HeKDisco

Healthcare Knowledge Discovery

Deliverable 5.2

ANALYSIS & DESIGN OF THE KNOWLEDGE DISCOVERY PLATFORM

WP5 – Knowledge Discovery Platform: Monitoring, Analytics and Search

Document Type	Document
Document Version	V1.0
Access Level	Confidential
Contractual Submission Date	April 2024
Actual Submission Date	June 2024
Editors	Mantis
Contributors	Bilbest/Mantis/ Caretronic

Document Contributors

Partner	Author	Role
Mantis	Behzad Naderalvojoud	Editor/Contributor
Mantis	Güven Köse	Contributor
Bilbest	Nurol Gençyılm	Contributor
Bilbest	Süheyla Türkyılmaz	Reviewer
Caretronic	Simona Brezar	Reviewer

Document History

Date	Version	Editors	Status
01/04/2024	0.0	Mantis	Table of Content
13/05/2024	0.5	Mantis	First draft
27/05/2024	0.6	Bilbest	Review
15/06/2024	0.7	Mantis	Second draft
22/06/2024	0.8	Caretronic	Review
29/06/2024	1.0	Mantis	Final

Glossary

Abbreviation	Meaning
BC	Breast Cancer
CM	Clinical Medicine
CDM	Common Data Model
EBM	Evidence-Based Medicine
EHR	Electronic Health Records
ETL	Extract Transform Load
LLM	Large Language Model
MeCSE	Meta Clinical Search Engine
ML	Machine Learning
MRI	Medical Imaging Technology
MS	Multiple Sclerosis
NLP	Natural Language Processing
OMOP	Observational Medical Outcomes Partnership
SSL	Secure Sockets Layer
TLS	Transport Layer Security
UI	User Interface

Table of Contents

List of Figures

EXECUTIVE SUMMARY

The HeKDisco project aims to revolutionize healthcare analytics by addressing the challenges of data heterogeneity and integration across multiple institutions. Leveraging the OMOP Common Data Model (CDM), HeKDisco proposes a platform to integrate diverse medical datasets, enabling continuous standardization, AI-driven knowledge extraction, and federated analysis. This deliverable outlines the system's design and key innovations, including scalable data extraction, monitoring, analysis, and semantic search by elastic query processing and dynamic data transformation. By bridging the gap in current federated analytics systems, HeKDisco provides a robust, interoperable solution for real-world clinical applications, fostering reliable evidence generation and advancing global healthcare decision-making.

1 Introduction

1.1 Project overview

The main purpose of the HeKDisco project is to reduce potential human mistakes in the medical care of patients. Traditional health care systems—clinical medicine (CM)—mainly rely on conservative methods to diagnose diseases and treat patients, depending on the individual knowledge and skills of physicians. On the contrary, evidence-based medicine (EBM) provides a workaround for poorly designed observational treatment that relies on physicians' personal experience with other patients. In this approach, evidence describes average results for groups of patients. HeKDisco, following EBM, aims to use the best (reliable) evidence in making decisions about the care of individual patients so that the clinician's experience, the patient's values and preferences, and the best empirical clinical guidelines are integrated.

In many diseases, especially infectious and chronic diseases, the same treatment may show different outcomes for different groups of patients. Therefore, physicians' ability to use reliable empirical evidence before any decision-making helps them select the best treatment option and decreases potential mistakes. According to a study by Johns Hopkins in 2016, more than 250,000 people in the U.S. die every year due to medical mistakes, making it the third leading cause of death after heart disease and cancer. In this line, HeKDisco proposes a novel knowledge discovery process for health care systems so as to provide physicians with reliable evidence on different treatment stages and clinical events, thereby reducing individual clinical errors. The overall idea of the HeKDico project is summarized in Figure 1. It includes five key objectives: (1) Perform multimodal knowledge discovery from medical images and clinical notes in order to extract high-level information, such as descriptive and predictive information, which is lacking in EHR systems. (2) Generate a knowledge base in conjunction with EHRs. (3) Develop a clinical search engine in order to provide a systematic analysis using a common data model as standardization. (4) The use of systematic analyses to create a computational ontology in order to recognize the relationships between the descriptive and predictive parameters. (5) Provide high-quality care.

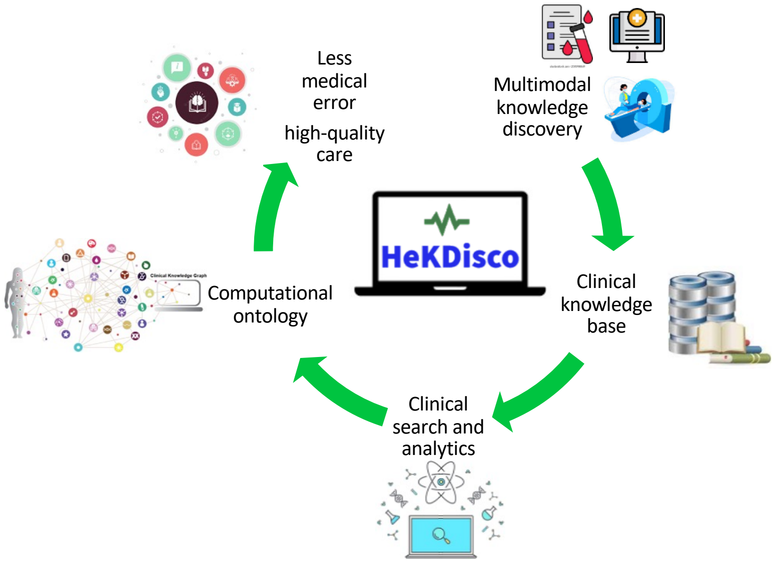


Figure 1. The overall view of the HeKDisco project

1.2 Knowledge discovery and clinical search

The integration of healthcare data across multiple sources and institutions poses significant challenges due to its inherent heterogeneity and diverse representations. The HeKDisco project aims to address these challenges by proposing a unified knowledge discovery platform that leverages AI-powered knowledge extraction and semantic search functionalities. This deliverable outlines the design and specifications of the proposed system, emphasizing its capabilities for data monitoring, analytics, and federated search. The platform adopts the OMOP Common Data Model (CDM) to standardize diverse medical data collected from electronic health records (EHRs), MRI scans, and pathology images/reports. The architecture supports a multi-site deployment, enabling seamless integration between local sites and central meta operations. Each site employs CDM mapping services and AI modules for extracting medical knowledge from raw clinical data, while a centralized search engine aggregates and visualizes results for population-level analyses. Key features of the system include continuous patient monitoring, AI-driven extraction of clinical parameters, and cross-site federated queries. By combining these elements, the platform facilitates reliable evidence generation for healthcare decision-making. The architecture ensures scalability and interoperability, making it a robust solution for advancing global healthcare analytics and fostering collaboration among diverse providers. This deliverable serves as a comprehensive guide to the technical design and operational framework of the HeKDisco knowledge discovery platform, highlighting its potential to transform healthcare data utilization through innovative AI and semantic search integrations.

2 HeKDisco Architecture

2.1 High-level design

Figure 2 illustrates the high-level architecture of the knowledge discovery and search platform developed in the HeKDisco project. This architecture seamlessly integrates AI-powered knowledge extraction modules designed to process diverse data sources, such as MRIs and pathology images, and transform them into structured clinical information. To ensure consistency and interoperability, all input data and outputs from the AI modules are standardized by mapping them to the OMOP Common Data Model (CDM). The standardized data is stored in a CDM-compliant database, which serves as the foundation for subsequent processing and clinical search.

A search data warehouse and indexing system organize the CDM-stored data, enabling efficient querying and semantic search based on a predefined search schema. The architecture is designed with modular data flows, ensuring that each step operates independently, enhancing system resilience against disruptions.

This architecture supports multi-site implementation with diverse clinical use cases. For example, in the HeKDisco project, two use cases are showcased: Multiple Sclerosis (MS) and Breast Cancer (BC). In the MS use case, knowledge extraction modules derive MRI parameters, while in the BC use case, pathology parameters are extracted from image data. All extracted information is standardized, stored in the CDM database, and indexed for search and analytics.

Additionally, high-level information, such as risk of diseases outcomes, can be derived from broader data, such as EHRs and image-based parameters, using AI-CDM modules. These modules exclusively operate on data mapped to the CDM, further emphasizing the platform’s commitment to standardization and interoperability.

Finally, the platform’s meta-search engine enables cohort-level search capabilities, distributing queries across multiple sites to aggregate, analyse, and visualize results in a federated manner. This approach ensures comprehensive and scalable analysis across diverse datasets while maintaining standardization and interoperability.

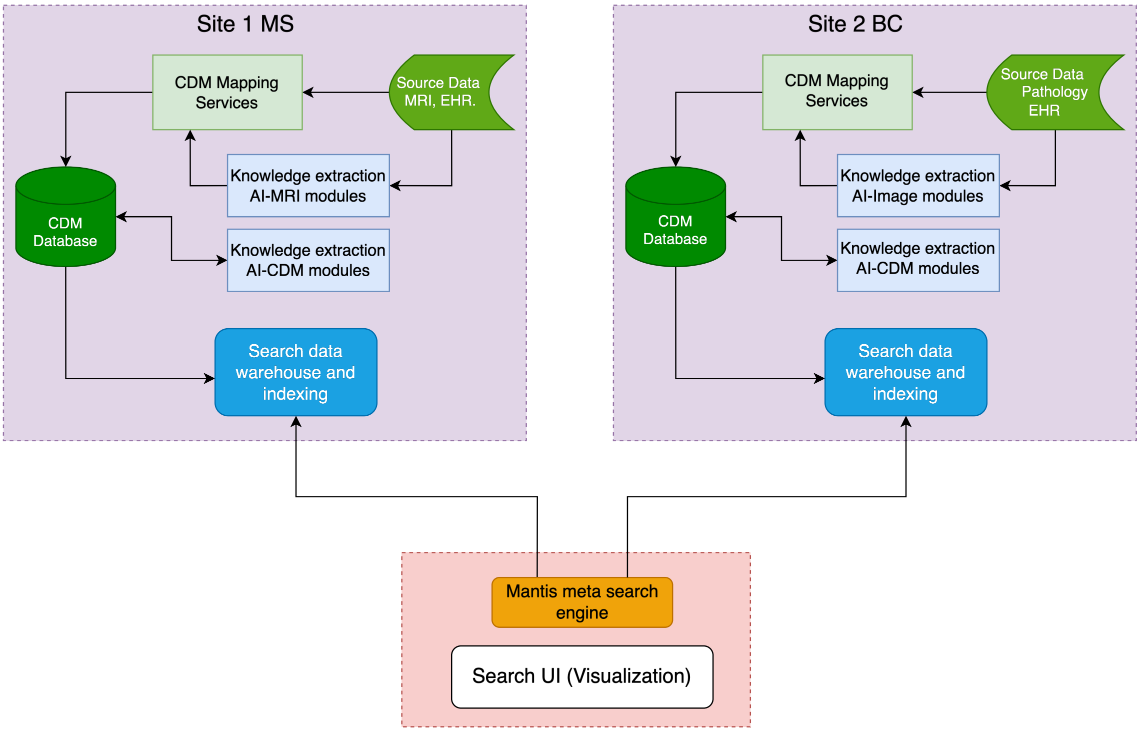

Site 1 MS
CDM Mapping Services
Source Data MRI, EHR.
Knowledge extraction AI-MRI modules
CDM Database
Knowledge extraction AI-CDM modules
Search data warehouse and indexing
Site 2 BC
CDM Mapping Services
Source Data Pathology EHR
Knowledge extraction AI-Image modules
CDM Database
Knowledge extraction AI-CDM modules
Search data warehouse and indexing
Mantis meta search engine
Search UI (Visualization)

Figure 2. The overall architecture of the knowledge discovery and search platform.

2.2 CDM mapping services

The CDM mapping services are a cornerstone of the HeKDisco architecture, enabling standardization and interoperability across the system. These services are tailored to accommodate the diversity of data sources and project use cases, ensuring seamless integration of heterogeneous datasets into the OMOP CDM. Developed as part of the project's third work package (see Deliverable D3.2 for details), these mapping services utilize a key-value structure, where the key represents the source code of the input data, and the value corresponds to its equivalent CDM concept ID. To address the complexity of clinical data, specialized services have been designed for various clinical domains, including diagnosis/conditions, measurements, procedures, observations, and drugs. Each service processes individual patient records from source sites, transforming them into CDM-compliant records. These records populate the standardized CDM tables, ensuring that all patient data across different domains are harmonized within a unified schema.

While the OMOP CDM provides open-source extract, transform, and load (ETL) tools for converting raw clinical data into the CDM format, these tools are not optimized for continuous data conversion. In the HeKDisco project, we addressed this limitation by developing services capable of performing seamless, small-scale, transformations for the project's use cases. To our knowledge, this represents the first instance of seamless clinical data transformation into the CDM, enabling immediate integration into analytics workflows.

This innovation has significant implications for real-world applications. By enabling seamless mapping, it facilitates the deployment of CDM-based AI models in clinical practice and supports federated analytics across distributed sites. The mapping services also provide APIs that allow external systems to submit raw clinical data, transform it into CDM-compliant records, and store it in the database. This approach bridges the gap between research and clinical practice, paving the way for scalable and interoperable AI-driven healthcare solutions.

2.3 AI-powered knowledge extraction modules

The second core component of the HeKDisco platform is the AI-powered knowledge extraction modules, which play a pivotal role in extracting clinical information from diverse data sources. These modules are designed for seamless integration into the platform through the previously described CDM mapping services. Developed independently, each module is tailored to process specific data types, such as MRI scans, pathology images, clinical reports, and unstructured clinical notes. Despite the variability in input formats, the output of these modules is standardized to predefined clinical parameters that can be mapped to the OMOP CDM and subsequently incorporated into the HeKDisco platform.

The flexibility of these AI modules lies in their algorithmic diversity. They leverage cutting-edge AI techniques, including computer vision for image data, natural language processing (NLP) for textual data, and hybrid models for multimodal datasets. For example, MRI-specific modules extract quantitative parameters such as lesion volume or contrast intensity, while pathology modules identify cellular features or biomarkers. Similarly, NLP modules extract key clinical details, such as diagnoses, medications, or symptoms, from unstructured clinical narratives. All extracted data is integrated with patient demographic information and stored in the CDM database via standardized tables, ensuring interoperability and uniformity.

The development of these modules was undertaken as part of the project’s fourth work package, with a focus on aligning module capabilities with the needs of various data sources and clinical domains (refer to Deliverable D4.2 for technical specifications). These modules not only process clinical data from EHRs but also enable direct extraction from raw data sources, such as medical images or textual notes, bridging the gap between raw data acquisition and structured data utilization.

In the HeKDisco project, the integration of AI-powered modules extends beyond routine data extraction to support real-world applications. By enabling the transformation of raw clinical information into actionable insights within the CDM framework, these modules enhance the system’s ability to handle complex, heterogeneous data environments. This architecture ensures that knowledge extraction is not only scalable but also adaptable to new data types and clinical use cases, such as those involving multiple sclerosis and breast cancer in the current implementation. This is in line with the project market and technology value chain described in the project proposal, as these modules can be used not only as part of the HeKdisco platform, but also as standalone products in the healthcare market.

2.4 Search data warehouse and indexing

The Search Data Warehouse and Indexing component is a foundational element of the HeKDisco platform, providing a unified infrastructure for big data analysis and real-time clinical search capabilities. Leveraging OpenSearch, a fully open-source search and analytics engine, this component is designed to support real-time monitoring, cohort-level search, and population-wide analytics. By integrating advanced search technology with standardized clinical data, it ensures fast, scalable, and efficient querying, enabling actionable insights for both research and clinical decision-making.

At the core of this component lies the use of a nested schema for indexing data in OpenSearch, which mirrors the structure of the OMOP CDM standardized tables. This alignment allows seamless integration between the CDM database and the indexing system, facilitating accurate and efficient translation of data for advanced search and analytics. Unlike traditional CDM-based relational databases or even big data platforms like BigQuery, which often require significant processing time for long cohort analyses, this nested indexing schema drastically reduces query execution times. Complex analyses that may take several minutes in other systems can now be performed in real-time, significantly enhancing the platform's utility for time-sensitive applications.

The indexing strategy supports complex relationships inherent to clinical datasets. Each indexed document in OpenSearch includes nested fields that preserve the hierarchical and relational nature of the original CDM data. For example, patient data spanning diagnoses, medications, procedures, and observations is stored in a structured format that enables precise filtering and aggregation, even for complex queries. This approach not only improves search accuracy but also optimizes performance for large-scale datasets spanning multiple sites.

The integration of OpenSearch extends beyond indexing to include robust analytics capabilities. Through its distributed architecture, OpenSearch handles diverse workloads efficiently, supporting both exploratory queries and predefined analytics pipelines. The platform's real-time processing capabilities make it ideal for monitoring key clinical metrics, tracking trends, and generating population-level insights. This is particularly relevant for the HeKDisco project's use cases, such as identifying clinical patterns in multiple sclerosis and breast cancer cohorts.

By coupling big data search and analytics with standardized healthcare data models, this component addresses the challenges of scalability, interoperability, and speed in federated health systems. The Search Data Warehouse and Indexing component provides the backbone for clinical intelligence in HeKDisco, empowering researchers and clinicians with tools to uncover insights quickly and effectively.

Further details on the nested schema and the specific types of queries supported by this component are discussed in the subsequent section, HeKDisco Clinical Search.

2.5 Meta search engine

The meta search engine is the final component of the HeKDisco platform, enabling secure and distributed query execution across multiple clinical sites while preserving patient privacy and ensuring data security. It facilitates federated search functionality by distributing user queries to local OpenSearch servers at participating sites and aggregating results for unified visualization and analysis. This approach complies with the patient privacy protocols established in the project's second work package Deliverable D2.2, ensuring that sensitive data remains under the control of its custodians. Instead of sharing raw data, the meta search engine transmits queries to local OpenSearch servers, where data is processed, and results are generated. These results are transmitted back to the meta search engine, which aggregates and visualizes them through an intuitive dashboard. This component manages complex queries across geographically dispersed datasets, translating user inputs into formats compatible with the local indexing schema. This ensures consistency and accuracy in search results.

The engine excels in real-time cohort-level searches, enabling users to query population metrics such as disease prevalence, treatment outcomes, and medication usage patterns across multiple sites. Its scalability and distributed nature accommodate large datasets and high query volumes, supporting clinical research and operations such as monitoring disease trends and identifying potential clinical trial cohorts. Additional features include user authentication, query management, and access controls, ensuring only authorized personnel can perform specific queries. These safeguards align with the platform's privacy and security requirements, promoting ethical and responsible data use.

The meta search engine bridges localized data storage with global healthcare insights, empowering collaboration among researchers, clinicians, and policymakers. By combining federated search with data aggregation, it fulfills HeKDisco's mission to advance knowledge discovery and clinical innovation while adhering to the highest standards of privacy issues.

3 HeKDisco Clinical Search

The architecture of the Meta Clinical Search Engine (MeCSE) is shown in Figure 3. MeCSE facilitates federated search and analytics across multiple distributed OpenSearch sites while maintaining patient privacy and data security. The components of this architecture are described below:

- **Users:** The end-users (clinicians, researchers, or analysts) interact with the system to execute queries and visualize results. They submit their queries via a user-friendly interface and receive insights derived from distributed data sources.

- **Query Analyzer:** It provides users with tools that facilitate creating a query. The Concept & Relation Extraction module extracts clinical concepts and relationships associated with user queries from the CDM Standardized Vocabulary and uses them in the query. The query generator converts user queries into structured OpenSearch queries based on the extracted concepts, ensuring compatibility with the data schema of distributed OpenSearch indices. These structured queries are distributed to decentralized OpenSearch databases, and the results are aggregated and visualized for the user through the MeCSE interface, which includes analytics and visualization capabilities.

- **OpenSearch Sites:** Multiple distributed OpenSearch instances (e.g., Site 1 to Site *K*) hold indexed clinical data. Each site corresponds to a different clinical facility or dataset. These sites remain decentralized, ensuring that raw data stays local and never leaves its origin.

- **Analytics & Visualization Module**: It provides tools for real-time cohort analytics and population-level insights. It aggregates and visualizes query results in an intuitive format, enabling users to make data-driven decisions effectively.

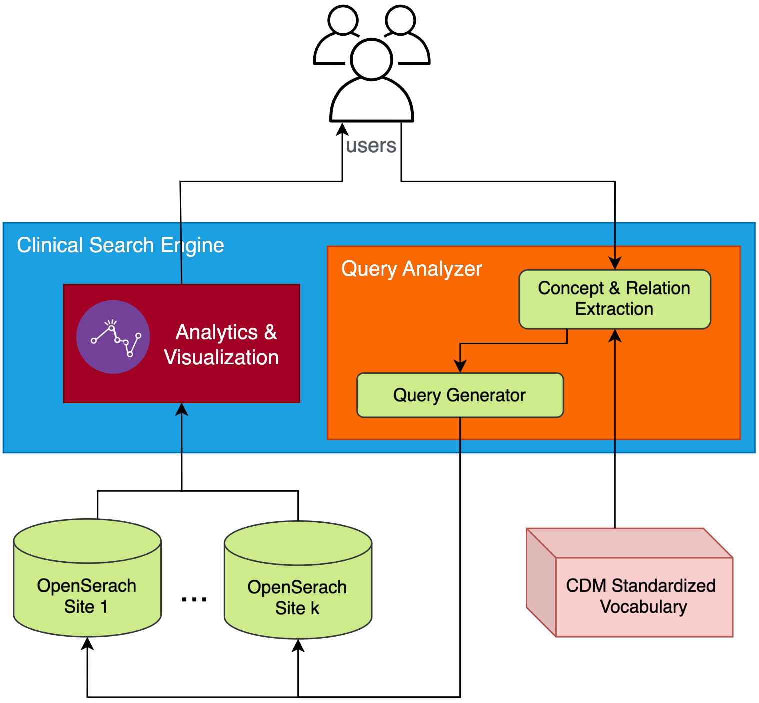


Figure 3. Architecture of the Meta Clinical Search Engine (MeCSE): A federated search framework enabling secure, real-time clinical cohort analysis across distributed OpenSearch sites.

3.1 Search indexing schema

Based on CDM's standardized clinical tables, shown in Figure 4, the following tables were used in the search indexing schema:

- **Person:** Used for demographic information including age, race and ethnicity.
- **Death:** Used for death dates.
- **Visit_occurrence:** Used for visit types such as inpatient, outpatient, and emergency room and their dates.
- **Condition_occurrence:** Used for diagnoses and their start dates.
- **Drug_exposure:** Used for medications and their start and end dates.
- **Procedure_occurrence:** Used for procedures and their dates.
- **Measurement:** Used for measurements, lab results and their dates.

- **Observation:** Used for observations, symptoms, any patient history inflation and their dates.

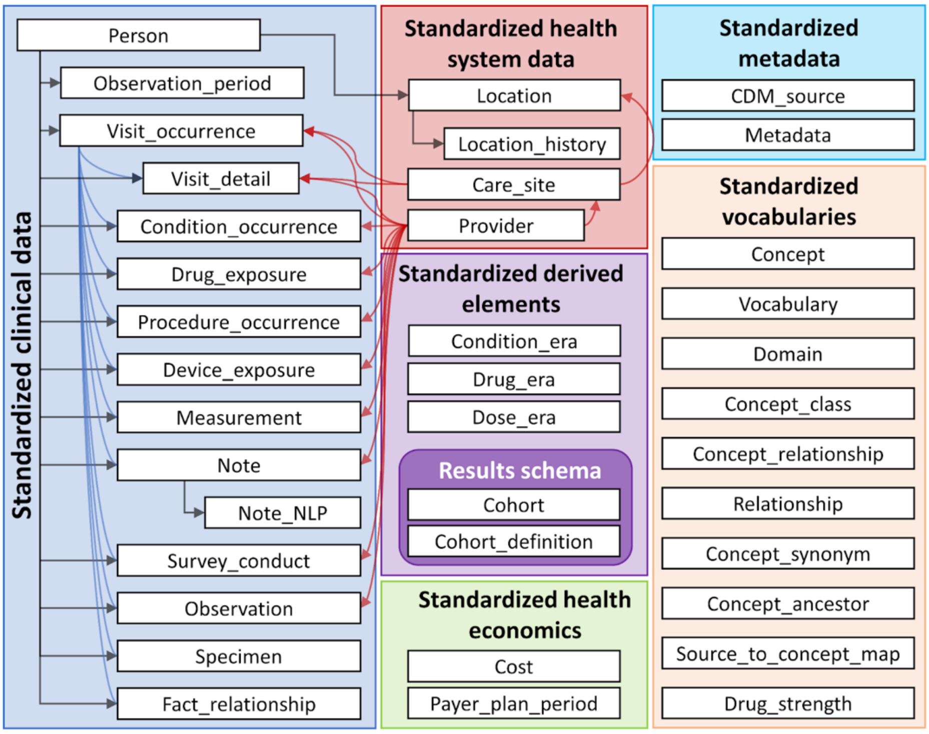


Figure 4. Common Data Model (CDM) schema.

The person table is the central table, to which all other clinical tables are linked. The Death table was merged with the Person table by adding the death date. This field can be updated when a person dies; otherwise, its value is null, indicating that there was no death. For other clinical tables, the corresponding concept ID and associated dates were considered in the indexing schema. For measurement and observation tables, three additional fields were considered for measurement or observation values (number, concept, or string) in the index schema. Given all this, the nested index schema was designed as follows:

Nested Schema

```
<Person_id> bigint
<gender_concept_id> int
<year_of_birth> int
<race_concept_id> int
<ethnicity_concept_id> int
<death_date> date

<condition> dict array: [
                          {
                            <occurrence_id> bigint
                            <concept_id> bigint
```

```
                           <start_date> date
                           <end_date> date
                          }
                       ]


<procedure> dict array: [
                         {
                           <occurrence_id> bigint
                           <concept_id> bigint
                           <start_date> date
                           <end_date> date
                          }
                       ]


<drug> dict array:     [
                         {
                           <occurrence_id> bigint
                           <concept_id> bigint
                           <start_date> date
                           <end_date> date
                          }
                       ]



<visit> dict array:   [
                         {
                           <occurrence_id> bigint
                           <concept_id> bigint
                           <start_date> date
                           <end_date> date
                          }
                       ]


<measurement> dict array: [
                                {
                                  <occurrence_id> bigint
                                  <concept_id> bigint
                                  <start_date> date

                                  <value_as_number> double
                                  <value_as_concept_id> bigint
                                  <value_as_string> string
                                  <unit_concept_id> bigint
                                 }
                           ]
```

```
<observation> dict array: [
                                     {
                                       <occurrence_id> bigint
                                       <concept_id> bigint
                                       <start_date> date

                                       <value_as_number> double
                                       <value_as_concept_id> bigint
                                       <value_as_string> string
                                       <unit_concept_id> bigint
                                      }
                               ]
```

3.2 Search query templates

The query template for the project's clinical search engine consists of two interconnected components designed to facilitate precise and flexible cohort selection querying across clinical data:

- **Event Query:** This query identifies primary clinical events associated with a specific domain (e.g., conditions, procedures, drugs, or measurements) and provides an index date for each event. The index date serves as a reference point for analyzing and contextualizing the event within the clinical timeline. For example, an event query might retrieve all diagnoses of a particular condition, each with its corresponding diagnosis date as the index date.

- **Inclusion/Exclusion Query:** This query filters the identified events by adding specific criteria relative to the event’s index date. These criteria can apply to various clinical attributes such as demographic factors, conditions, procedures, or medications. For instance, inclusion criteria might select events where patients had a particular procedure within a certain timeframe before or after the index date, while exclusion criteria might filter out events involving a specific medication.

Together, these components create a robust cohort selection query framework that allows users to retrieve clinically relevant events/patients and refine their results by leveraging the temporal and contextual relationships provided by the event’s index date.

The second query type depends entirely on the results of the first, meaning that all filters in the inclusion/exclusion criteria must be applied to every record from the event query. This dependency often results in computationally expensive queries, as complex criteria can require multiple database joins, leading to significant query times and costs.

To address this challenge, the project introduces an innovative search solution that redefines the process. This solution allows all required filters to be executed within a single query, dramatically reducing the computational burden. This advancement enables real-time online cohort selection and analysis—transforming a traditionally time-intensive process into a seamless and efficient search capability. To the best of our knowledge, this platform is the first to make clinical cohort selection possible as an online query in real-time, paving the way for groundbreaking advancements in clinical research efficiency and accessibility.

3.2.1 Event query template

The event query identifies all patients with a specific gender or age and a **<set of concept IDs>** in one of the clinical tables within a **<specific date interval>.** Multiple filters can be added for the event query based on different clinical tables.

The date in the first clinical filter is considered the index date in the event query for each patient, and all subsequent inclusion and exclusion filter queries are conducted based on it.

Event query examples:

1) All patients with **breast cancer diagnoses** within **2020 and 2024**
2) All patients with **lumpectomy procedures** within **2020 and 2024**
3) All patients with grade 1 (measurement) tumors within **2020 and 2024**
4) All patients who took the Gabapentin drug between **2020 and 2024**

3.2.2 Inclusion/exclusion query template

This query template selects patients who have at least one occurrence of a <set of particular concept IDs> in one of the clinical tables within a <specific date interval> relative to the index date.

Inclusion/exclusion examples:

By considering the first event query noted above, the following inclusion and exclusion query example are defined:

Example 1:

- **Events:** All female patients aged 30-70 with breast cancer diagnoses within 2020 and 2024
- **Inclusion/exclusion:** Patients with lumpectomy procedures within 30 to 180 days of the index date (diagnosis date)

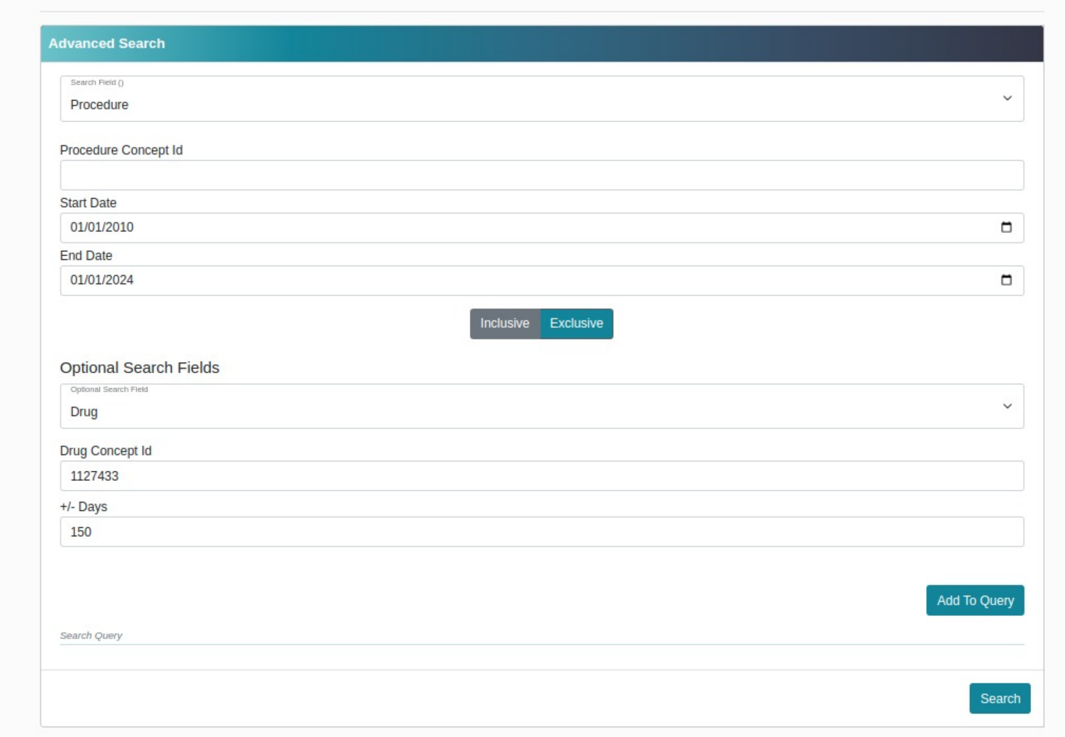
Example 2:

- **Events:** All female patients aged 30-70 with breast cancer diagnoses within 2020 and 2024
- **Inclusion/exclusion:**
  - Patients with grade 4 (measurement) tumors within 90 days before or 90 days after the index date (diagnosis date) AND
  - Patients who took the anastrozole drug (a hormone therapy) within 30 to 180 days of the index date (diagnosis date)

Multiple inclusion/exclusion queries can be combined using the OR and AND operators.

3.2.3 User interface for generating and running a search query

The user interface (UI) displayed in Figure 5 has been designed to facilitate query generation for clinical searches. It allows users to specify query parameters such as the type of clinical event (e.g., procedure), concept IDs, and date ranges (start and end dates). Additionally, users can add optional search criteria, such as related drug concept IDs and temporal constraints (e.g., ± days from the index date). The UI supports both inclusive and exclusive filtering options, enabling clinicians to refine their search with clinical criteria. Once all parameters are defined, the “Add to Query” and “Search” functionalities generate and execute the search query.



Search Field ()
Procedure

Procedure Concept Id

Start Date
01/01/2010

End Date
01/01/2024

Optional Search Fields

Optional Search Field
Drug

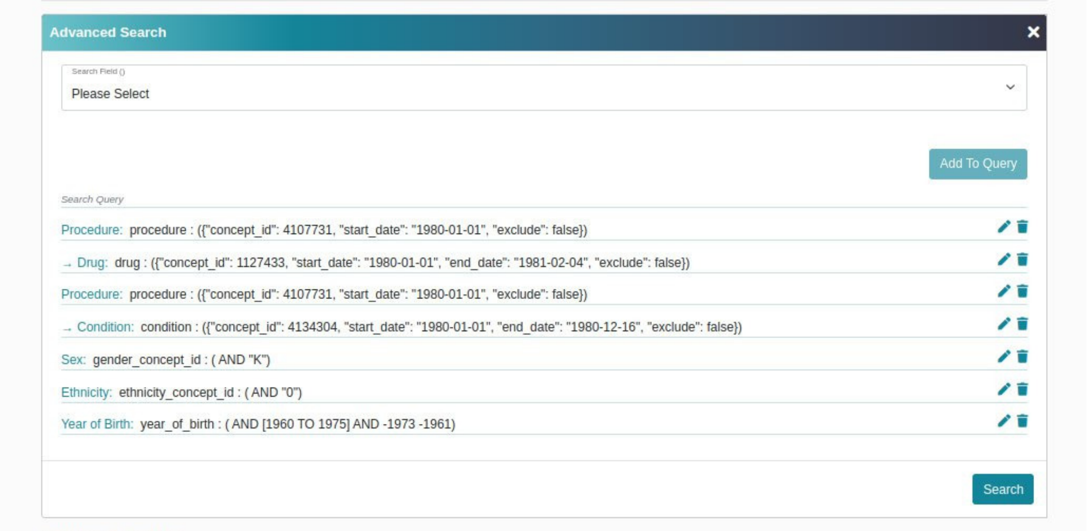
Drug Concept Id
1127433

+/- Days
150

Search Query

Figure 5. Advanced query generation interface for HeKDisco clinical search.

After receiving the search parameters via both query types, a summary of all query templates is displayed (Figure 6), and the user can review and edit them again. Finally, by clicking the “Search” button, a single search query is generated and executed across multiple OpenSearch instances in various clinical sites.



Procedure: procedure : ({"concept_id": 4107731, "start_date": "1980-01-01", "exclude": false})

→ Drug: drug : ({"concept_id": 1127433, "start_date": "1980-01-01", "end_date": "1981-02-04", "exclude": false})

Procedure: procedure : ({"concept_id": 4107731, "start_date": "1980-01-01", "exclude": false})

→ Condition: condition : ({"concept_id": 4134304, "start_date": "1980-01-01", "end_date": "1980-12-16", "exclude": false})

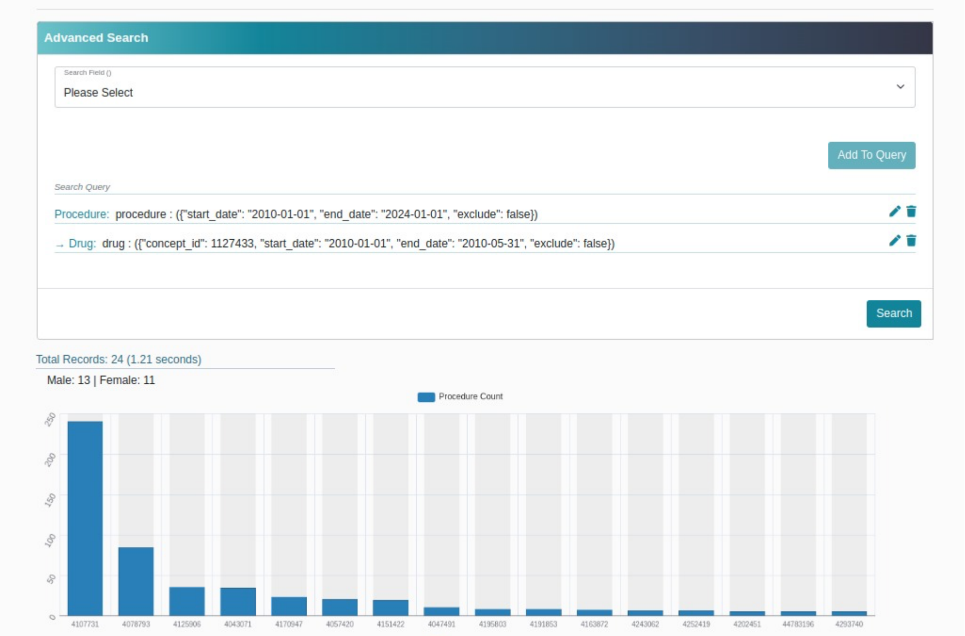
Sex: gender_concept_id : (AND "K")

Ethnicity: ethnicity_concept_id : (AND "0")

Year of Birth: year_of_birth : (AND [1960 TO 1975] AND -1973 -1961)

Figure 6. User-generated search query with advanced search UI.

After running the search query, the results are displayed in a stratified format by male and female, as well as relevant statistics based on the query parameters. Figure 7 depicts the simple query results, which search for all patients who underwent any procedure between 2010 and 2024 and used a specific drug.



Procedure: procedure : ({"start_date": "2010-01-01", "end_date": "2024-01-01", "exclude": false})

→ Drug: drug : ({"concept_id": 1127433, "start_date": "2010-01-01", "end_date": "2010-05-31", "exclude": false})

Total Records: 24 (1.21 seconds)

Male: 13 | Female: 11

Figure 7. Example of search query results.

4 System Security

We integrated several security solutions into the architecture of the HeKDisco platform and divided them into two major levels:

4.1 Secure data handling and storage

We implemented role-based access control at all layers, including the CDM database, OpenSearch, and meta search engine, to ensure that only authorized users can access certain data or features. Only necessary clinical information is stored and indexed, reducing the risk of sensitive data disclosure. Finally, we deidentified patients' IDs while they were indexed in OpenSearch using specific hash functions.

4.2 Secure communication across component

At this level, two main security solutions were considered: end-to-end encryption and API security. We used secure protocols (e.g., HTTPS, SSL/TLS) to communicate with all components, including AI modules, CDM mapping services, search engines, and the meta search engine. To prevent misuse, all APIs were authenticated using mechanisms such as OAuth 2.0 or API keys.

5 Technical Innovations of the Proposed HeKDisco Architecture

The HeKDisco platform introduces several technical innovations that address critical gaps in the current literature on federated clinical analytics and cohort search systems. Figure 8 summarizes these innovations.

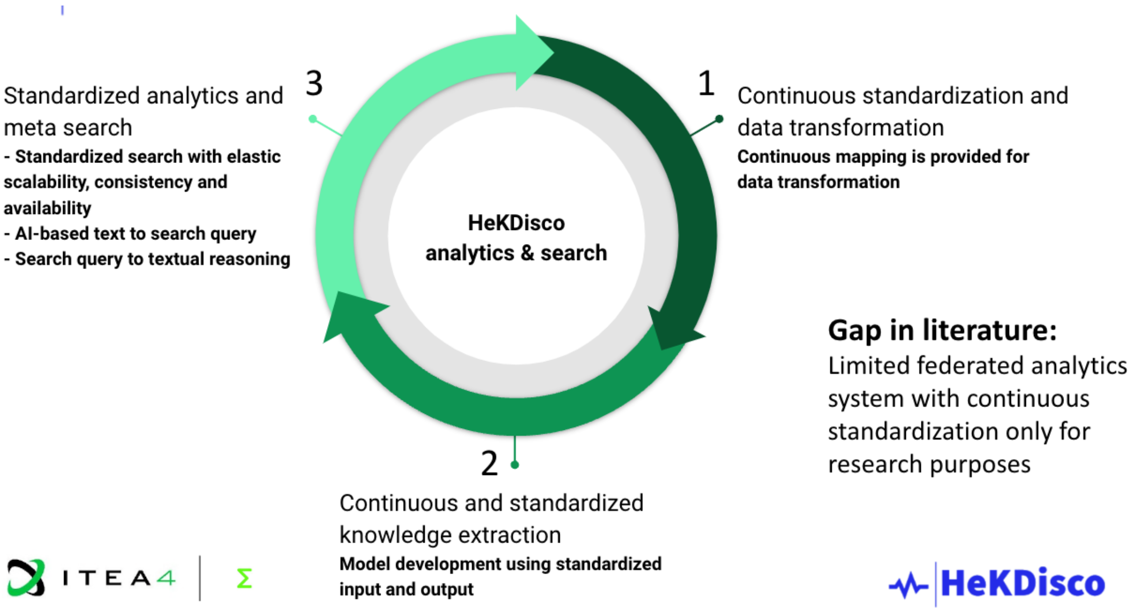

3
Standardized analytics and meta search
- Standardized search with elastic scalability, consistency and availability
- AI-based text to search query
- Search query to textual reasoning
1
Continuous standardization and data transformation
Continuous mapping is provided for data transformation
HeKDisco analytics & search
Gap in literature:
Limited federated analytics system with continuous standardization only for research purposes
2
Continuous and standardized knowledge extraction
Model development using standardized input and output
ITEA4
Σ
HeKDisco

Figure 8. Technical innovation in HekDisc's analytics and search platform.

5.1 Standardization and data transformation

The platform provides a robust and continuous data mapping mechanism to enable seamless standardization across diverse data sources. This ensures that all data is harmonized for federated analytics, thereby addressing the challenges posed by heterogeneous datasets.

5.2 Continuous and standardized knowledge extraction

HeKDisco integrates advanced models with standardized output frameworks to enable consistent and reliable knowledge extraction that can be subsequently used in federated analyses across clinical sites. This standardization facilitates the development of AI models for data-driven research, bridging the gap between clinical data and actionable insights.

5.3 Standardized analytics and meta search

The platform employs a standardized search framework with elastic scalability, ensuring consistency and availability for large-scale, distributed data queries. Clinical cohort analysis is inherently time-intensive due to its reliance on complex queries that combine events with index dates and apply detailed inclusion/exclusion criteria. These processes often involve multiple joins across large, heterogeneous datasets, requiring precise filtering of every event record to satisfy the inclusion/exclusion criteria tied to the index date. Existing analytics platforms struggle with the computational demands of these operations, particularly when processing large volumes of distributed clinical data.

The HeKDisco meta-search engine addresses these challenges by enabling real-time, scalable cohort searches, transforming what were traditionally time-intensive queries into a seamless, efficient process. In addition to its scalability, the platform provides a standardized infrastructure to support federated analysis across multiple clinical sites, fostering consistency and collaboration in clinical research. Moreover, HeKDisco holds significant potential for integration with emerging large language model (LLM) technologies. This integration can enable AI-powered text-to-query and query-to-reasoning capabilities, paving the way for intuitive and intelligent analytics tailored to diverse clinical and research needs.

6 Conclusion and Future Directions

The proposed architecture and implementation of the HeKDisco platform represent a groundbreaking advancement in clinical data search and cohort selection. By integrating innovative components such as the Query Analyzer and Meta Search Engine, the platform enables real-time and scalable analytics for multi-site clinical studies. The introduction of nested indexing schemas, OpenSearch-based infrastructure, and a unified query framework significantly reduces computational costs and query time, addressing the complexity of inclusion and exclusion criteria in clinical research. The user-friendly interface further empowers researchers to define and execute queries efficiently, transforming cohort selection from a traditionally time-intensive task into an interactive and immediate process. To the best of our knowledge, this is the first platform to support real-time cohort selection across distributed data sources, with enormous potential for streamlining clinical research and improving data-driven healthcare outcomes. In the future, we hope to integrate it with AI-powered text-to-query and query-to-reasoning tools, allowing for intuitive and intelligent analytics for diverse research needs.