking for a particular International Classification of Diseases (ICD) code or disease keyword ("lyme") is not enough

    - You need to consult clinical practice guidelines
    - Sometimes there's an initial diagnosis that's wrong – you may need to look for lab results (ex. 2 separate diagnoses of diabetes within 6 months of each other, and an HbA1c level)
    - So, you might write a query that looks for patients with three Type 2 diabetes codes within a 6-month period, and two HbA1c levels above a certain threshold, rather than simply looking for patients with one Type 2 diabetes code
    - When possible, consult physicians and clinicians when devising how to identify patients with a particular disease or phenotype, peer-reviewed publications are also a good source of information
    - For example – NIH N3C specific codes for COVID-19: https://tinyurl.com/4dwruasx

Checklist

Item	Comments
Check if Patient IDs are deleted, merged, orphan, or substituted, where possible check against MPI	MPI = master patient index
When selecting codes, consult a clinician or use peer-reviewed literature	
Do not use keyword searches on vocabularies	Ex. "stroke" vs. "cerebral ischemia"
Use mapping/ vocabulary browser tools when necessary: Athena, Bioportal, RxNav, etc.	Remember: Epic IDs are licensed, so you won't find public mappings for Epic IDs vs. other coding systems
When constructing your digital phenotype, use clinical practice guidelines to inform the development of your phenotype	Get a clinician to help you understand a clinical practice guideline

Thank You