John Snow LABS

Medical Language Models for Data Scientists

October 2025

Building Patient Journeys & Cohorts

Clinical Data is Spread in Multiple Sources

Multi-Modal

Unstructured

Not Normalized

Not Consistent

Not Certain

Continuously Updating

Some Clinical Data is Only Available in Unstructured Text

> J Med Internet Res. 2022 Mar 23;24(3):e27210. doi: 10.2196/27210

A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports

"Accuracies for predicting group-level site and histology codes were 93.5% and 97.6% respectively."

> *npj Digital Medicine* vol. 7, no. 6 (2024)

Large language models to identify social determinants of health in electronic health records

"Our models identified 93.8% of patients with adverse SDoH, while ICD-10 codes captured 2.0%."

> PHUSE/FDA 2025 Computational Science Symposium (CSS).

The Importance of Information Extraction from Unstructured Clinical Data in Pharmacoepidemiology

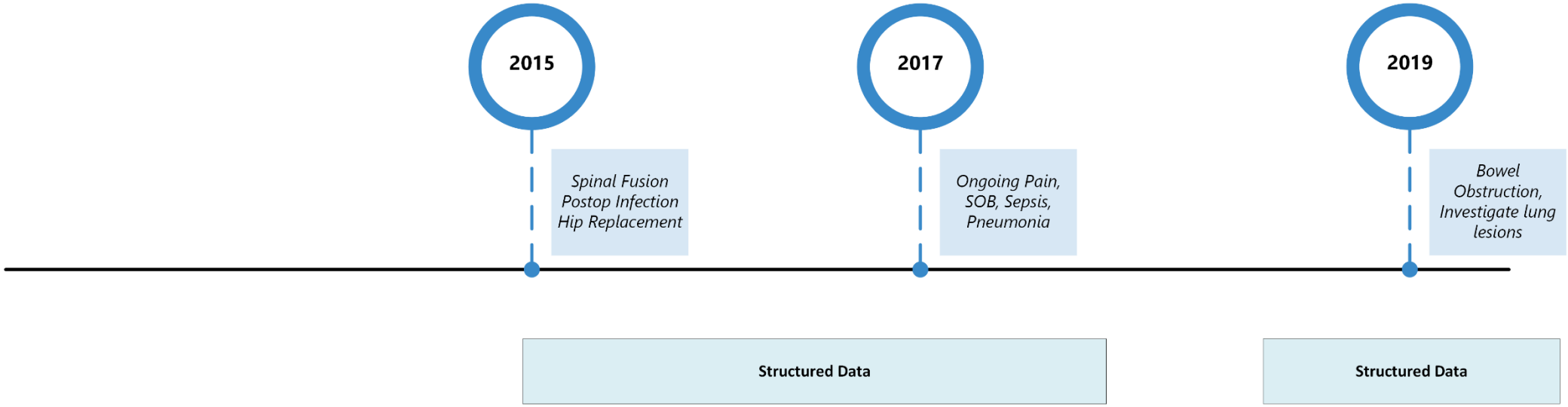
"the number of observed suicidality and self-harm events doubled with the addition of unstructured EHR data."

> AMIA Annu Symp Proc. 2015 Nov 5;2015:2035–2042.

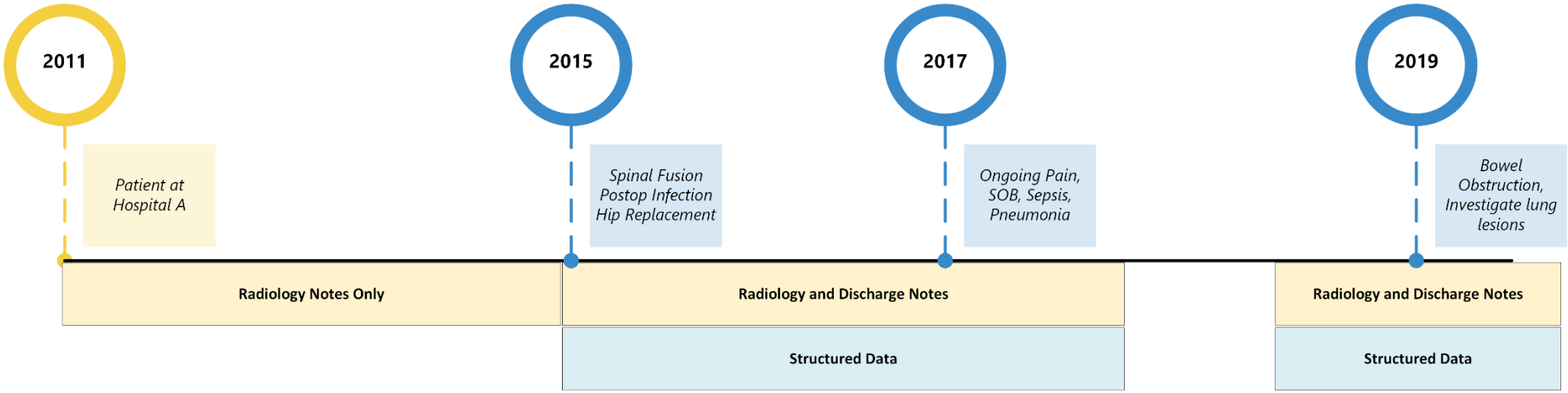
An Assessment of Family History Information Captured in an Electronic Health Record

58.7% of the observations from the Neurology Admission Note contained family history, versus only 5.2% of structured records.

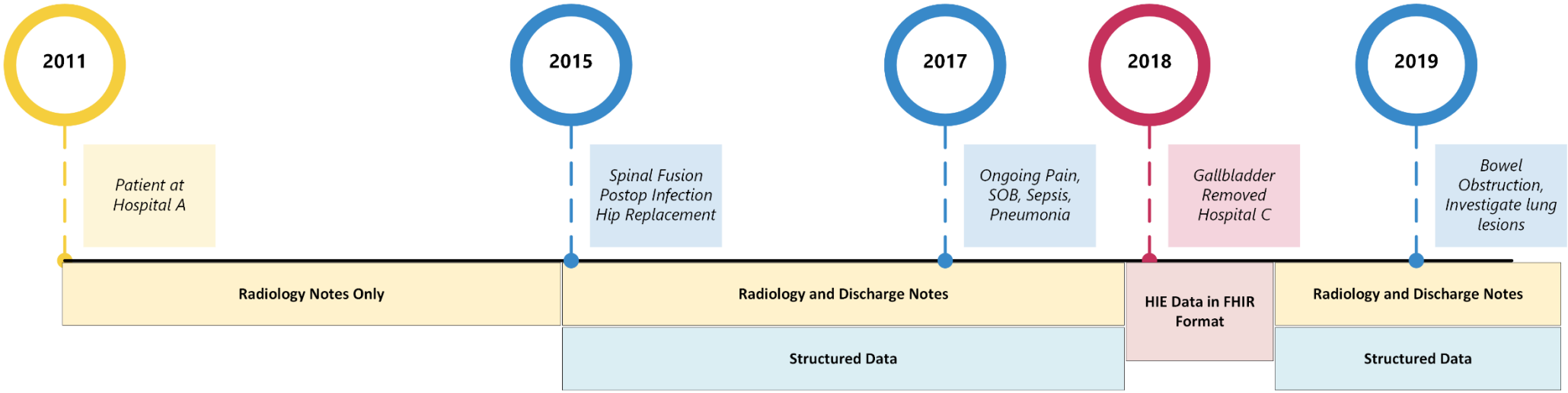
One Patient's Journey: Structured EHR Data



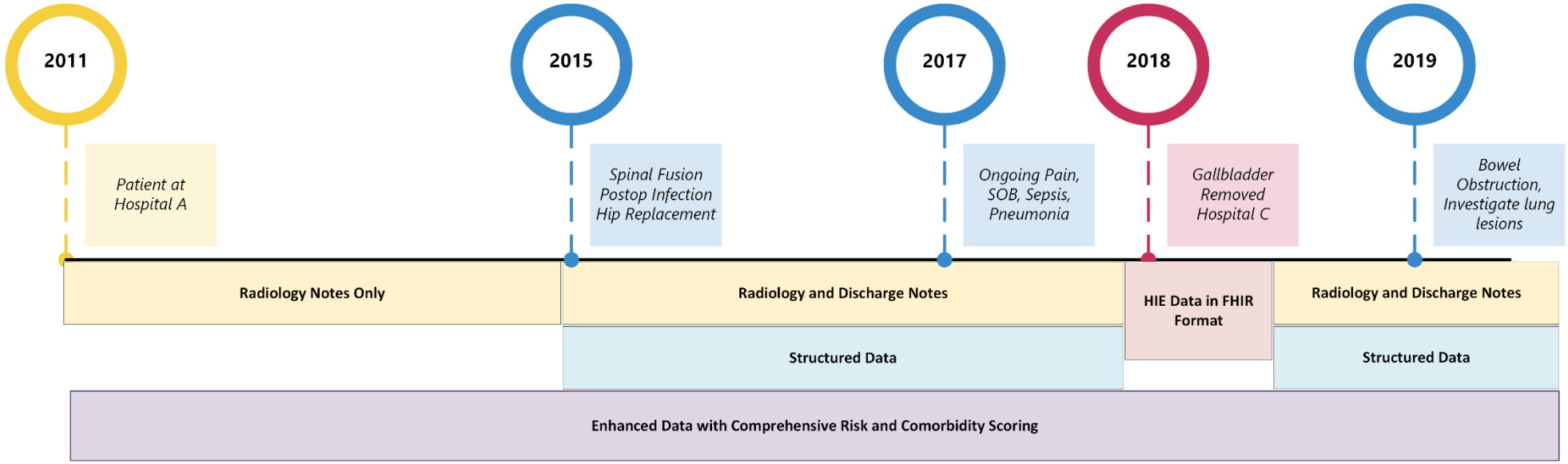
Adding Radiology & Discharge Notes



Adding FHIR Resources



Enriching the Data by Calculating Risk Scores



Current LLMs Can’t Work on Complex Queries

Q: “Find patients diagnosed with back pain that have had spinal fusion.”

- **RAG** can’t find relevant information
- **Text2SQL** hallucinates in real-world DB settings, or build **queries that fail**
- Lack of **consistency**
- **Democratize** Cohort creation

```
WITH
-- Identify tuberculosis diagnosis concepts
tb_diagnosis_concepts AS (
    SELECT c.concept_id
    FROM concept c
    JOIN concept_ancestor ca ON ca.descendant_concept_id = c.concept_id
    WHERE ca.ancestor_concept_id IN (
            -- Add the root concept ID for tuberculosis and its descendants
            SELECT concept_id
            FROM concept
            WHERE concept_name = 'Tuberculosis Family'
            -- Ensure you replace 'Tuberculosis Family' with the correct name if different
        )
        AND c.standard_concept = 'S'
),

-- Identify tuberculosis drug concepts
tb_drug_concepts AS (
    SELECT c.concept_id
    FROM concept c
    JOIN concept_ancestor ca ON ca.descendant_concept_id = c.concept_id
    WHERE ca.ancestor_concept_id IN (
            -- Add the root concept ID for tuberculosis treatments and its descendants
            SELECT concept_id
            FROM concept
            WHERE concept_name = 'Tuberculosis Treatment'
            -- Ensure you replace 'Tuberculosis Treatment' with the correct name if
different
        )
        AND c.standard_concept = 'S'


), ● ● ●
```

Our Approach

Enterprise Grade Design

John Snow LABS

Compliance

Designed for PHI
Zero Data Sharing
Air-Gap Execution

Accuracy

Healthcare LLMs
Consistent Answers
Explainable Results

Operations

Easy to Operate
Easy to Integrate
Keep Up to Date

AI + BI

Chatbots
Dashboards
Notebooks

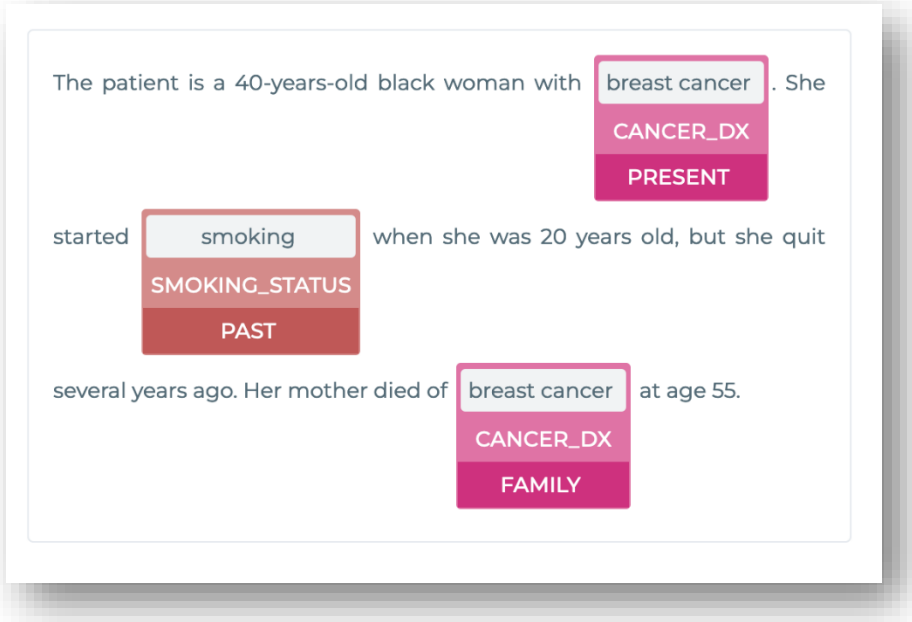
Scalability

Billions of Documents
Millions of Patients
Thousands of Users

Open Data Model

Industry Standard
Toolset Ecosystem

Semantic Information Extraction (Healthcare NLP)



begin	end	entity_type	assertion	confidence
47	59	Cancer_Dx	Present	0.9992
74	80	Smoking_Status	Past	0.9310
160	172	Cancer_Dx	Family	1.0000

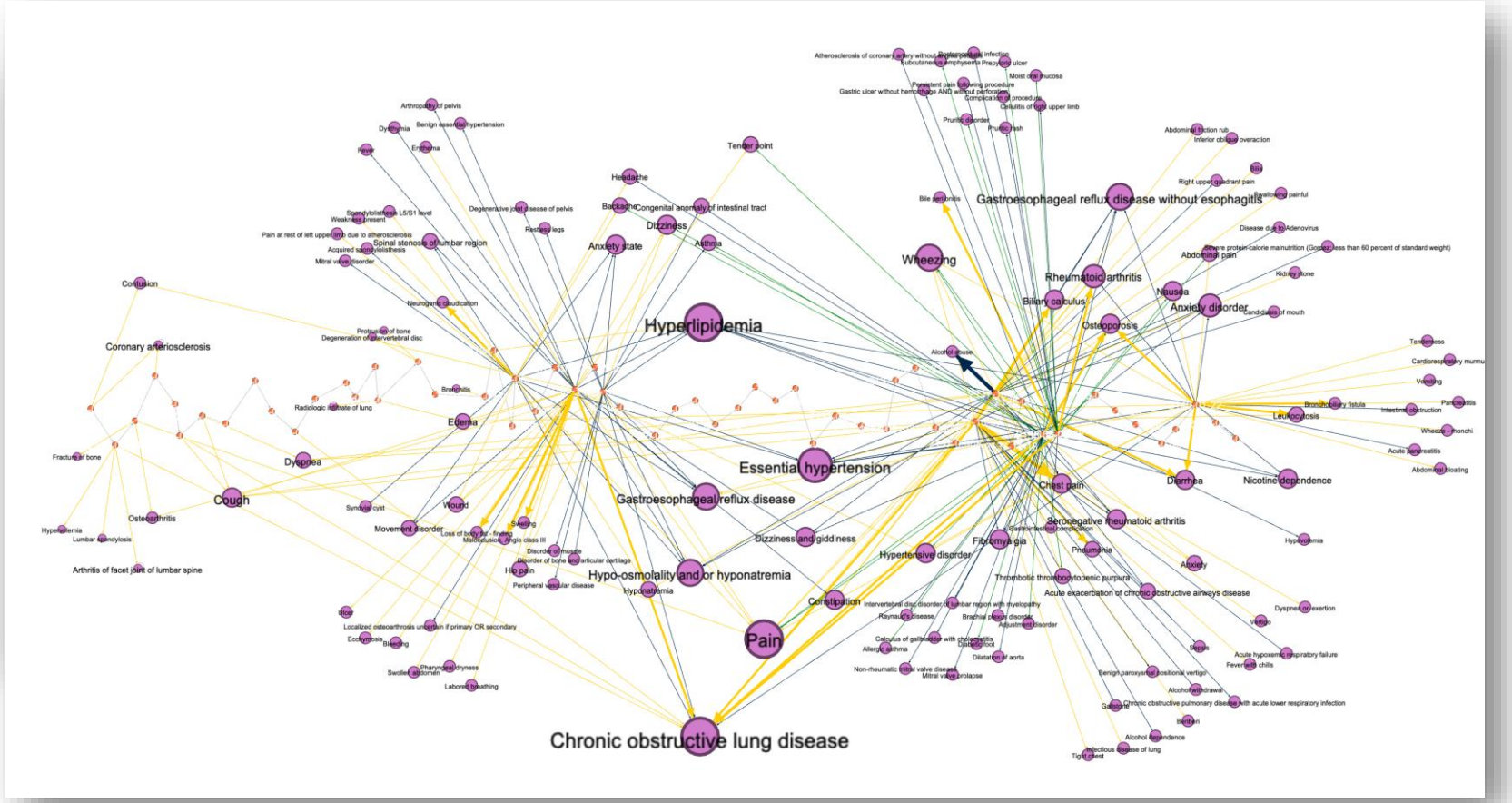
Terminology Server: Resolving to Standard Codes


```
{
  "url": "https://fhir/Mutation#assessed.gene",
  "valueCodeableConcept": {
    "coding": [
      {
        "system": "http://ncit.nci.nih.gov",
        "code": "C17757",
        "display": "EGFR"
      }
    ]
  }
},
{
  "url": "https://fhir/Mutation#assessed.referenceSeq",
  "valueCodeableConcept": {
    "coding": [
      {
        "system": "http://ncbi.nlm.nih.gov/CCDS",
        "code": "5514.1",
        "display": "CCDS 5514.1"
      }
    ]
  }
},
{
  "url": "https://fhir/Mutation#assessed.variant",
  "valueCodeableConcept": {
    "coding": [
      {
        "system": "http://www.hgvs.org/mutnomen",
        "code": "c.2369C>T",
        "display": "c.2369C>T"
      },
      {
        "system": "http://www.hgvs.org/mutnomen",
        "code": "p.T790M",
        "display": "T790M"
      }
```

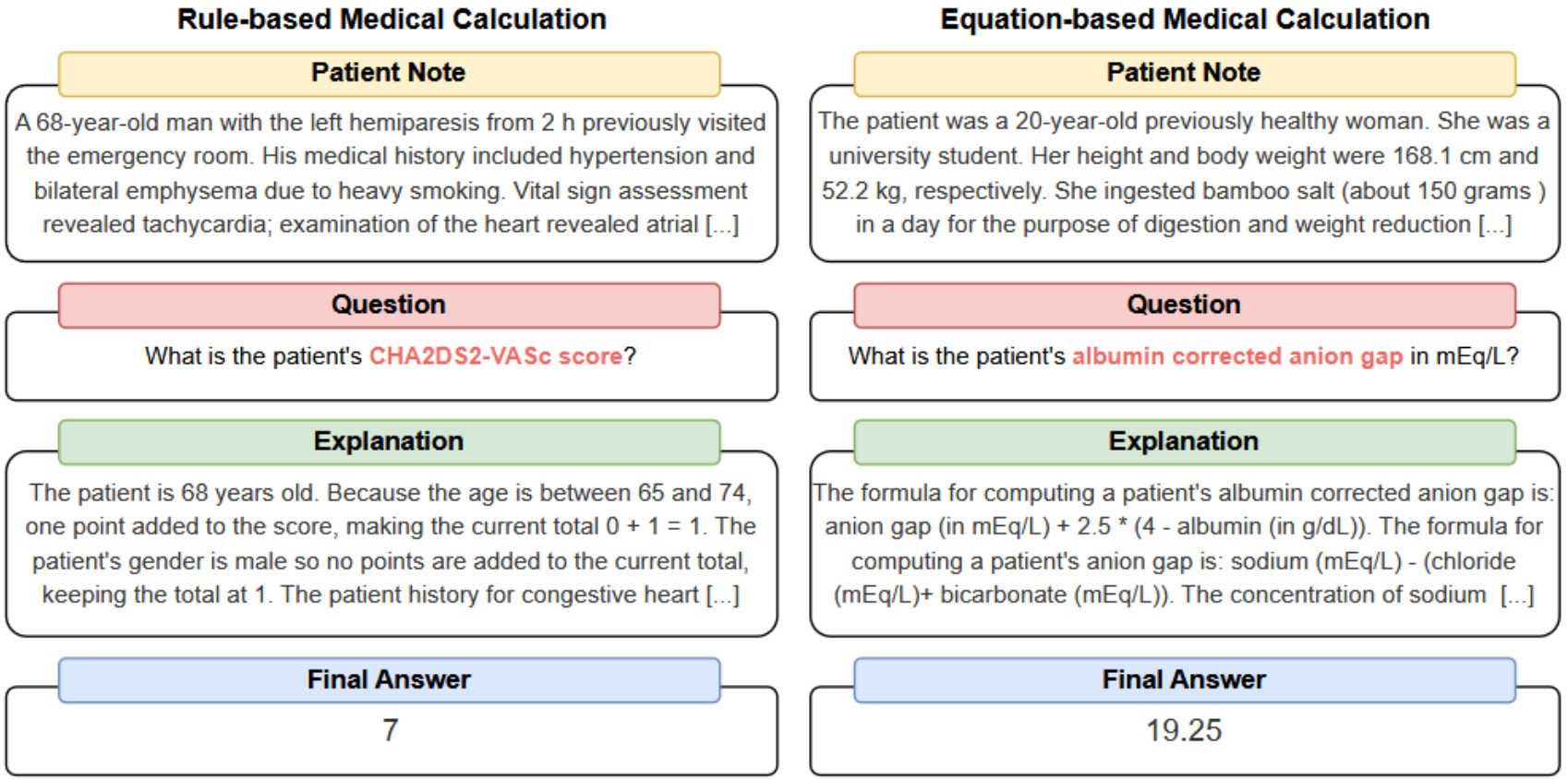

ner_chunk	entity	snomed_code	resolved_text
Catheterization of left heart	Procedure	67629009	Catheterization of left heart
selective coronary angiogram	Test	33367005	Coronary angiography
common femoral angiogram	Test	4701000087107	Angiography of right femoral artery
StarClose closure of right common femoral artery	Procedure	310621009	Patch repair of femoral artery

Merging & Deduplicating Facts to Build a Patient Graph

4150125 - Persistent pain following procedure	EHR billing record
77670 - Chest pain	EHR billing record
198263 - Right upper quadrant pain	EHR billing record
77670 - Chest pain	NLP derived
4329041 - Pain	NLP derived
200219 - Abdominal pain	NLP derived
4329041 - Pain	NLP derived
4329041 - Pain	NLP derived
4170554 - Hip pain	NLP derived
4329041 - Pain	NLP derived
761856 - Pain at rest of left upper limb due to ather...	NLP derived
4329041 - Pain	NLP derived
4170554 - Hip pain	NLP derived
4329041 - Pain	NLP derived
77670 - Chest pain	NLP derived
4147218 - Swallowing painful	NLP derived
4329041 - Pain	NLP derived
200219 - Abdominal pain	NLP derived
4329041 - Pain	NLP derived
4329041 - Pain	NLP derived



Making Clinical Inferences and Calculations



MedCalc-Bench: Evaluating Large Language Models for Medical Calculations

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, Aidan Gilson, Maxwell B Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, Zhiyong Lu

The OMOP CDM

Observational Medical Outcomes Partnership - Common Data Model

Enhancing Healthcare through Data, since 2009

Foundation: Part of the Observational Health Data Sciences and Informatics (OHDSI) initiative.

Objective: Utilize open-source data solutions to improve human health via large-scale analysis.

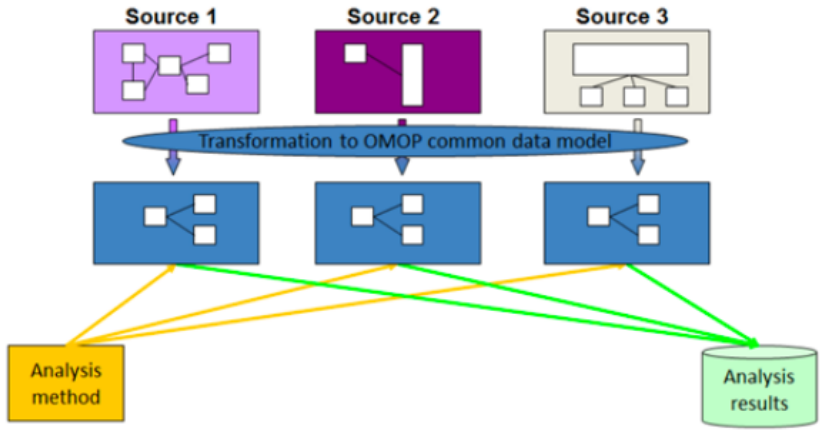
Purpose: Standardize the structure and content of observational healthcare data.

Methods:

- Through pseudonymisation and common data quality assessments, the OMOP-CDM provides a robust framework for converting complex EMR data into a standardised format.
- By securely sharing de-identified and aggregated data and conducting analyses across multiple OMOP-converted databases, patient-level data is securely firewalled within its respective local site.

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

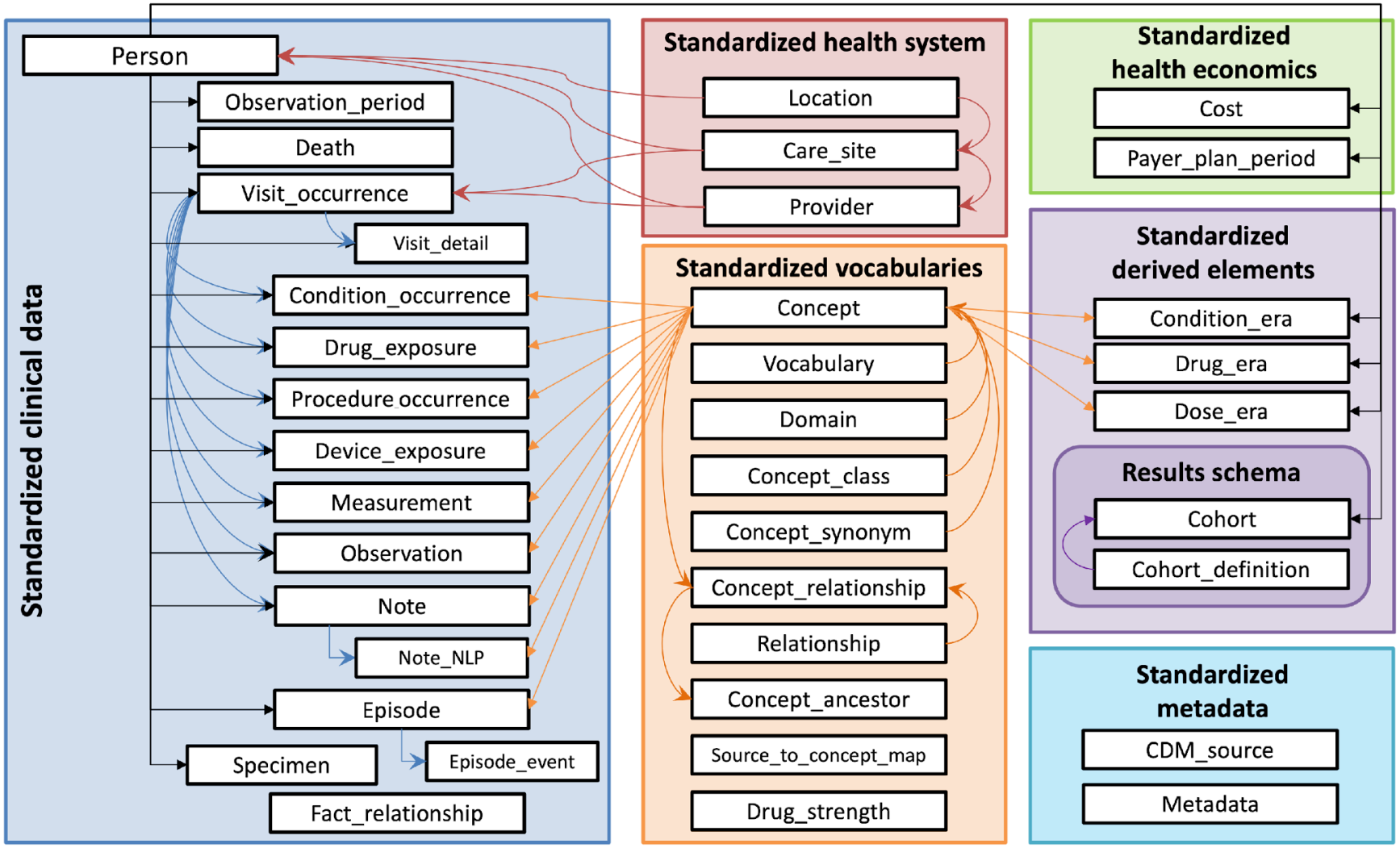
"an international collaborative whose goal is to create and apply open-source data analytic solutions to a large network of health databases to improve human health and wellbeing"



Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM, Feb 2024.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10882353/

The OMOP CDM

Observational **M**edical **O**utcomes **P**artnership - **C**ommon **D**ata **M**odel

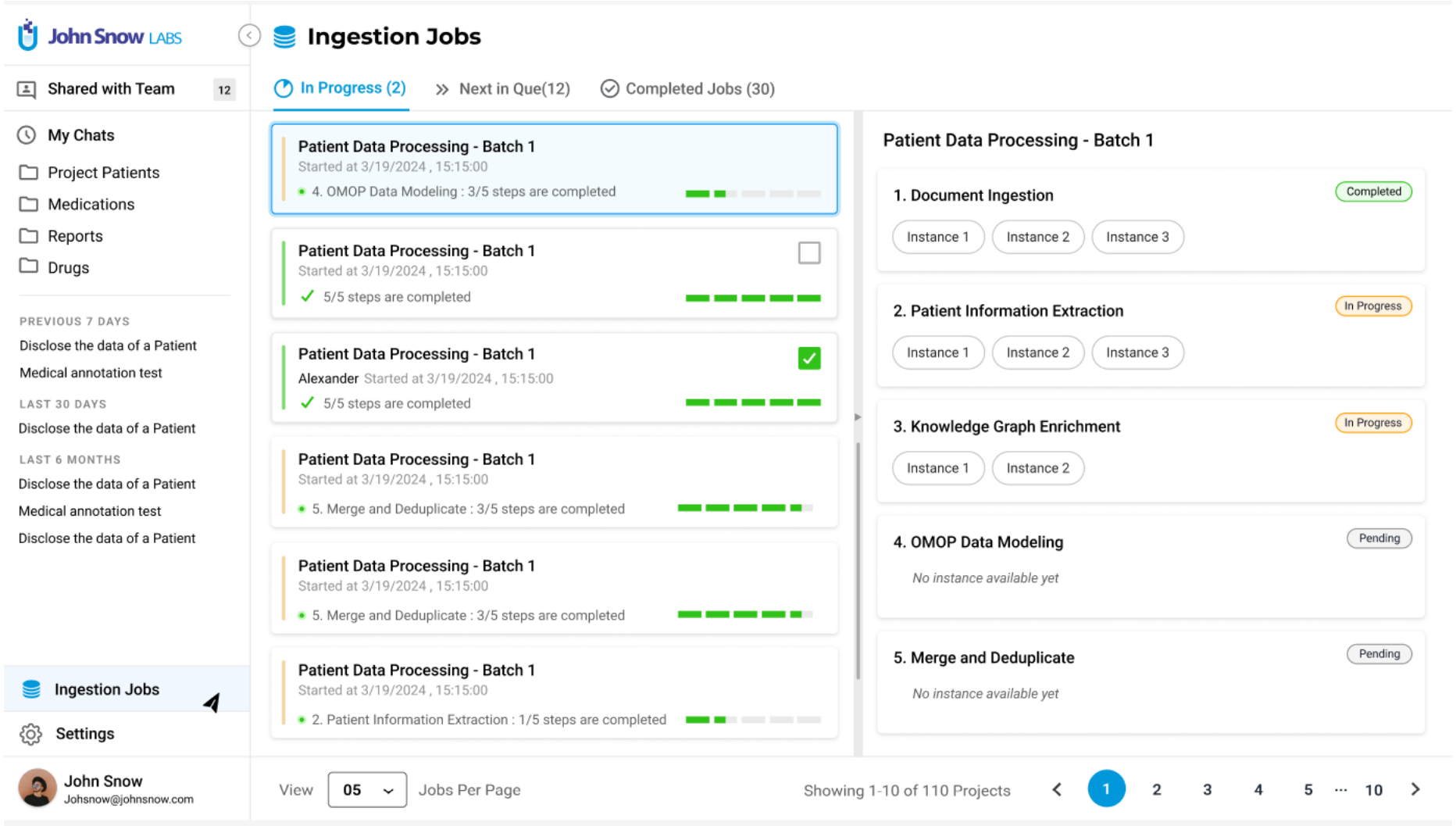


https://ohdsi.github.io/CommonDataModel/cdm54.html

CDM v5.4

- 39 tables
- 433 fields
- 7 categories

Data Ingestion



Cohort Building from Prompts

What would you like to know?

Show patients who were diagnosed with back pain and who have had spinal fusion

	condition_name	condition_source	procedure_name
0	Backache	EHR billing record	Lumbar and lumbosacral fusion of the posterior column
1	Backache	EHR billing record	Lumbar and lumbosacral fusion of the posterior column
2	Backache	EHR billing record	Fusion of 2 or more Lumbar Vertebral Joints with Autologo
3	Backache	EHR billing record	Fusion of 2 or more Lumbar Vertebral Joints with Synthetic
4	Backache	EHR billing record	Fusion of Lumbosacral Joint with Autologous Tissue Subst
5	Backache	14260897-DS-5	Arthrodesis
6	Backache	14260897-DS-5	Lumbar and lumbosacral fusion of the posterior column
7	Low back pain	EHR billing record	Fusion of Lumbar Vertebral Joint with Interbody Fusion De
8	Low back pain	EHR billing record	Lumbar and lumbosacral fusion by anterior technique
9	Low back pain	EHR billing record	Lumbar and lumbosacral fusion of the posterior column

Behind the scenes: Multi-agent system

Find patients diagnosed with back pain that have had spinal fusion

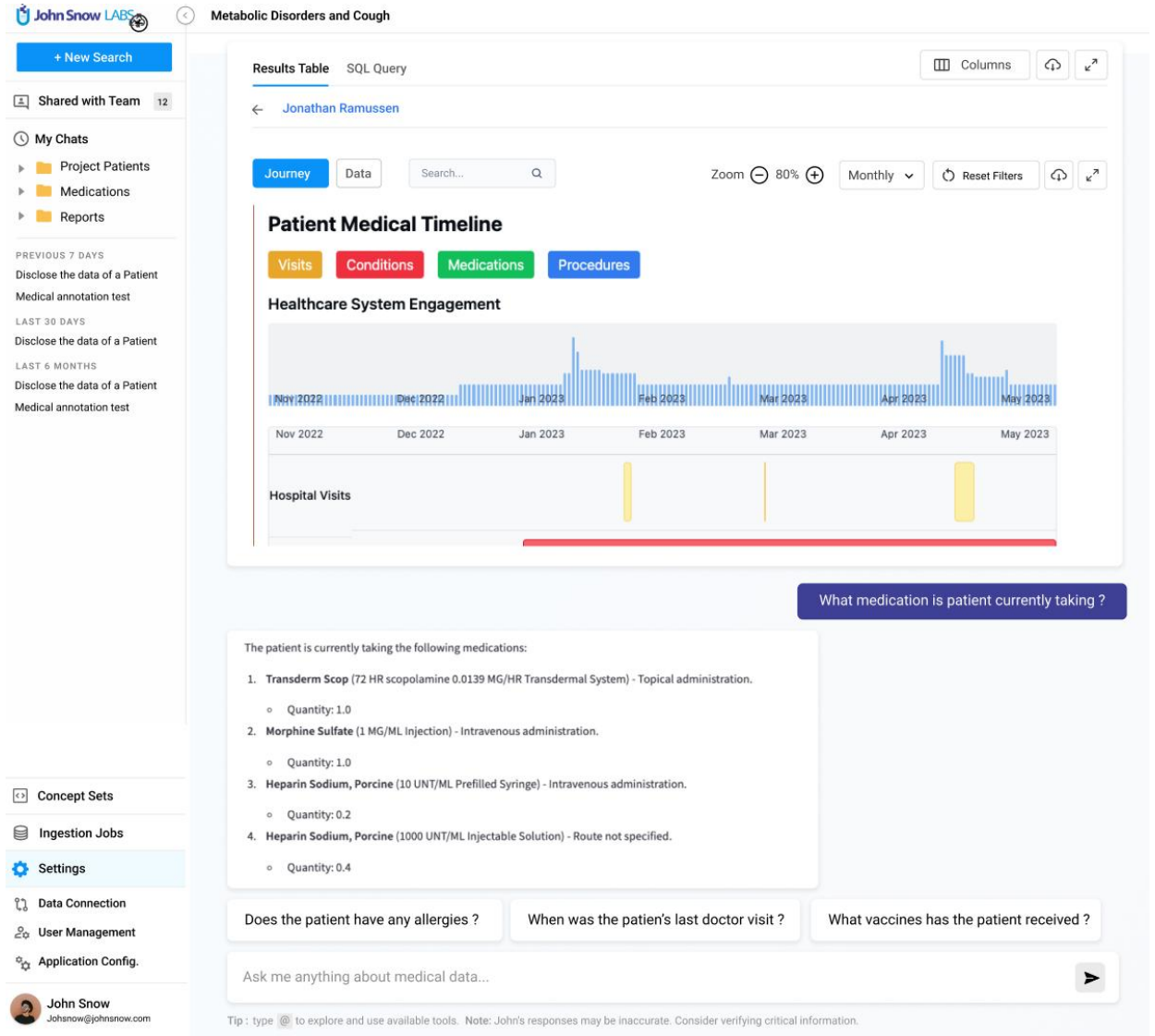
Concept resolver: Find concept id for given entity.

Back pain: Condition (SNOMED 194133)
Spinal fusion: Procedure (SNOMED 4177164)

Build query for OMOP CDM

Retrieve records and make reply

Patient Level Analysis: Chat With Your Data

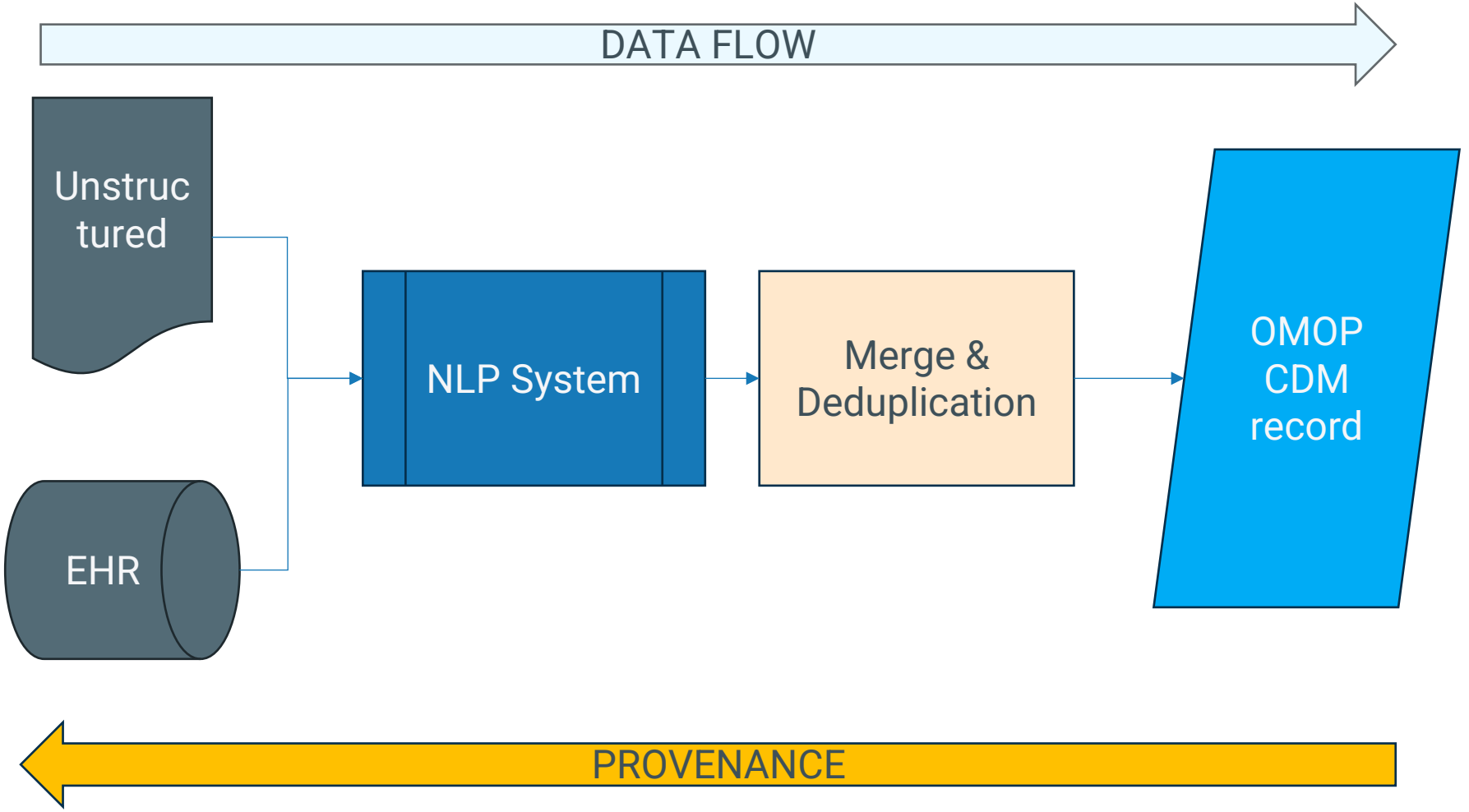


Track Provenance of Information

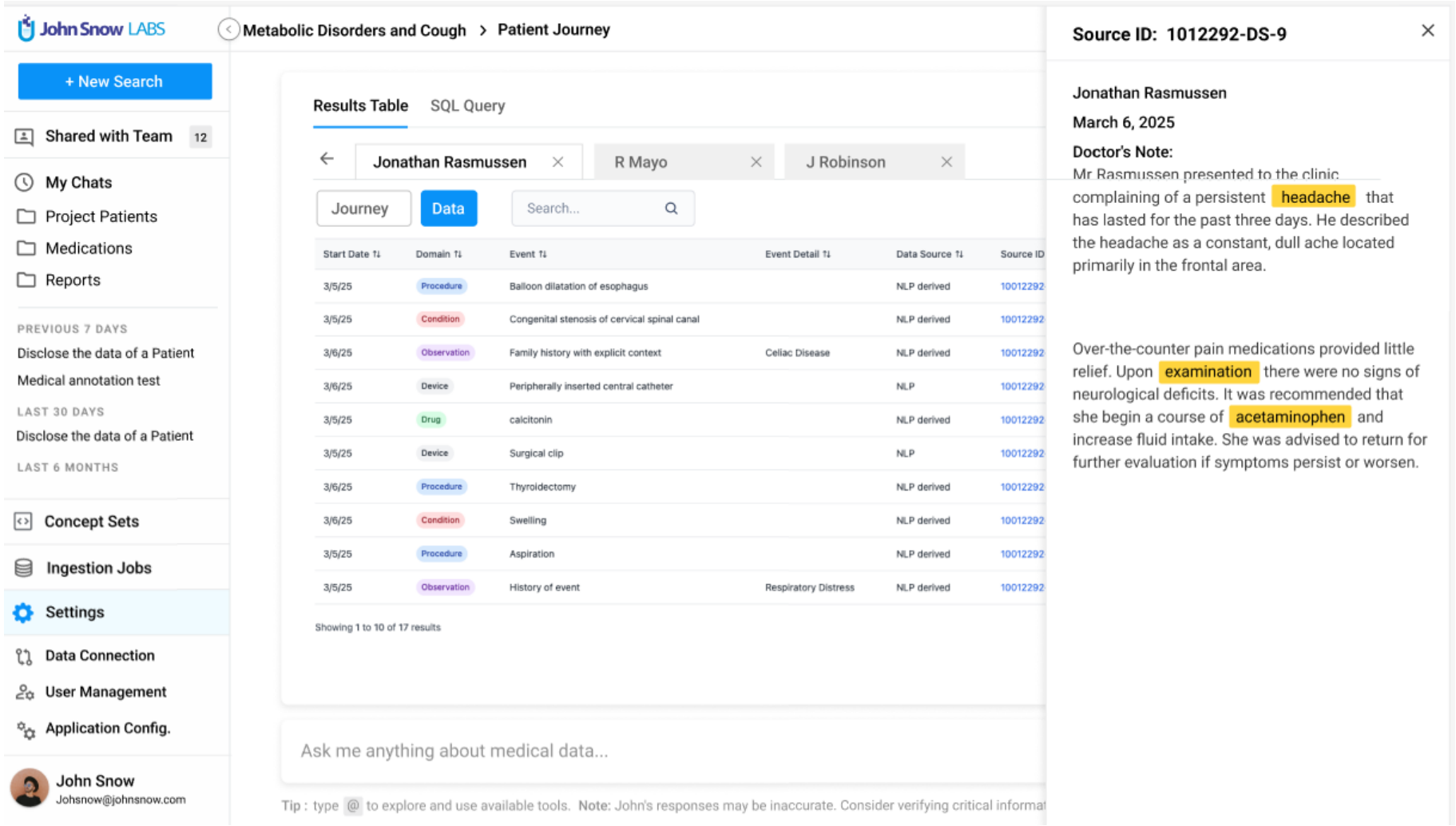
Discharge Diagnosis:

PRIMARY: Dyspnea, chest pain
SECONDARY: chronic obstructive pulmonary disease, coronary
artery disease, hyponatremia, hyperkalemia, hypertension

condition_occurrence_id [PK] integer	person_id integer	condition_concept_id integer	condition_start_date date
8	10	316866	2142-10-21



Track Provenance of Information



- EHR System
- FHIR document
- Raw clinical notes
- Other

OMOP CDM (PostgreSQL)

Records for one patient

table_name name	row_count bigint
note_nlp	5852
observation	211
visit_occurrence	151
note	151
condition_occurrence	138
measurement	76
person	11
procedure_occurrence	11
drug_exposure	8

OMOP CDM (PostgreSQL)

Can query the DB directly

```
SELECT
    condition_occurrence_id, person_id, visit_occurrence_id, condition_concept_id, condition_source_value
FROM
    condition_occurrence
ORDER BY
    condition_occurrence_id
```

Output Messages Notifications

SQL

condition_occurrence_id [PK] integer	person_id integer	visit_occurrence_id integer	condition_concept_id integer	condition_source_value character varying (50)
1	1	151	4195039	osteopenia
2	2	150	4195039	Osteopenia
3	3	132	197236	fibroids
4	5	113	197676	Hepatomegaly
5	5	113	197676	Hepatomegaly
6	6	94	253796	pneumothorax
7	6	94	253796	pneumothorax
8	6	94	4145627	Cholelithiasis
9	2	70	254761	Cough
10	8	120	4061577	lethargy
11	6	94	261880	atelectasis
12	3	111	441408	vomiting
13	3	151	4195039	Osteopenia
14	2	106	75860	constipation
15	2	106	254761	cough
16	6	151	254761	cough
17	2	138	254761	cough
18	2	70	254761	Cough