Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

Sariga Kakkamani

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. ir. Liesbet PEETERS
dr. ir. ing. Joeri VERBIEST

SUPERVISOR :
Frederic JUNG

2024
2025

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

Sariga Kakkamani
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. ir. Liesbet PEETERS
dr. ir. ing. Joeri VERBIEST

SUPERVISOR :
Frederic JUNG

Acknowledgment

I would like to sincerely acknowledge all those who have supported me throughout my Master's journey.

My deepest gratitude to my internal supervisors, Prof. Liesbet Peters and Dr. Joeri Verbiest, for your unwavering guidance, insightful feedback, and immense support throughout all stages of this thesis. I am extremely grateful for your time and availability to review my work and provide continuous feedback. Working under your supervision was a great experience for me.

To my external supervisor at VITO, Frederic Jung, I cannot thank you enough. Your constant support, practical guidance, and commitment allowed me to face the challenges with confidence. I am especially grateful for your help in preparing and submitting the abstract for the OHDSI symposium, which was a significant milestone for me.

I also would like to thank all the professors, seniors and my classmates for their kindness and encouragement. Your presence made this academic experience truly enriching.

Last, but with boundless gratitude, I thank my partner, Priyesh. Your constant support and deep understanding carried me through this Master's journey. I am truly grateful.

Abstract

Feature extraction from Electronic Health Records (EHR) data is one of the crucial steps in observational studies. This requires translating high-level clinical concepts into queries compatible with standard terminologies. Observational health data are often standardised to the commonly used OMOP-CDM standards. This enables us to carry out efficient analyses that can generate reliable evidence. However, understanding these standards and vocabulary terms requires medical knowledge, particularly for users without domain expertise. Defining and extracting relevant features from structured EHRs remains a key challenge. This thesis proposes an RAG-enhanced LLM pipeline to extract required features from the OMOP-CDM concepts of SNOMED CT vocabulary. When the user inputs the query or the feature, the input is encoded and compared against pre-generated embeddings - OMOP concepts stored in a vector database. The top-k most semantically similar matches are retrieved and passed to the LLM using a structured prompt. The LLM generates context-aware concept IDs as suggestions. This workflow has been successfully validated in the context of the REALM project, where it supports the generation of standardised AI feature sets from natural language cohort definitions. This framework enables the semantic mapping of natural language cohort definitions to standardise queries compliant with the OMOP-CDM, thus improving the accuracy of feature mapping. The end-to-end automation of this process makes it accessible to users, even those without expertise in the medical field. In the future stage, this workflow will be integrated into the REALM testing environment, where the AI model developer can directly get the recommendations of the concept names while submitting the cohort requirements. The proposed RAG-LLM pipeline focuses on helping AI model developers map cohort features to OMOP-CDM standards, to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare professionals.

Acronyms

API Application Programming Interface

COPD Chronic Obstructive Pulmonary Disease

GPU Graphics Processing Unit

HPC High-Performance Computing

JSON Java Script Object Notation

Llama Large Language Model Meta AI

LLM Large Language Model

MIMIC Medical Information Mart for Intensive Care

NLP Natural language processing

OMOP-CDM Observational Medical Outcomes Partnership- Common Data Model

OHDSI Observational Health Data Sciences and Informatics

PostgreSQL Relational database management system for Structured Query Language (SQL)

RAG Retrieval Augmented Generation

REALM Real-world-data Enabled Assessment for heaLth regulatory

RIANA REALM Intelligent Analytics Dashboard

RWD Real World Data

SNOMED Systematized Nomenclature of Medicine

SBERT Sentence-Bidirectional Encoder Representations from Transformers

VITO Vlaamse Instelling voor Technologisch Onderzoek

Contents

1 Introduction

1.1 Background

Real-world data (RWD) in healthcare includes routinely collected records from various sources, including electronic health records (EHR), wearable devices, claims, registries, etc. The advancement in digital data-driven technologies has widely attracted researchers to use RWD for different analyses. One of the biggest challenges for people working in health data science is data standardization Ziletti and D'Ambrosi (2024). Across the world, different hospitals, registries, etc., use different standards or medical coding dictionaries for data capture. Using a unified data model is extremely difficult or impossible.

The Observational Medical Outcome Partnership Clinical Data Model (OMOP-CDM) transforms the diverse structure of these data into a common data model, making it globally interoperable OHDSI OMOP-CDM (2025). This allows researchers or data users to make use of healthcare data seamlessly Wilkinson et al. (2016). Data scientists or statisticians can utilize the vast amount of real-world data to make predictions and inferences that can generate reliable evidence. This often requires high-quality data that is representative of a broader distribution or population. Thereby translating high-level clinical concepts into queries compatible with the OMOP-CDM standard terminologies. However, understanding these standards and vocabulary terms requires medical knowledge, along with OMOP-CDM expertise. This makes feature extraction crucial, particularly for users without domain expertise Yang et al. (2022). Defining and extracting relevant features from structured Electronic Health Records (EHRs) that adhere to OMOP-CDM standards remains a key challenge. Currently, Observational Health Data Sciences and Informatics (OHDSI) tools such as Athena and Usagi are used to search and help users map vocabulary following OMOP concepts OHDSI Tools (2025). However, these tools come with their own limitations and do not meet the exact contextual requirements.

This thesis proposes a concept mapping pipeline using strategies different from the traditional NLP (Natural Language Processing) techniques and searching algorithms used by the OHDSI tooling (e.g. Fuzzy and Lucene search). The Retrieval Augmented Generation - Large Language Model (RAG-LLM) pipeline is proposed to create an automated vocabulary mapping for required features. To ensure real-world applicability, the proposed pipeline was applied to the use cases of the **Real-world-data Enabled Assessment reguLatory decision-Making (REALM)** Realm-EU (2025) project. REALM aims to provide a powerful sandbox environment for the future evaluation of AI as a medical device that goes on the EU market. REALM brings together EU regulatory authorities, data-driven software developers, healthcare professionals, and policy makers to create and evaluate Artificial Intelligence (AI) models for the direct benefit of patients and healthcare professionals. This, in turn, is adopted in clinical practices to ensure accurate diagnosis and personalized treatment pathways. This thesis focuses on mapping the REALM use-case features to the OMOP vocabulary, which is based on an extended AI model description or model cards. Beyond the use cases of REALM, this solution holds strong potential for a wide range of real-world standardisation challenges where efficient, accurate vocabulary mapping is essential.

Research context: "This thesis explores how LLMs can enhance or automate the semantic mapping of natural language cohort definitions to standardise queries, compliant with the OMOP-CDM, thus improving the accuracy of feature mapping. The goal is to develop a RAG enhanced LLM pipeline that efficiently processes medical concepts, accurately maps them to standard vocabularies, and optimises retrieval for improved feature extraction."

Contributions: This thesis aims to provide an end-to-end pipeline that makes it accessible to users, even those without expertise in the medical field, to directly get the recommendations of concept names while submitting the cohort requirements. The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy and usability for the direct benefit of patients and healthcare practitioners.

1.2 State of the art

Extracting features from standardized health care data from commonly used standards such as OMOP-CDM remains a key challenge. This will require a clear understanding of the data structure and standards followed by the OMOP common data models. The OHDSI community is responsible for OMOP-CDM and its associated standardised vocabularies. Athena and Usagi are a searchable engine or databases maintained by the OHDSI community (see Annexe C). Athena is a web application for browsing and downloading the Standardized Vocabularies used in OMOP-CDM. Using Athena, a researcher can search for the feature of interest and their corresponding standard concepts. Usagi is the OHDSI tool, which was designed to aid code mapping between local coding systems and OMOP standard concepts; however, these tools are more beneficial for someone who wants to map their source data to OMOP-CDM. Moreover, these tools are developed with string matching algorithms such as fuzzy match and lucene search, cannot capture the semantic meaning of the searching terms, thereby requiring manual input from the users and fail to meet the exact contextual requirements Mitchell-White et al. (2024). Even though Athena contains an extensive range of vocabulary lists within OMOP CDM, users often get overwhelmed with the volume of search results being generated with a single search term. In addition to this, users are required to review the search results manually to identify the exact matching terms they are interested in. Usagi, on the other hand, is a semi-automated mapping tool; someone who wants to benefit from it must have a clear idea about the source data. In the context of REALM use cases, when an AI model developer or regulatory authorities are trying to obtain the required data to evaluate and benchmark the AI models, we do not expect them to have information on the source data; hence, Usagi may not be a suitable option for them.

With the application of LLM, semantic mapping has become much more efficient and reliable Tekumalla and Banda (2024). Hence, by incorporating LLM into the mapping pipeline between the users and the vocabulary terms, the process can be automated by excluding manual intervention, almost entirely omitted. LLMs are pretrained on massive and diverse terminologies across different domains, which enables them to capture the semantic meaning of the natural language text and map it to the required concepts. However, it sometimes generates hallucinated or falsified responses when the model is not trained on the required data, in this case, the OMOP-CDM vocabularies. To overcome this challenge, this thesis proposes an RAG-LLM pipeline using OMOP-CDM standards. This will enable us to feed the LLM with relevant information to be generated, thereby limiting the LLM from generating false results. This approach will allow LLM to focus on fine-tuning the results retrieved through RAG with the context-relevant information Lewis et al. (2020).

REALM Use cases and Demonstrators, this thesis focuses on evaluating five REALM use-cases, developed by different demonstrators. These demonstrators include AI models from different European partners covering various healthcare domains. They provide critical insights into usability, functionality, and requirements for capturing AI model claims effectively. The use cases include ASCOPD, STAR, COPowereD, DuneAI and PGx2P.

- **ASCOPD - Traqbeat:** AI model that detects the inpatient risk stratification of patients suffering from asthma and chronic obstructive pulmonary disease (COPD).
- **COPowereD - Comunicare**: They aim to predict hospitalisation or acute exacerbations in patients with COPD by using medical AI algorithms that include patient-reported outcomes.
- **STAR - University of Liege**: Project aims to develop an AI model-based decision-support system to enable precise regulation of blood glucose levels in intensive care unit (ICU) patients.
- **DuneAI - Maastricht University** DuneAI is a use case that involves evaluating an automated segmentation software for detecting and segmenting non-small-cell lung cancer tumours in CT scans.
- **PGx2P - Vito**: Pharmacogenomics Passports to Practice (PGx2P) is a use case that aims to implement preventive pharmacogenetics testing of a panel of genetic variants approved for clinical.

Model cards explain the intended use cases, targeted population and evaluation metrics of the AI model, hence the model can be benchmarked and evaluated accordingly (Mitchell et al. (2019)). The evaluation section of the model card refers to the data or the cohort group for which the model's needs can be evaluated. The required inclusion-exclusion criteria can be defined from this section, so the cohort can be identified and used to extract the features. The example model card for one of the use cases, ACOPD, including the detailed descriptions, can be seen in the Fig. 1. For detailed description and model cards for all the other use cases, refer to Appendix E.

Model Card for ASCOPD

AI models for COPD and ASTHMA inpatient risk stratification (ASCOPD) (TRAQBEAT) is a use case that proposes a solution for the inpatient risk stratification of patients suffering from asthma and chronic obstructive pulmonary disease. The solution includes a Traqbeat sensor and biomarkers measuring methods for predicting inpatient mortality, acute respiratory failure, and ventilator dependence.

  - **Model Details:**
  - **Date:**
  - **Version:**
  - **License:**
  - **Owner: traqbeat**

- **Intended Use:** The proposed solution is aimed at reducing hospitalization duration and cost and increasing the quality of life of patients suffering from these diseases.
- **Metrics:** Accuracy, Precision, Recall, F1 Score
- **Evaluation Data:** Age, sex, pH, diastolic pressure, systolic pressure, respiratory rate, GCS, pneumonia, pulse rate, smoking pack years
- **Limitations:**
- **Ethical Considerations:**

Figure 1: Model card for REALM use case: AI model ASCOPD from Traqbeat, for COPD and Asthma inpatient risk stratification

REALM OMOP Data Catalogue is a PostgreSQL database repository where patient data is stored, which can be utilised for research purposes. These include open-source or synthetic data generated within the REALM framework. The structure of the data follows the OMOP-CDM standardisations.

RIANA: REALM Intelligent Analytics Dashboard is a framework for defining complex patient groups or cohort definitions, specifically designed to capture the intent of the AI model while leveraging OHDSI methodologies and OMOP-CDM standards. RIANA is a user-friendly dashboard that enables the user to create the target patient group required to evaluate the AI model. It can also be used to check the availability of relevant patient data in the REALM data catalogue, ensuring that AI models can be evaluated against real-world datasets. RIANA can be used to retrieve a high-level summary of statistics of the available patient population. Despite these usability the dashboard provides OMOP-CDM auto-complete fields streamlining the selection of standardised concepts and ensuring interoperability with OMOP-based datasets. The automation of this usability through an RAG-enhanced LLM pipeline, which is the main focus of this project, will be explained in the methodology section of this report. Fig. 2. The integration of the pipeline to the RIANA dashboard will not be implemented as part of this thesis; however, it is explained in detail with a workflow diagram in the later section for readers to understand the use cases and the complete integration and impacts.

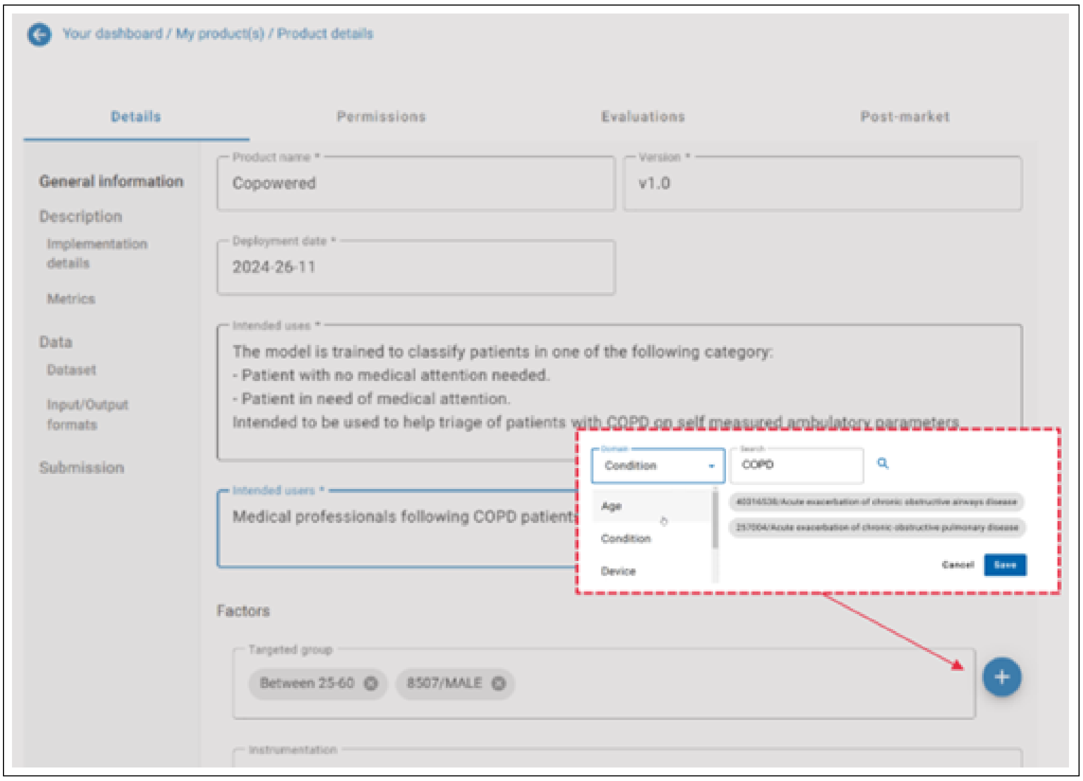


Figure 2: RIANA dashboard interface, where AI model developers can search for the required features in the search tab by specifying the domain from the drop-down list, and the suggestions will appear as the output, which are allowed to be saved

This thesis focuses on the REALM use cases, which use plain structured text files (CSV, TSV, TXT, etc.) as input/features. Some of the use cases use imaging data or genomic data; they do have OMOP-CDM standards, but this project excluded features from those use cases.

2 Data Description

This section provides a detailed description of the data used throughout the pipeline. The data included in the whole pipeline comes from OHDSI's standard vocabularies, which are open source and downloaded from OHDSI's vocabulary page OHDSI Vocabulary (2025) and stored in-house as SQL tables.

2.1 OHDSI Standardized Vocabularies

An OHDSI standardised vocabulary serves as the central component of OMOP-CDM. The vocabulary contains different concepts and terminologies from multiple standards The Book Of OHDSI, 2025. Each standard is referred to as vocabulary with its corresponding vocabulary_id's, for example, SNOMED, ICD-10, RxNorm Extension, LOINC and OSM, which are stored in standardised vocabulary tables of OMOP-CDM OHDSI Vocabulary, 2025. The ontologies, such as ICD, SNOMED or LOINC, follow their standard codes and descriptions. OMOP-CDM unified all those identification systems into unique concept IDs that allow for identifying all the terms regardless of the ontology they come from. The OMOP-CDM follows a complex hierarchical structure that allows the user to keep track of the source ontology they are derived. Whenever concepts overlap between ontologies, the standard one is allowed to be used by default (Appendix A.2). Each of these terms is assigned a unique concept ID within the OMOP-CDM, which is used as the unique identifier for each concept. This analysis uses the Systematic Nomenclature of Medicine - Clinical Terms (SNOMED-CT Version: 2024-02-01 International Edition) vocabulary from OMOP-CDM of v5.4 (see Annexe A.2). SNOMED-CT is a collection of medical terms with its standard codes, terms, synonyms, and definitions used in clinical documentation and reporting. The standard terms covered in it include clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices, and specimens OHDSI Vocab. SNOMED-CT (2025).

SNOMED-CT: No of concepts in each domains (n)	
Domain	**Count (n)**
Conditions	112051
Procedures	50305
Measurements	19639
Drug	211743
Device	14837
Observation	113401
Total	**521976**

Figure 3: The table contains the number of concepts available in each of the six domains with SNOMED-CT vocabulary. These concepts serve as the input data to the RAG pipeline.

Each clinical entities or term are categorised into different domains, which define the ontology of that specific concept within the standardized vocabulary. This analysis takes into account the six major domains within OMOP-CDM, including CONDITION, PROCEDURE, OBSERVATION, MEASUREMENT, DRUG and DEVICE.

- **CONDITION:** Major diagnosis or diseases and clinical findings.
- **PROCEDURE:** Any manual or robotic manipulation on a patient.
- **OBSERVATION:** Observational finding or assessment, and observable entity.
- **MEASUREMENT:** Measure of an analyte or entity, including the assessment scale.
- **DRUG:** Drug product or Vaccines.
- **DEVICE:** Includes medical device, contrast media, blood product, dietary supplement

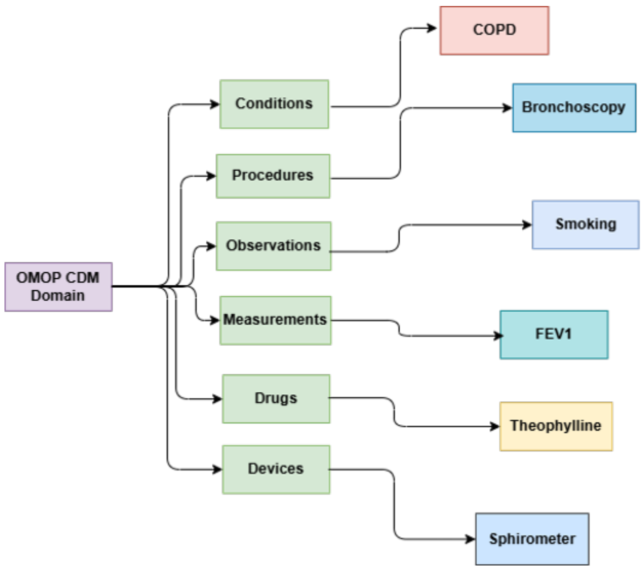


Figure 4: The figure demonstrates an example of concept or terminology included in six major domains of OMOP-CDM with an example use case of COPD.

In Fig. 3 summarizes the number of concepts available within the SNOMED vocabulary of six major domains. In Fig. 4 illustrate an example use case COPD, with condition as COPD, procedure done for COPD as bronchoscopy, smoking as one of the observations reported for COPD patient, FEV1 as measurement clinically taken, Theophylline common drug for treating COPD and spirometer as a device which is commonly used to measure lung capacity.

2.2 Extracted features from Model Cards

The extracted features define the population/cohort of the six REALM use-cases. These are variables or data points required for evaluating the AI models, which are then categorized into six OMOP-CDM domains: conditions, measurements, observations, devices, procedures and drugs. Figure 5 includes the extracted features, used in evaluating the proposed pipeline.

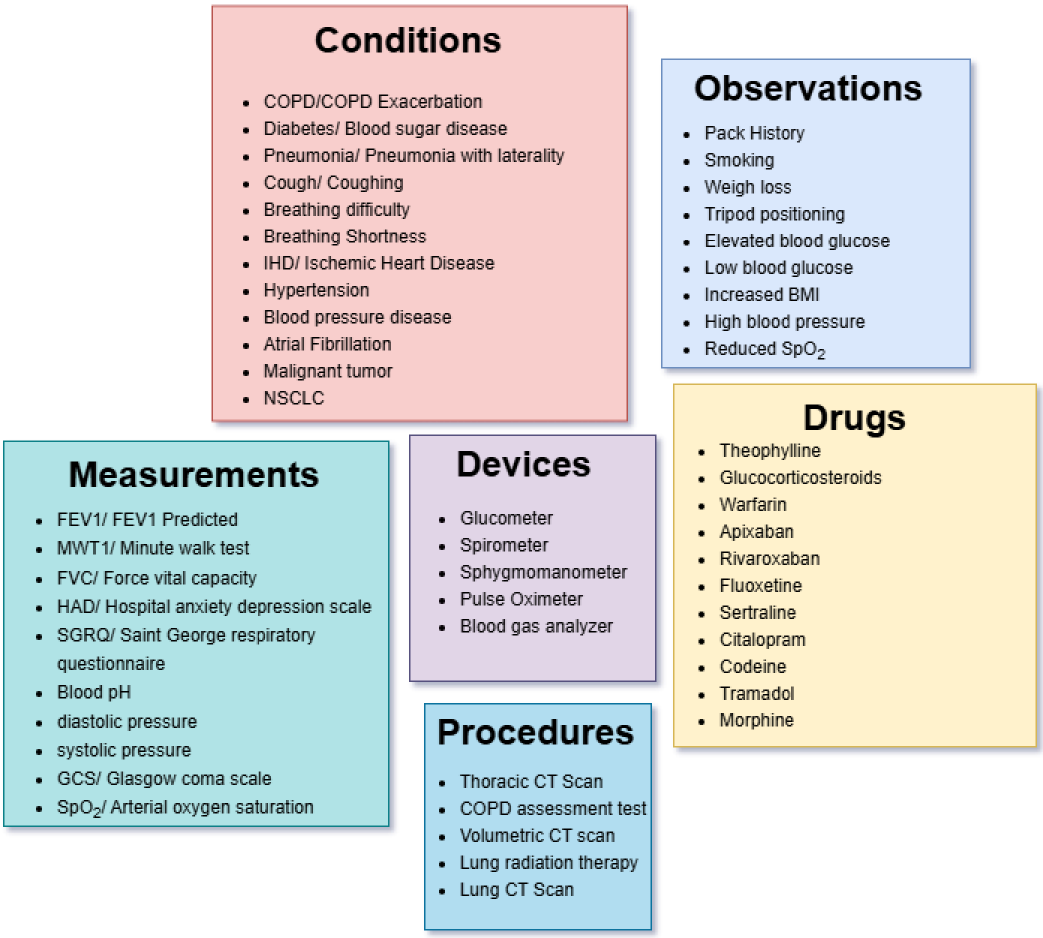


Figure 5: The extracted features from the model card separated into six main OMOP-CDM domains

3 Methodology

3.1 Retrieval Augmented Generation

RAG was first introduced in 2020 by Lewis et al. (2020). A classical RAG workflow involves Indexing, Retrieval, Augmentation, and Generation Gao et al. (2023). RAG combines the pre-trained retriever with a pre-trained seq2seq model for generating the output in Natural Language Processing (NLP) contexts. RAG is primarily useful for tasks that involve the generation of output from a defined data source. RAG get its prominence due to the limitations of LLM, such as generating hallucinated outputs. RAG can enhance the LLMs by finding the relevant and most plausible responses through the similarity search, and then the LLM can use this to generate the context-relevant response as the user output.

Indexing: In the initial indexing stage, the data index for the source data, which is SNOMED-CT vocabulary concepts, is generated and stored in a vector database for obtaining quick and easy search results in the further steps.

Vector Storage and Embeddings Generation: A Vector database, which is often referred to as a knowledge base, contains and stores the vectors. For the text format of input concept_name, embeddings are generated by transforming to numerical values called vectors. These embeddings capture the exact semantic meaning of the text. In this analysis, dense vector embeddings are generated with two models, one with 'all-MiniLM-L6-v2' from Sentence Transformers and bge-large-en-v1.5 from Hugging Face Wolf et al. (2020) as in Fig. 6 .

Retrieval Step: The retriever can access the embeddings stored in the Chroma database and retrieve the relevant matching concepts as per the user query through similarity search. The user query is encoded as a vector through the same embedding process that is used to generate vectors of the concept names. The query vector is compared to the concept vector to identify the top-k similar or matching concepts to the query.

For a given query *'q'*, to retrieve a top k relevant concept *'C'* from a large corpus V Gai et al. (2024)

$$c = c_1, c_2, c_3, \ldots\ldots\ldots\ldots, c_k$$

The similarity or the distance metrics is computed with cosine similarity using the equation,

$$\text{cosine\_similarity}(\mathbf{q}, \mathbf{c}) = \frac{\mathbf{q} \cdot \mathbf{c}}{||\mathbf{q}||\,||\mathbf{c}||}$$

$$sim(q,c) = \langle Encoder_q(q), Encoder_c(c) \rangle$$

Retrieve the top k matching concepts,

$$C = topk_{d \in C}(sim(q,c))$$

Hence, this ensures only the matching results are retrieved and passed to the next step Lewis et al. (2020).

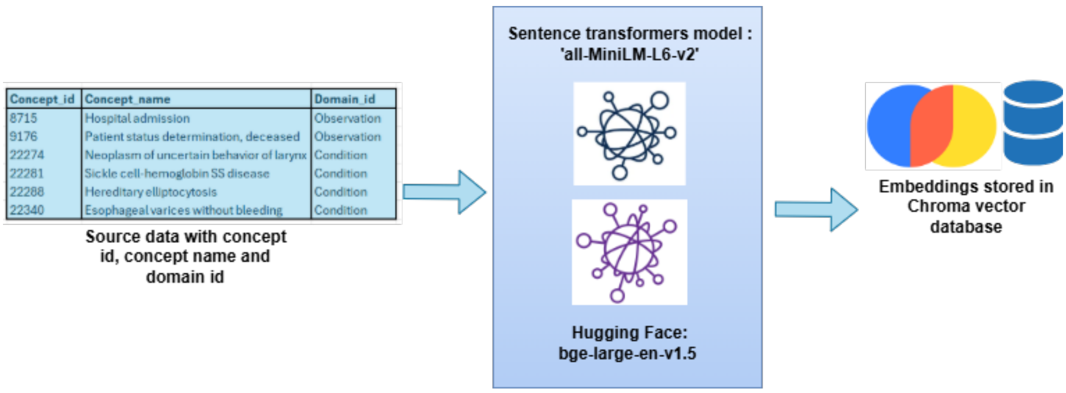


Figure 6: The embeddings for the OMOP-CDM standard concepts from SNOMED-CT vocabulary are generated using 'all-MiniLM-L6-v2' or 'bge-large-en-v1.5' models and then stored in the Chroma vector database.

Augmentation Step: In the augmentation step, the input query is concatenated with the retrieved concepts and passed to the generation step to produce the contextually relevant output with the help of LLM.

$$x_i = concat(q, c_i)$$

for i=1,...,k

Generation Step: In the generation step, the context-relevant output is generated with the help of LLM. In addition to the top k concept names and IDs, the external relevant information is added to the output.

For the input query q, the probability of retrieving the response which is y is computed by marginalising over the relevant retrieved concepts k,

$$P(y \mid q) = \sum_{n=1}^{k} P(y \mid q, c_i) \cdot P(c_i \mid q)$$

$P(y|q)$ is the probability of retrieving relevant concepts k, given query q computed with cosine similarity.

$P(y|q,d_1)$ is the probability of retrieving relevant concepts k, given query q and the concepts Lewis et al. (2020).

Loss Function: RAG can simultaneously train the retrieval process and output generation by marginal likelihood objective, backpropagation updates both components together in the workflow.

$$\mathscr{L} = \sum_{(x,y)} \log \sum_{q} P(q \mid c) \cdot P(y \mid q, c)$$

3.2 Large Language Model (LLM)

Large language models are trained on a large amount of data, marking a significant advancement in the natural language processing field. LLM has the ability of deeper semantic search, which can enhance the quality of the concept mapping pipeline Yan et al. (2024).

Large Language Model Meta AI: LlaMa

This thesis utilize the capability of a widely used large language model from Meta AI, 'llama3.3:70b -instruct-q4_K_M' Touvron et al. (2023). Llama is an open-source model, and can be self-hosted by the project owners without the need for a third-party interface, making it well-suited for health care applications by retaining data privacy. LlaMa outperforms other open source state-of-the-art models in terms of scalability and cost effectiveness, along with its high performance Huang et al. (2024). Like any other LLM, LlaMa is built on transformer-based architecture. The attention mechanism within transformers enables them to generate human-like responses by understanding the contextual meaning in the text. Even though LLMs are trained on a vast amount of data, they still struggle with factual corrections of the response being generated.

3.3 Prompt Engineering

Prompting is a very crucial step in crafting well-specified inputs for generating the desired output. There are different ways of prompting, including zero-shot prompting without including any instructions or examples, few-shot prompting with fewer instructions, instruction-based prompting with well-defined instructions, contextual prompting by providing relevant context, and bias mitigation prompting commonly used for debugging purposes Marvin et al. (2023). Since the source data from which the response is expected will already be fed to LLM through the RAG pipeline, zero-shot prompting will be utilised in this analysis at the beginning Wang et al. (2019). Depending on the retrieved results, few-shot prompting or instruction-based prompting can be introduced further in the workflow.

3.3.1 Zero-Shot Prompt

A simple prompt starting with *'What is the concept ID for'* as in Fig. 7 is used as input to the RAG pipeline. The extra explanations and suggestions are retained as is at this stage. Further fine-tuning of the LLM response can be performed when the pipeline is fully connected to the RIANA dashboard, which is outside the scope of this thesis.

```
Question: What is the concept ID for COPD?
Answer: The concept ID for COPD (Chronic Obstructive Pulmonary Disease) is not explicitly stated in the provided lis
t with that exact name. However, a closely related concept is listed as:

- Concept ID: 255573, Concept Name: Chronic obstructive lung disease, Domain ID: Condition

This can be considered synonymous with COPD for many purposes, as both terms refer to conditions characterized by ai
rflow limitation and are often used interchangeably in clinical contexts.

If you're looking for a more specific or direct reference to COPD, it might be worth noting the following options wh
ich are closely related but specify different stages or aspects of the condition:

- Concept ID: 4209097, Concept Name: Severe chronic obstructive pulmonary disease
- Concept ID: 46284163, Concept Name: Chronic obstructive pulmonary disease resolved
- Concept ID: 36685454, Concept Name: GOLD (Global Initiative for Chronic Obstructive Lung Disease) 2017 group D
- Concept ID: 4193588, Concept Name: Moderate chronic obstructive pulmonary disease
- Concept ID: 4196712, Concept Name: Mild chronic obstructive pulmonary disease
- Concept ID: 44791725, Concept Name: Very severe chronic obstructive pulmonary disease
```

Figure 7: The figure represents the example of zero-shot prompt used in the pipeline for retrieving the relevant OMOP-CDM concepts

3.4 Project Pipeline: Architecture Overview

The proposed tool utilises the RAG pipeline to create an automated vocabulary mapping for OMOP-CDM concepts. The pipeline is implemented using the Python programming language version 3.16. The features that are required to be mapped are expected to be extracted through the RIANA dashboard. The output of this application will be a list of features that describe the patient target group. This is the same as what is available in the evaluation data section of the model card. Once the AI model developer inputs the required features, it is fed as input to the pipeline through an API call. The concept name for each corresponding concept id, concept name and domain id were extracted. The RAG pipeline was generated by creating embeddings for text-based concept names and storing them in a vector database. The RAG was then connected to the LLM through the direct prompt injection method. When a user inputs a query, embeddings will be generated for the input query with the same method, through the similarity search between the query embeddings and vocabulary embeddings. Then, top 'k' matches are created for the concept name with its corresponding concept ID. This information is then sent to the LLM and is retrieved as an output, which provides suggestions for concept names and their corresponding IDs. A well-defined, structured prompt was designed to give a clear, standardised query as input and to retrieve meaningful responses from the pipeline.

Figure 8 provides an overview of the RAG-LLM Semantic Mapping Pipeline. This end-to-end pipeline integrates a Retrieval-Augmented Generation (RAG) architecture with a Large Language Model (LLM) to support the semantic mapping of clinical features to OMOP-CDM concepts. There are three main sections in this pipeline:

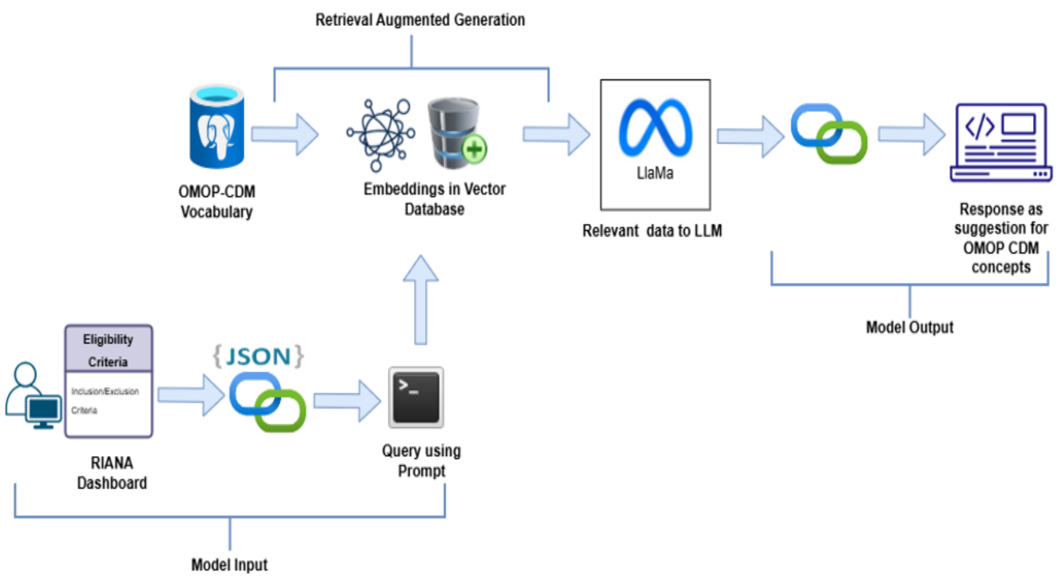


Figure 8: The overall pipeline for mapping features to OMOP-CDM

Section 1: User inputs cohort features as natural text in the RIANA dashboard, which is transmitted through an API in JSON format (Appendix D). As in Fig. 9, the user can select the required domain through the drop-down section of the RIANA dashboard and type the feature or the concept to be mapped in the search box. At the backend, this searched concept is converted into a query to the LLM prompt.

Section 2: Input is encoded and compared against pre-generated embeddings (OMOP concepts stored in a vector database). This is the main section where the features are semantically mapped to the standard OMOP-CDM concept. The top-k most semantically similar matches

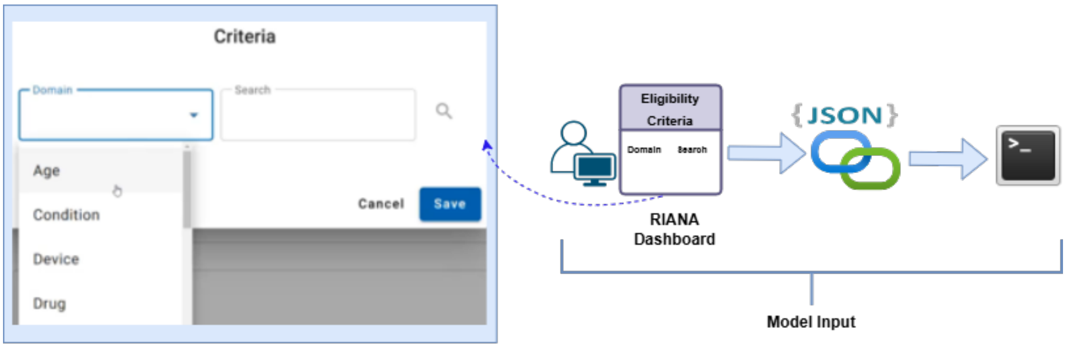


Figure 9: Input of the pipeline generated through the RIANA dashboard

are retrieved and passed to the LLM using a structured prompt.

Section 3: The LLM generates context relevant concept_ids as suggestions at the RIANA user interface. As in Fig. 10, the mapped features according to the input search term appear as suggestions. The user will be allowed to select the required terms from the available list and save.

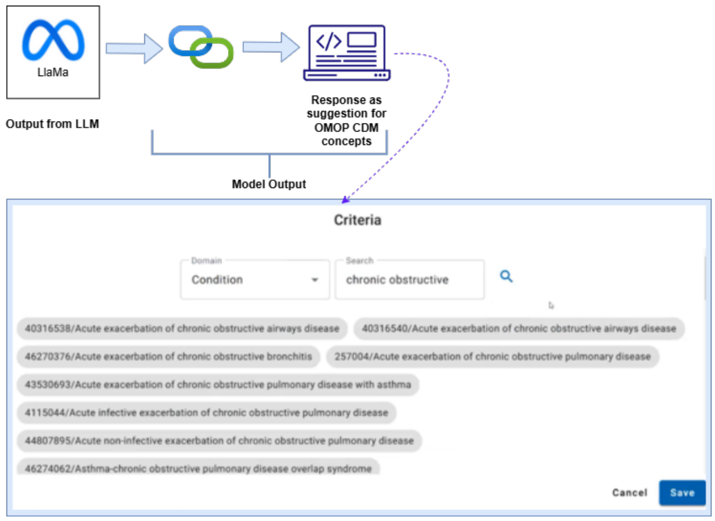


Figure 10: Output of the pipeline fed to the RIANA dashboard.

3.5 Evaluation Metrics

Once the responses have been received from the pipeline, the output is compared with the concepts generated using the Athena OHDSI tool and evaluated against the concepts derived by an OMOP-CDM expert, which is considered the ground truth. The ground truth includes the concepts derived from the six REALM use-cases, mapped to the concept IDs of OMOP-CDM standard terminologies of SNOMED vocabulary by a domain expert. This was time-consuming and required knowledge of medical terminology along with OMOP expertise. The same terms that serve as the input to the pipeline will be used in Athena as the search term by applying filters, relevant domain, concept standard, vocabulary as 'SNOMED', and validity as 'valid'. The top k:(5, 10) results of vector search and Athena are compared. The evaluation metrics, such as precision, accuracy, recall and F1 scores, were used to validate the retrieved result Sawarkar et al. (2024).

- **Accuracy:** Measures the accuracy of the retrieved result among all searched concepts.

$$\text{Accuracy} = \frac{\text{Number of correctly mapped concepts}}{\text{Total number of concepts}}$$

- **Precision:** Measures the proportion of relevant concepts among the retrieved concepts.

$$\text{Precision} = \frac{\text{Number of correctly mapped concepts}}{\text{Total number of mapped concepts in predicted and ground truth}}$$

- **Recall:** Measures the proportion of relevant concepts which are successfully retrieved

$$\text{Recall} = \frac{\text{Number of correctly mapped concepts}}{\text{Total number of mapped concepts in ground truth}}$$

- **F1 Score:** Measures the harmonic mean between precision and recall, which provides a balance between precision and recall.

$$\text{F1 Score} = 2.\frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

Furthermore, the results obtained from the RAG and that from the LLM will be compared to investigate the usefulness. The human evaluation will also serve as a crucial metric to evaluate the contextual appropriateness for a qualitative assessment. The computation time and cost will be evaluated to assess the feasibility of using the proposed pipeline over the existing one.

3.6 Experimental Setup

The entire setup for the integration and testing of the pipeline is achieved through the Vlaamse Instelling voor Technologisch Onderzoek (VITO) infrastructure Vito, BE 2025.

VITO is a large contributor to the REALM project and has a crucial role in data standardisation and regulatory methodologies Vito-REALM, BE 2025.

The RAG-LLM pipeline code was implemented and tested within the VITO infrastructure as shown in Fig. 11, using the VITO hardware (VITO HPC) and software (Code written on JupyterLab notebooks and version-controlled on Bitbucket) with the following specifications:

- **Hardware Specification:** A GPU platform, for high-performance computing (HPC), is used to connect the LLM interface. Testing environment with 2 CPU, and 4 GB of memory facilitated the pipeline integration with LLM, which was running in a separate machine within the Thunder server of 256 GiB of System memory.
- **Software Specifications:** The software setups with the Python frameworks were used. Sentence transformers from the SBERT with the Python module were used to generate embeddings. Chroma DB for storing vectors. The source dataset stored in a separate machine of a PostgreSQL database, which includes OMOP 5.4 Database together with Athena SNOMED Vocabulary (v20240830), was retrieved to the HPC interface.

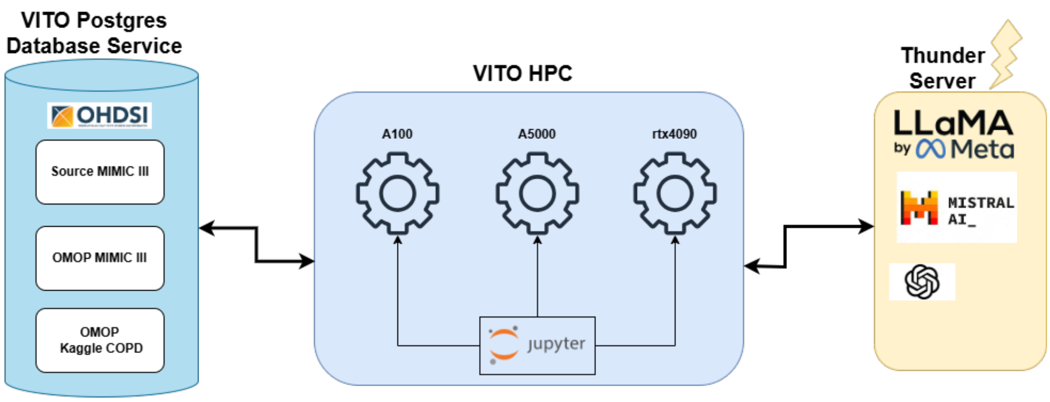


Figure 11: Experimental setup within VITO infrastructure

4 Results

The analysis of the experimental pipeline was conducted in three steps. During the initial stage, a separate pipeline was introduced for each domain. The results were then compared using different parameters and methods to proceed with the next stage. In the second set of evaluation, a single pipeline is built by creating a vector of concepts from all six domains into one single vector database and thereby querying one single RAG pipeline for retrieving concepts irrespective of the specific domain. This was mainly to compare the quality of embeddings generated during both steps and their impact on the retrieval process. In the later stage, a more advanced RAG is implemented by introducing query routing within the pipeline before the retrieval process. Finally, this was chosen as the optimal approach out of three. These will be discussed separately in each section.

To ensure the generalizability and broader application Gabín and Parapar (2025), the pipeline is evaluated for search terms including synonyms, acronyms, abbreviations, false terms and even with natural language terms without proving the actual medical terminology alone. This involved 'diabetes: blood sugar disease', 'thermometer: temperature measuring device', etc. On average, 15 to 20 test cases were evaluated for each domain.

Evaluation metrics such as accuracy, precision, recall and F1 score were assessed at each stage of the evaluation. The retrieved results were evaluated against the ground truth concepts generated by an OMOP-CDM expert for the five use cases of the REALM project. The same search terms are used to search in Athena with applied filters, and top-k results are compared to the pipeline.

4.1 Initial Prototype with Domain: CONDITION

As an initial step, a simple pipeline was created with the standard OMOP-CDM concepts from the domain CONDITION. The embeddings were generated using two models, Model 1: 'all-MiniLM-L6-v2' and Model 2: 'bge-large-en-v1.5'. The top-k retrieved results were compared for k = 5 and 10. The results were compared against the results generated with the OHDSI tool Athena. The concepts mapped by the domain expert served as the ground truth. The detailed evaluation metrics for this pipeline are added to the tab. 1.

	Top-5			Top-10		
Metrics	**Model 1**	**Model 2**	**Athena**	**Model 1**	**Model 2**	**Athena**
Accuracy	0.75	0.54	0.45	0.78	0.66	0.45
Precision	0.85	0.62	0.45	0.86	0.68	0.45
Recall	0.80	0.58	0.48	0.89	0.70	0.48
F1-Score	0.82	0.60	0.46	0.88	0.69	0.46

Table 1: Evaluation metrics comparison for Top-5 and Top-10 retrievals.

The experimental pipeline showed a significantly better result for mapping the cohort features to OMOP-CDM concepts compared to the mapping generated by Athena. The pipeline generated using vector embeddings of Model 1 showed promising results compared to Model 2. There is also a gradual improvement in the evaluation metrics when moving from k=5 to k=10. The Model 1 with top k = 10 achieved an accuracy of 78%, Precision of 86%, a recall of 89% and an F1 score of 88%. Hence, the pipeline with Model 1 and top k = 10 was chosen for further analysis.

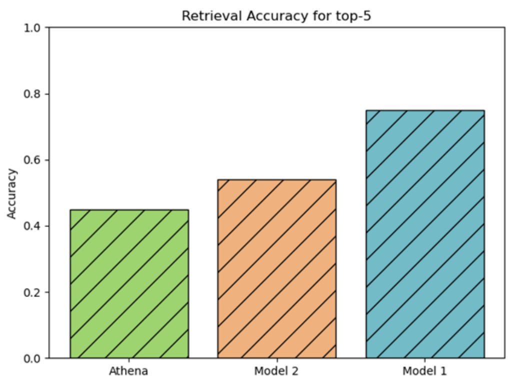


Figure 12: Retrieval accuracy for k = 5

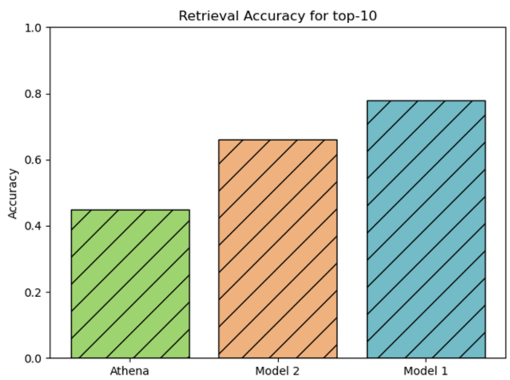


Figure 13: Retrieval accuracy for k = 10

Fig. 14 shows the result generated by the LLM prompt for the search term 'pneumonia with laterality'. The pipeline correctly maps the concepts for each laterality, left or right, along with each zone, upper and lower. It also provides a clear explanation on choosing the relevant concept ID, which is highly useful for someone who lacks medical expertise. These additional explanations or context-relevant guidance provided by the LLM avoid the additional research required while performing the concept mapping for a non-domain expert. As the focus of this thesis is to suggest the concepts rather than providing the exact matching terms alone, these extra explanations can be useful.

In Fig. 15, the mapped concepts of Athena for the same keyword did not show even a single relevant concept in the results k = 10. This ensured the reliability of the pipeline to continue with the next steps in a more intricate approach.

```
Question: What is the concept_id for Pneumonia with laterality?
Answer: The concept_id for Pneumonia with laterality can be broken down into several options based on the provided information:

- **Left lower zone pneumonia**: Concept ID is 4116488.
- **Right lower zone pneumonia**: Concept ID is 4117114.
- **Left upper zone pneumonia**: Concept ID is 4114030.
- **Right upper zone pneumonia**: Concept ID is 4114031.

These concept IDs represent pneumonia with specific laterality (left or right) and location (upper or lower zone). If you are l
ooking for a general concept of pneumonia that includes the idea of laterality without specifying the exact location within the
lung, none of the provided options directly offer a "Pneumonia with laterality" concept ID. However, based on the closest match
from the given list which generally refers to pneumonia and includes an aspect of laterality (even though it specifies zones),
you might consider:

**Concept ID: 4116488**, **Concept ID: 4117114**, **Concept ID: 4114030**, or **Concept ID: 4114031** depending on the specific
zone and side.

But if we're strictly interpreting "Pneumonia with laterality" as a need for a concept that represents pneumonia in general whi
le acknowledging laterality without specifying zones, none of the given options perfectly fit this description. The closest gen
eral term provided is:

- **Concept ID: 255848**, Concept Name: Pneumonia.

However, this does not inherently include the concept of laterality. For a direct answer to "Pneumonia with laterality," it see
ms there might be an expectation for a specific concept that combines these ideas directly, which isn't explicitly listed in th
e provided options. Given the information and looking for a concept ID related to pneumonia and considering aspects of location
(laterality), you would have to choose based on the specificity needed (e.g., zone and side).
```

Figure 14: Output from LLM prompt: the concepts mapped by the RAG-LLM pipeline (domain:CONDITION) for the search term 'Pneumonia with laterality'

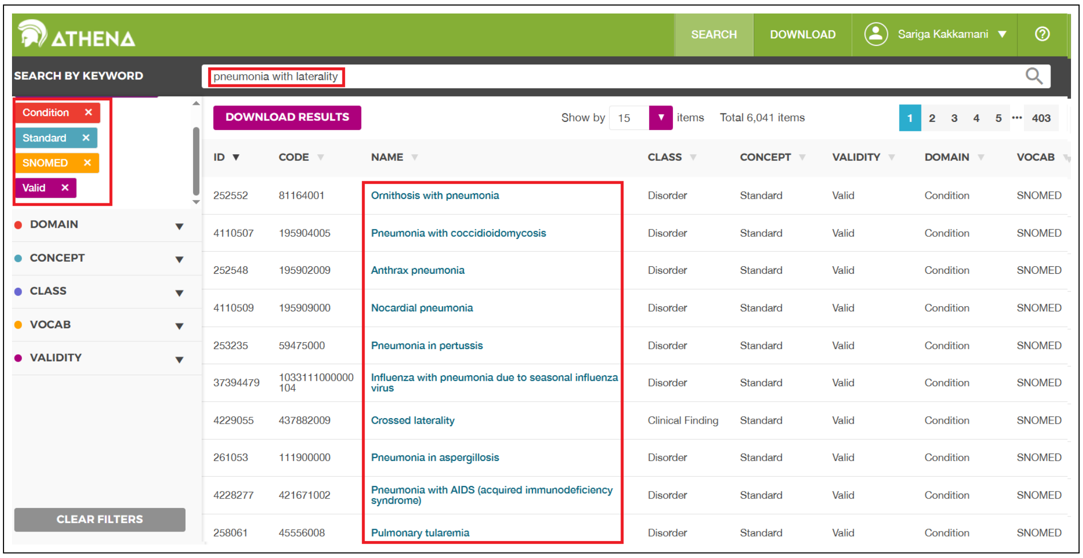


Figure 15: Concept search result example on Athena - the concepts search results mapped by Athena's built-in search algorithm for the term 'Pneumonia with laterality' and filtered by SNOMED terms, for domain CONDITION and valid standard concepts.

4.2 Separate pipeline for each domain

Based on the significant and relevant results generated using the initial pipeline, separate models for each remaining five domains, which include OBSERVATION, MEASUREMENT, PROCEDURE, DRUG and DEVICE, were implemented with the same setup. The embeddings were generated using 'all-MiniLM-L6-v2' from Sentence Transformers and

’llama3.3:70b -instruct-q4_K_M’ as LLM; the top k = 10 results were evaluated against the ground truth and compared with the concepts mapped using Athena. The detailed evaluation metrics are summarised in the tables below.

Table 2: Performance metrics comparison between RAG-LLM and Athena

Domain: Measurements		
Metrics	**Model 1**	**Athena**
Accuracy	0.63	0.50
Precision	0.77	0.61
Recall	0.70	0.55
F1-Score	0.73	0.57

Domain: Procedure		
Metrics	**Model 1**	**Athena**
Accuracy	0.91	0.37
Precision	0.90	0.30
Recall	0.95	0.45
F1-Score	0.92	0.36

Domain: Observation		
Metrics	**Model 1**	**Athena**
Accuracy	0.86	0.73
Precision	0.84	0.69
Recall	0.91	0.75
F1-Score	0.88	0.72

Domain: Drug		
Metrics	**Model 1**	**Athena**
Accuracy	0.92	0.50
Precision	0.91	0.58
Recall	0.92	0.54
F1-Score	0.93	0.50

Domain: Condition		
Metrics	**Model 1**	**Athena**
Accuracy	0.78	0.45
Precision	0.86	0.45
Recall	0.89	0.48
F1-Score	0.88	0.46

Domain: Device		
Metrics	**Model 1**	**Athena**
Accuracy	0.93	0.87
Precision	0.92	0.85
Recall	0.86	0.79
F1-Score	0.89	0.81

From the results, it is evident that the proposed pipeline was able to achieve a comparable or higher precision and accuracy compared to the commonly used tool, Athena. Of the six domains, the highest accuracy is achieved for the domains PROCEDURE (91%), DRUG (92%), and DEVICE (93%). This can be inferred as these domains consist of more standardised terms compared to the other three domains. The high diversity of terms in the CONDITION domain and OBSERVATION domain increases the complexity of the mapping process. Another major findings were noted from the domain MEASUREMENT. Although the number of concepts was comparatively lower in the measurement domain, it was unable to achieve higher accuracy. This could be due to the acronyms or similar terms present in the source data. The pipeline failed to correctly map the concepts such as ’HAD scale’ with ’Hospital anxiety and depression scale’ or ’SGRQ score’ with ’Saint George’s respiratory questionnaire score’. However, these were correctly recognised by the LLM, but no matching embeddings were generated in the top-10 results of the retrieval step.

Figure 16: Comparison of retrieval accuracy between RAG-LLM and Athena on six vocabulary domains: CONDITION, PROCEDURE, OBSERVATION, MEASUREMENT, DEVICE, DRUG

As in Fig. 16, the retrieval accuracy of both methods is somewhat similar for the domains OBSERVATION, MEASUREMENT, and DEVICE. However, there is a significant difference in the accuracy between the two methods for domains CONDITION, PROCEDURE and DRUG

```
Question: What is the concept id for theophylline tablet?
Answer: The concept IDs for theophylline tablet are:

* 4191531 (Theophylline anhydrous 300 mg oral tablet)
* 4132840 (Theophylline 300 mg oral tablet)
* 4132231 (Theophylline anhydrous 200 mg oral tablet)
* 4133425 (Theophylline 250 mg oral tablet)
* 4130499 (Theophylline 200 mg oral tablet)
* 4188932 (Theophylline 100 mg oral tablet)
* 4163321 (Theophylline anhydrous 400 mg oral tablet)

Note that there are also concept IDs for modified-release tablets:

* 46102554 (Theophylline 300mg modified-release tablets 56 tablet)
* 46102553 (Theophylline 300mg modified-release tablets 100 tablet)
* 46102565 (Theophylline 200mg modified-release tablets 56 tablet)

If you are looking for a specific concept ID, please let me know and I can help you narrow it down.
```

Figure 17: Output from LLM prompt: Retrieved result for search term ’theophylline’ from the model with the separate pipeline for the domain DRUG

One of the significant advantages of the pipeline is that relevant terms for the medication with varying dosages are retrieved with proper suggestions to choose from. For instance, with a single keyword search for the term ’*theophylline*’, all the relevant medications with varying dosages are retrieved, as in Fig. 17. This is beneficial when we are unaware of all the possible options to choose from.

4.3 Single RAG-LLM pipeline with all six domains

Using a separate RAG-LLM pipeline for each different domain is not usually advisable. Apart from the six main domains, there are other domains like visit, geography, currency, cost, specimen, etc. This will, in turn, lead to scalability issues when we start implementing the pipeline for multiple domains across multiple vocabularies. Maintaining a separate connection with LLM for each domain will also increase the computation cost and time. However, the separate pipeline can be advantageous in situations where each domain needs to be segregated without having any interconnection between the retrieving information. In this case scenario, at the end, once the pipeline is fully functional by incorporating it with the RIANA dashboard, we would require all domains to be added as a single pipeline. Fortuitously, the setup of the RIANA dashboard we can filter on the right domain along the search term or the feature to be mapped. This will allow the user to select the domain first and then search for the required concepts. Hence, maintaining a single pipeline for all domains is more advantageous and maintainable here.

In the second stage of this thesis, vector embeddings were generated for the concept names from SNOMED vocabulary, this time by including a total of 521976 concepts into a single vector database. The embeddings were generated using 'all-MiniLM-L6-v2' from Sentence Transformers and 'llama3.3:70b -instruct-q4_K_M' as LLM. The concepts were mapped using the same parameters as before (top k=10). The keywords used for the evaluation of separate pipelines were still used here to generate the mappings against the natural language like terms to concept_ids and concept_names of OMOP-CDM standards. The results generated were compared against the previous results generated by the pipeline.

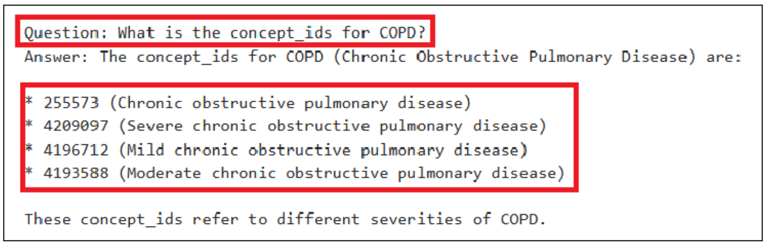


Figure 18: Output from LLM prompt: Model with separate pipeline for CONDITION domain for the search term COPD.

Figure 18 contains the retrieved result from the pipeline with a single domain (condition) as the input embeddings in the vector database. The search term 'COPD' correctly maps to the relevant ground truth with top k=10 retrieved terms. Fig. 19 contains the retrieved result from the second pipeline with all six domains in a single vector database. The same search term as 'COPD' is used for mapping and without specifying the domain in the input query. The pipeline failed to match the exact or relevant terms with the ground truth. Rather, it is mapped to the concept from the OBSERVATION domain.

```
Question: What is the concept id for COPD?
Answer: To find the concept ID specifically for COPD itself and not related concepts such as management, monitoring, or exceptio
ns from quality indicators, we look through the provided list. Most of the given Concept IDs relate to specific actions, measure
ments, or exceptions concerning COPD rather than the condition itself.

However, none of the listed Concept IDs directly state "COPD" without additional context (like management, monitoring, etc.), wh
ich makes it challenging to pinpoint a single ID that represents COPD in its simplest form.

Given the options and looking for the most direct reference to COPD:
- **Concept ID: 42689543** refers to "Number of hospital admissions due to COPD (chronic obstructive pulmonary disease) in past
year," which directly involves COPD as a condition.
- Other IDs involve actions or exceptions related to COPD, not the condition itself.
```

Figure 19: Output from LLM prompt: Model single pipeline for all six domains without specifying the domain in the query for COPD use case.

The search term ’COPD’ was queried again in the same pipeline by specifying the domain (CONDITION) in the query as in Fig. 20. This time, none of the terms were retrieved. LLM respond to the query with an explanation that the top-10 matched terms are from the Observation domain, which is irrelevant for the requested query and semantically do not match the condition of ’COPD’ directly.

```
Question: What is the concept id for COPD from condition domain?
Answer: To find the concept ID for COPD from the condition domain, we need to look through the provided list and identify which
concept directly represents Chronic Obstructive Pulmonary Disease (COPD) as a condition rather than an observation related to it
s management or monitoring. However, upon reviewing the list, it appears that all concepts provided are categorized under the "O
bservation" domain, which relates to observations or assessments made about patients, treatments, or disease management, rather
than conditions themselves.

Given this context, none of the provided concept IDs directly correspond to a condition domain for COPD but rather relate to var
ious aspects of COPD management, monitoring, or exceptions in quality indicators. Therefore, based on the information given, the
re isn't a direct answer that matches a concept ID specifically labeled as being from the "condition" domain for COPD.
```

Figure 20: Output from LLM prompt: Single pipeline for all six domains mapped terms by specifying the domain in the query for COPD use cases.

This indicates that using a single vector database is only useful when searching for an exact term. There are more chances of semantic overlap between the concepts when the number of terms increases. Adding more terms to the vector database can lead to embedding space saturation, failing to identify the differences between more similar concepts. Even though it returns similar concepts, it is not necessarily semantically correct or relevant to the input query. This also leads to increased search time due to the density of the vector created. Due to this low-quality performance of the pipeline, no further fine-tuning was performed on this setup.

4.4 Advanced RAG with Query Routing

During the third and final stage of the analysis, a more advanced but very simple implementation of the pipeline was adopted. This includes using six separate vector databases for each different domains and adding a static query routing between the user query and the retrieval step. This will allow to use of a single pipeline but with multiple context-specific vector databases altogether.

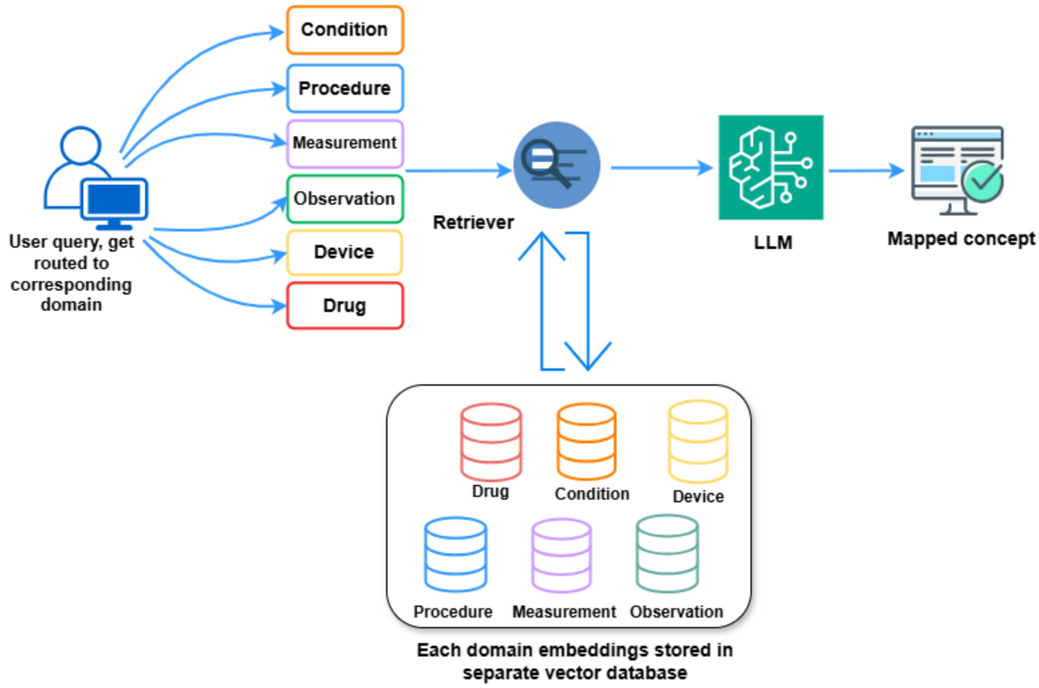


Figure 21: RAG-LLM pipeline with query routing by including all domains into a single setup

As in the Fig. 21, once the user enters the query for example '*What is the concept_id for COPD from domain condition?*', the keyword 'condition' is identified from the query and the searching is routed to the vector database which stores embeddings of concepts from domain CONDITION. A predefined rule-based query routing is implemented here. The semantically similar concepts are identified for the corresponding search term. This set-up works similarly to the separate pipeline created for each domain, as in the initial step. This will avoid the higher ambiguity generated from the second approach, also by avoiding calling LLM multiple times for each different domain.

```
Question: What is the concept_id for lung biopsy from domain procedure?
Answer: The concept_id for lung biopsy from the domain procedure is 4303062.
```

Figure 22: Output from LLM prompt: Query routing pipeline with example use case for specifying the domain in the query

```
Question: What is the concept_ids for COPD from domain condition?
Answer: The concept_ids for COPD from the domain condition are:

* 255573 (Chronic obstructive pulmonary disease)
* 4196712 (Mild chronic obstructive pulmonary disease)
* 4209097 (Severe chronic obstructive pulmonary disease)

Note that these three concept IDs all refer to different forms or severity levels of Chronic Obstructive Pulmonary Disease (COPD).

Additionally, the following concept ID is also related to COPD:

* 46274062 (Asthma-chronic obstructive pulmonary disease overlap syndrome)
```

Figure 23: Output from LLM prompt: Query routing pipeline, for COPD use case with specifying the domain in the query

Figure 23 shows that the COPD use case is now queried by specifying the domain, which resulted in a similar response received with the separate pipeline for the condition domain. The results generated through this query routing approach are the same as those produced through the single pipeline. Hence, the evaluation metrics and the benchmarking process remain the same as the initial one. The only difference is in the way it is implemented and queried.

5 Discussion

This thesis utilises a RAG-enhanced LLM framework to enable the semantic mapping of natural language cohort definitions to standardise queries compliant with OMOP-CDM, to improve the accuracy of feature mapping. The analysis was carried out in three stages, iteratively, and the final one by incorporating query routing to the RAG-LLM seems to be more efficient, compared to the first two. The sentence-transformed based embedding model outperformed the other benchmarking OHDSI's tool, Athena. However, it is important to note that the intention here is not to build a method that outperforms available tools; rather, we are trying to implement a pipeline that can be integrated into the REALM framework by providing suggestions for concept mapping, which is beyond the scope of standalone OHDSI tools. The pipeline achieved improved performance over the OHDSI tool Athena. Especially with the proposed approach, in addition to the suggestions on matching concepts, we get an explanation on which concepts are more appropriate as per the input query. This is advantageous over the available state-of-the-art methods, which focus merely on the concepts Zhou et al. (2025). The usage of RAG and query routing makes the pipeline more practical for use in real-world use cases. It allows us to choose among the vocabulary or domain as per our requirement, which can be easily implemented within the pipeline. Beyond REALM, the pipeline has broad applicability for real-world semantic mapping tasks, allowing non-expert users to get assistance in the semantic mapping process. This project contributes to the ongoing research of integrating AI in automating the semantic mapping process, specifically by using RAG and LLM.

This project is a small step towards a rather big goal: OMOP-CDM mapping is one of the biggest challenges within the OHDSI community. While Athena is available, it is highly complicated and has limited flexibility. The proposed tool has the potential to overcome these issues through AI-based semantic mapping. The same is well appreciated by the OHDSI community and hence has been accepted as a poster presentation and lightning talk in the upcoming OHDSI symposium.

Implications of the pipeline within REALM

REALM aims to develop a powerful sandbox environment for the future evaluation of AI as a medical device that goes on the EU market. In this context, we initiated a large effort for the mapping of the REALM use-case features to the OMOP vocabulary, which is initially based on an extended AI model card. Using these well-documented use cases, we can provide context-based features as a search query to the LLM and get suggestions of OMOP-CDM standard concepts. Figure 24 illustrates the high-level workflow where the interconnection between the model card, RIANA dashboard, and RAG LLM pipeline occurs to achieve efficient mapping. The mapping of the concepts happens at step (2). RIANA helps users to define patient cohorts, map input features and report performance metrics using standardised vocabularies and clinical concepts from the OMOP Common Data Model. This thesis plays an important role in mapping the input features to OMOP-CDM concepts. By mapping correctly to OMOP-CDM concepts, it not only allows for the identification of the standard concept, but it can also be applied to the data structure to extract the source data and then use it for evaluating the AI models. This structured approach ensures that each AI model is linked to a well-defined and varied target population, streamlining preparation for CE marking and compliance with emerging EU regulatory frameworks.

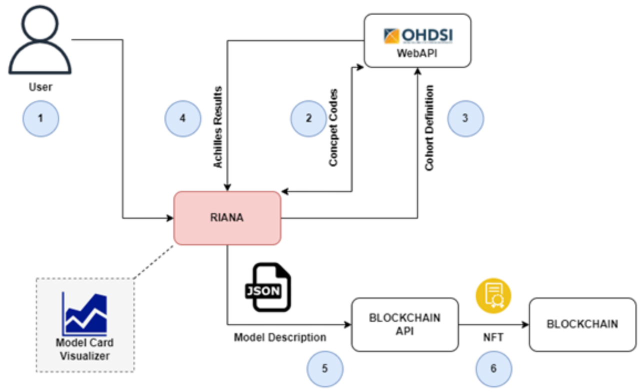


Figure 24: High-level user workflow in the REALM sandbox environment: (1) the user entering key information and usefulness of the AI model. (2) The user is guided through selecting appropriate standardised concept codes, including the definition of a clear patient target group using OMOP vocabularies. Based on this input, the user can (3) generate a cohort definition. A (4) high-level summary of the results is returned to the dashboard following Heracles analysis. These outputs are then (5) compiled into a structured JSON model card, which is (6) finally published to the blockchain, ensuring transparency, traceability, and long-term reproducibility of the model evaluation.

Limitations and Future Research

The major limitation of the proposed methods is that it is only focused on the SNOMED vocabulary with six commonly used domains. In the real world scenario, researchers use vocabularies such as ICD-10, LOINC, etc., with extended domains such as gender, visits, etc. However, this can be easily achieved by extending the pipeline with the required domain and vocabularies. Another limitation, which could be considered as future work, includes fine-tuning the embeddings to improve the accuracy, mainly in the condition and measurement domains. Additionally, it may also be possible to experiment with other LLMs; however, it may not impact the results much, as the pipeline primarily relies on embeddings. At this stage of the thesis, the RAG-LLM pipeline is not fully integrated into the RIANA dashboard, which is out of the scope of this thesis. Achieving this goal will occur over the coming months, allowing users seamless access through an intuitive interface. The current implementation of our RAG-enhanced LLM pipeline is operational within the VITO infrastructure, allowing interactive testing and refinement through prompt-based queries. The future work would include the complete integration of the pipeline to the RIANA dashboard, making it fully functional within the REALM environment. A proper investigation on the impact of computational cost and time will also be required to answer the open questions over the feasibility of the proposed approach.

6 Conclusion

This project presents a practical OMOP-CDM concepts mapping framework utilising recent technologies: RAG and LLM. This framework enables the semantic mapping of natural language cohort definitions to standardise queries with the help of LLM prompts, thus improving the accuracy of feature mapping. This was achieved in three stages: first by mapping the concepts from a single domain and then by combining all domains into one framework. The final design, by incorporating query routing to the RAG-LLM, was found to be more efficient compared to the first two. The sentence-transformed-based embedding model outperformed the OHDSI's tool, Athena. The end-to-end automation of this process makes it accessible to users, even those without expertise in the medical field. The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare practitioners. Core components, such as the embedding model, vector database, and LLM, can be easily changed to accommodate different model setups. The workflow has already been validated against the requirements of the REALM project and will soon be integrated into the RIANA dashboard, offering a user-friendly interface for AI model developers. Beyond REALM, this solution holds strong potential for a wide range of real-world standardisation challenges where efficient, accurate vocabulary mapping is essential.

Ethical Considerations

This study is fully in line with the standards and principles of Vito infrastructure. No Protected health information or Electronic health records are utilised in this analysis.

Project Code

The code used in this project is available at: `https://github.com/Sarigakakkamani/RAG-LLM-Pipeline.git`

References

Jorge Gabín and Javier Parapar. Leveraging retrieval-augmented generation for keyphrase synonym suggestion. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, pages 311–327, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-88711-6.

Zhenhua Gai, Lianxin Tong, and Quan Ge. Achieving higher factual accuracy in llama llm with weighted distribution of retrieval-augmented generation. 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

Donghao Huang, Zhenda Hu, and Zhaoxia Wang. Performance analysis of llama 2 among other llms. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1081–1085. IEEE, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer, 2023.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

James Mitchell-White, Reza Omdivar, Esmond Urwin, Karthikeyan Sivakumar, Ruizhe Li, Andy Rae, Xiaoyan Wang, Theresia Mina, John Chambers, Grazziela Figueredo, et al. Llettuce: An open source natural language processing tool for the translation of medical terms into uniform clinical encoding. *arXiv preprint arXiv:2410.09076*, 2024.

OHDSI OMOP-CDM, 2025. URL `https://ohdsi.github.io/CommonDataModel/`.

OHDSI Tools, 2025. URL `https://ohdsi.org/software-tools/`.

OHDSI Vocab. SNOMED-CT, 2025. URL `https://github.com/OHDSI/Vocabulary-v5.0/wiki/Vocab.-SNOMED`.

OHDSI Vocabulary, 2025. URL `https://athena.ohdsi.org/vocabulary/list`.

OHDSI Vocabulary, 2025. URL `https://github.com/OHDSI/Vocabulary-v5.0`.

Realm-EU. Realm – real-world-data enabled assessment for health regulatory decision-making, 2025. URL `https://realm-ai.eu/`.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024*

IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), pages 155–161. IEEE, 2024.

Ramya Tekumalla and Juan M Banda. Towards automated phenotype definition extraction using large language models. *Genomics & Informatics*, 22(1):21, 2024.

The Book Of OHDSI, 2025. URL `https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vito, BE 2025. URL `https://realm-ai.eu/about/partners/vito`.

Vito-REALM, BE 2025. URL `https://vito.be/en/projects/realm`.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (2):1–37, 2019.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3 (1):1–9, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

Yeqing Yan, Peng Zheng, and Yongjun Wang. Enhancing large language model capabilities for rumor detection with knowledge-powered prompting. *Engineering Applications of Artificial Intelligence*, 133:108259, 2024.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Xinyu Zhou, Lovedeep Singh Dhingra, Arya Aminorroaya, Philip Adejumo, and Rohan Khera. A novel sentence transformer-based natural language processing approach for schema mapping of electronic health records to the omop common data model. In *AMIA Annual Symposium Proceedings*, volume 2024, page 1332, 2025.

Angelo Ziletti and Leonardo D'Ambrosi. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. *arXiv preprint arXiv:2403.09226*, 2024.

Appendices

A OHDSI Symposium Europe 2025

An abstract submitted to the OHDSI Europe Symposium 2025, based on the work in this project, was accepted for both a poster presentation and a lightning talk. The accepted abstract and the work-in-progress poster are included below.

A.1 OHDSI Abstract

RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

Sariga Kakkamani[1] (sariga.kakkamani@student.uhasselt.be), Frederic Jung[2] (*frederic.jung@vito.be*), Joeri Verbiest[1,3,4] (*joeri.verbiest@uhasselt.be*), Liesbet Peeters[1,3,4] (*liesbet.peeters@uhasselt.be*)

[1]Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium
[2]VITO, Vlaamse Instelling voor Technologisch Onderzoek, Mol, Belgium
[3]Biomedical Research Institute (BIOMED), Hasselt University, Diepenbeek, Belgium
[4]University MS Center (UMSC), Hasselt-Pelt, Belgium

Background

Feature extraction from Electronic Health Records (EHR) data is crucial in real-world evidence analysis[1]. This requires translating high-level clinical concepts into queries compatible with the standard terminologies. Observational health data are often standardized to the OMOP Common Data Models (CDM) which are widely used standards. This enables us to carry out efficient analyses that can generate reliable evidence. However, understanding these standards and vocabulary terms requires medical knowledge, particularly for users without domain expertise. Defining and extracting relevant features from structured EHRs remains a key challenge[2]. At the moment, OHDSI tools such as Athena and Usagi are used to search and assist user to map vocabulary following OMOP concepts. However, these tools come with their own limitations and fail to meet the exact contextual requirements (Figure 1).

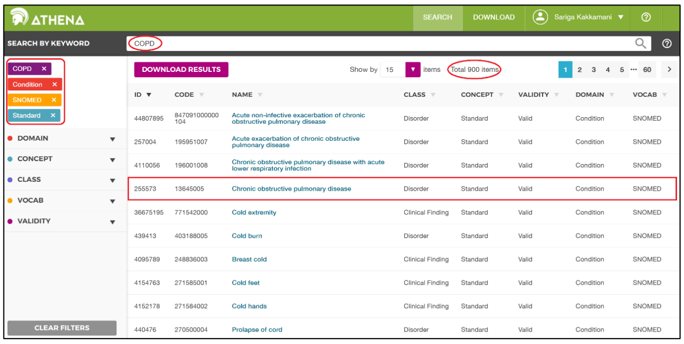


*Figure 1: **Limitations of Concept Search in OHDSI ATHENA** - This screenshot illustrates the search results in the ATHENA vocabulary browser for the term "COPD" with filters applied for SNOMED vocabulary, Condition domain, and Standard concepts. Despite these filters, the most relevant concept appears only as the fourth result and gives a full list of over 900 suggestions. This highlights the challenge in identifying the correct mapping when using non-standard terminology (e.g. SNOMED, LOINC), and when using acronyms*

We have investigated innovative techniques to propose better concept mapping and using different strategies than the traditional NLP (Natural Language Process) techniques and searching algorithms used by the OHDSI tooling[3] (e.g. Fuzzy, Lucene search). The Retrieval Augmented Generation- Large Language Model (RAG-LLM) pipeline we propose in this abstract is based on our latest research in identifying technologies for assisting the concept mapping process. To give it a bit more of real-world context, we applied this method to the use cases of the Real-world-data Enabled Assessment reguLatory decision-Making (REALM) project (EU-funded project 101095435)[4]. This project aims to provide a powerful sandbox environment for the future evaluation of AI as a medical device that goes on the EU market. In this context, we initiated a large effort for the mapping of the REALM use-cases features to the OMOP vocabulary which is initially based on an extended AI model description. Using these well documented use cases, can provide some context to the LLM and get encouraging results.

Methods

The proposed tool uses the RAG pipeline for creating an automated vocabulary mapping for OMOP CDM concepts[5]. The vocabularies, including SNOMED (2024-02-01 SNOMED CT International Edition), ICD-10(2021), RxNorm Extension (20240701), LOINC (2.77) and OSM (Release 2019-02-21), which are stored in standardized vocabulary tables of OMOP CDM V 5.4[6] are used in this analysis. The concept name for each corresponding concept id, concept name and domain id were extracted. The RAG pipeline (Figure 2) was generated by creating embeddings for text-based concept names and stored in a vector database. The RAG was then connected to the LLM through the Direct Prompt Injection method.

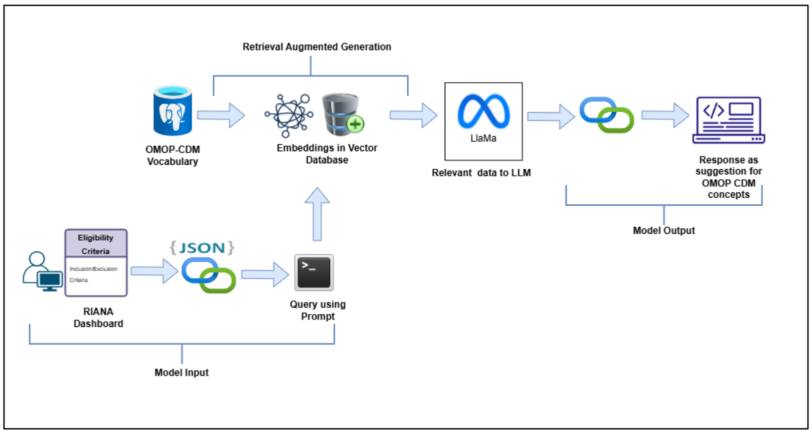


*Figure 2: **Overview of the RAG-LLM Semantic Mapping Pipeline** - This end-to-end pipeline integrates a Retrieval-Augmented Generation (RAG) architecture with a Large Language Model (LLM) to support semantic mapping of clinical features to OMOP-CDM concepts. (1) user inputs vocabulary in our RIANA dashboard and transmitted through an API. (2) input is encoded and compared against pre-generated embeddings (OMOP concepts stored in a vector database). (3) The* top-k *most semantically similar matches are retrieved and passed to the LLM using a structured prompt. The LLM generates context-aware* concept_ids *suggestions.*

The features that we want to map to the OMOP CDM were extracted through a web application called the (REALM Intelligent Analytics) RIANA dashboard. The output of this application gives us a well-documented list of described features that comes with a large and extended description of the patient target group. We are getting this list using an API call. The embeddings will be generated for the input query with the same method, through the similarity search between the query embeddings and vocabulary embeddings top 'k' matches are created for concept name and concept id. This relevant information is then sent to the LLM and is retrieved as an output, which provides suggestions for concept names and their corresponding ids. A well-defined structured prompt was designed to give a clear, standardized query as the input and to retrieve meaningful responses from the pipeline. An open-source LLM 'llama3.3:70b -instruct-q4_K_M' was chosen for this analysis. The vector embeddings are generated using 'all-MiniLM-L6-v2' `from` 'Sentence Transformers' specifically optimized for semantic mapping.

Results

The current implementation of our RAG-enhanced LLM pipeline is operational within the VITO infrastructure, allowing interactive testing and refinement through prompt-based queries (Figure 3). This setup enabled us to iteratively improve the workflow and achieve high-quality results in semantic mapping tasks. We were able to guide the LLM toward accurate vocabulary suggestions based on user-defined feature descriptions called model cards. Our pipeline demonstrated strong performance in mapping natural language clinical features to standardized OMOP-CDM concepts. In the use case of Chronic Obstructive Pulmonary Disease (COPD) severity classification, a common categorical feature of our AI models that we successfully mapped. Finally, we managed to get comparable mapping results using our pipeline rather than the concept identified manually.

This workflow has been successfully validated in the context of the REALM project, where it supports the generation of standardized AI feature sets from natural language cohort definitions. Integration into the RIANA dashboard is currently ongoing, enabling seamless access for users through an intuitive interface. Beyond REALM, the pipeline has broad applicability for real-world semantic mapping tasks, allowing non-expert users to get assistance in the semantic mapping process.

```
Question: What is the concept ID for COPD?
Answer: The concept ID for COPD (Chronic Obstructive Pulmonary Disease) is not explicitly stated in the provided lis
t with that exact name. However, a closely related concept is listed as:

- Concept ID: 255573, Concept Name: Chronic obstructive lung disease, Domain ID: Condition

This can be considered synonymous with COPD for many purposes, as both terms refer to conditions characterized by ai
rflow limitation and are often used interchangeably in clinical contexts.

If you're looking for a more specific or direct reference to COPD, it might be worth noting the following options wh
ich are closely related but specify different stages or aspects of the condition:

- Concept ID: 4209097, Concept Name: Severe chronic obstructive pulmonary disease
- Concept ID: 46284163, Concept Name: Chronic obstructive pulmonary disease resolved
- Concept ID: 36685454, Concept Name: GOLD (Global Initiative for Chronic Obstructive Lung Disease) 2017 group D
- Concept ID: 4193588, Concept Name: Moderate chronic obstructive pulmonary disease
- Concept ID: 4196712, Concept Name: Mild chronic obstructive pulmonary disease
- Concept ID: 44791725, Concept Name: Very severe chronic obstructive pulmonary disease
```

Figure 3: ***Prompt-Based Semantic Mapping of COPD Concepts Using the RAG-LLM Pipeline*** *- This figure shows an example of the interactive prompt and the corresponding output generated by the RAG-LLM pipeline for the use case of identifying COPD-related concept IDs. The input prompt describes the clinical feature (COPD and its severity levels), and the model returns a list of standardized SNOMED concept codes within the Condition domain.*

Table 1: ***Concept Mapping of COPD severity used as AI features*** *- This table illustrates the manual mapping of COPD severity levels. This example was taken based on COPD use case where the feature COPD_severity was used as a categorical variable, and each modality needs to be mapped to one concept. This manual mapping was supervised by data analysts and required expertise from the model developers and clinicians who originally captured the data used as a training set.*

Code	Domain	Condition
255573	Condition	Chronic obstructive lung disease
4196712	Condition	Mild chronic obstructive pulmonary disease
4193588	Condition	Moderate chronic obstructive pulmonary disease
4209097	Condition	Severe chronic obstructive pulmonary disease
44791725	Condition	Very severe chronic obstructive pulmonary disease

Conclusion

We designed an innovative automated framework that utilizes RAG enhanced LLM. This framework enables the semantic mapping of natural language cohort definitions to standardize queries compliant with the OMOP CDM, thus improving the accuracy of feature mapping. The end-to-end automation of this process makes it accessible to users, even those without expertise in the medical field. In the future stage, this workflow will be integrated into the REALM testing environment, where the AI model developer can directly get the recommendations of concept names while submitting the cohort requirements. The approach will be benchmarked for five different use cases of REALM. The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare practitioners.

Importantly, our pipeline is highly modular and adaptable. Core components, such as the embedding model, vector database, and LLM, can be easily changed or fine-tuned to accommodate different technical setups. The workflow has already been validated against the requirements of the REALM project and will soon be integrated into the RIANA dashboard, offering a user-friendly interface for AI model developers. Beyond REALM, this solution holds strong potential for a wide range of real-world standardization challenges where efficient, accurate vocabulary mapping is essential.

References

1. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for Electronic Health Records. npj Digital Medicine. 2022 Dec 26;5(1). doi:10.1038/s41746-022-00742-2
2. Ziletti A, DAmbrosi L. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. Proceedings of the 6th Clinical Natural Language Processing Workshop. 2024;47–53. doi:10.18653/v1/2024.clinicalnlp-1.4
3. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI standardized vocabularies—a large-scale centralized reference ontology for International Data Harmonization. Journal of the American Medical Informatics Association. 2024 Jan 4;31(3):583–90. doi:10.1093/jamia/ocad247
4. Realm – Real-world-data Enabled Assessment for heaLth regulatory decision-Making [Internet]. Realm-ai.eu. 2025 [cited 2025 Mar 30]. Available from: https://realm-ai.eu/
5. OMOP Common Data Model; [Internet], [cited 2025 Mar 28]. Available from: https://ohdsi.github.io/CommonDataModel/cdm54.html
6. OMOP CDM v5.4 ERD [Internet] 2022 [cited 2025 Mar 24]. Available from: https://ohdsi.github.io/CommonDataModel/cdm54erd.html

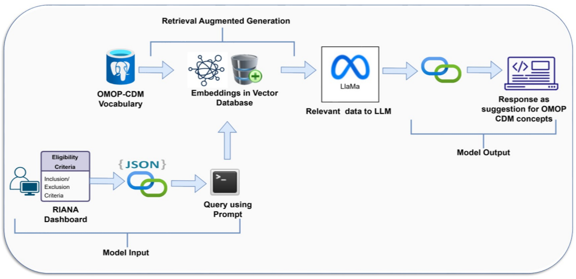
A.2 OHDSI Poster

Retrieval Augmented Generation - Large Language Model (RAG-LLM) pipeline to create an automated OMOP-CDM vocabulary mapping for improved features extraction.

Title: RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

Background: Observational health data are often standardized to the commonly used OMOP-CDM standards. This enables us to carry out efficient analyses that can generate reliable evidence. However, understanding these standards and vocabulary terms requires medical knowledge, along with OMOP-CDM expertise. This makes feature extraction crucial, particularly for users without domain expertise.

Overview of the RAG-LLM Semantic Mapping Pipeline



Method: User input is encoded and compared against pre-generated embeddings (OMOP concepts stored in a vector database). The top-k most semantically similar matches are retrieved and passed to the LLM. Context relevant suggestions appear as output.

Embedding Model: Sentence Transformers-all-MiniLM-L6-v2
LLM: Llama - llama3.3:70b -instruct-q4_K_M
Vector DB: Chroma Database

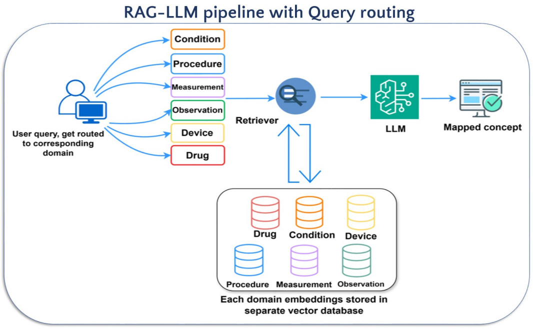
Results: The pipeline achieved improved performance over the OHDSI tool Athena. Especially with the proposed approach in addition to the suggestions on matching concepts we get explanation on which concepts are more appropriate as per the input query.

Prompt

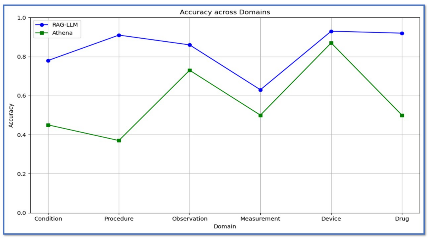
< / >

What is the concept ID for COPD from ***CONDITION*** *domain ?*

RAG-LLM pipeline with Query routing



Mapping accuracy across six domains



Conclusion: The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare practitioners.

OHDSI

Sariga Kakkamani[1], Frédéric Jung[2], Joeri Verbiest[1,3,4], Liesbet Peeters[1,3,4]

[1]Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium
[2]VITO, Vlaamse Instelling voor Technologisch Onderzoek, Mol, Belgium
[3]Biomedical Research Institute (BIOMED), Hasselt University, Diepenbeek, Belgium
[4]University MS Center (UMSC), Hasselt-Pelt, Belgium

Funded by the European Union

REALM

B OMOP Common Data Model (CDM)

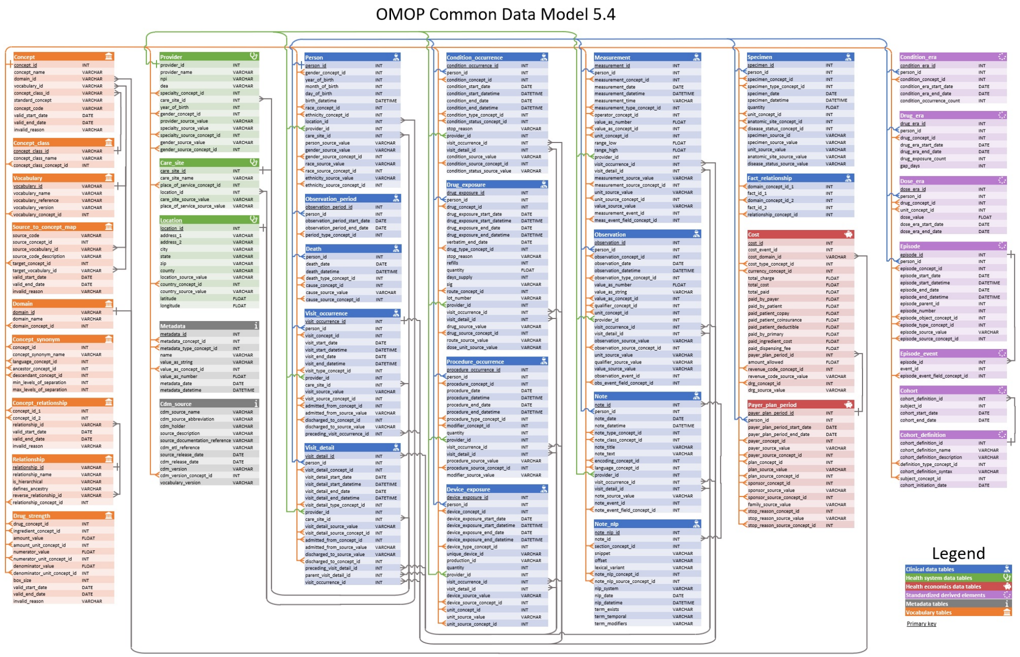


Figure 25: Database structure of the OMOP Common Data Model (CDM) version 5.4. The tables are grouped by domain and source, illustrating the standardized data schema adopted by the OHDSI community. Picture credit: Martijn Schuemie and Renske Los.

C OHDSI Tools

C.1 Athena

Fig. 26 shows the default homepage for Athena. The users can choose their domain of choice and add the search term in the search box. Additional filters, such as VOCAB, VALIDITY, CLASS, etc., can be chosen at the following steps.

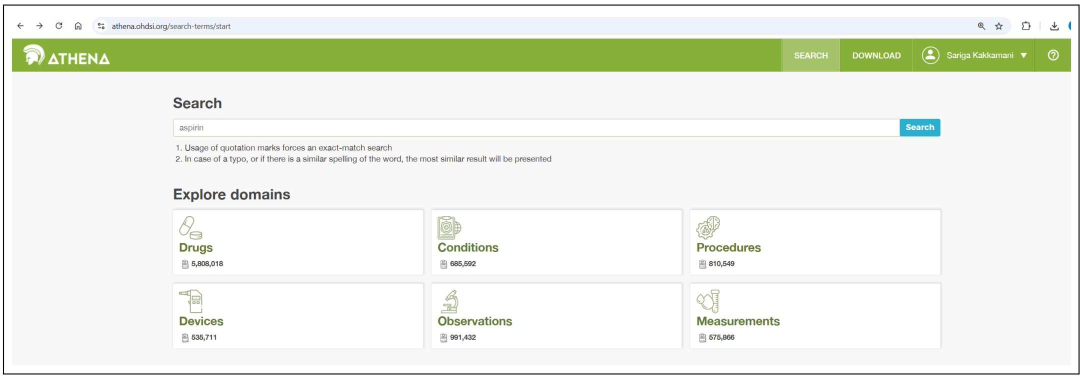


Figure 26: Default interface of the OHDSI tool Athena

C.2 Usagi

Usagi generates the OMOP-CDM mappings for the source codes, which are imported into the system. Suggestions can be either approved or unchecked, as shown in the figure. 27 OHDSI Tools (2025)

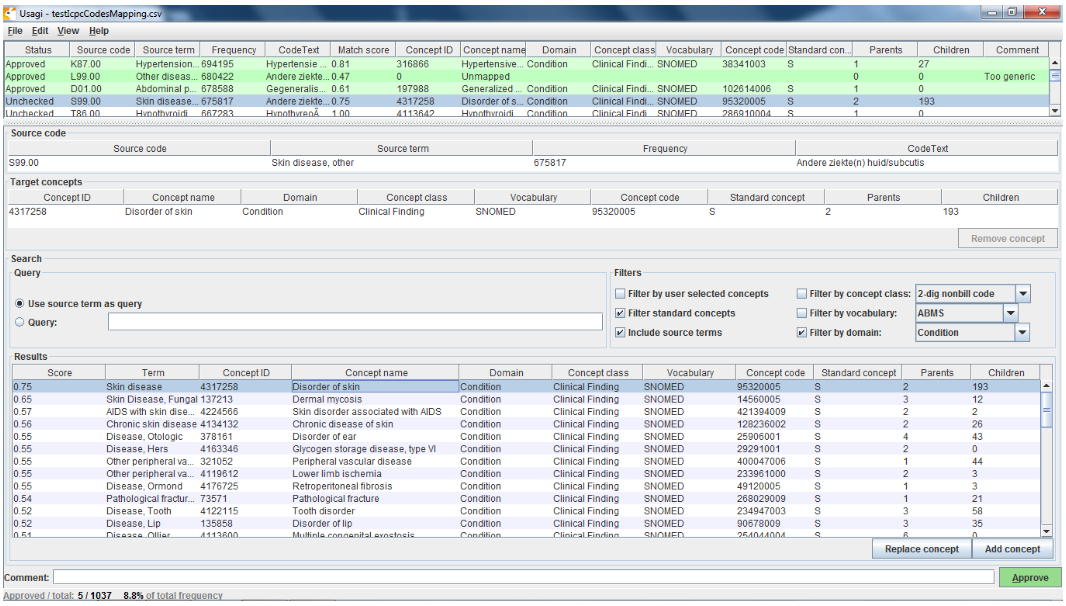


Figure 27: Mapping of source concepts to OMOP-CDM standards through a similarity approach

D RIANA Dashboard

The format of the input and output JSON code in the RIANA dashboard.

Model Card Section	Representative Examples

Factors Includes demographic or phenotypic groups, environmental conditions, and other stratification factors. These are defined as cohort definitions in JSON format, allowing a structured representation of patient groups.	"InclusionRules": [{
	"expression": {
	"Gender":
	{
	"CONCEPT_ID": 8507,
	"CONCEPT_NAME": "MALE",
	}
	},
	"ConditionOccurrence": {
	"CodesetId": 12,
	"Age": {
	"Value": 22,
	"Op": "gt"
	}

Features Provides the feature mapping between the input dataset and the OMOP CDM. For each feature and its modalities, the mapping ensures alignment with standardized OMOP concept.	{ "name": "COPDSEVERITY", "source_values": ["SEVERE", "MODERATE",], "target_values": [4209097, 4193588,], "type": "categorical", "domain": "condition_occurrence", },

Evaluation Dataset Provides details on the dataset used for quantitative analyses in the model card. This includes Heracles results per OMOP data sources along with patient counts, generation timestamps, summarizing demographics and clinical distributions	"datasets": [
	{
	"name": "MIMICIII",
	"datasetID": "1234",
	"summary": {
	"PersonDistribution": {
	"conceptName": "MALE",
	"conceptId": 8507,
	"countValue": 95
	},
	"ConditionDistribution": [{
	"conceptId": 4193588,
	"conceptPath": Moderate COPD",
	"numPersons": 39},
	{
	"conceptId": 4209097,
	"conceptPath": Severe COPD",
	"numPersons": 27
	},[…]

Metrics Reflects the potential real-world impact of the model by capturing performance measures. These metrics are stored alongside the evaluation dataset and are linked using the same *datasetID*.	"metrics": [{
	"datasetID": "1234",
	"name": "Accuracy",
	"score": "0.85",
	},
	{
	"datasetID": "1234",
	"name": "Recall",
	"score": "0.92",
	}
	],

Figure 28: A simplified example of the expected JSON structure for specific sections of the model card.

E Model Cards

The example model cards for the REALM use cases.

COMUNICARE **Model Card for COPowereD**

COPowereD (COM) aims to predict hospitalization or acute exacerbations in patients with chronic obstructive pulmonary disease by using medical AI algorithms that include patient-reported outcomes. The solution is a part of the COMUNICARE application, which is a generic solution for patient empowerment and self-efficacy in chronic care and post-hospitalization contexts.

- **Model Details:**
- **Date:**
- **Version:**
- **License:**
- **Owner:** COMUNICARE

- **Intended Use:** The goal is to reduce the number of hospital admissions for chronic obstructive pulmonary disease exacerbations.
- **Metrics:** Accuracy, Precision
- **Evaluation Data:** Age, sex, height(cm), weight(cm), BMI, COPD gold stage, short of breath, cough, sputum, heart rate, pulse Ox
- **Limitations:**
- **Ethical Considerations:**

Figure 29: Model card for REALM use-case COPowereD

LIÈGE université **Model Card for STAR**

STAR project (UoL) aims to develop a model-based decision-support system to enable precise regulation of blood glucose levels in intensive care unit (ICU) patients. The project involves selecting 2-5 ICU clinicians and 30 ICU nurses from Hungary, New Zealand, and Belgium to evaluate the system's usability in a clinical setting.

- **Model Details:**
- **Date:**
- **Version:**
- **License:**
- **Owner:** STAR

- **Intended Use:** The goal is to assess the usability of STAR in a clinical setting, comparing results from experienced users versus naïve clinical users.
- **Metrics:** Accuracy, Precision
- **Evaluation Data:** Diabetic status, Blood glucose, device, insulin rate, enteral nutritional rate, parenteral nutritional rate, other IV drugs and solutions, carbohydrates concentration, glucocorticosteroids
- **Limitations:**
- **Ethical Considerations:**

Figure 30: Model card for REALM use-case STAR

Maastricht University

Model Card for DuneAI

DuneAI (UM) is a use case that involves evaluating an automated segmentation software for detecting and segmenting non-small-cell lung cancer tumors in CT scans. This deep learning technique-based software has been validated on a large dataset and has shown high sensitivity and specificity in addition to fast segmentation times when compared to human experts.

- **Model Details:**
- **Date:**
- **Version:**
- **License:**
- **Owner:** UM

- **Intended Use:** The purpose of this evaluation is to assess the software's usability and usefulness in a clinical setting and to improve its integration into clinicians' workflow.
- **Metrics:** Accuracy, Precision
- **Evaluation Data:** Age, sex, height(cm), weight(cm), BMI, lung CT, NSCLC, thoracic scan
- **Limitations:**
- **Ethical Considerations:**

Figure 31: Model card for REALM use-case DuneAI

vito

Model Card for PGx2P

Pharmacogenomics Passports to Practice (PGx2P) is a use case that aims to implement preventive pharmacogenetics testing of a panel of genetic variants approved for clinical use by the Dutch Pharmacogenetics Working Group guidelines.

- **Model Details:**
- **Date:**
- **Version:**
- **License:**
- **Owner:** Vito

- **Intended Use:** The project aims to facilitate stratified prescription of commonly used medications, resulting in a >25% reduction of overall adverse drug responses and >20% better adherence to prescriptions in PGx2P tested individuals.
- **Metrics:** Accuracy, Precision
- **Evaluation Data:** Age, sex, height(cm), weight(cm), genetic mutation, CYP2D metabolizer
- **Limitations:**
- **Ethical Considerations:**

Figure 32: Model card for REALM use-case PGx2P