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

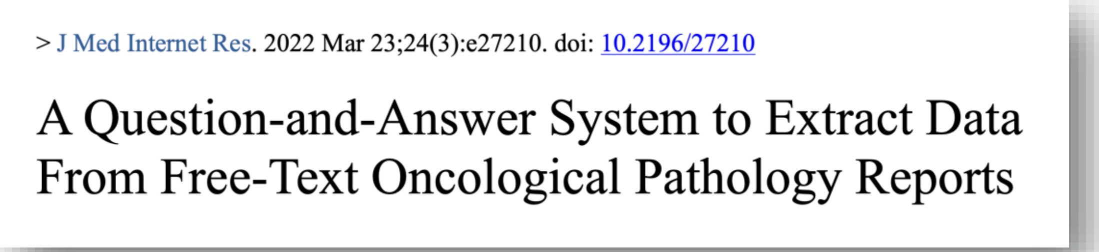
Not Consistent

Not Certain

Continuously Updating

TXT
ComputerHope.com
EHR
PDF
CSV
FHIR
API
DICOM
Digital Imaging and Communications in Medicine
HL7

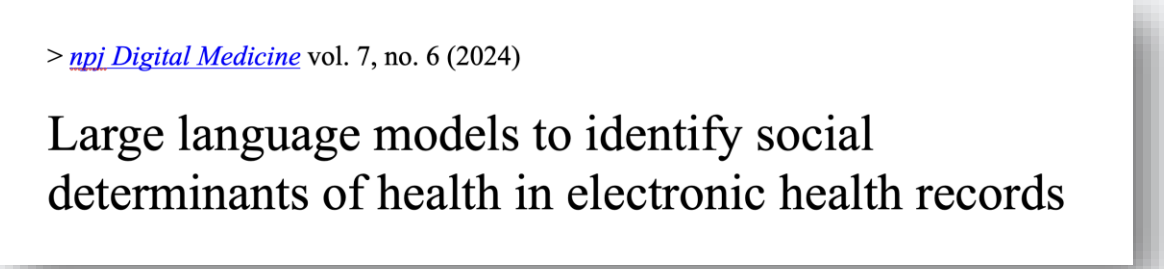
Some Clinical Data is Only Available in Unstructured Text



> 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

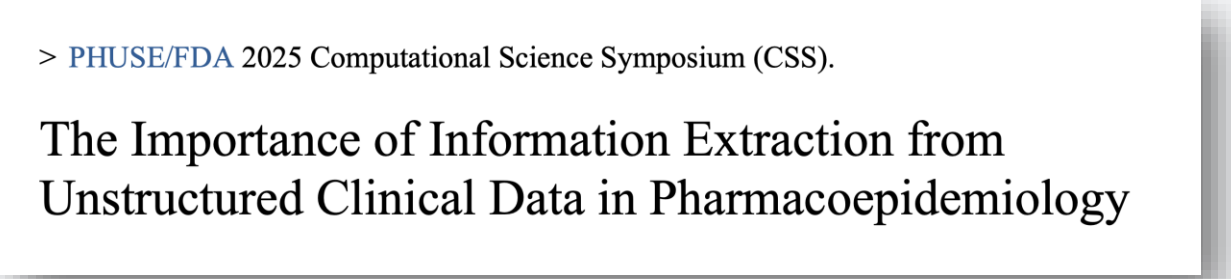
“Accuracies for predicting group-level site and histology codes were 93.5% and 97.6% respectively.”



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

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

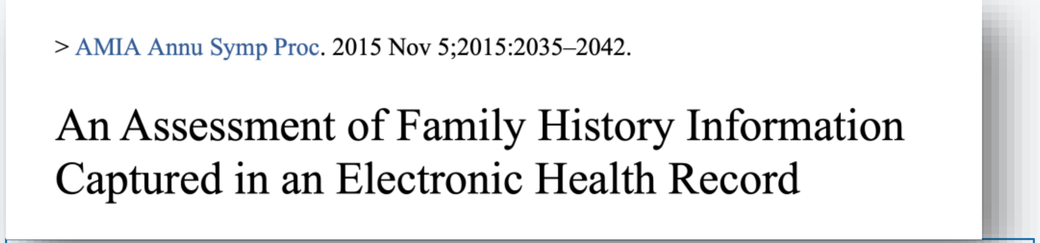
“Our models identified 93.8% of patients with adverse SDoH, while ICD-10 codes captured 2.0%.”



> 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.”

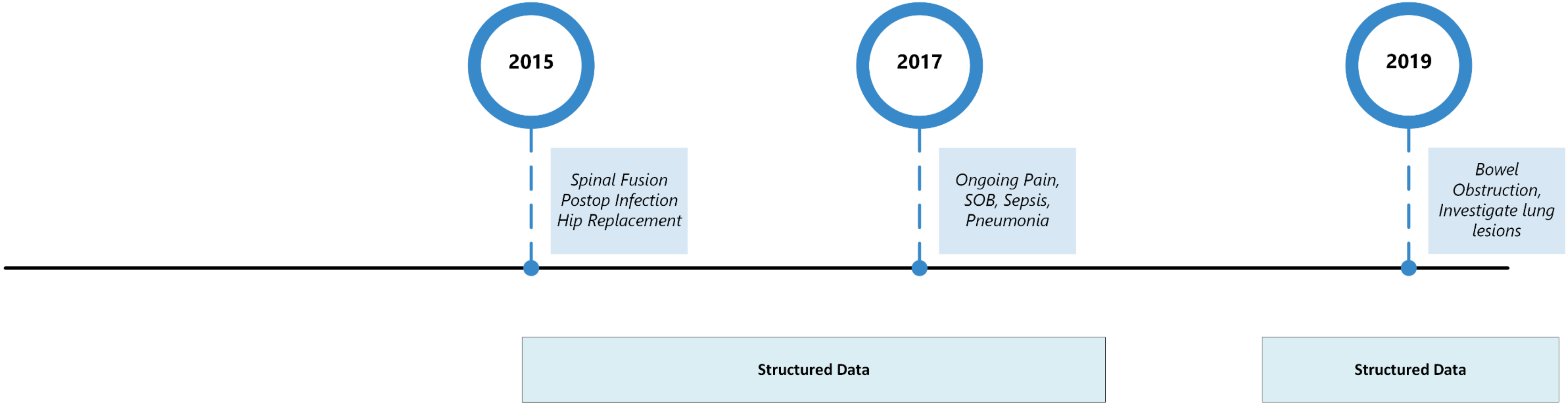


> AMIA Annu Symp Proc. 2015 Nov 5;2015:2035–2042.

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



2015

Spinal Fusion
Postop Infection
Hip Replacement

2017

Ongoing Pain, SOB, Sepsis, Pneumonia

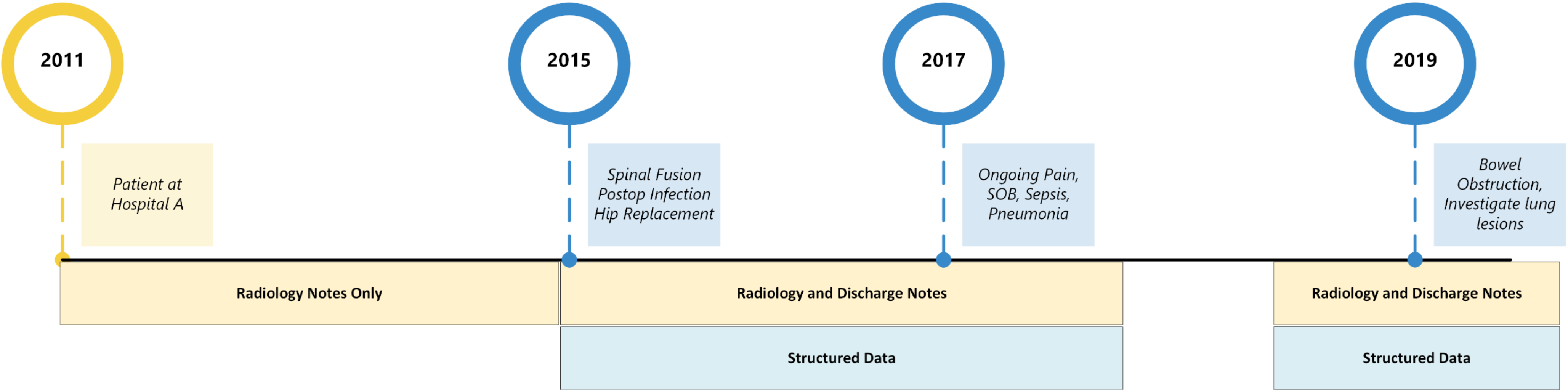
2019

Bowel Obstruction, Investigate lung lesions

Structured Data

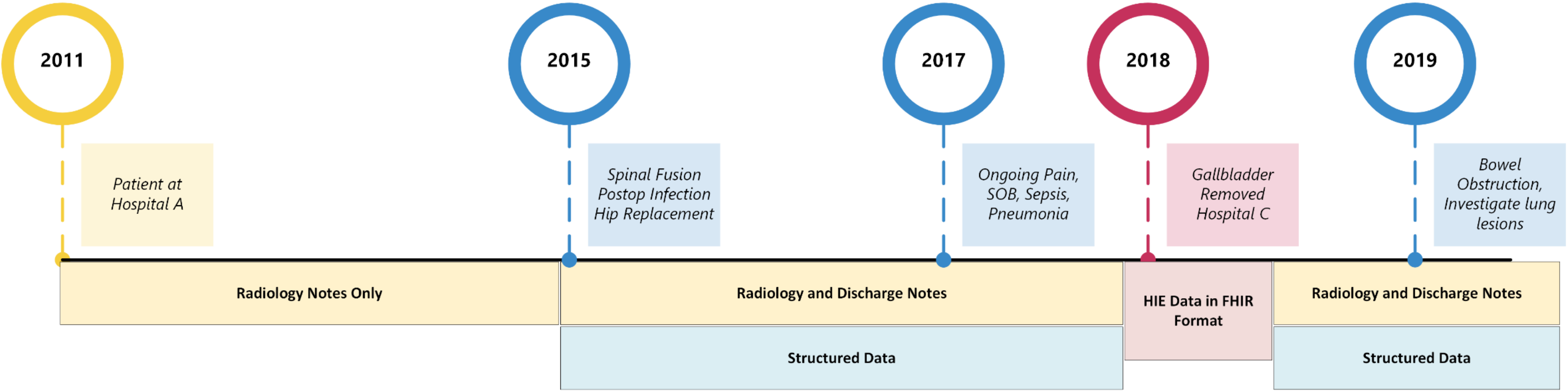
Structured Data

Adding Radiology & Discharge Notes



2011
Patient at Hospital A
2015
Spinal Fusion Postop Infection Hip Replacement
2017
Ongoing Pain, SOB, Sepsis, Pneumonia
2019
Bowel Obstruction, Investigate lung lesions
Radiology Notes Only
Radiology and Discharge Notes
Structured Data
Radiology and Discharge Notes
Structured Data

Adding FHIR Resources



2011

Patient at Hospital A

2015

Spinal Fusion Postop Infection Hip Replacement

2017

Ongoing Pain, SOB, Sepsis, Pneumonia

2018

Gallbladder Removed Hospital C

2019

Bowel Obstruction, Investigate lung lesions

Radiology Notes Only

Radiology and Discharge Notes

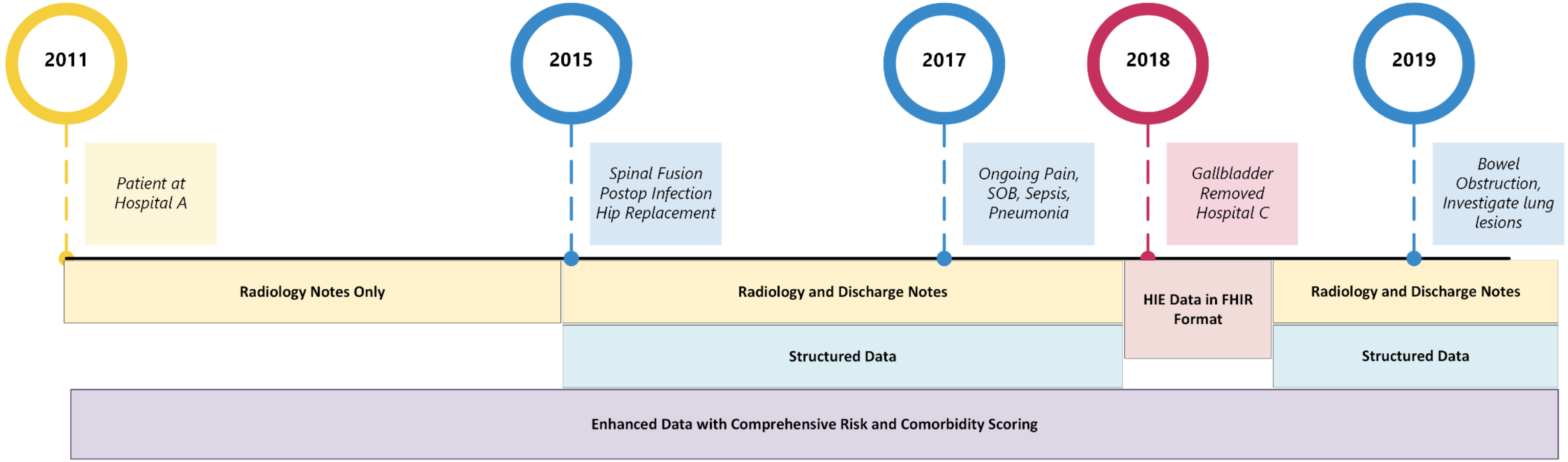
Structured Data

HIE Data in FHIR Format

Radiology and Discharge Notes

Structured Data

Enriching the Data by Calculating Risk Scores


2011
Patient at Hospital A
2015
Spinal Fusion Postop Infection Hip Replacement
2017
Ongoing Pain, SOB, Sepsis, Pneumonia
2018
Gallbladder Removed Hospital C
2019
Bowel Obstruction, Investigate lung lesions
Radiology Notes Only
Radiology and Discharge Notes
HIE Data in FHIR Format
Radiology and Discharge Notes
Structured Data
Structured Data
Enhanced Data with Comprehensive Risk and Comorbidity Scoring

Current LLMs Can’t Work on Complex Queries

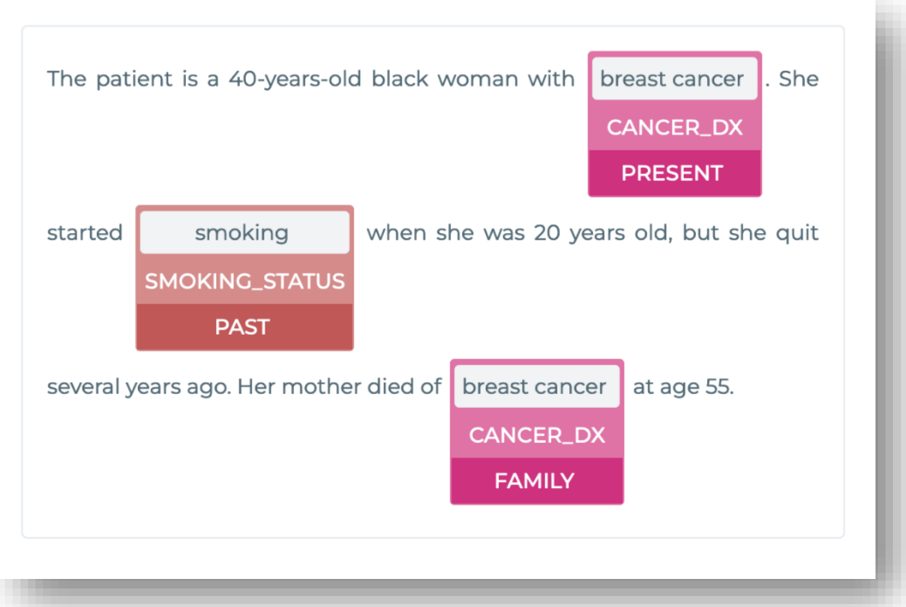
Q: “Find patients diagnosed with back pain that have had spinal fusion.”

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

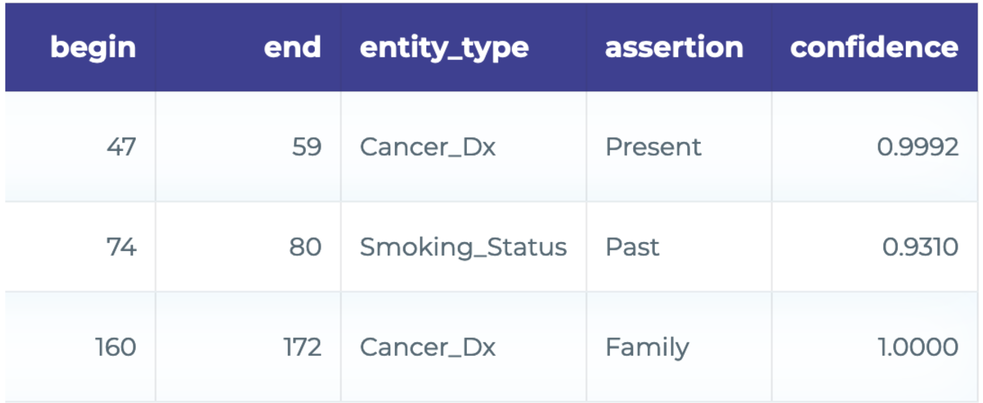
```
WITH
-- Identify tuberculosis diagnosis concepts
tb_diagnosis_concepts AS (
   SELECT c.concept_id
   FROM concept c
   JOIN concept_ancestor ca ON ca.descendant_concept_id = c.concept_id
   WHERE ca.ancestor_concept_id IN (
          -- Add the root concept ID for tuberculosis and its descendants
          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

Semantic Information Extraction (Healthcare NLP)

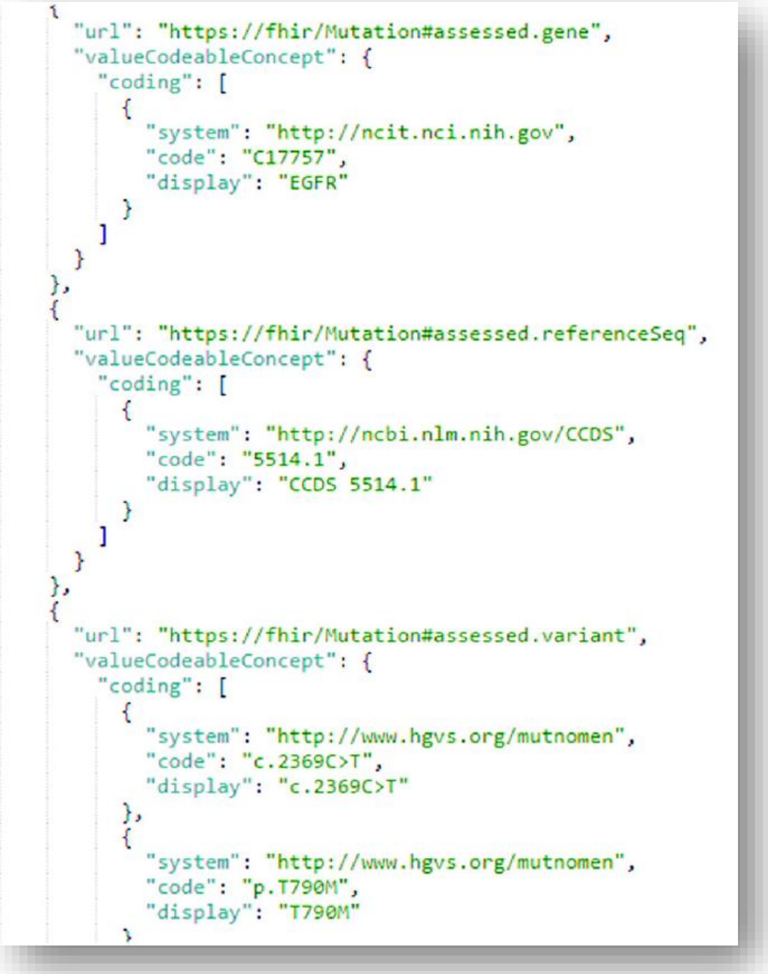


The patient is a 40-years-old black woman with breast cancer. She started smoking when she was 20 years old, but she quit several years ago. Her mother died of breast cancer at age 55.

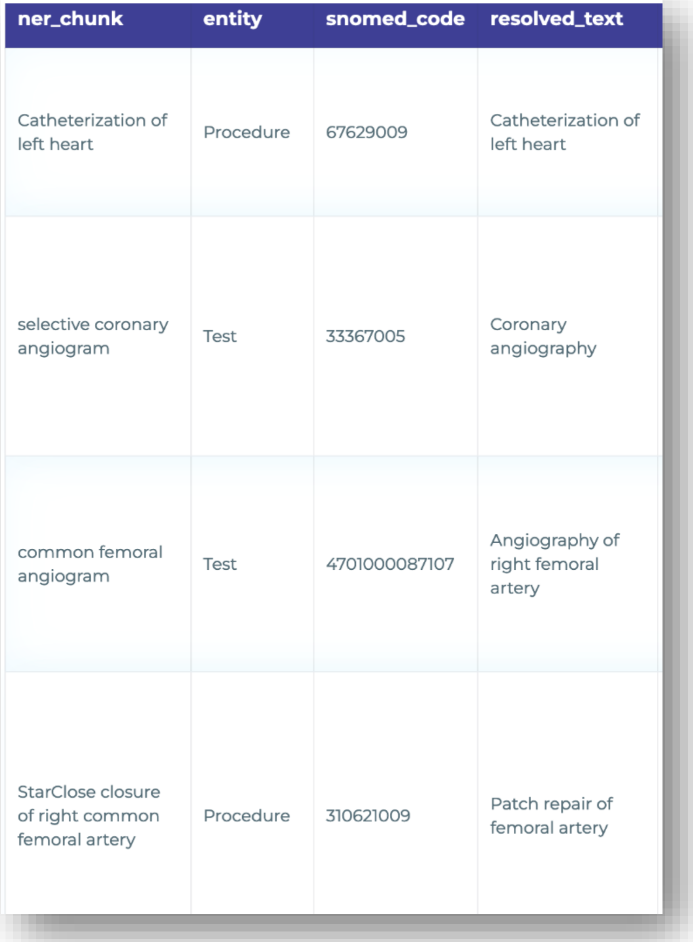


begin	end	entity_type	assertion	confidence
47	59	Cancer_Dx	Present	0.9992
74	80	Smoking_Status	Past	0.9310
160	172	Cancer_Dx	Family	1.0000

Terminology Server: Resolving to Standard Codes

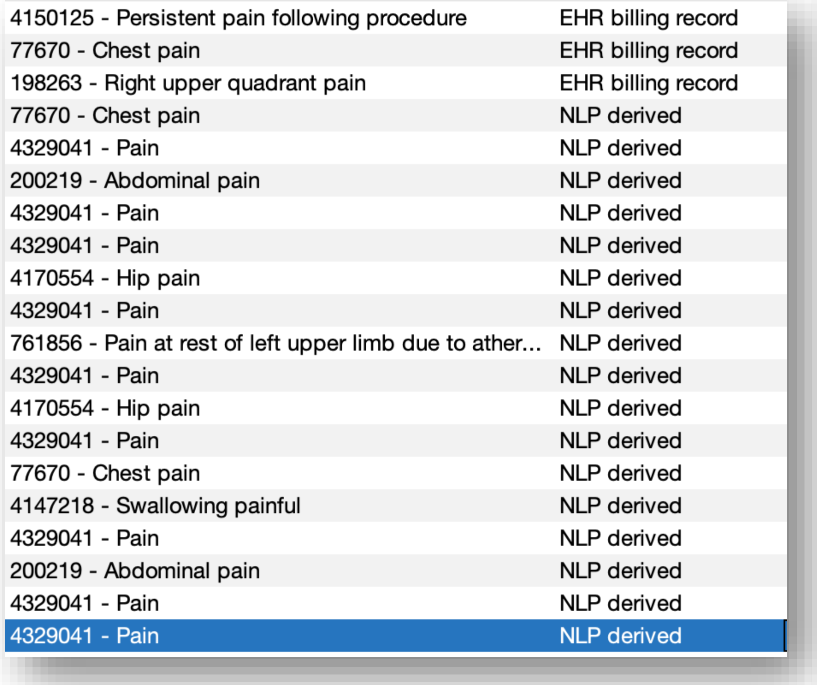


```
  {
    "url": "https://fhir/Mutation#assessed.gene",
    "valueCodeableConcept": {
      "coding": [
        {
          "system": "http://ncit.nci.nih.gov",
          "code": "C17757",
          "display": "EGFR"
        }
      ]
    }
  },
  {
    "url": "https://fhir/Mutation#assessed.referenceSeq",
    "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"
```

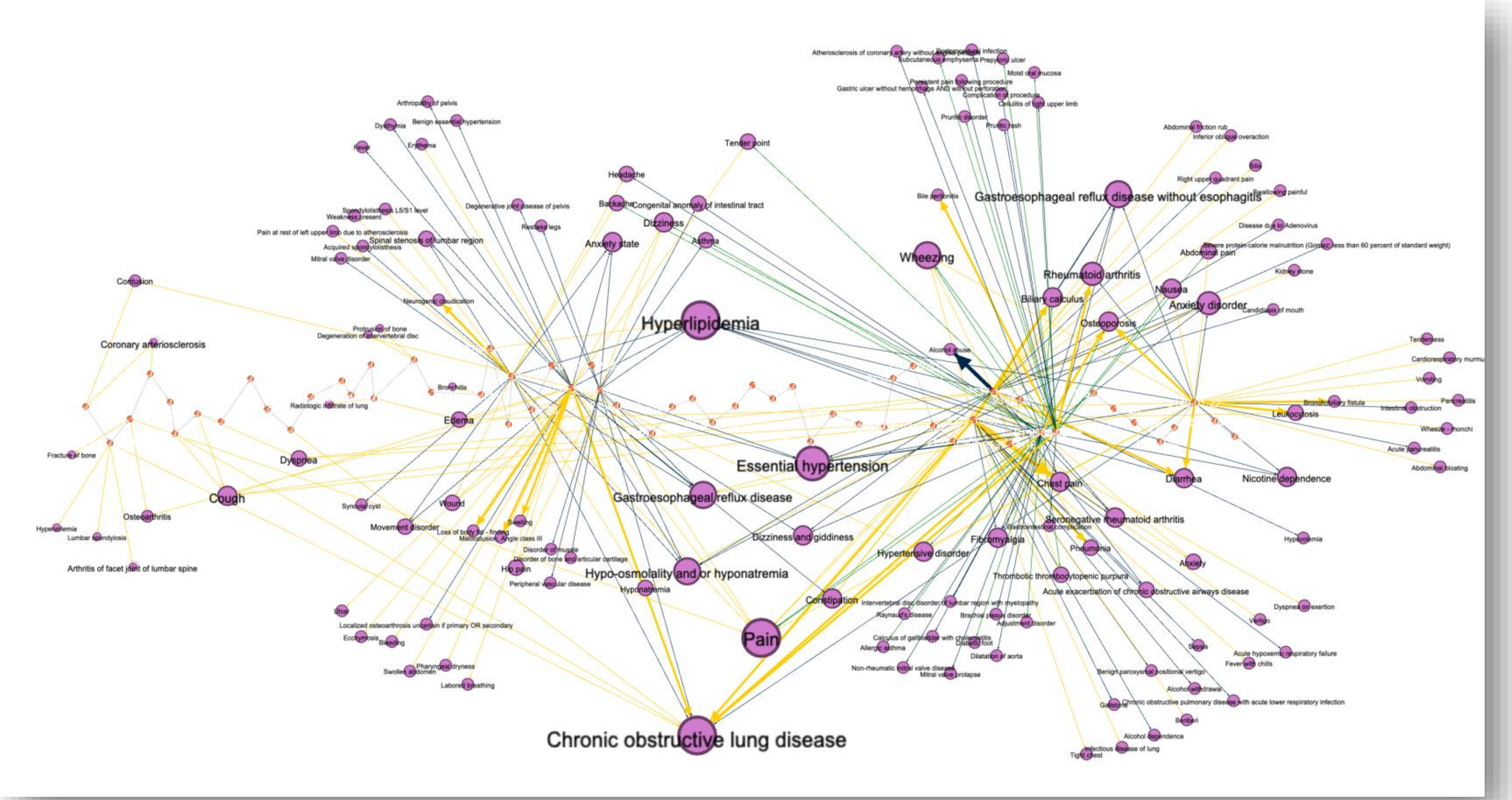


ner_chunk	entity	snomed_code	resolved_text
Catheterization of left heart	Procedure	67629009	Catheterization of left heart
selective coronary angiogram	Test	33367005	Coronary angiography
common femoral angiogram	Test	4701000087107	Angiography of right femoral artery
StarClose closure of right common femoral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

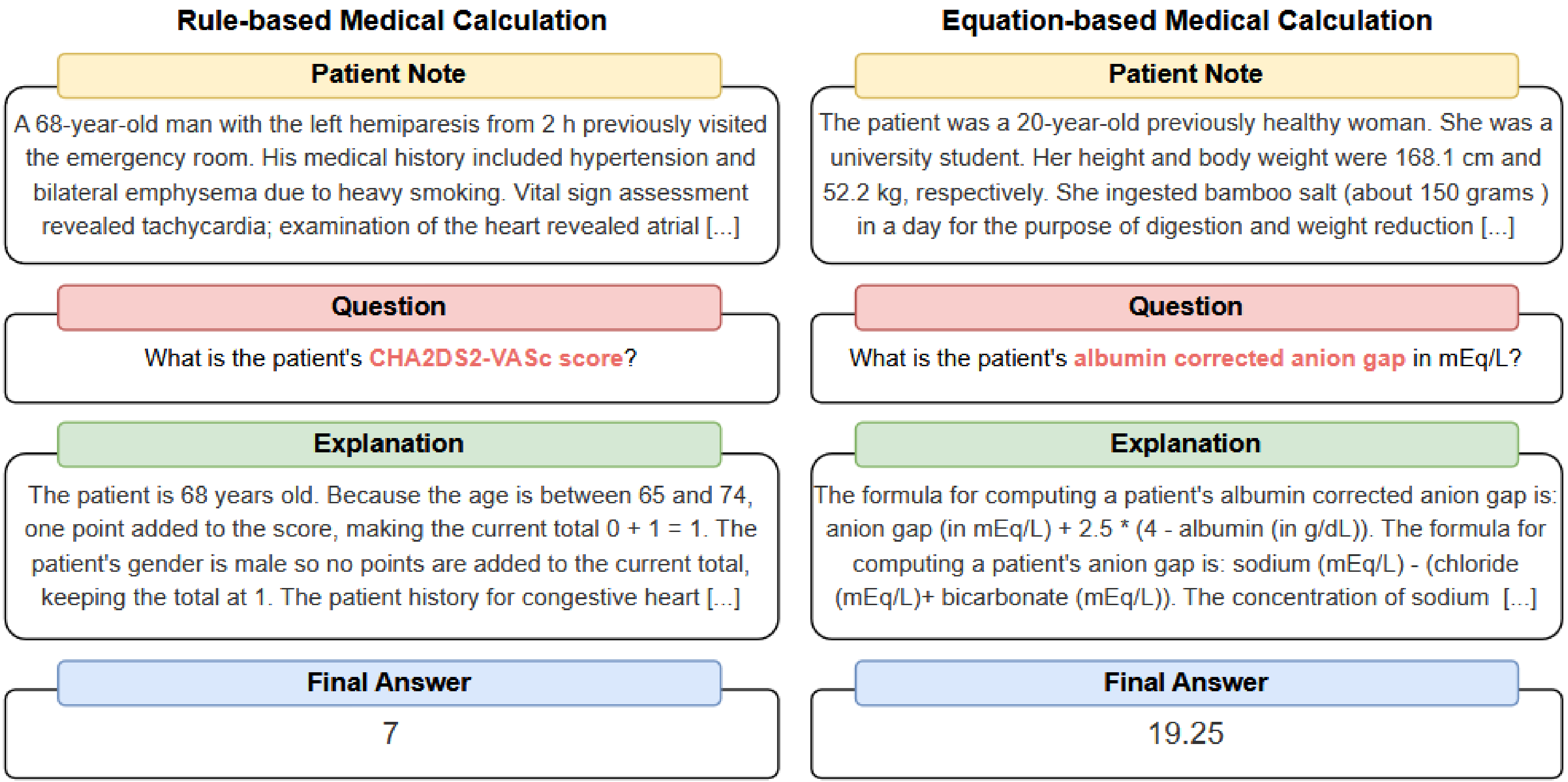
Merging & Deduplicating Facts to Build a Patient Graph



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


Atherosclerosis of coronary artery without angina pectoris
Postprocedural infection
Subcutaneous emphysema
Prepyloric ulcer
Moist oral mucosa
Persistent pain following procedure
Gastric ulcer without hemorrhage AND without perforation
Complication of procedure
Cellulitis of right upper limb
Pruritic disorder
Pruritic rash
Arthropathy of pelvis
Dysthymia
Benign essential hypertension
Fever
Erythema
Tender point
Headache
Degenerative joint disease of pelvis
Backache
Congenital anomaly of intestinal tract
Spondylolisthesis L5/S1 level
Weakness present
Restless legs
Dizziness
Pain at rest of left upper limb due to atherosclerosis
Spinal stenosis of lumbar region
Acquired spondylolisthesis
Anxiety state
Asthma
Mitral valve disorder
Abdominal friction rub
Inferior oblique overaction
Bile peritonitis
Gastroesophageal reflux disease without esophagitis
Bilis
Right upper quadrant pain
Swallowing painful
Disease due to Adenovirus
Severe protein-calorie malnutrition (Gomez: less than 60 percent of standard weight)
Abdominal pain
Kidney stone
Wheezing
Rheumatoid arthritis
Nausea
Biliary calculus
Anxiety disorder
Candidiasis of mouth
Osteoporosis
Contusion
Neurogenic claudication
Protrusion of bone
Degeneration of intervertebral disc
Hyperlipidemia
Coronary arteriosclerosis
Alcohol abuse
Tenderness
Cardiorespiratory murmur
Vomiting
Bronchitis
Radiologic infiltrate of lung
Bronchobiliary fistula
Intestinal obstruction
Pancreatitis
Leukocytosis
Edema
Wheeze - rhonchi
Fracture of bone
Acute pancreatitis
Dyspnea
Abdominal bloating
Essential hypertension
Chest pain
Diarrhea
Nicotine dependence
Cough
Gastroesophageal reflux disease
Synovial cyst
Wound
Osteoarthritis
Swelling
Seronegative rheumatoid arthritis
Gastrointestinal complication
Movement disorder
Loss of body fat - finding
Malocclusion, Angle class III
Hyperlipidemia
Lumbar spondylosis
Dizziness and giddiness
Fibromyalgia
Pneumonia
Hypovolemia
Disorder of muscle
Disorder of bone and articular cartilage
Hypertensive disorder
Anxiety
Arthritis of facet joint of lumbar spine
Hip pain
Hypo-osmolality and or hyponatremia
Peripheral vascular disease
Thrombotic thrombocytopenic purpura
Hyponatremia
Acute exacerbation of chronic obstructive airways disease
Ulcer
Constipation
Intervertebral disc disorder of lumbar region with myelopathy
Dyspnea on exertion
Localized osteoarthrosis uncertain if primary OR secondary
Raynaud's disease
Brachial plexus disorder
Vertigo
Adjustment disorder
Ecchymosis
Bleeding
Pain
Calculus of gallbladder with cholecystitis
Diabetic foot
Sepsis
Allergic asthma
Acute hypoxemic respiratory failure
Dilatation of aorta
Fever with chills
Pharyngeal dryness
Swollen abdomen
Non-rheumatic mitral valve disease
Mitral valve prolapse
Benign paroxysmal positional vertigo
Labored breathing
Alcohol withdrawal
Gallstone
Chronic obstructive pulmonary disease with acute lower respiratory infection
Beriberi
Chronic obstructive lung disease
Alcohol dependence
Infectious disease of lung
Tight chest

Making Clinical Inferences and Calculations



Rule-based Medical Calculation

Patient Note

A 68-year-old man with the left hemiparesis from 2 h previously visited the emergency room. His medical history included hypertension and bilateral emphysema due to heavy smoking. Vital sign assessment revealed tachycardia; examination of the heart revealed atrial [...]

Question

What is the patient's **CHA2DS2-VASc score**?

Explanation

The patient is 68 years old. Because the age is between 65 and 74, one point added to the score, making the current total 0 + 1 = 1. The patient's gender is male so no points are added to the current total, keeping the total at 1. The patient history for congestive heart [...]

Final Answer

7

Equation-based Medical Calculation

Patient Note

The patient was a 20-year-old previously healthy woman. She was a university student. Her height and body weight were 168.1 cm and 52.2 kg, respectively. She ingested bamboo salt (about 150 grams) in a day for the purpose of digestion and weight reduction [...]

Question

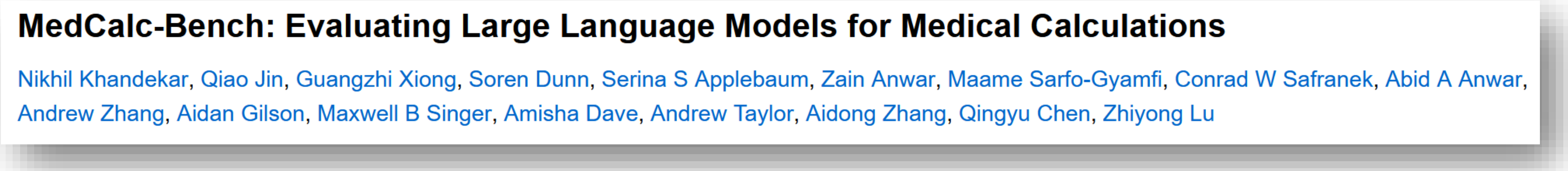
What is the patient's **albumin corrected anion gap** in mEq/L?

Explanation

The formula for computing a patient's albumin corrected anion gap is: anion gap (in mEq/L) + 2.5 * (4 - albumin (in g/dL)). The formula for computing a patient's anion gap is: sodium (mEq/L) - (chloride (mEq/L)+ bicarbonate (mEq/L)). The concentration of sodium [...]

Final Answer

19.25



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.

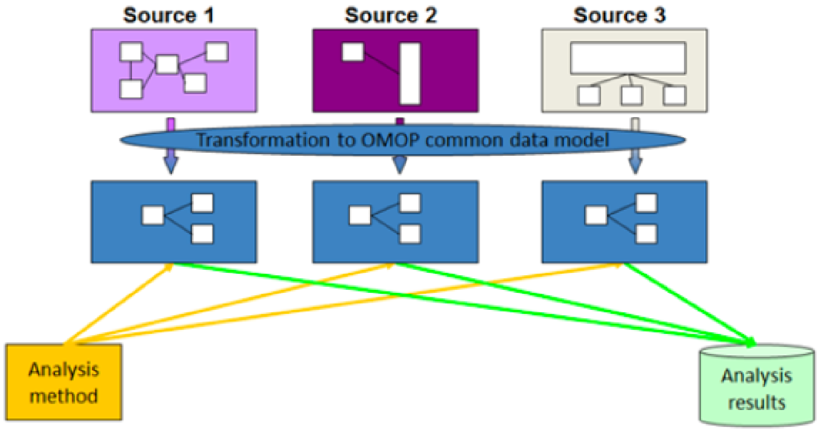
Methods:

- 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.

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

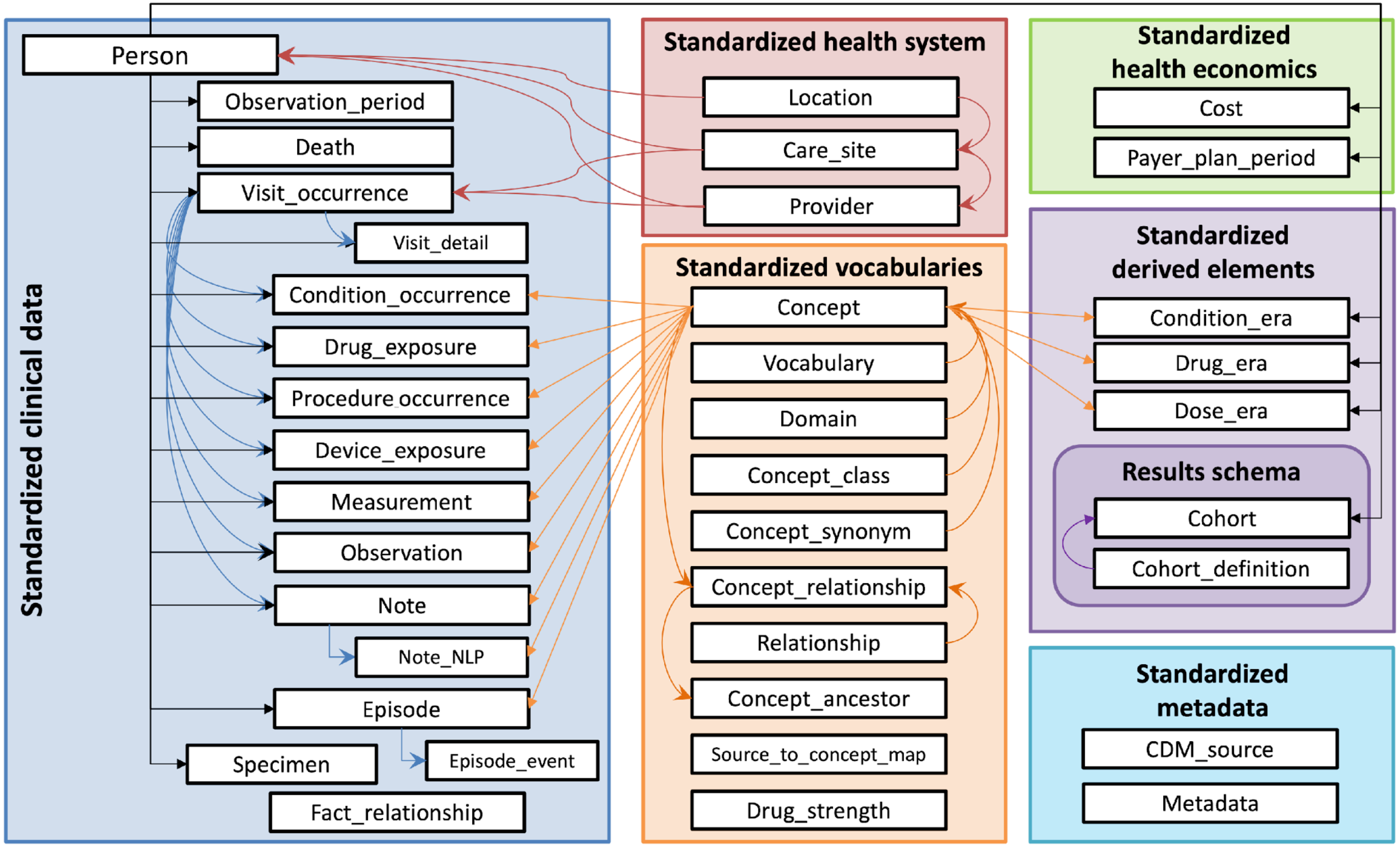
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

"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"



The OMOP CDM

Observational Medical Outcomes Partnership - Common Data Model



Standardized clinical data

Person

Observation_period

Death

Visit_occurrence

Visit_detail

Condition_occurrence

Drug_exposure

Procedure occurrence

Device_exposure

Measurement

Observation

Note

Note_NLP

Episode

Specimen

Episode_event

Fact_relationship

Standardized health system

Location

Care_site

Provider

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_synonym

Concept_relationship

Relationship

Concept_ancestor

Source_to_concept_map

Drug_strength

Standardized health economics

Cost

Payer_plan_period

Standardized derived elements

Condition_era

Drug_era

Dose_era

Results schema

Cohort

Cohort_definition

Standardized metadata

CDM_source

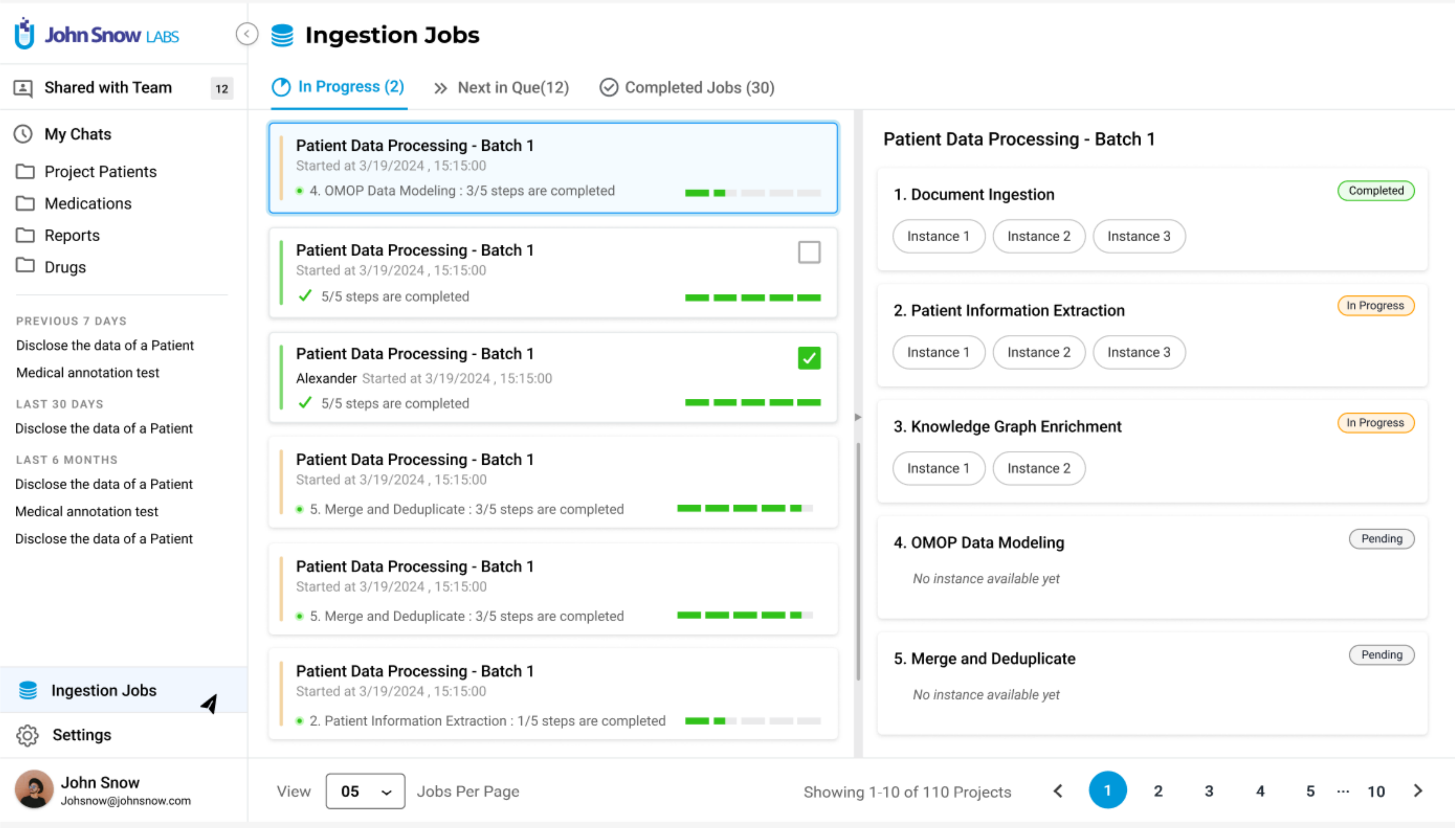
Metadata

CDM v5.4

- ➢ 39 tables
- ➢ 433 fields
- ➢ 7 categories

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

Data Ingestion



John Snow LABS

Shared with Team 12

My Chats

Project Patients

Medications

Reports

Drugs

PREVIOUS 7 DAYS

Disclose the data of a Patient

Medical annotation test

LAST 30 DAYS

Disclose the data of a Patient

LAST 6 MONTHS

Disclose the data of a Patient

Medical annotation test

Disclose the data of a Patient

Ingestion Jobs

Settings

John Snow
Johsnow@johnsnow.com

Ingestion Jobs

In Progress (2) | Next in Que(12) | Completed Jobs (30)

Patient Data Processing - Batch 1
Started at 3/19/2024 , 15:15:00
● 4. OMOP Data Modeling : 3/5 steps are completed

Patient Data Processing - Batch 1 □
Started at 3/19/2024 , 15:15:00
✓ 5/5 steps are completed

Patient Data Processing - Batch 1 ☑
Alexander Started at 3/19/2024 , 15:15:00
✓ 5/5 steps are completed

Patient Data Processing - Batch 1
Started at 3/19/2024 , 15:15:00
● 5. Merge and Deduplicate : 3/5 steps are completed

Patient Data Processing - Batch 1
Started at 3/19/2024 , 15:15:00
● 5. Merge and Deduplicate : 3/5 steps are completed

Patient Data Processing - Batch 1
Started at 3/19/2024 , 15:15:00
● 2. Patient Information Extraction : 1/5 steps are completed

Patient Data Processing - Batch 1

1. Document Ingestion — Completed
Instance 1 | Instance 2 | Instance 3

2. Patient Information Extraction — In Progress
Instance 1 | Instance 2 | Instance 3

3. Knowledge Graph Enrichment — In Progress
Instance 1 | Instance 2

4. OMOP Data Modeling — Pending
No instance available yet

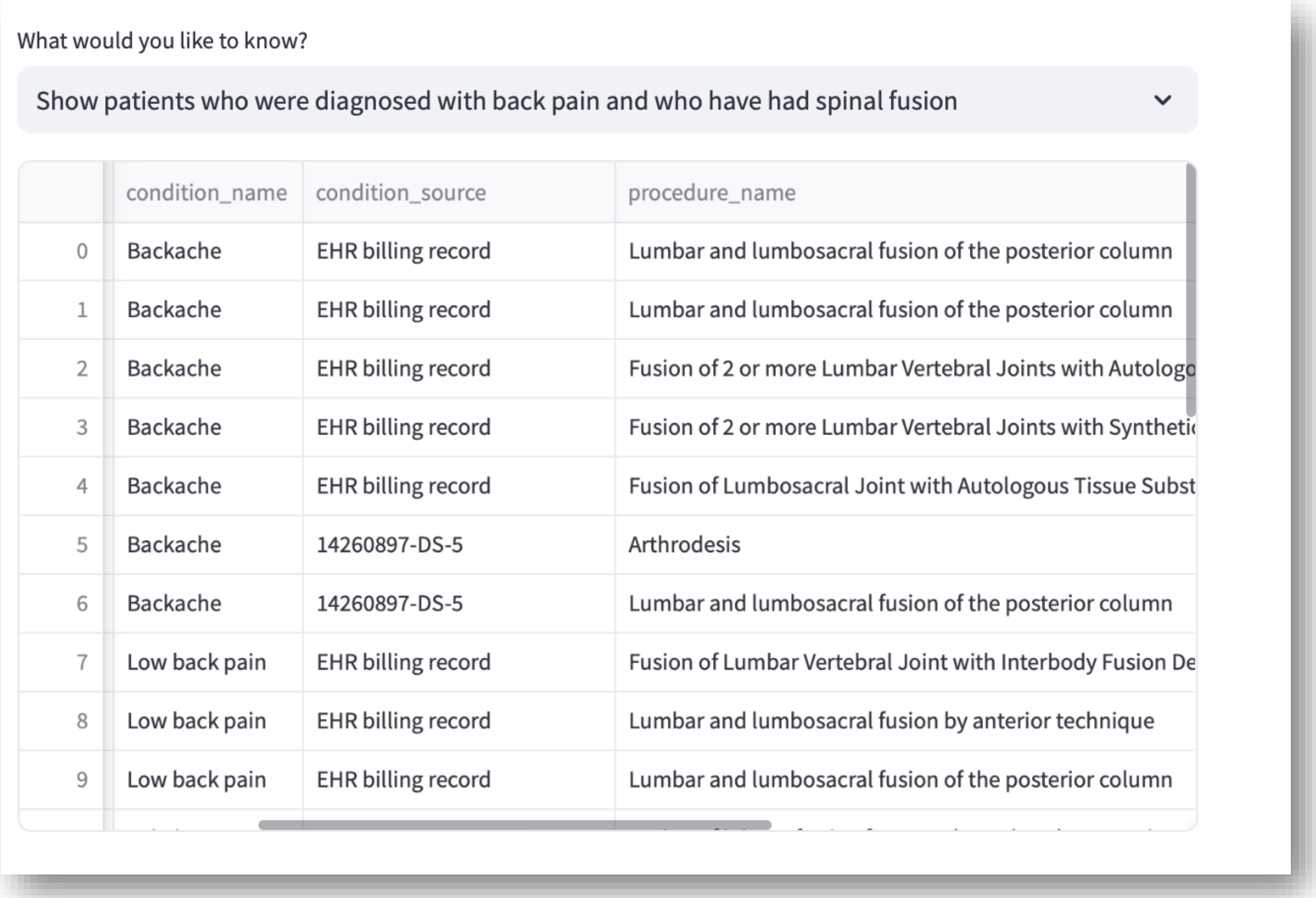
5. Merge and Deduplicate — Pending
No instance available yet

View 05 Jobs Per Page

Showing 1-10 of 110 Projects

< 1 2 3 4 5 ··· 10 >

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 Syntheti
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

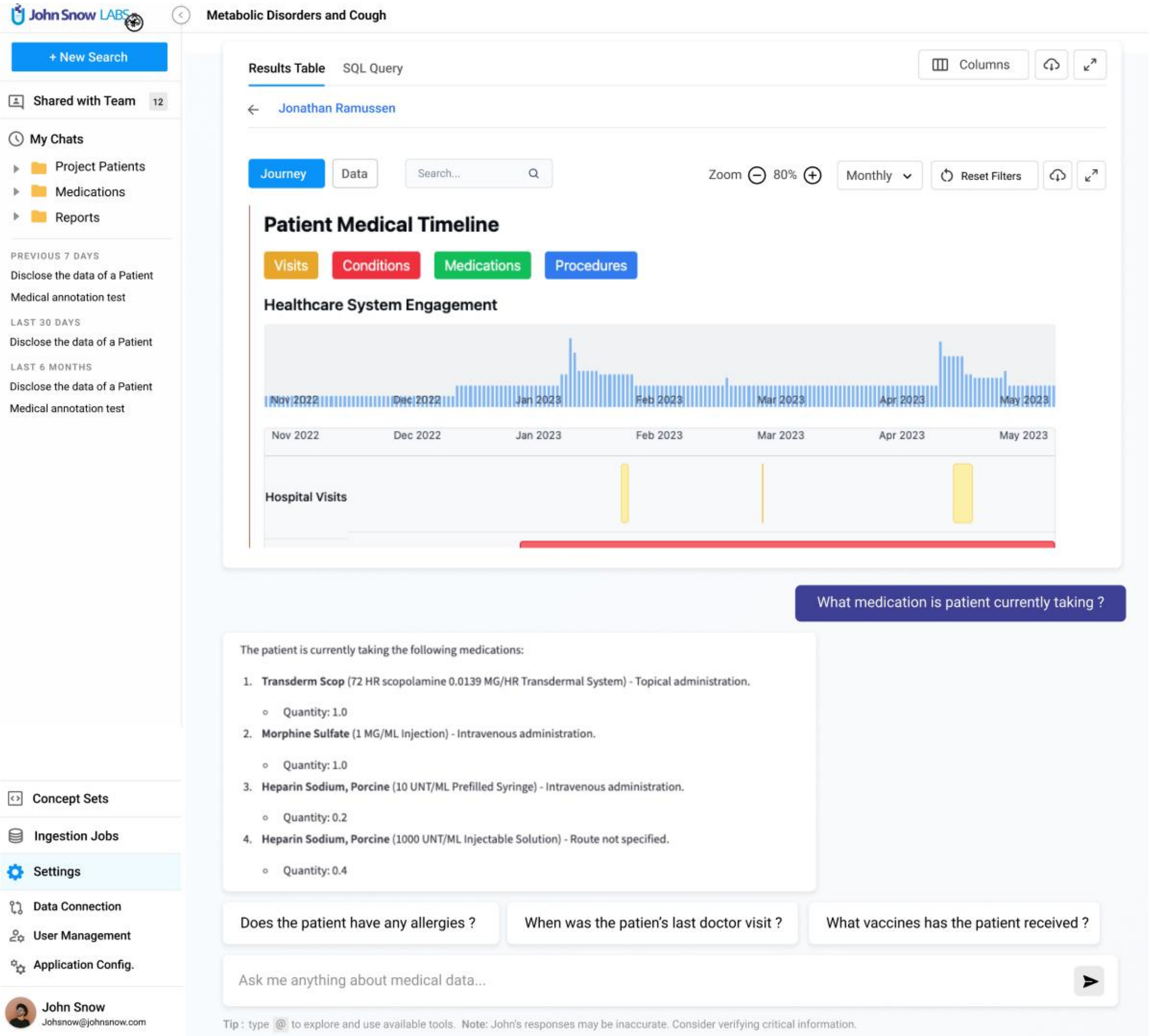
Concept resolver: Find concept id for given entity.

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



What medication is patient currently taking ?

The patient is currently taking the following medications:

1. **Transderm Scop** (72 HR scopolamine 0.0139 MG/HR Transdermal System) - Topical administration.
   - Quantity: 1.0
2. **Morphine Sulfate** (1 MG/ML Injection) - Intravenous administration.
   - Quantity: 1.0
3. **Heparin Sodium, Porcine** (10 UNT/ML Prefilled Syringe) - Intravenous administration.
   - Quantity: 0.2
4. **Heparin Sodium, Porcine** (1000 UNT/ML Injectable Solution) - Route not specified.
   - Quantity: 0.4

Does the patient have any allergies ?

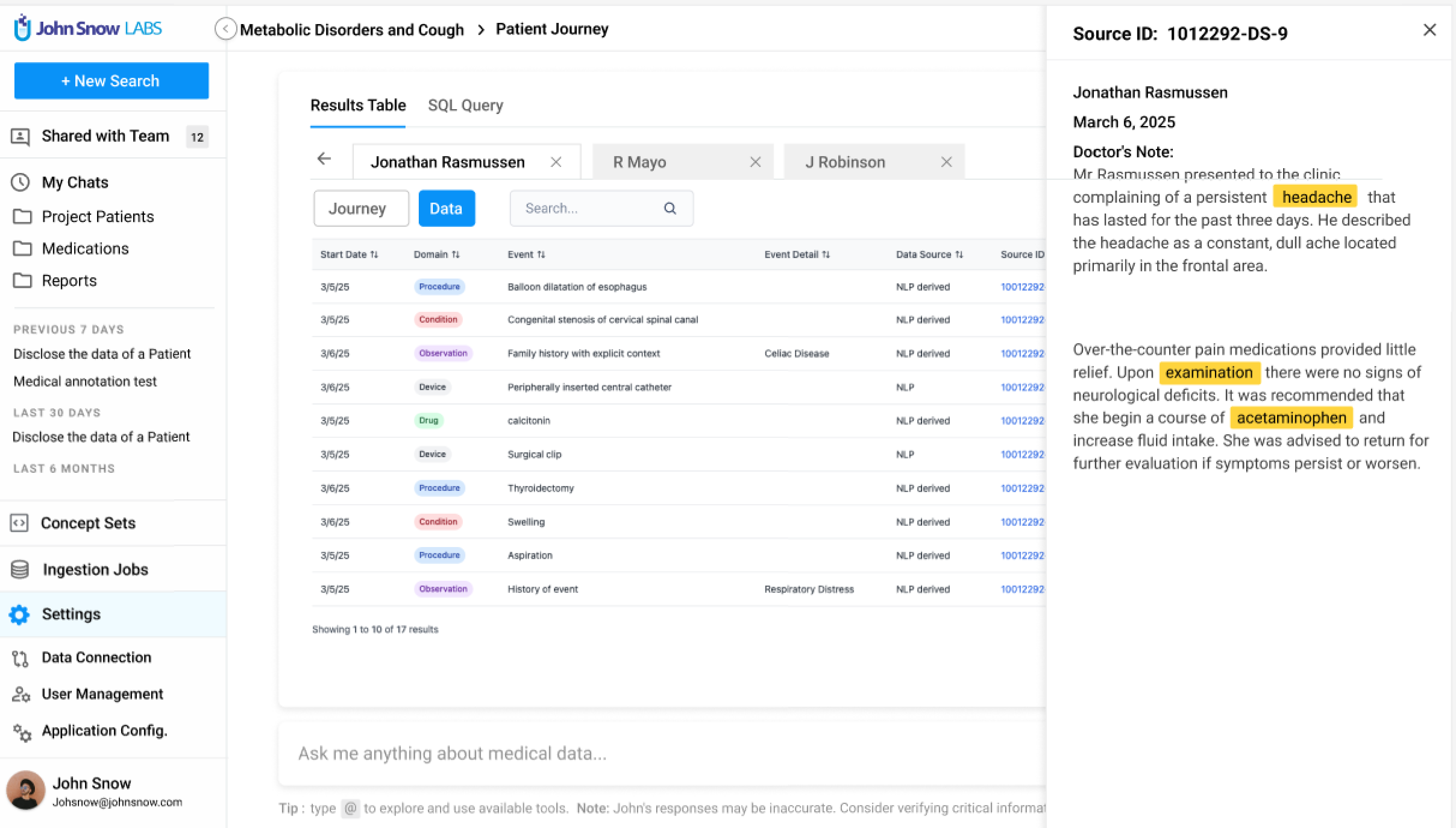
When was the patien's last doctor visit ?

What vaccines has the patient received ?

Ask me anything about medical data...

Tip : type @ to explore and use available tools. **Note:** John's responses may be inaccurate. Consider verifying critical information.

Track Provenance of Information



Start Date ⇅	Domain ⇅	Event ⇅	Event Detail ⇅	Data Source ⇅	Source ID
3/5/25	Procedure	Balloon dilatation of esophagus		NLP derived	10012292
3/5/25	Condition	Congenital stenosis of cervical spinal canal		NLP derived	10012292
3/6/25	Observation	Family history with explicit context	Celiac Disease	NLP derived	10012292
3/6/25	Device	Peripherally inserted central catheter		NLP	10012292
3/5/25	Drug	calcitonin		NLP derived	10012292
3/5/25	Device	Surgical clip		NLP	10012292
3/6/25	Procedure	Thyroidectomy		NLP derived	10012292
3/6/25	Condition	Swelling		NLP derived	10012292
3/5/25	Procedure	Aspiration		NLP derived	10012292
3/5/25	Observation	History of event	Respiratory Distress	NLP derived	10012292

Showing 1 to 10 of 17 results

Ask me anything about medical data...

Tip : type @ to explore and use available tools. Note: John's responses may be inaccurate. Consider verifying critical informat

Source ID: 1012292-DS-9

Jonathan Rasmussen

March 6, 2025

Doctor's Note:

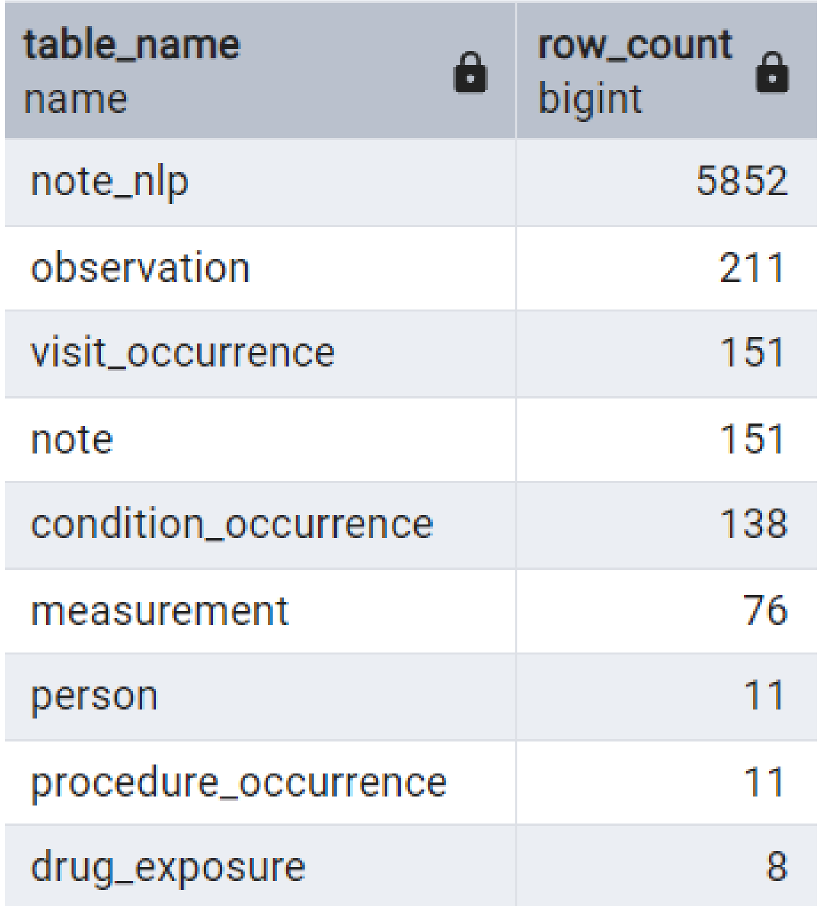
Mr Rasmussen presented to the clinic complaining of a persistent headache that has lasted for the past three days. He described the headache as a constant, dull ache located primarily in the frontal area.

Over-the-counter pain medications provided little relief. Upon examination there were no signs of neurological deficits. It was recommended that she begin a course of acetaminophen and increase fluid intake. She was advised to return for further evaluation if symptoms persist or worsen.

- 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