

João António Assis Reis  
98474

✉ joaoreis16@ua.pt



# A conversational query builder on medical databases

Mestrado em Engenharia Informática

4th July 2024

Under the guidance of:

**Advisor** João Rafael Almeida  
**Co-advisor** José Luís Oliveira  
**Collaborator** Tiago Almeida





# Use case for a medical researcher

**Task:** Conduct a study about COVID-19.

**Procedure:**

- 1) Definition of the study protocol.
- 2) Contact data owners interested in collaborating in the study.
- 3) Conducting the study.
- 4) Aggregate the results.
- 5) Publish the findings.

**Complex and time-consuming process**

**Why?**

**Motivation**

Problem

Proposed Solution



**Motivation**

## Problem

## Proposed Solution





Motivation

**Problem**

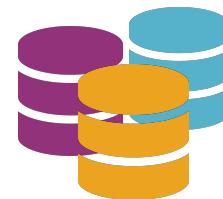
Proposed Solution



EUROPEAN HEALTH DATA &amp; EVIDENCE NETWORK



New entry on  
database catalogue



Database Catalogue



Databases of interest  
for research question



Uploading metadata



Network Dashboards

**Difficult and  
time-consuming**

**Data Custodian**

**Researcher**



## Motivation

## Problem

## Proposed Solution

**Cohort #1788542**  
created by anonymous on 2024-01-13 18:53, modified by anonymous on 2024-01-14 8:22

Example cohort definition

Definition | Concept Sets | Generation | Samples | Reporting | Export | Versions | Messages

Enter a cohort definition description here

### Cohort Entry Events

Events having any of the following criteria:

- a drug exposure of ACE INHIBITORS (example) + Add attribute... Delete Criteria
- a condition era of Any Condition + Add attribute... Delete Criteria

with continuous observation of at least 365 days before and 0 days after event index date  
Limit initial events to earliest event per person.

Restrict initial events to:  
having all of the following criteria:  
+ Add criteria to group...

Limit initial events to earliest event per person.  
Remove initial event restriction

### Inclusion Criteria

New inclusion criteria

- has hypertension diagnosis in 1 year prior to treatment
- Has no prior antihypertensive drug exposures in medical history
- Is only taking ACE as monotherapy, with no concomitant combination treatments
- Unnamed Criteria
- 9519
- Unnamed Criteria

Limit qualifying events to earliest event per person.



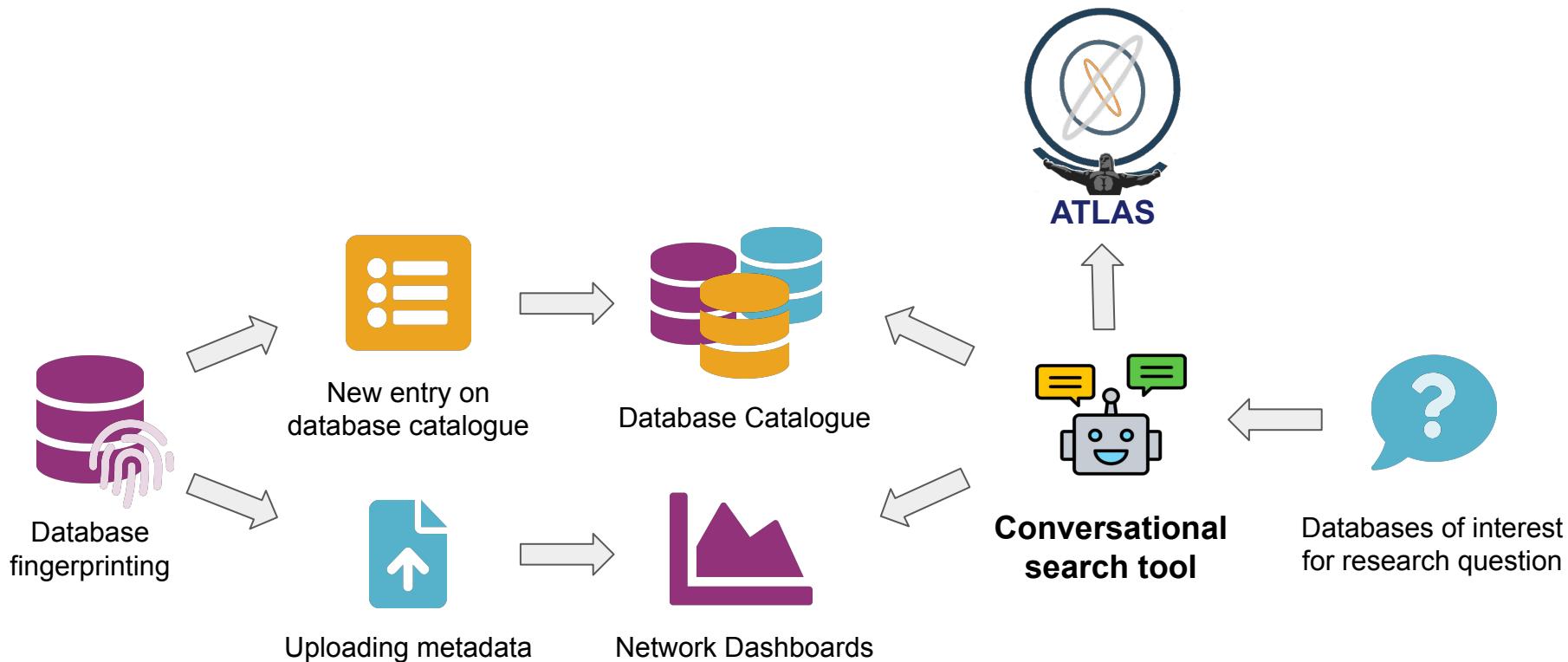
## Cohort Definition

- Query that **defines a set of persons** who meet certain criteria over a specified duration



Motivation

Problem

**Proposed Solution**

**Information Retrieval**

Large Language Models

Conversational User Assistants

LLM integration

**Information  
Retrieval**

Retrieve relevant information from collections of unstructured data, such as documents

- **Traditional methods**

- Term Frequency - Inverse Document Frequency (TF-IDF)
- Best Matching 25 (BM25)

- **Neural methods**

- Interaction-based method

**Techniques**

Focuses on providing a single and specific answer to a question posed in natural language

- Uses **Natural Language Processing (NLP)** and **Information Retrieval (IR)**

**Question  
Answering**



## Definition

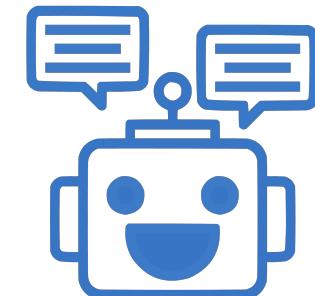
Designed to **comprehend and generate text** that is coherent and contextually relevant, engaging in human language interactions.

- These models process and **predict patterns** with accuracy.

## Comparison

Model	Provider	Fine-tuneability	Open-source
GPT-4	OpenAI	No	No
LLaMa 2	Meta	Yes	Yes
PaLM 2	Google	No	No
Falcon	TII	Yes	Yes
Mistral	Mistral AI	Yes	Yes

Due to the use of  
**sensitive and private data**





Types of chatbots based on their **response generation**:

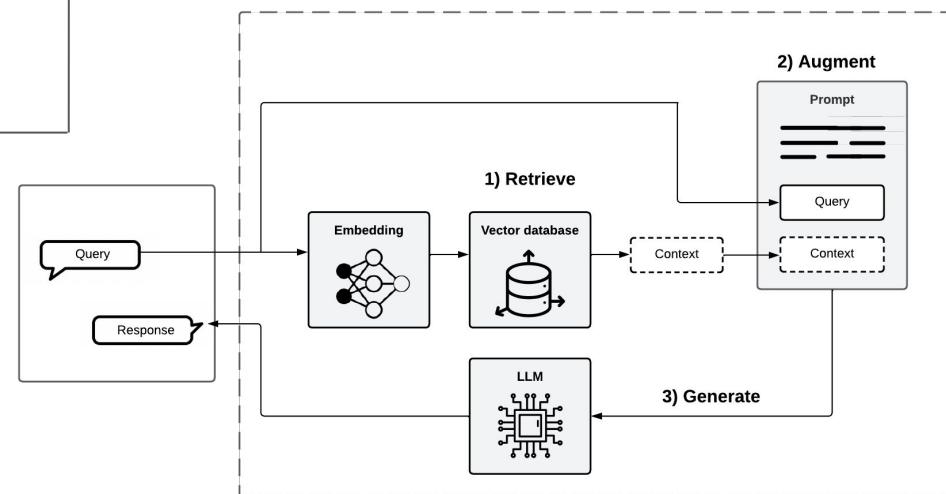
- Rule-based chatbot
- Retrieval-based chatbot
- **Generative-based chatbot**

improve and adapt the LLM

**Reinforcement Learning from Human Feedback (RLHF)**

overcoming limitations of the LLM, such as hallucination

### Retrieval-Augmented Generation (RAG)





Low-code tools to build customized LLM orchestration flow  
and AI agents





## Technology used

Architecture

Database discovery

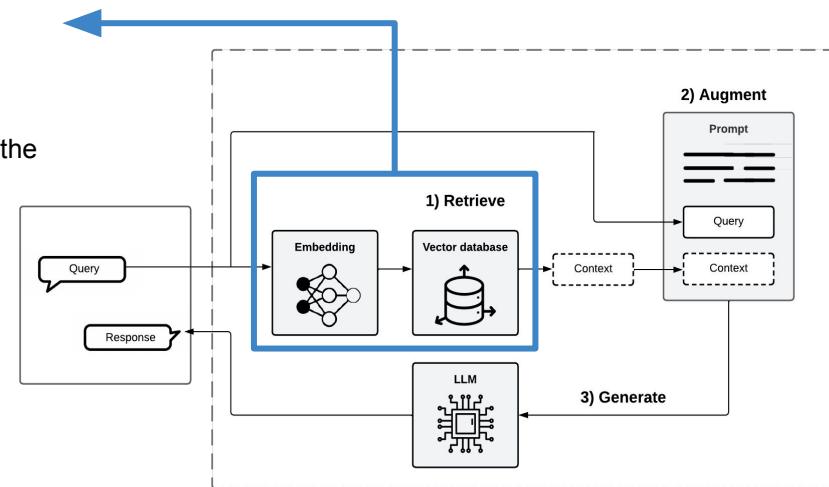
Query Builder

## Information Retrieval

- **Model:** BM25
- Used to retrieve the most suitable databases
  - Based on the **databases concepts** extracted from the **Network Dashboards** tool

## Large Language Model

- **Model:** Nous Hermes 2 Mixtral 8x7B
- Used as a NLP tool
  - To extract information from user's messages
  - To generate responses to the user
- Also used to create a fluid conversation



## Retrieval-Augmented Generation (RAG)

Generate a response to retrieve the best databases to the user



## Technology used

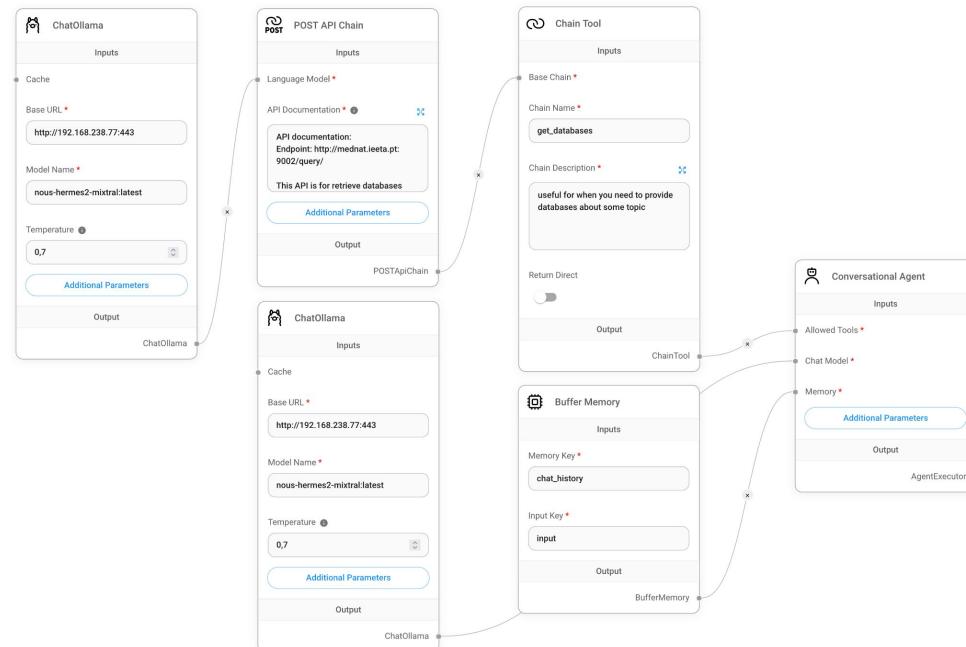
Architecture

Database discovery

Query Builder



To **orchestrate the interactions** between the LLM, the user interface and the information retrieval component



- FlowiseAI proved to be more **promising** than Langflow
- Development of a chatbot to discover medical databases using FlowiseAI
- However, it not addresses the **complexity of the cohort definition requirements**

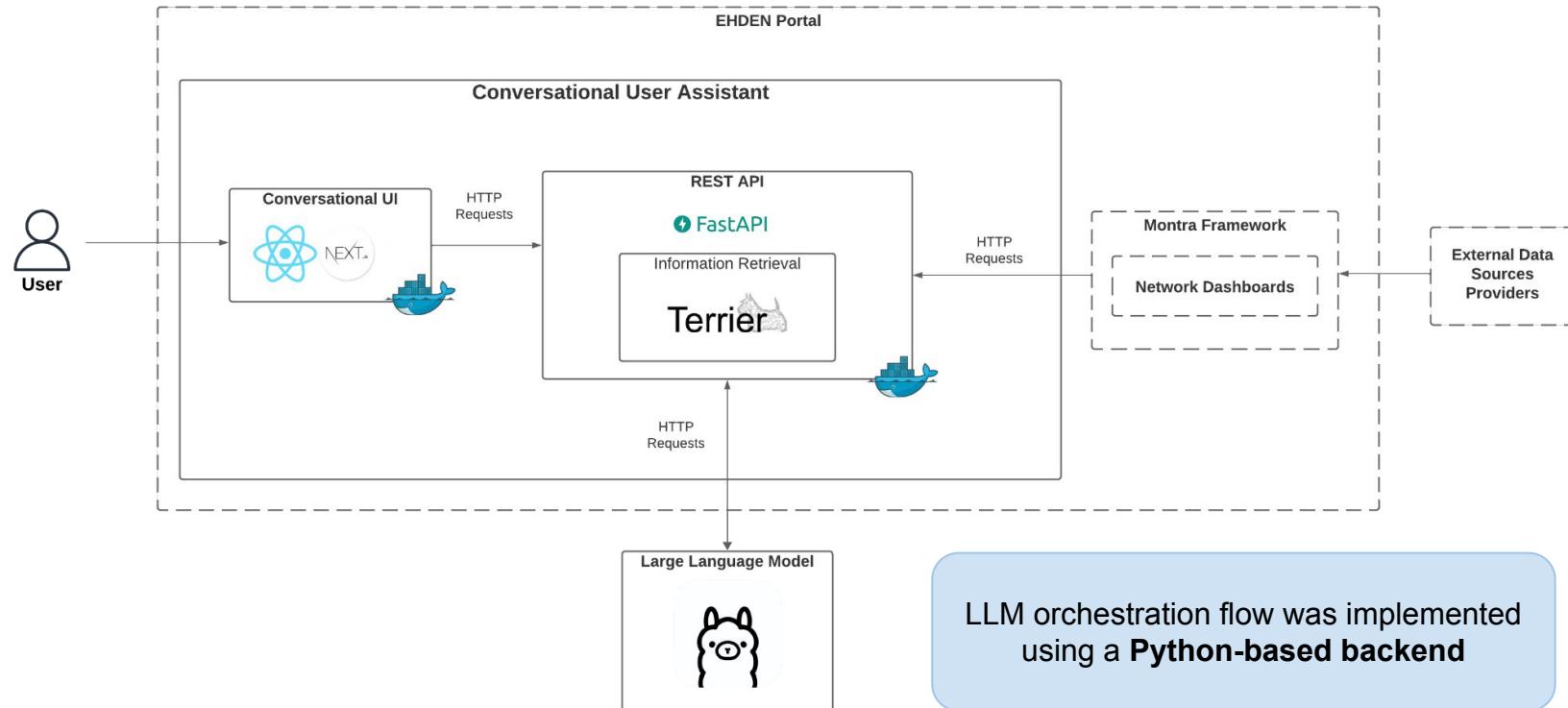


Technology used

**Architecture**

Database discovery

Query Builder





EHDEN PORTAL

Free text search EHDEN

Jobo

HOME

CATALOGUE

DASHBOARD

ACADEMY

EHDEN

PUBLICATIONS

STATUS

CHATBOT

MANAGE 180

API INFO

PORTAL

ABOUT

GET STARTED

FEEDBACK

PROFILE

DEVELOPERS

SIGN OUT

ADMIN

ADMIN

IMPORT

EXPORT

Hippocrates

I'm here to help you find the best OMOP CDM databases for your observational studies. Please provide me with questions or specific details about your study needs so I can identify the most suitable databases for you.

Hello, could you please describe the scope of your study?

How can I help you? ➤

The screenshot shows the EHDEN Portal interface with the 'CHATBOT' section selected. On the left is a sidebar with various navigation links. In the center, a chat window is open with a message from 'Hippocrates' asking for a description of the user's study scope. At the bottom, there is a text input field with the placeholder 'How can I help you?' and a send button.

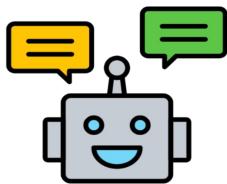


# Goal

Enhance the conversational system to support the cohort definition for observational studies.



Researchers interested in conduct a study



Conversational system



ATLAS

**New Cohort Definition**

Cohort Definition Example

Definition Concept Sets Generation Samples Reporting Export Versions Messages 1

Enter a cohort definition description here

**Cohort Entry Events**

Events having any of the following criteria:

+ Add Initial Event... ▾

with continuous observation of at least 0 days before and 0 days after event index date  
Limit initial events to: earliest event per person.

Restrict initial events

**Inclusion Criteria**

New inclusion criteria

Limit qualifying events to: earliest event per person.

**Cohort Exit**

Event Persistence:  
Event will persist until: end of continuous observation

Censoring Events:  
Exit Cohort based on the following criteria:

+ Add Censoring Event... ▾

No censoring events selected.

**Cohort Eras**

- Specify era collapse gap size: 0 days
- [add trimming options...](#)



Technology used

Architecture

Database discovery

Query Builder

**Concept Set**Collection of **medical concepts** used to define clinical elements, like diseases, drugs, procedures

Concept Set Name

Concept Test

ID	Concept	Select ?	Excluded ?	Descendants ?	Mapped ?
710158	COVID-19	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
756061	Asymptomatic COVID-19	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
35894915	COVID-19 vaccine	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
37310268	Suspected COVID-19	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
703431	COVID-19 excluded	<input type="checkbox"/>			

Submit

The concept set was successfully defined. You can either download it or send it to your ATLAS instance.

Download

Send to ATLAS



```
{
  "id": 0,
  "name": "<CONCEPT_SET_NAME>",
  "expression": {
    "items": [
      {
        "concept": {
          "CONCEPT_ID": 231256,
          "CONCEPT_NAME": "Covid-19",
          "DOMAIN_ID": "Condition",
          "VOCABULARY_ID": "ASGG7HR",
          "CONCEPT_CLASS_ID": "ASG67 code",
          "CONCEPT_CODE": "953635",
          "VALID_START_DATE": "2000-01-01",
          "VALID_END_DATE": "2099-12-31"
        },
        "isExcluded": false,
        "includeDescendants": true,
        "includeMapped": false
      }
    ],
    "...": ...
  }
}
```



## Cohort Definition

### Hippocrates

I'm here to help you find the best OMOP CDM databases for your observational studies. Please provide me with questions or specific details about your study needs so I can identify the most suitable databases for you.



[Hippocrates] Cohort 2024-06-15

Help us defining the cohort?

Definition Concept Sets Generation Samples Reporting Export Versions Messages

Hippocrates custom cohort definition

7 days is ok

365 days

#### Cohort Entry Events

Events having any of the following criteria:

+ Add Initial Event... ▾

a condition occurrence of **Concept Test** ▾

+ Add attribute... ▾ Delete Criteria

with continuous observation of at least **7** days before and **365** days after event index date

Limit initial events to: earliest event ▾ per person.

Restrict initial events

18

The screenshot shows the 'Cohort Definition' section of the Hippocrates platform. At the top, there's a navigation bar with tabs for 'Definition', 'Concept Sets', 'Generation', 'Samples', 'Reporting', 'Export', 'Versions', and 'Messages'. Below this, a title bar says '[Hippocrates] Cohort 2024-06-15' and includes a help icon. A message 'Help us defining the cohort?' is displayed. The main area is titled 'Hippocrates custom cohort definition'. It contains a 'Cohort Entry Events' section with a blue header. Inside, it says 'Events having any of the following criteria:' followed by '+ Add Initial Event... ▾'. Below this, there's a row: 'a condition occurrence of' followed by a dropdown menu set to 'Concept Test' (which is highlighted with a red box), and another dropdown menu set to '365 days' (also highlighted with a red box). Further down, it says 'with continuous observation of at least 7 days before and 365 days after event index date'. At the bottom, there's a limit for initial events: 'Limit initial events to: earliest event ▾ per person.' and a 'Restrict initial events' button.



```
{  
  "ConceptSets": "Are there any other concepts you'd like to add? If yes, please add them.  
  If no, simply respond with 'no'.",  
  
  "PrimaryCriteria": {  
    "ObservationWindow": {  
      "PriorDays": "Regarding the observation window, what is the minimum number of days  
      needed before the continuous observation? You must choose from 0, 1, 7, 14, 21, 30, 60,  
      90, 120, 180, 365, 548, 730 or 1095.",  
      "PostDays": "How many days after event index date? You must choose from 0, 1, 7,  
      14, 21, 30, 60, 90, 120, 180, 365, 548, 730 or 1095."  
    }  
  }  
}
```

**Processed by the LLM until it returns an acceptable answer**



## Database Discovery Tool

## Query Builder

**EHDEN Database Catalogue**

**EHDEN Network Dashboards**

**EHDEN PORTAL**

**Hippocrates**

I'm here to help you find the best OMOP CDM databases for your observational studies. Please provide me with questions or specific details about your study needs so I can identify the most suitable databases for you.

Hello, could you please describe the scope of your study?

How can I help you?

The diagram illustrates the flow from the Database Discovery Tool and Query Builder to the EHDEN Network Dashboards, with arrows pointing from the left side towards the right side. A large arrow also points down from the Database Discovery Tool section to the Hippocrates chatbot interface.



Database Discovery Tool

## Query Builder

**ATLAS**

- Home
- Data Sources
- Search
- Concept Sets
- Cohort Definitions
- Characterizations
- Cohort Pathways
- Incidence Rates
- Profiles
- Estimation
- Prediction
- Reusables
- Jobs
- Configuration
- Feedback

Apache 2.0  
open source software  
provided by  
**OHSI**  
data-for-healthcare

Concept set definition process in **ATLAS**

Search

Type your search here

Advanced Options

**EHDEN PORTAL**

Hippocrates

I'm here to help you find the best OMOP CDM databases for your observational studies. Please provide me with questions or specific details about your study needs so I can identify the most suitable databases for you.

Concept Set Name

ID	Concept	Select	Excluded	Descendants	Mapped
710158	COVID-19	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
756061	Asymptomatic COVID-19	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
35894915	COVID-19 vaccine	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
37310268	Suspected COVID-19	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
703431	COVID-19 excluded	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Submit

How can I help you?

© Bioinformatics.UA.UA            

Concept set definition process using the **chatbot**

**Conclusions**

Future Work

Contributions



**Streamline the process  
of medical database  
discovery**



**Impact on the speed  
and quality of research  
efforts**



**Innovation and  
advancement of the  
medical research field**



- Enhance the query builder feature in order to provide the **full cohort definition**
- Creating **synthetic labeled dataset** to test and validate the Information Retrieval methods
- Enhance the Information Retrieval methods to **produce better search results**



## Conferences

- “Using Flowise to Streamline Biomedical Data Discovery and Analysis”
  - 22nd IEEE Mediterranean Electrotechnical Conference (MELECON), 2024
- “HealthDBFinder: a question-answering task for health database discovery”
  - 37th IEEE International Symposium on Computer-Based Medical Systems (CBMS), 2024
- “A chatbot-like platform to enhance the discovery of OMOP CDM databases”
  - 34th Medical Informatics Europe Conference (MIE), 2024
- “BIT.UA at BioASQ 12: From Retrieval to Answer Generation” (submitted)
  - 15th Conference and Labs of the Evaluation Forum (CLEF), 2024

## Poster / Software Demo

- “A Chatbot to help discover OMOP DCM databases within EHDEN Network”
  - OHDSI Europe Symposium, 2024.

# Thank you!