

Medical Language Models for Data Scientists

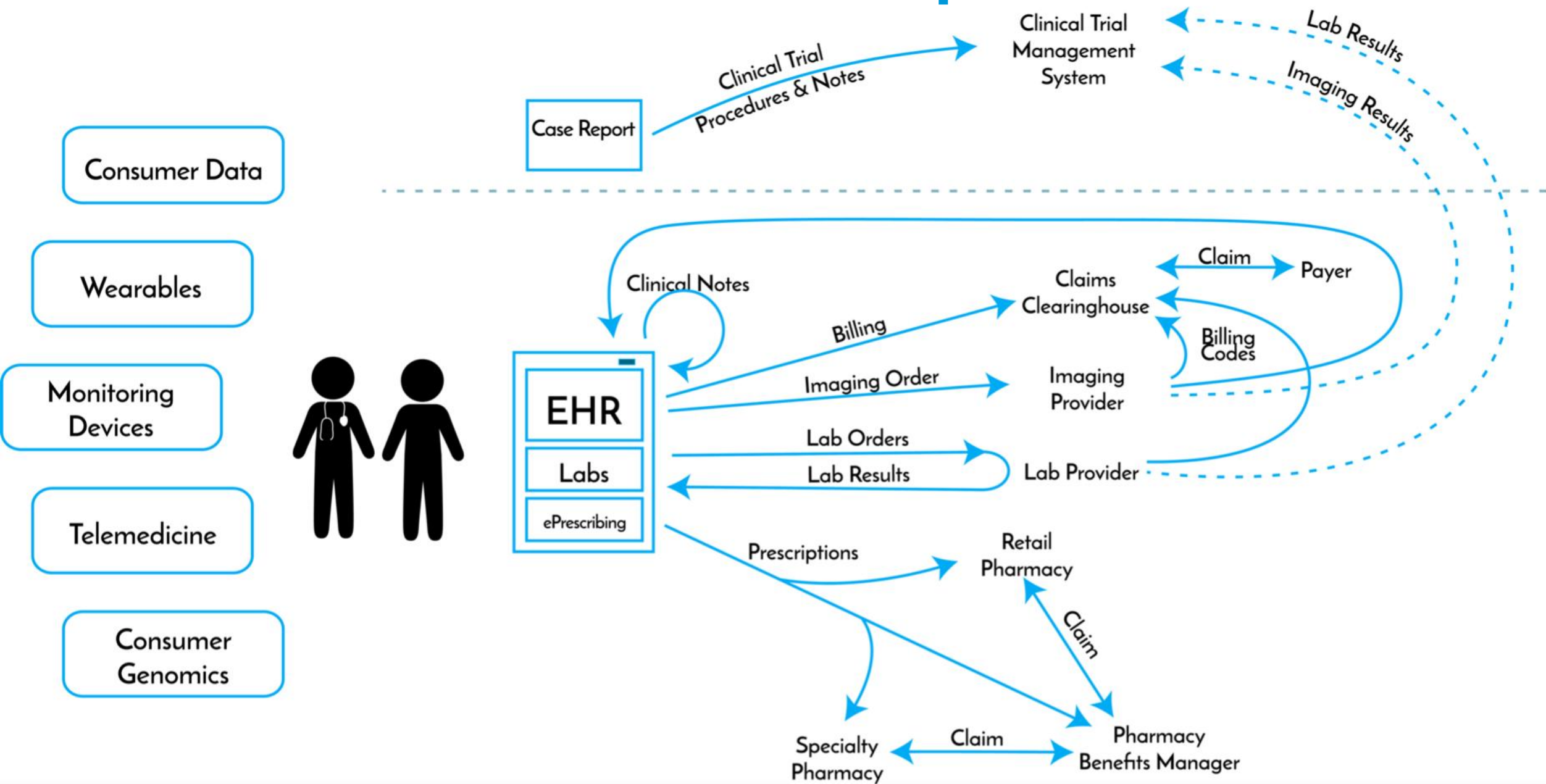
April 2025

Welcome - We have a lot of things ahead of us

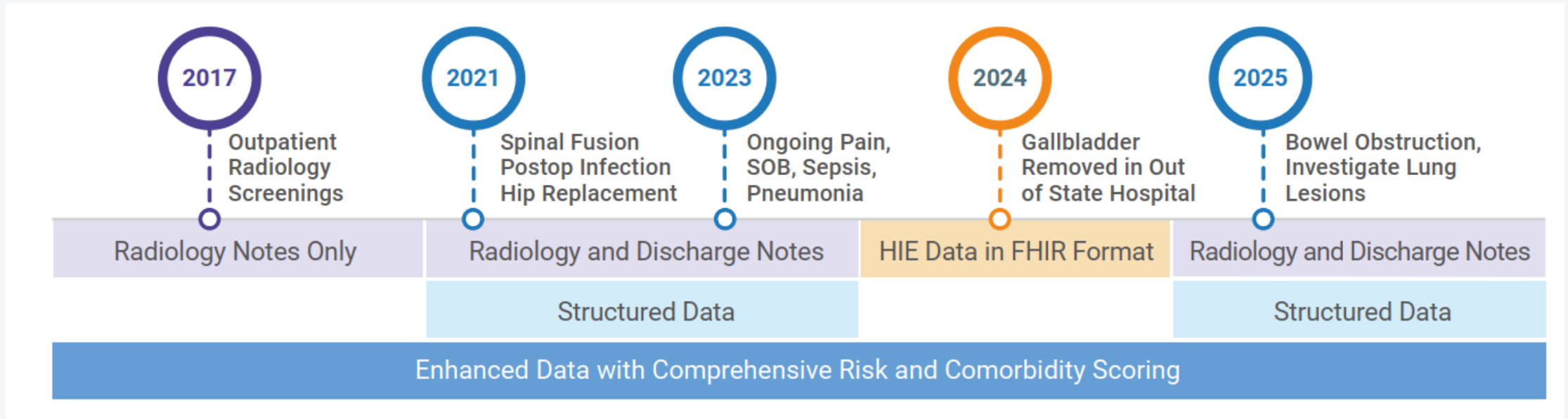
| | | |
|--|--------|---|
| | 50 min | Building Patient Journeys & Cohorts |
| | 10 min | Break |
| | 50 min | Medical Terminology Server |
| | 10 min | Break |
| | 50 min | Oncology research and treatment with Healthcare |
| | 10 min | Break |
| | 50 min | De-identification of Medical Images in DICOM Format |

Building Patient Journeys & Cohorts

Healthcare Data Comes From Multiple Sources



Building Patient Journeys



Integrate Multimodal Data

Natural Language Q&A

Build Patient Cohorts

Understand Timelines

Make Clinical Inferences

Explain & Discuss

Keep Provenance of Information

Using Structured Data Only

The JSON data provided does not explicitly list any details related to preventative care or screenings for the patient. There are no specific entries for routine vaccinations, screenings like mammograms, colonoscopies, or other preventative measures typically logged in EHRs, such as blood pressure checks or diabetes screenings.

Adding Unstructured Notes

Routine Imaging and Screenings:

- There is a mention of a 'Diagnostic mammogram and ultrasound panel' which typically serves as a screening tool for breast cancer. This reflects an engagement in preventative health measures related to potentially identifying malignancies early.

Lung and Respiratory Health Monitoring:

- The repeated references to 'pulmonary emphysema,' 'bronchiectasis,' and 'chronic obstructive lung disease' suggest ongoing monitoring efforts for respiratory conditions, which can be preventative in nature, particularly to manage the progression and complications associated with chronic lung diseases.

Skin Assessments:

- Several entries indicate monitoring of skin nodules and conditions that may correlate with skin cancer screenings or evaluations:: 1105-1111.

Blood Tests:

- There are a variety of routine blood tests mentioned which can provide information necessary for preventative care regarding metabolic health, and cardiovascular risk factors. Screening for conditions such as hyperlipidemia relates to cardiovascular prevention strategies

Adding FHIR Resources

Chronic Obstructive Lung Disease (COPD) Management: The patient has numerous instances indicating chronic obstructive lung disease with an emphasis on ongoing symptom management, including the use of inhalers such as **tiotropium**. Regular monitoring and use of inhaled medications are part of the standard preventative care for COPD patients to manage symptoms and reduce exacerbations.

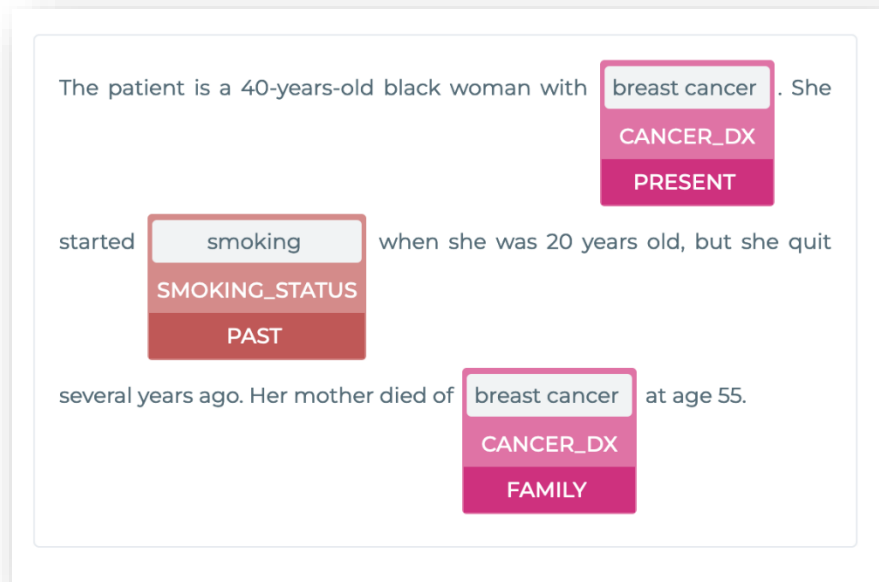
Screening for Lung Health: There are several measurements related to lung capacity, including **Exploratory Reserve Volume** and **Diffusion Capacity** noted within the entries. These tests are critical in assessing lung function and can be part of routine screenings for patients with respiratory conditions.

Disease Management for Cardiovascular Health: There are references to **Coronary Arteriosclerosis** and treatments like **atorvastatin**, which suggest routine cholesterol screening and treatment to mitigate cardiovascular risk factors is part of the patient's preventative care regimen.

Screening for Psychosocial Issues: The records also suggest continuous monitoring of mental health through various scales that may be used for evaluating conditions like anxiety and depression as part of preventative mental health care.

Routine Lab Work: Frequent entries related to blood counts, renal function tests, and other metabolic panels indicate an emphasis on preventative health measures through regular laboratory assessments, which are essential for early detection of any abnormalities that may require intervention.

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 |

Medical Terminology: Resolving to Standard Codes

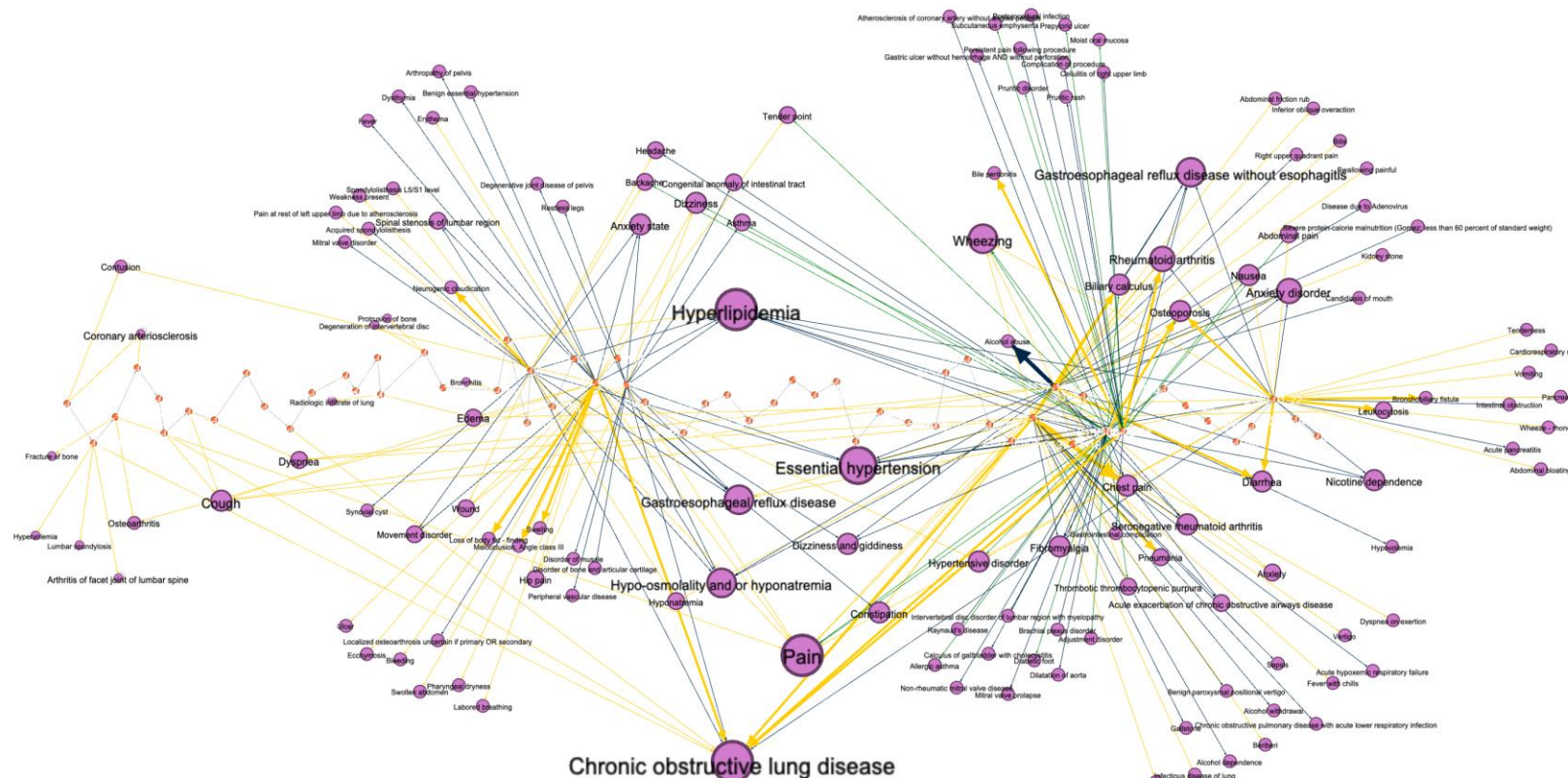
```
{
  "url": "https://fhir/Mutation#assessed.gene",
  "valueCodeableConcept": {
    "coding": [
      {
        "system": "http://ncit.nci.nih.gov",
        "code": "C17757",
        "display": "EGFR"
      }
    ]
  }
},
{
  "url": "https://fhir/Mutation#assessed.referenceSeq",
  "valueCodeableConcept":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

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: $\text{anion gap (in mEq/L)} + 2.5 * (4 - \text{albumin (in g/dL)})$. The formula for computing a patient's anion gap is: $\text{sodium (mEq/L)} - (\text{chloride (mEq/L)} + \text{bicarbonate (mEq/L)})$. 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

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

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.

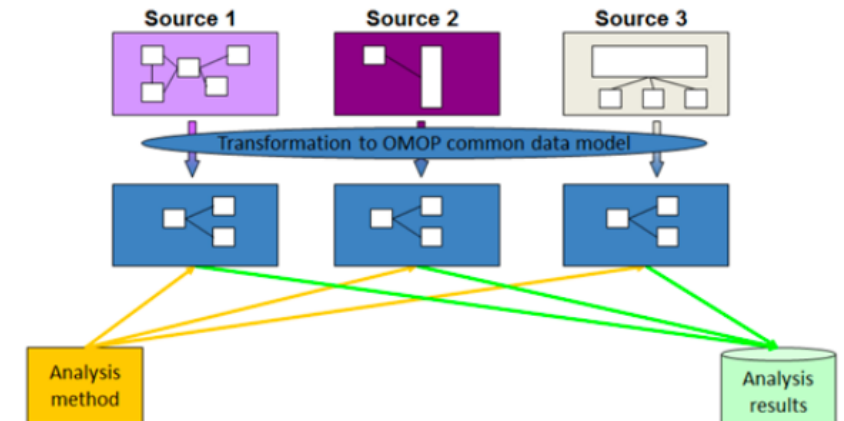
Methods:

- 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/>

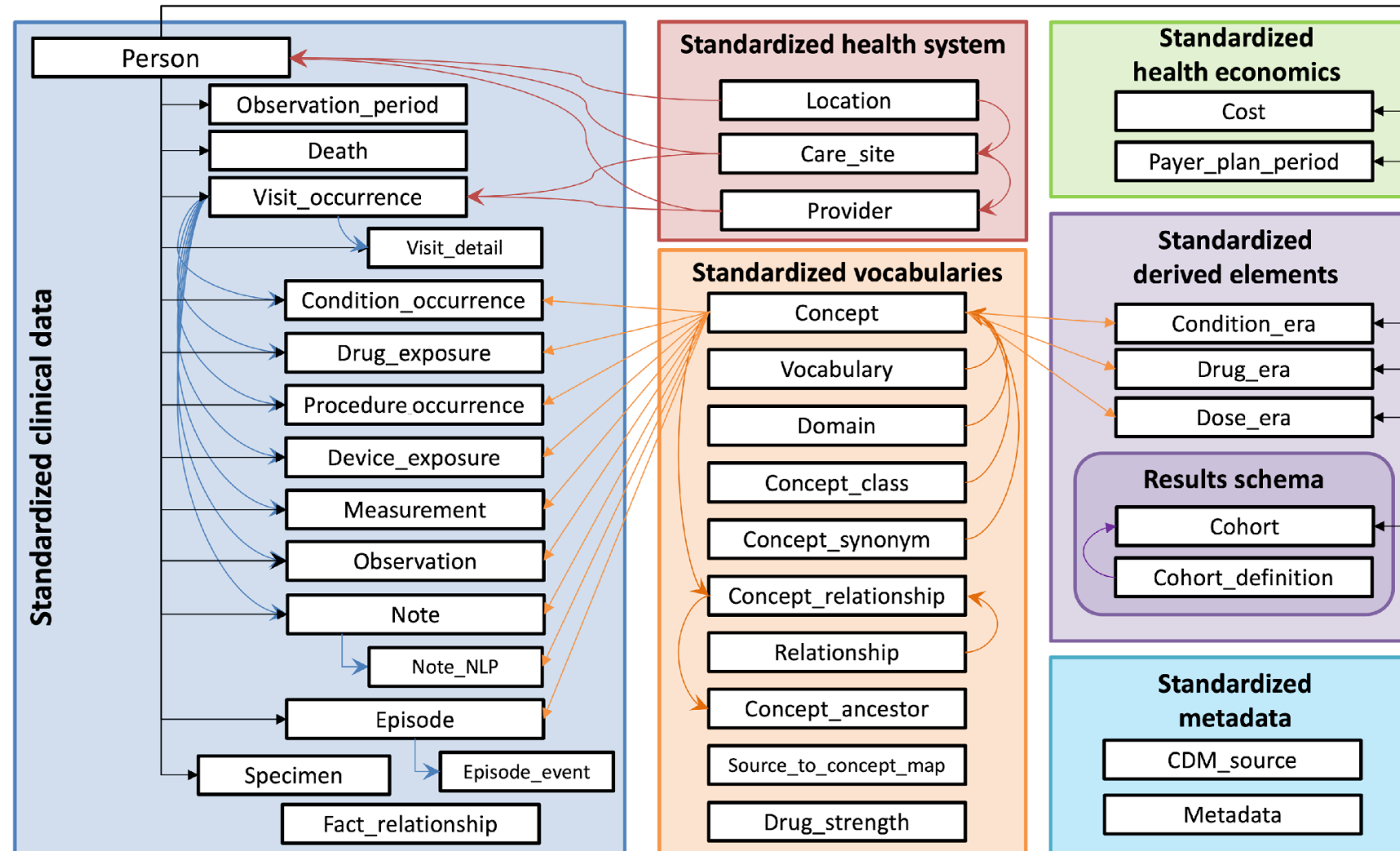


"an international collaborative whose goal is to create and apply open-source data analytic solutions to a large network of health databases to improve human health and wellbeing"



The OMOP CDM


Observational Medical Outcomes Partnership - Common Data Model



CDM v5.4

- 39 tables
- 433 fields
- 7 categories

Data Ingestion



Shared with Team12

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
Johnsnow@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 IngestionCompleted

Instance 1Instance 2Instance 3

2. Patient Information ExtractionIn Progress

Instance 1Instance 2Instance 3

3. Knowledge Graph EnrichmentIn Progress

Instance 1Instance 2

4. OMOP Data ModelingPending

No instance available yet

5. Merge and DeduplicatePending

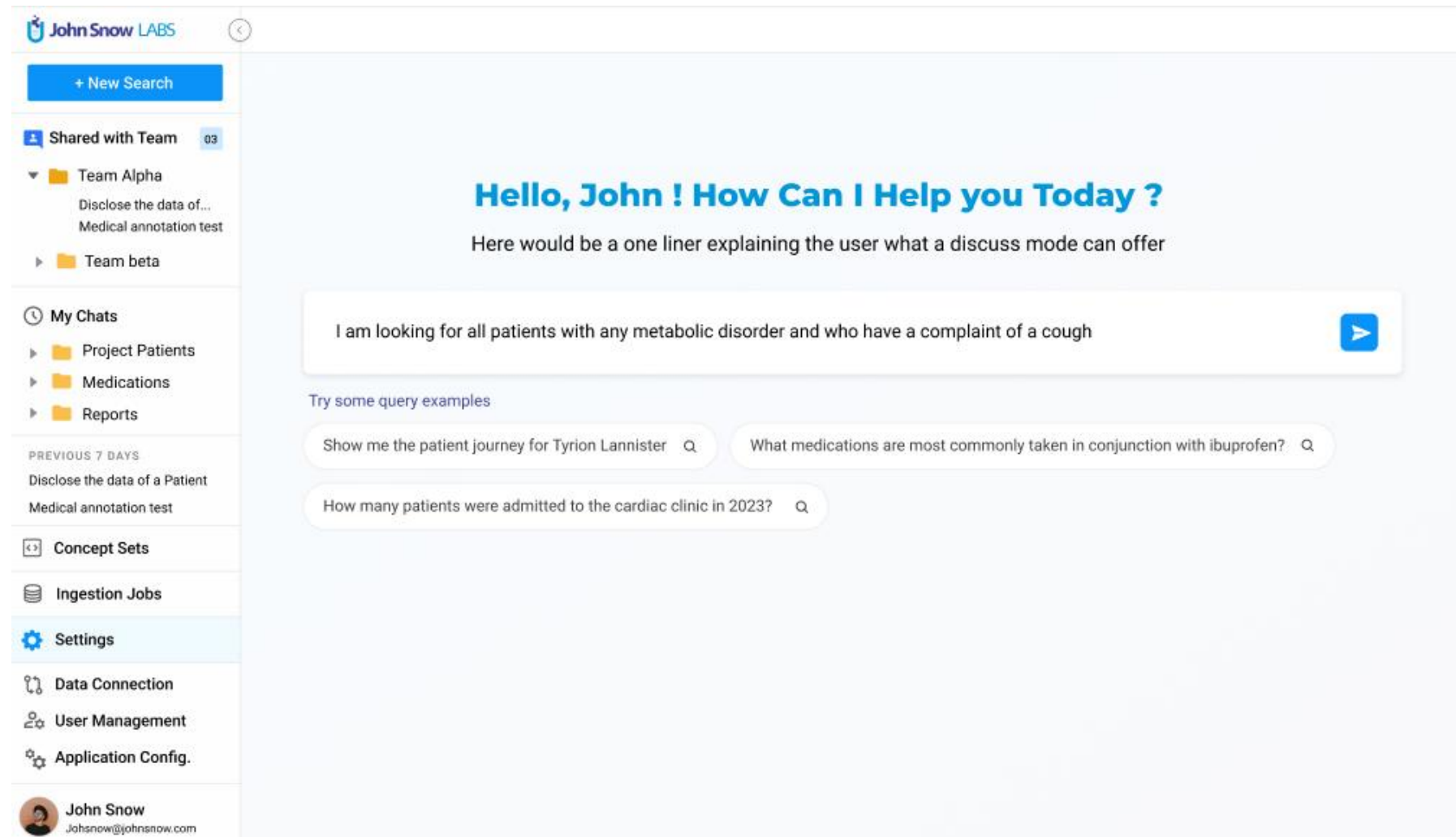
No instance available yet

View05Jobs Per Page

Showing 1-10 of 110 Projects

12345...10

Natural Language in Cohort Level



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

Flexible Cohort Review

 John Snow LABS

+ New Search

Shared with Team

12

My Chats

Project Patients

Medications

Reports

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

Concept Sets

Ingestion Jobs

Settings

Data Connection

User Management

Application Config.

John Snow

Johnsnow@johnsnow.com

Metabolic Disorders and Cough

I am looking for all patients with any metabolic disorder and who have a complaint of a cough

Do you mean Asthma and Bronchitis patients as well?

No, only conditions specifically about coughing

OK, I'll search for patients with a cough and metabolic disorders...

Here are the concept sets for

metabolic disorders

 and

cough

 Click to view or modify these concept sets

Results Table

SQL Query

Columns

| <input type="checkbox"/> | Last Name ↓ | MRN | ↑ | OMOP ID ↑ | DOB | ↓ | Gender ↓ | Race | ↑ |
|--------------------------|------------------------------------|----------|---|-----------|----------|---|----------|----------|---|
| <input type="checkbox"/> | Johnathan Ramussen | 10012292 | | 4 | 5/2/1989 | | Male | White | |
| <input type="checkbox"/> | Mayo | 10012293 | | 1 | 5/2/1989 | | Male | Hispanic | |
| <input type="checkbox"/> | Robinson | 10014765 | | 2 | 5/2/1989 | | Male | Hispanic | |
| <input type="checkbox"/> | Katherine | 10023708 | | 4 | 5/2/1989 | | Female | White | |

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.

Natural Language in Patient Level

John Snow LABS

+ New Search

Shared with Team 12

My Chats

- Project Patients
- Medications
- Reports

PREVIOUS 7 DAYS

Disclose the data of a Patient

Medical annotation test

LAST 30 DAYS

Disclose the data of a Patient

Medical annotation test

LAST 6 MONTHS

Disclose the data of a Patient

Medical annotation test

Concept Sets

Ingestion Jobs

Settings

Data Connection

User Management

Application Config.

John Snow
Johnsnow@johnsnow.com

Metabolic Disorders and Cough

Results Table SQL Query

Columns

Jonathan Ramussen

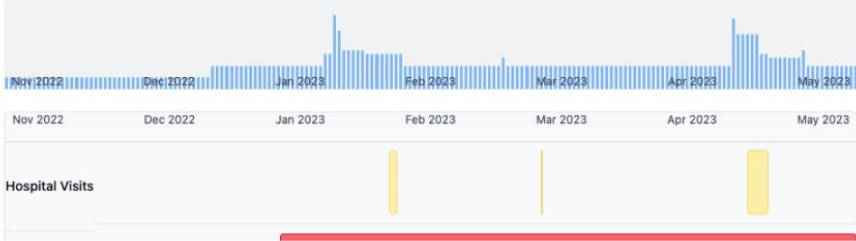
Journey Data Search...

Zoom 80% Monthly Reset Filters

Patient Medical Timeline

Visits Conditions Medications Procedures

Healthcare System Engagement



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 patient'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.

Fine-Grained View of Patient Journeys

John Snow LABS

+ New Search

Shared with Team 12

My Chats

- Project Patients
- Medications
- Reports

PREVIOUS 7 DAYS

- Disclose the data of a Patient
- Medical annotation test

LAST 30 DAYS

- Disclose the data of a Patient

LAST 6 MONTHS

Concept Sets

Ingestion Jobs

Settings

Data Connection

User Management

Application Config.

John Snow
johnsnow@johnsnow.com

Metabolic Disorders and Cough > Patient Journey

Results Table SQL Query

Columns

Jonathan Rasmussen R Mayo J Robinson

Journey Data

Search...

Monthly

Reset Filters

| Start Date T1 | Domain T1 | Event T1 | Event Detail T1 | Data Source T1 | Source ID T1 | Event Concept ID T1 | Detail Concept ID T1 |
|---------------|-------------|--|----------------------|----------------|----------------|---------------------|----------------------|
| 3/5/25 | Procedure | Balloon dilatation of esophagus | | NLP derived | 10012292-DS-10 | 4305073 | |
| 3/5/25 | Condition | Congenital stenosis of cervical spinal canal | | NLP derived | 10012292-RR-66 | 36674692 | |
| 3/6/25 | Observation | Family history with explicit context | Celiac Disease | NLP derived | 10012292-DS-10 | 4210989 | 1341215 |
| 3/6/25 | Device | Peripherally inserted central catheter | | NLP | 10012292-DS-11 | 4161814 | |
| 3/5/25 | Drug | calcitonin | | NLP derived | 10012292-DS-9 | 42900359 | |
| 3/5/25 | Device | Surgical clip | | NLP | 10012292-RR-69 | 4300940 | |
| 3/6/25 | Procedure | Thyroidectomy | | NLP derived | 10012292-DS-10 | 4030107 | |
| 3/6/25 | Condition | Swelling | | NLP derived | 10012292-RR-66 | 4275722 | |
| 3/5/25 | Procedure | Aspiration | | NLP derived | 10012292-RR-41 | 4033232 | |
| 3/5/25 | Observation | History of event | Respiratory Distress | NLP derived | 10012292-RR-69 | 1340204 | 1234987 |

Showing 1 to 10 of 17 results

Does the patient have any allergies ?

When was the patient'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

John Snow LABS

Metabolic Disorders and Cough > Patient Journey

+ New Search

Shared with Team 12

My Chats

- Project Patients
- Medications
- Reports

PREVIOUS 7 DAYS

- Disclose the data of a Patient
- Medical annotation test

LAST 30 DAYS

- Disclose the data of a Patient

LAST 6 MONTHS

Concept Sets

Ingestion Jobs

Settings

Data Connection

User Management

Application Config.

John Snow
Johnsnow@johnsnow.com

Results Table SQL Query

Jonathan Rasmussen

R Mayo

J Robinson

Journey

Data

Search...

| Start Date T1 | Domain T1 | Event T1 | Event Detail T1 | Data Source T1 | 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 information

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 (NLP)

OMOP CDM (PostgreSQL)

Records for one patient

| table_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

| 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 |