Pilot 7: Lung Cancer

1. Key Information

1.1 Involved Partners

- National Center for Scientific Research “Demokritos” (DEM)
- Servicio Madrileno de Salud (HUMPH)

- Athens Technology Center SA (ATC)
- Leibniz Universitat Hannover (LUH)
- Universidad Politecnica de Madrid (UPM)

1.2 Involved Countries

- Spain
- Greece
- Germany

1.3 Keywords

- Length of Hospital Stay
- Assessment of People at Risk of developing Lung Cancer
- Toxic interaction of oncological & non-oncological drugs
- Drug adverse effects prediction

1.4 Task Description

The aim of the Lung-Cancer pilot is to improve the management of patients with cancer during their treatment, follow-up, and during their last period of life. The pilot utilizes Big Data to improve not only patients' experience, satisfaction, and main outcomes, but also to save substantial costs to the healthcare budget. The pilot addresses these shortcomings, by adopting a pipeline that starts with medical data (open and patient records), performs pattern extraction and ends up in a knowledge graph that captures essential correlations in the Lung-Cancer treatment.

Figure 1 presents the different data sources and processing steps in the lung cancer pilot. Data from electronic health records and publications as preprocessed applying NLP methods, such as named entity recognition and relation extraction and then normalized to a structured vocabulary. Extracted information is then stored as triplets in a *knowledge graph*. On this graph further methods are applied to detect patterns, e.g., among potential drug-drug interactions, toxicities, and visits to the patients to the hospital.

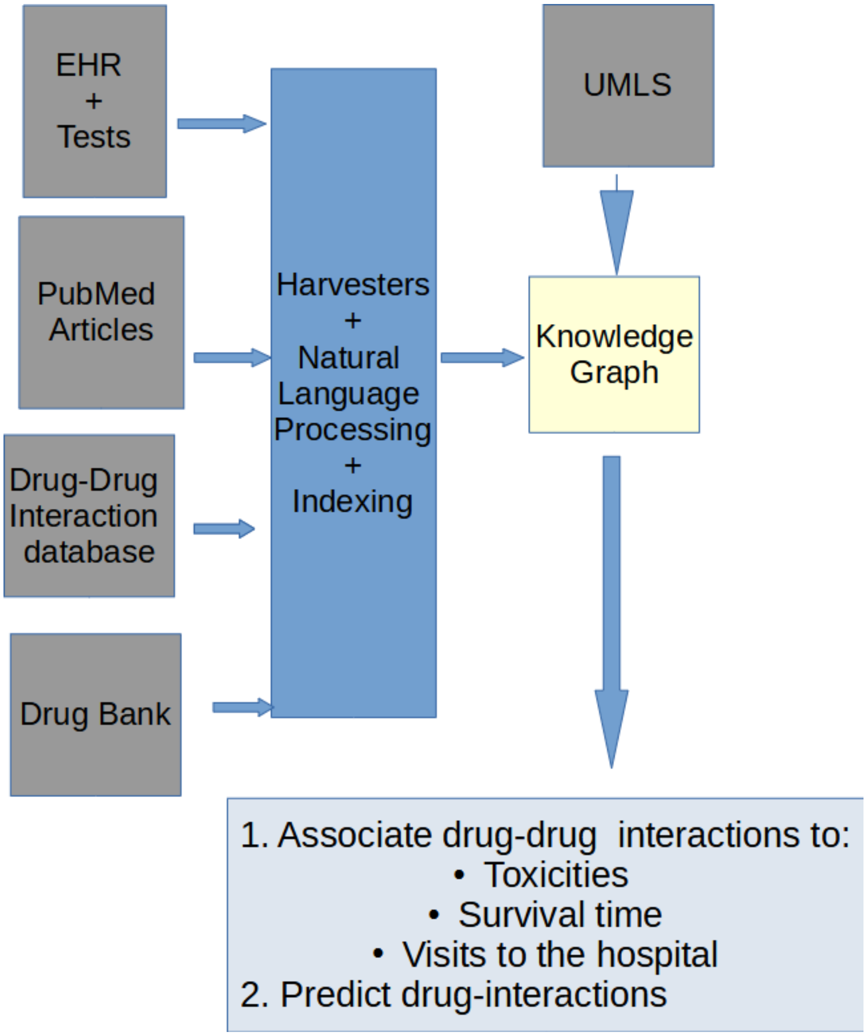

EHR
+
Tests
PubMed Articles
Drug-Drug Interaction database
Drug Bank
Harvesters
+
Natural Language Processing
+
Indexing
UMLS
Knowledge Graph
1. Associate drug-drug interactions to:
• Toxicities
• Survival time
• Visits to the hospital
2. Predict drug-interactions

Figure 1: Conceptual Lung-Cancer Architecture

2. Building Blocks

2.1 Architecture

2.1.1 System Architecture

The high-level architecture of the Lung-Cancer pilot follows a layered architecture pattern. Figure 2 presents a high-level overview of the pilot's software architecture.The arrows inside the diagram represent the flow of information between components. The system's modules are divided into multiple distinct layers, each responsible for a specific set of functionalities. By convention, the layers interact with each other in a top-down manner, with each layer being able to access all layers below. The layers are described as follows:

- Hardware Layer: Docker is used as a packaging and deployment mechanism for the pilot components.
- Resource Management Layer: It provides mechanisms for the management of the Docker containers.
- Data Layer: is a logical set where all the heterogeneous data sources are archived.
- Extraction Layer: basic analysis of the disparate data sources exploited in the pilot is happening in order to extract knowledge from them.
- Platform Layer & Semantic Layer: Platform layer consists of the big data technologies/tools/frameworks that the various pilot's components use to perform their processing. Part of the platform layer is the Semantic Layer that semantically integrates all the heterogeneous data sources and the knowledge extracted in the Extraction layer and gives the opportunity to generate information that is rich, auditable and reliable. The result of semantically integrating the heterogeneous data source is a knowledge graph, where the meaning of the integrated data is described using a unified schema.
- Analytics Layer: This is the layer where the high-level analysis of all the heterogeneous sources of information integrated in the knowledge graph is happening; a process of inspecting, cleaning, transforming and modelling data with the goal of discovering useful information, patterns, suggesting conclusions and supporting decision-making for the users of the pilot.
- Presentation Layer: Contains the User Interfaces and Visualization Modules that help the user to get all the information that the pilot offers in a user-friendly way.
- Support Layer: Support layer is a vertical layer that offer utilities in more than one layers and components.
- Access Control Layer: This is the second vertical layer responsible to cover security-related functionalities that need to be enforced by the pilot components according to an agreed component-specific security policy. It is responsible for providing a centralized authentication and authorization service so that all components can communicate in a secure and reliable way. Furthermore, some components of the pilot are physically hosted in remote places. These components are considered to as external components and communicate with the rest of the platform over REST APIs.

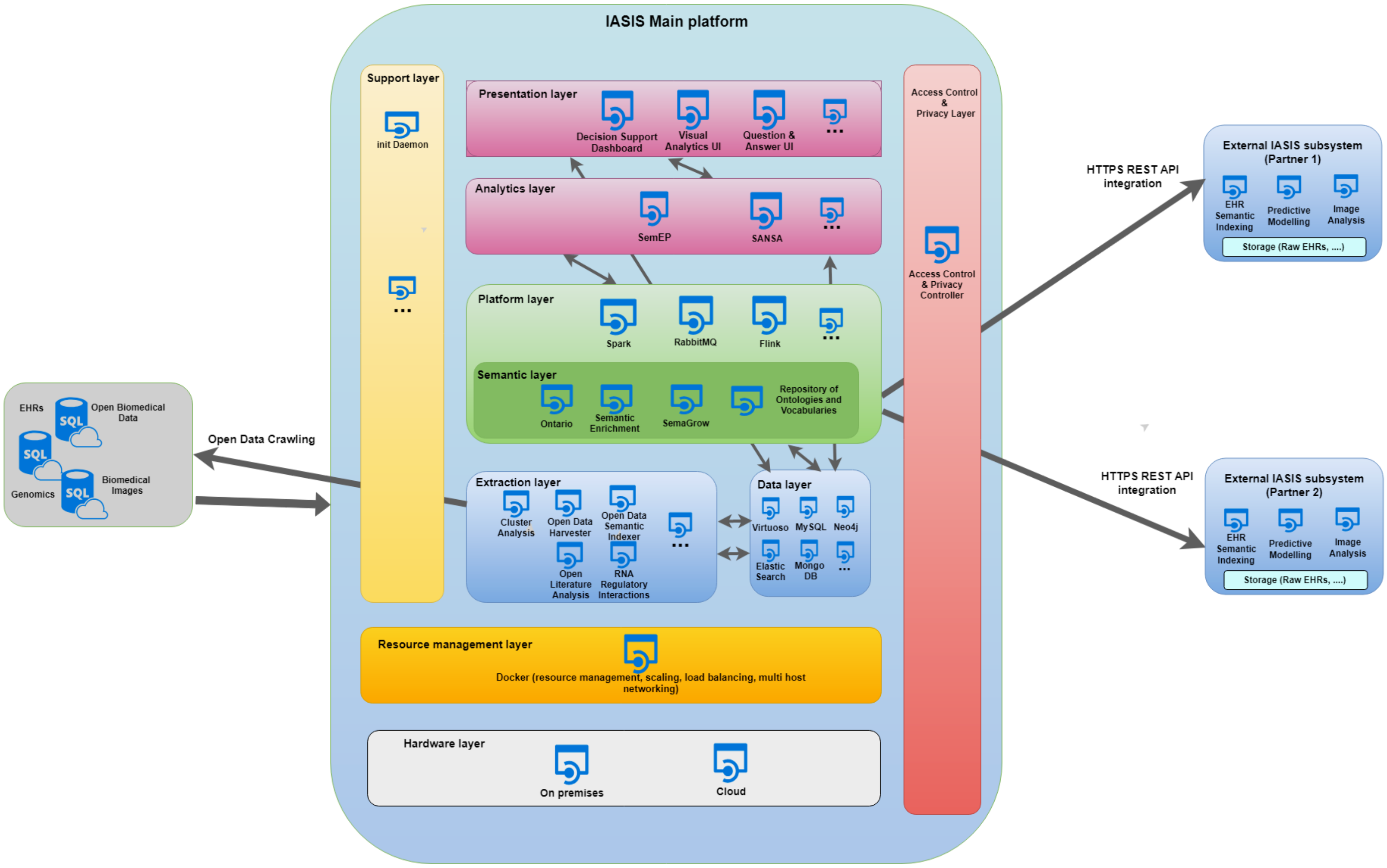


Figure 2: Lung Cancer pilot, software architecture

The physical deployment diagram is presented in Figure 3.
In the Lung Cancer pilot, we have three physical nodes, each accomplishing different tasks for the needs of the pilot.
1. The Analysis node, which is based in Madrid (UPM) and its responsibility, is to hold raw data which processes it in order to extract useful information. Part of this processed dataset is sent to the Data enrichment node, through a secured channel. The processed information is also made available through a series of web services and is consumed by the central node components.
2. The Data enrichment node, which is located in Hannover (LUH) and is responsible to combine information from various sources and semantically annotate them, forming a knowledge graph. The knowledge graph then is sent to the Central node again through a secured channel.
3. The Central node, which is located in Athens (NCSR & ATC) and is the node where most of the pilot's components reside. Its role is not only to provide the infrastructure for the components, and to orchestrate the communication between them in order to retrieve more information, but also to present this information to the users based on their needs and interactions with the system.

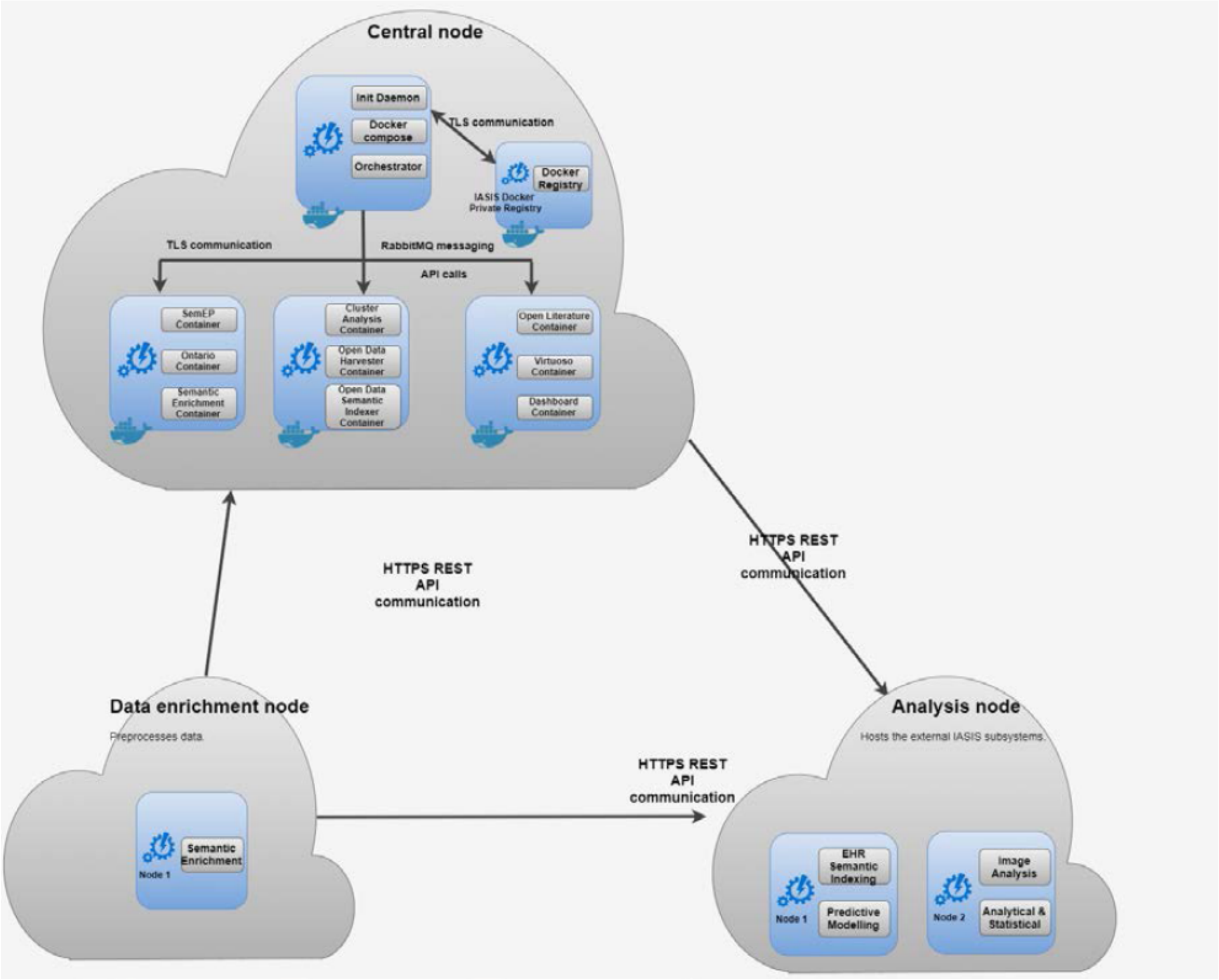


Figure 3 Physical deployment of the Lung Cancer pilot

2.1.2 Data Flow & Interoperability of services

In this subsection we provide an overview of the communication between the integrated components that participate in the off-line and real time workflows of the system. With the off-line workflows the system achieves the gathering of new information that is added to the knowledge graph, while the real time processes are the realization of the use case scenarios.

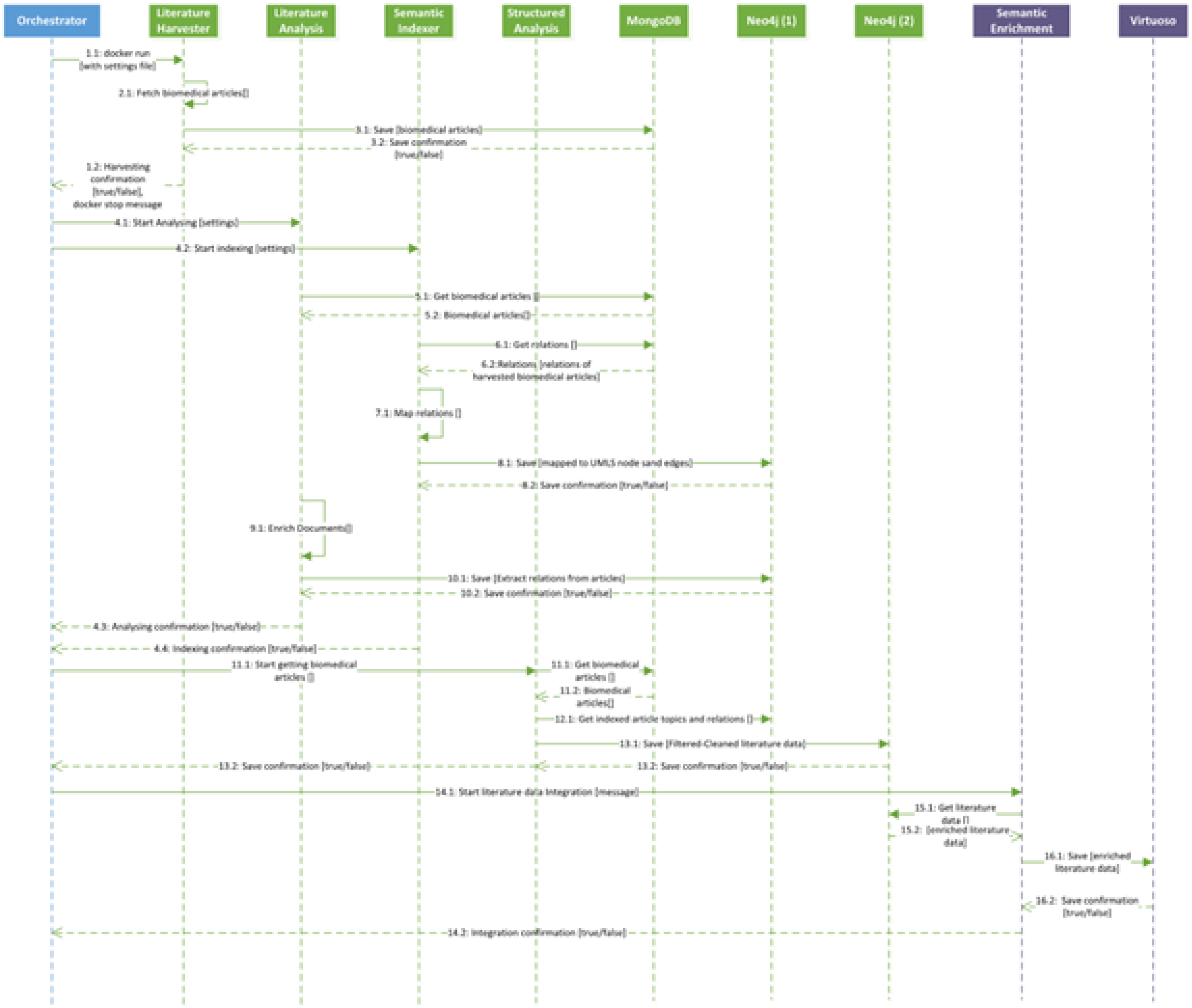
The main user functionalities supported currently in the pilot, in summary are:

- Search based on population criteria (statistics, kpis)
- Search in the knowledge graph (publications, drug interactions, side effects)
- Get answers to questions
- Ask for experts opinion (chat)

In the following sequence diagrams we present the interactions between the components for the completion of the first three functionalities, as well as, the completion of the data gathering process which is an offline process.

Literature harvesting process

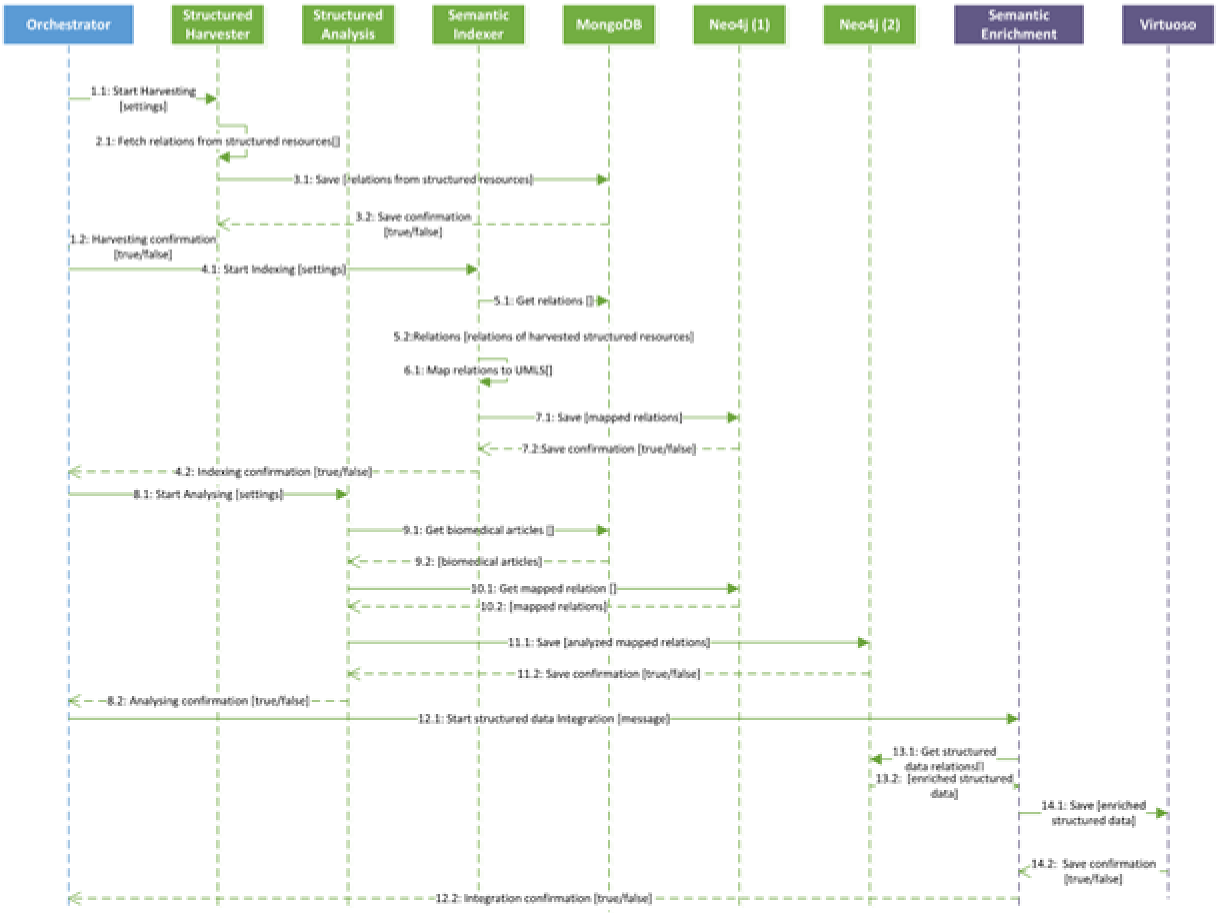
The next diagrame depicts the sequence diagram followed by the components for integrating knowledge extracted from the open publications. First, the components for knowledge extraction are contacted; they perform literature analysis and annotate the extracted knowledge using controlled vocabularies and ontologies. Then, the results of the semantic annotation and indexing of the publications, are uploaded into MongoDB and Neo4j. Next, the semantic enrichment component is called and it executes several queries over Neo4j in order to collect the data to be uploaded in the knowledge graph. Once the data is semantically enriched, the corresponding RDF triples are stored in Virtuoso.


Orchestrator
Literature Harvester
Literature Analysis
Semantic Indexer
Structured Analysis
MongoDB
Neo4j (1)
Neo4j (2)
Semantic Enrichment
Virtuoso
1.1: docker run [with settings file]
2.1: Fetch biomedical articles[]
3.1: Save [biomedical articles]
3.2: Save confirmation [true/false]
1.2: Harvesting confirmation [true/false], docker stop message
4.1: Start Analysing [settings]
4.2: Start indexing [settings]
5.1: Get biomedical articles []
5.2: Biomedical articles[]
6.1: Get relations []
6.2 Relations [relations of harvested biomedical articles]
7.1: Map relations []
8.1: Save [mapped to UMLS node sand edges]
8.2: Save confirmation [true/false]
9.1: Enrich Documents[]
10.1: Save [Extract relations from articles]
10.2: Save confirmation [true/false]
4.3: Analysing confirmation [true/false]
4.4: Indexing confirmation [true/false]
11.1: Start getting biomedical articles []
11.1: Get biomedical articles []
11.2: Biomedical articles[]
12.1: Get indexed article topics and relations []
13.1: Save [Filtered-Cleaned literature data]
13.2: Save confirmation [true/false]
13.2: Save confirmation [true/false]
14.1: Start literature data integration [message]
15.1: Get literature data []
15.2: [enriched literature data]
16.1: Save [enriched literature data]
16.2: Save confirmation [true/false]
14.2: Integration confirmation [true/false]

Literature Harvesting process

Structured Harvesting process

The process of harvesting data from structured sources is similar to the Literature Harvesting process and is depicted in the next sequence diagram.


Orchestrator
Structured Harvester
Structured Analysis
Semantic Indexer
MongoDB
Neo4j (1)
Neo4j (2)
Semantic Enrichment
Virtuoso
1.1: Start Harvesting [settings]
2.1: Fetch relations from structured resources[]
3.1: Save [relations from structured resources]
3.2: Save confirmation [true/false]
1.2: Harvesting confirmation [true/false]
4.1: Start Indexing [settings]
5.1: Get relations []
5.2:Relations [relations of harvested structured resources]
6.1: Map relations to UMLS[]
7.1: Save [mapped relations]
7.2:Save confirmation [true/false]
4.2: Indexing confirmation [true/false]
8.1: Start Analysing [settings]
9.1: Get biomedical articles []
9.2: [biomedical articles]
10.1: Get mapped relation []
10.2: [mapped relations]
11.1: Save [analyzed mapped relations]
11.2: Save confirmation [true/false]
8.2: Analysing confirmation [true/false]
12.1: Start structured data Integration [message]
13.1: Get structured data relations[]
13.2: [enriched structured data]
14.1: Save [enriched structured data]
14.2: Save confirmation [true/false]
12.2: Integration confirmation [true/false]

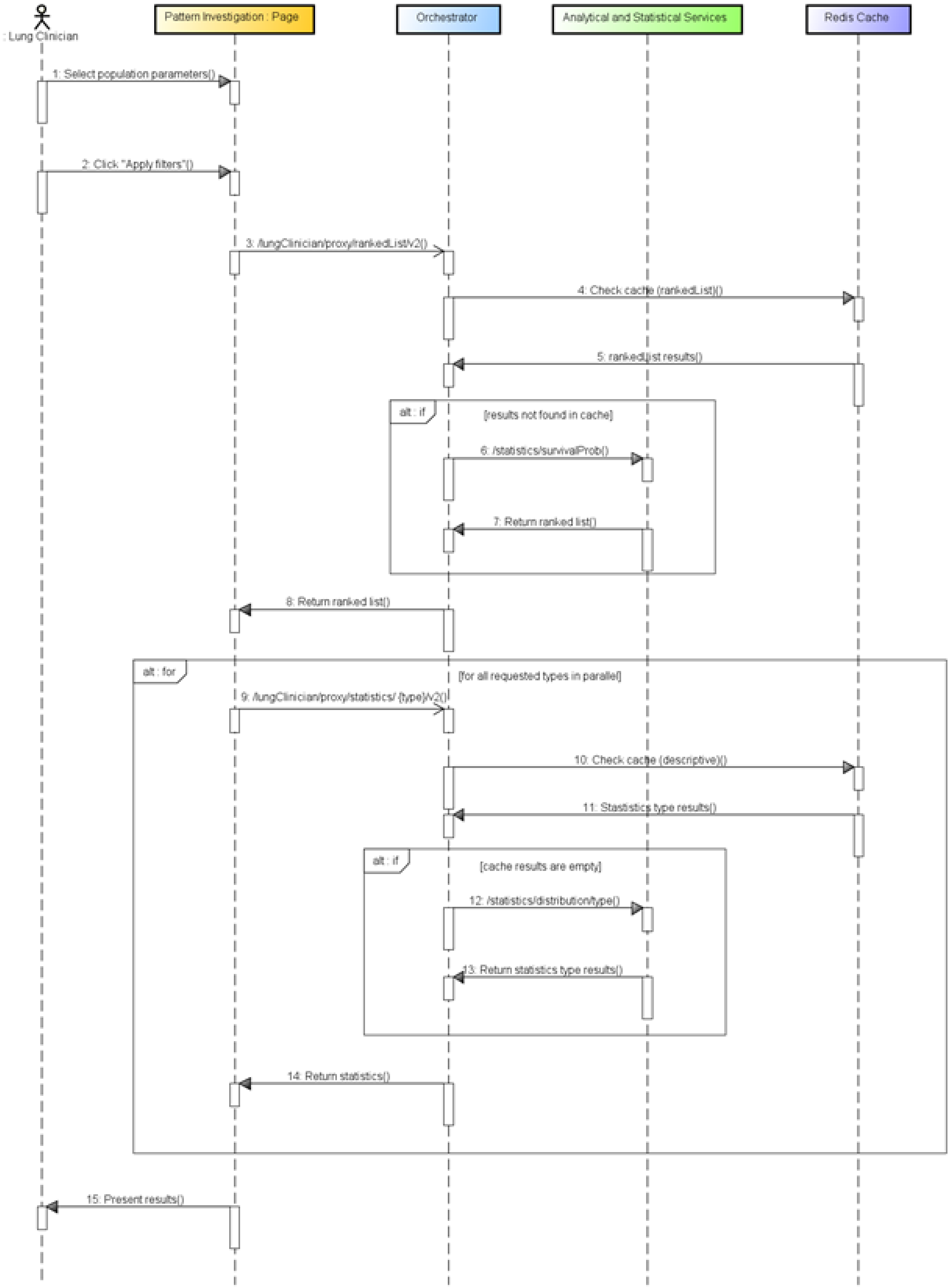
Structured Harvesting process

Search based on population criteria

The sequence diagram of the actions followed when a user searches based on selected population criteria is depicted on next diagram. For simplicity we present the flow for the statistics, but the same is for the data related to the KPIs and the survival probabilities.

First, a user enters the dashboard and most specifically the “New Pattern Investigation” page and selects the search criteria. This allows for the definition of a population to be studied. Then, the Orchestrator checks first if the results for the specific search values are already available in the cache. If not, it contacts the Analytical & Statistical Services component to retrieve the statistical information.

The Analytical & Statistical component analyses the criteria sent, identifies the resulting population and calculates the necessary statistics which it returns to the Orchestrator. The Orchestrator then contacts the dashboard for visualizing the results.

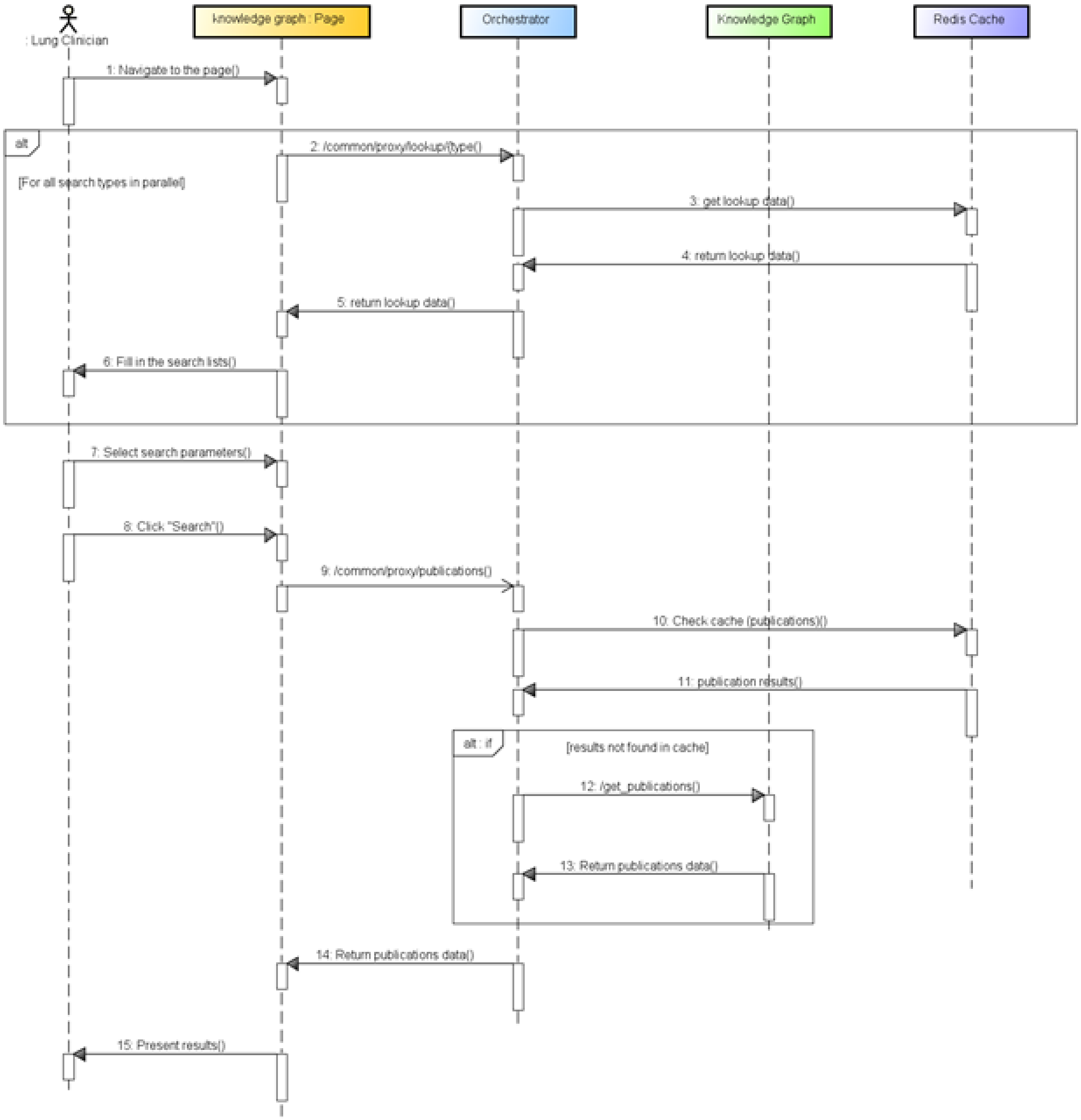

: Lung Clinician
Pattern Investigation : Page
Orchestrator
Analytical and Statistical Services
Redis Cache
1: Select population parameters()
2: Click "Apply filters"()
3: /lungClinician/proxy/rankedList/v2()
4: Check cache (rankedList)()
5: rankedList results()
alt : if
[results not found in cache]
6: /statistics/survivalProb()
7: Return ranked list()
8: Return ranked list()
alt : for
[for all requested types in parallel]
9: /lungClinician/proxy/statistics/ {type}/v2()
10: Check cache (descriptive)()
11: Stastistics type results()
alt : if
[cache results are empty]
12: /statistics/distribution/type()
13: Return statistics type results()
14: Return statistics()
15: Present results()

Search based on population criteria

Search in the knowledge graph

In this scenario, a clinician can search in the knowledge graph for publications, drug interactions and side effects. For simplicity, we present the search for publications, since the flow and the components involved are basically the same.

Initially the user navigates to the “knowledge graph” page. Every time the page is initialized, the Orchestrator fetches the lookup data that populate the lists of the parameters from the cache. The user selects the search criteria and searches for publications. The Orchestrator again checks if the results of the specific search values are present in the cache. If not, the Orchestrator retrieves the results from the Knowledge graph component and updates the cache. The results are then returned to the UI and presented to the user.

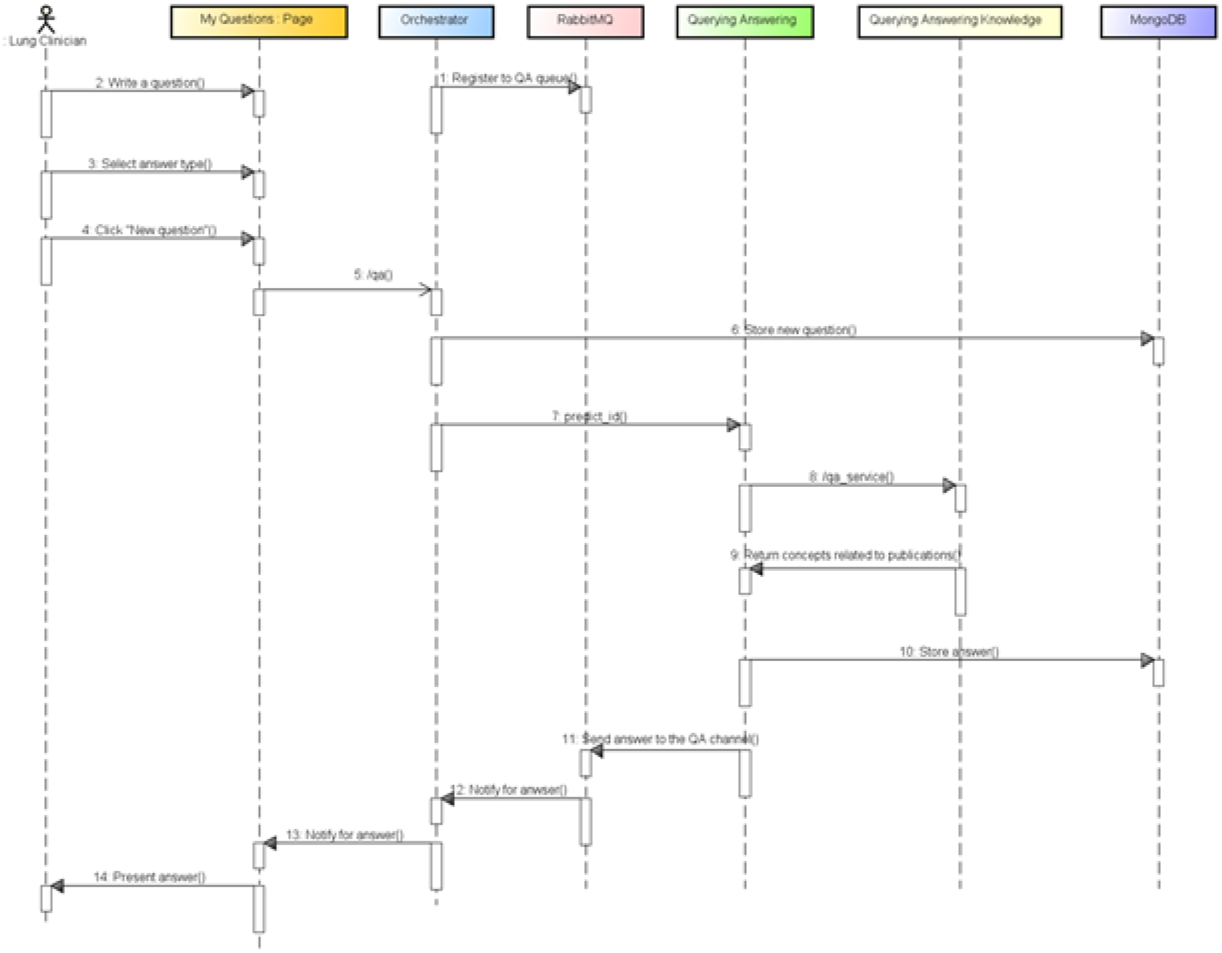

: Lung Clinician
knowledge graph : Page
Orchestrator
Knowledge Graph
Redis Cache
1: Navigate to the page()
alt
[For all search types in parallel]
2: /common/proxy/lookup/{type}()
3: get lookup data()
4: return lookup data()
5: return lookup data()
6: Fill in the search lists()
7: Select search parameters()
8: Click "Search"()
9: /common/proxy/publications()
10: Check cache (publications)()
11: publication results()
alt : if
[results not found in cache]
12: /get_publications()
13: Return publications data()
14: Return publications data()
15: Present results()

Search in the knowledge graph

Get answers to questions

The steps that need to be performed by a clinician in order to submit his/her questions and receive answers from the system, are presented in the next diagramme. When the Orchestrator is initialised for the first time, it registers to a RabbitMQ queue in order to listen for messages related to questions and answers. In order for a user to ask the system a question, first the user navigates to the My Questions page and then writes the question in natural language and selects the type of answer. The Orchestrator stores the question and notifies the Querying

Answering component that there is a new question. The Querying Answering component finds the answer and a set of relevant publications and then contacts the Querying Answering Knowledge component. This component queries the knowledge graph in order to find relevant concepts to the publications found and returns them back to the Querying Answering component, which in turn notifies, through the RabbitMQ queue, the Orchestrator.



Get answers to questions

2.1.3 Necessary Hardware

With the current set up of the pilot components the minimum hardware required is:
3 physical nodes or VMs with quad core CPU, 64GB of memory and 2TB disk each.

2.1.4 Software Components

This section provides an overview of the software components developed in the context of the Lung Cancer pilot.

Orchestrator

The Orchestrator module comprises the core of the platform. This module acts as a mediator that enables the communication between the different components of the platform and the exchange of information among them in an orchestrated distributed environment, thus it provides a set of key features towards the integration of components into the platform and the implementation of the interoperability requirements and business cases of the platform

Open Data Harvester

The Open Data Harvester is responsible for collecting and transforming data found in various ontologies and selected resources and making them available to the Knowledge graph for further analysis. It comprises two subcomponents, the Structured Harvester and the Literature Harvester.

Open Literature Analysis

This module processes the text data that have been collected from the Literature Harvester and enriches them with the biomedical entities and relations found. In order to do so, it currently relies on SemRep[1], a third-party software that conforms to the UMLS lexicon and semantic network

Open Data Semantic Indexer

This module transforms the data, meaning the ontologies and other structured sources that have been collected from the Structured Harvester, and maps them to the UMLS lexicon. Moreover, it coverts them to an appropriate format and inserts them as nodes and edges in a Neo4j database. In order to do the mapping, the module currently relies on the REST API offered by the UMLS.

Open Data Analysis

This module collects, filters and analyses structured, semantically indexed data available in a neo4j graph produced both from semantic indexing and literature analysis. The Structured Analysis module handles details relevant to the extraction of knowledge from the resources of each relation (e.g. specific articles, ontologies etc) and produces semantic predications relevant to the use cases annotated with confidence scores and provenance details available for integration in the Knowledge Graph.

Question Answering Component

This module provides answers and contextual information to questions posed by the end-users through the platform. More specifically, a user poses a question using free text and denotes the type of the question by selecting between ‘yesno’, ‘factoid’ and ‘list’ types. The question is posed to the module, which acknowledges the task with an immediate response and asynchronously processes the request. Once the process is completed, the module saves the results in a pre-defined MongoDB and notifies the platform through the dedicated RabbitMQ using the question’s unique identifier, provided with the initial request.

Ontario

Ontario is a federated SPARQL query engine able to process SPARQL queries against a federation of SPARQL endpoints. In this component, a federation of SPARQL endpoints provides access to the knowledge graph and the external knowledge graphs that are linked to it, for example, DBpedia, DrugBank, among others.

Semantic Enrichment

The Semantic Enrichment component creates a knowledge graph from Big data sources and links it to external knowledge graphs. The Semantic Enrichment receives as input a set of mappings and Big heterogeneous datasets, e.g. electronic health records, genomic data, open datasets, medical images, and it outputs the knowledge graph.

SemEP

SemEP implements unsupervised machine learning methods able to identify patterns and unknown associations among entities in the Knowledge Graph. It combines techniques for graph partitioning with semantic similarity measures to discover communities of subgraphs and predict associations in the Knowledge Graph.

EHR Semantic Indexing

The EHR Semantic Indexing component analyses Electronic Health Records and gets information that is provided to the Knowledge graph via REST API in order to enhance the collected information.

Knowledge Graph

The Knowledge Graph is explored to return the publications which have annotations related to the provided input (relations HAS_TOPIC and MENTIONED_IN). The publications are ranked based on the average of the confidence score of each annotation (Jaccard index). The publications can be filtered based to reduce the number of unrelated publications. The filtering is done by considering only the CUI ids of concepts which have a relation to a lung cancer drug. Similarly, the Knowledge Graph component also provides information about side effects and drug interactions.

Querying Answering Knowledge

This component is called from the main Querying Answering module to provide some related knowledge graph data, based on a free text question. For this purpose, it receives a list of specific CUI/publication ids and responds with related annotations from the knowledge graph.

Analytical & Statistical Services

The Analytical & Statistical Services component provides Web services which perform statistical and predictive analysis for a given population selection criterion. This component is deployed on the UPM node and communicates with the Orchestrator through secured API calls. Through this component, the Orchestrator retrieves information on the estimated distribution of survival and statistics of confidence.

---

[1] https://semrep.nlm.nih.gov/

2.3 Data Processing

2.3.1 Processing of large structured / unstructured data sources

2.3.1.1 Data Sources

Literature related data sources include the PubMed (about 28 million articles) and the PubMed Central (PMC), with about 4.5 million articles In the first, only the abstract can be accessed for free, whereas on the second there can be accessed the whole article. Apart from keyword search PubMed also supports search based on MeSH terms.

Structured databases include the DrugBank that contains information about drug-to-drug interactions. DrugBank contains more than 10,000 drug entities, with more than 200 fields of information per drug. There are about 300,000 drug-to-drug interactions. DrugBank data are described in the XML format.

Biomedical ontologies are an important source of domain knowledge.

● The Disease Ontology (DO) integrates disease terms and identifiers from different resources. It includes 10,000 concepts with 6,000 mappings to UMLS.

● The Gene Ontology (GO) contains 24,000 concepts representing biological processes, molecular functions and cellular components. GO is mapped to UMLS.

● MeSH is a hierarchical controlled vocabulary developed by NLM to semantically index biomedical articles primarily based on their topics. It provides more than 28,000 descriptors organized in sixteen tree structures that cover a broad spectrum of knowledge, from chemicals and organisms to humanities and geographical locations. All MesH descriptors are linked to specific concepts, which are integrated in the UMLS.

Electronic Health Records of Patients (EHRs):

- The data provided by Hospital Universitario Puerta de Hierro de Madrid in terms of Electronic Health Records refer to 749 patients that are suffering or have suffered from lung cancer.
- The total number of documents (clinical notes and reports) that belong to the 749 patients is 199,846. We also have 452 clinical notes from the call service that correspond to 133 patients.
- In terms of non-scheduled visits, a separate file has been provided by the hospital. This file contains two columns: 1) the date of the visit and 2) the ID of the patient. The current number of unscheduled visits is 139 for 84 patients.

Literature data harvester: A PubMed harvester has been built to retrieve articles based on MeSH terms. The articles are returned in XML format containing more than 100 types of hierarchically related entities. Then the harvester keeps only the abstract and the topics of the article and converts them to JSON format. The PMC harvester is similar to the PubMed harvester, but it returns the full text of the article. The drugbank harvester access the DrugBank and converts the results to JSON format. The results of the 3 harvesters are mapped to the Unified Medical Language System (UMLS).

Parallel execution of the harvesters

Some experiments have been performed to increase the performance of the data harvesters on fetching data from literature sources by running multiple instances of each harvester simultaneously. Experiments were also performed to increase the performance of literature analysis (entity recognition and relation extraction), which in the serial version takes about a week. In addition, experiments were performed on a parallel dumping the results of literature analysis on a graph database (Neo4j). Finally, the graph needs to cleaned-up for duplicate triplets but also to compute a confidence factor for each triplet.

Data Source	**Description**	**Acquisition**	**Characteristic (Size, Patients, Years, Origin/Region)**
PubMed	Biomedical arcticles	Software Harvester	160,000 abstracts from PubMed and 10200 articles form PubMed Centreal
Electroni Health Records	Electronic Health Records of Lung-Cancer patients from Puerta De Hierro Majadahonda	Software pipeline	More than 300,000 documents comprising

			clinical notes, reports, etc from around 1,000 lung cancer patients.

multiple sources	integration to data warehouse	data access	data stored in cloud	multi-party architecture	secure environment	transform raw / unstructured data
yes	yes	UPM access hospital data from HUMPH; the data are de-identified and stored on a secure server	yes	no	yes	yes

2.3.1.2 De-Identification and anonymisation

2.3.1.3 Acquisition

Raw data are acquired from public structured databases (e.g. drug bank) and public unstructured databases (e.g. PubMed and PubMed Central). Data harvesters have been built to acquire the data.

In order to ensure that the clinical data is transferred in a secure way and respecting the data privacy and access control regulations. Knowledge about anonymised electronic health records of lung cancer patients is extracted using EHR Text Analysis; this is conducted by Universidad Politécnica de Madrid (UPM). The extracted knowledge is annotated with terms from the UMLS control vocabulary and made available from UPM to the partners of the Leibniz University of Hannover (LUH) via a REST API. The access of the API requires the authentication of a username and password which are encoded in Base64 following the TLS protocol, i.e. the communication established to download the data is encrypted respecting a security protocol. Data downloaded using the API is stored in a server at LUH; this server is not accessible via the Internet and only the partners from LUH have accounts and access it from the Intranet. The description of the semantified version of the downloaded clinical data states that HUPHM only grants the permission to UPM to Read (R), Merge (M), Storage (S), and Distribution (D), and to LUH for Read (R), Merge (M), and Storage (S); only aggregated properties can be Read (R) and Merge (M) for the rest of the partners. LUH has the rights to Read (R), Merge (M), and Storage (S) this data, and only distribute the results of the analysis.

Two types of connections are established to download data from the partners.

- Secure REST API: The data set Processed LC_EHR is retrieved using a REST API that is contacted via a REST client, e.g. Insomnia[1]. As a response, the data is obtained in a JSON file which includes anonymised records extracted from the clinical notes; this file is stored in an LUH server and encrypted by encryption methods provided by the operating system, e.g. GnuPG or mCrypt.
- Access from Private Clouds: Open data is collected; PubMed publications annotated with CUIs from UMLS are accessed from private clouds.

Once the data is collected the following steps are conducted:

1. Preprocessing: The collected data is processed to extract aggregated values, e.g. lists representing the evolution of patient conditions, tumours, and mutations according to treatments; entity recognition and linking are also conducted over textual attributes.
2. Secure Storing: The results of the data extraction and processing process are stored and encrypted in the LUH servers; these servers can only be accessed in the intranet of the Scientific Data Management group at LUH.
3. Data Integration and Semantic Enrichment: The aggregated data is used by the Semantic Enrichment process to populate the BigMediliytics knowledge graph. RML mapping rules are utilised for guiding this process. In the BigMediliytics knowledge graph, the aggregated data is only integrated with public and open data that includes drugs, mutations, publications, or side effects; no further personal data about the anonymised records is included or integrated in the BigMedilytics knowledge graph.

Knowledge Graph Exploration: The knowledge graph can be accessed via the APIs enabling the access of data in the knowledge graph that is collected from open data sources. Clinical partners can execute queries against the knowledge graph according to access regulations.

2.3.1.4 Cleansing

Regarding the analysis of open source data, we discard some relations that are not important in the lung-cancer, also we remove duplicate triplets. Also, some data items might be excluded based on the quality or the impact of the source.

Regarding the EHR of patients, data are collected that conform to a database schema that has been predetermined.

[1] https://insomnia.rest/

2.3.1.5 Data Integration

Each raw data stream is analysed, information is extracted, and then it is structured as triplets (if unstructured), finally it is mapped to biomedical ontologies (e.g. UMLS, SNOMED). At a second stage the triplets are mapped to the knowledge graph, that is represented in the RDF format.

2.3.1 Multi-velocity processing of heterogeneous data streams

Stream name	Contents of stream	Stream velocity
Open source unstructured data	Full text of biomedical articles from PubMed Central, and Abstracts of biomedical articles from PubMed that are related to lung-cancer and related UMLS concepts	The methods that process the data are incremental.
Open source structured data	Drug-Bank entries with selected relations among drugs.	The methods that process the data are incremental.
Medical health records	Structured and Unstructured information of medical records enriched with transcripts of phone logs	The stream is updated when a new lung cancer patient is admitted to the hospital, or when there is an update to the EHR. The hospital sends the new data to UPM when a reasonable amount of new data is available.

Combinations of different streams
First, we need to associate information in the medical records, and bibliographic data. At a later stage data mining will be performed on the integrated knowledge graph to make predictions about possible adverse effects of drugs.

Challenges of different streams
First, the data streams contain unstructured information, thus prediction algorithms cannot process data in raw format. Moreover, the unstructured information (i.e. text) is in English and in Spanish. Second, the streams are not semantically related, thus the integration is hindered. Finally, there are issues related to the data privacy of the EHRs.

Solution

The unstructured information, typically text, is analysed so that structure is extracted from it. We use different NLP tools for publications (which are typically in English), and for EHR (which are typically in Spanish). Regarding the issue of integration, the entities that are detected are mapped to biomedical ontologies (SNOMED, UMLS, MeSH) to facilitate semantic integration. Finally, all the data coming from the streams is represented as triplets, and integrated in a knowledge graph in RDF format. The knowledge graph can query and machine learning algorithms can be applied to it, so as to detect interesting patterns (e.g. patients that respond similarly to a drug treatment) or to predict possible adverse effects. Finally, regarding privacy issues: the raw EHR are stored in the same country they originate (i.e. Spain). Regarding the sensitive data that are stored in the knowledge graph, a federated query engine is used (MUDLER++) which can enforce the satisfaction of privacy constraints.

Streams mixed	**Technology used**	**Purpose of the process**	**Result (name if is a stream)**
Open source unstructured and structured data Medical Health Records	MongoDB, Neo4j, semantic technologies	Integrate data for statistical analysis, data mining and information retrieval	Knowledge graph

Stream combination	**Difficulties**	**Solutions**
Analytics input	Originally the streams are not semantically related Quality of information on open data	Each stream is represented as a collection of triplets, and thus they can be combined by semantic web technologies (e.g. RDF and existing biomedical ontologies e.g. UMLS, SNOMED) Associate metadata related to the provenance of information (e.g. impact of the journal, type of publication etc).

2.4 AI Components

2.4.1 Deep learning for multilingual NLP and image analytics

2.4.1.1 Natural Language Processing

Language	How will NLP support your pilot?	How will NLP help you to reduce costs?
Spanish	Enormous amounts of information is written in natural language in clinical texts. This unstructured information can be used to enrich structured data and thus significantly increase the amount and richness of patient data to improve the knowledge that we have about the patients and their associated processes in the hospital.	The retrieval of the EHR data along with the used in other sources could allow to have an improvement of the KPIs, which is translated in a reduction of the economic costs associated to the hospital.
English	Named entity recognition in biomedical text to recognize UMLS concepts. Extraction of semantic predications where the concepts are from the UMLS Metathesaurus and the relation from the UMLS semantic network	Thousands of articles related to Lung cancer cannot be manually processed (at a reasonable cost) to extract useful information about drugs, interactions and patients. Thus, automated methods offer a cost advantage.
English	A knowledge driven framework for transforming data in unstructured and structured formats into instances of a knowledge graph	Integration of data collected from different data sources from open data and clinical notes, into the knowledge graphs.

Which NLP tasks do you address?	Method	Software frameworks	Vocabularies/corpus used	Describe your method in a few sentences	Describe your corpus/training data?
Named Entity	Deterministic	Apache	UMLS repository	Rule-based system that deterministically	The Unified Medical Language System (UMLS)

Recognition	ic rule-based system	UIMA		classifies tokens in previously defined entities in UMLS repository. It relates a token with a concept and its semantics.	is a compendium of many controlled vocabularies in the biomedical sciences (created 1986). It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. The UMLS was designed and is maintained by the US National Library of Medicine and it is updated quarterly
Event (Concept-Date Relation) Recognition	Cyclic Bidirectional Dependency Network	Stanford Core NLP	AnCora Spanish 3.0, DEFT Spanish Treebank V2 (LDC2015E66)	Bidirectional dependency network approximate the joint distribution over a set of random variables with a set of local conditional probability distributions that are learned independently.	This corpus consists of about 17,000 sentences, drawn from Spanish (Spain) newswire and from an older balanced Castilian Spanish corpus (3LB). The DEFT Spanish Treebank V2 (LDC2015E66). This corpus contains the full International Spanish Newswire Treebank and the full Latin American Spanish Discussion Forum Treebank (roughly 5,000 sentences in total).
Part Of Speech Recognition	MLP-network	Apache UIMA with OpenNLP models	Trained on conll02 shared task data	Multilayer Perceptron Network using 5-grams (2 forward, 3 backward).	Reuters Corpus, Volume 2, Multilingual Corpus, 1996-08-20 to 1997-08-19 (Release date 2005-05-31, Format version 1, correction level 0). RCV2 from Reuters Corpora
Named Entity recognit	Proprietary	Metamap	Applied on PubMed, PubMed	Named entity recognition in biomedical text, it	Pretrained (https://metamap.nlm.nih.gov/)

ion			Central	recognizes UMLS concepts in text	
Relation Extraction	Proprietary	Semrep	Applied on PubMed, PubMed Central	Relation extraction in biomedical text. Extracts semantic predications where the concepts are from the UMLS Methathesaurus and the relation from the UML semantic network	Pretrained (https://semrep.nlm.nih.gov/)
Entity and Relation Extraction	Proprietary	FALCON	Applied to short text in open data sources	Relation extraction in biomedical text and linking to terms in UMLS and MEDRA	https://labs.tib.eu/info/en/project/falcon/ The corpus is the alignments among labels and URIs from DBepdia, Wikidata, and UMLS

2.4.1.1.1 Evaluation

The associations that follow were deemed as clinically useful for the oncologists and they were extracted from structured clinical data, that initially stem from raw data

- Risk of developing lung cancer associated to:
    - medical services used before diagnosis
    - and months passed between last service visit by the patient and their diagnosis date for: *cardiology, pneumology* & *emergencies* departments
- Services, lines of treatment, cancer stage associations:
    - Number of patients per *hospital service during & between* lines (1-7) of treatment
    - Number of patients per *hospital service* per *comorbidity during & between* lines (1-7) of treatment
    - Number of patients per *cancer stage (I-IV) during & between* lines (1-7) of treatment
    - Number of patients during & between lines (1-7) of treatment vs number of patients using *medical services*
- Distribution of the number of emergency services used: by gender, cancer stage, smoking habit (current, former, no), age, number of comorbidities, distance from hospital, for:
    - The patient cohort
    - Time from diagnosis to first emergency
    - Number of patients that die at the emergency
- Drug-Drug interactions associations:
    - Average number of services during/between lines of treatment
        - with/ without Omeprazole
        - with Carboplatin and with/without Omeprazole
    - Survival probability in (0-120) months using combinations of Non-Oncological & Oncological Drugs, i.e. Omeprazole, Vinorelbine, Cisplatin

The associations that have been instrumental in the detection of patterns regarding the suspicion of cancer, usage of services, toxicities and drug usage, and toxicities and use of services. This has led to an initial development of 5 medical intervention in the Hospital Universitario Puerta de Hierro de Madrid.

- *Cardiology, pneumology & emergencies* are the most frequent services used from a year up to 3 months before the diagnosis of lung cancer.
- There is a significant difference in survival time for patients in stage III who take: Vinorelbine, Cisplatin, and Omeprazole (higher survival time) and those taking only Vinorelbine, Cisplatin, and no Omeprazole (lower survival time)
- Carboplatin is not effective as the other oncological treatments but is less effective when it is combined with Omeprazole
- The number of services visited by the patients depends by the line and type of treatment

Interactions between oncological and non-oncological drugs seem to impact the number of services used by the patients

2.4.2 Prediction Algorithms

2.4.2.1 Task

Prediction of (1) Long Term Survivors, (2) Correlation between Commodities and Toxicities, and (3) Prediction of Drug Interactions.

Diverse methods have been used to compute interactions between drugs; these methods and a visualisation of the overlap between the drug-drug interactions discovered by each method is presented in https://drug-interactions.herokuapp.com/. The drug-drug interactions discovered by the methods have been utilised to identify the number of patients that have been prescribed with at least a pair of drugs that interact - according to the corresponding method.

2.4.2.2 Data, Data Modelling

2.4.2.3 Features

Embeddings of patients, PCA, and SVM and Logistics regression

- Patient embeddings are computed for representing the patient features.
- PCA techniques are used over patient embeddings for feature selection
- K-Fold cross-validation is followed in order to classify the patients according to survival time; the relevant features are considered during this classification process.

Embeddings of patients, feature selection methods, correlation matrices

- Patient embeddings are computed for representing the patient features.
- Correlation matrices/coefficients to compute the scores of the features

Community detection algorithms and drug embeddings

- A network of drugs and their interactions is used to create a bipartite graph. Drugs are related based on values of similarity and communities are computed based on these values. Links are predicted between drugs in the same community.
- Drug embeddings are used to compute missing links.

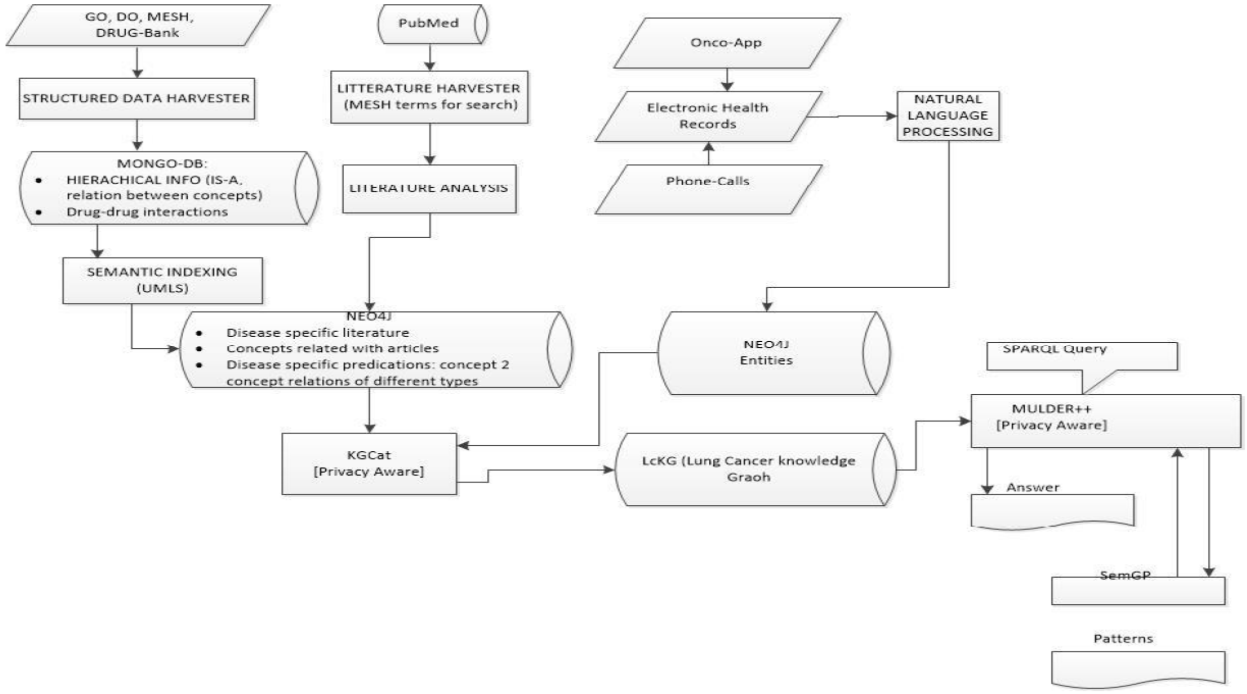
2.4.2.4 Model

The aim of the Lung cancer pilot is to improve the management of patients with cancer during their treatment, follow-up and during their last period of life through Big Data in order to improve not only their experience and satisfaction, and main outcomes, but also save substantial costs to the health budget. The pilot will try to address the said shortcomings, by adopting a pipeline that starts with medical data (open and patient records), performs pattern extraction and ends up in a knowledge graph that captures essential correlations in Lung-Cancer treatment.

Pattern discovery/Prediction, Machine Learning

1. Pattern discovery is performed on public data to extract interesting correlations between drugs, treatments and side effects.
2. Pattern discovery is performed on medical health records, logs of phone calls and on data furnished by a mobile application to detect important entities regarding the medical history of the patient (e.g., antecedents, diagnose, stage, performance status, treatment).
3. The patterns detected in stages 1 & 2 end up in a knowledge graph, on which pattern discovery is performed to detect toxicity, drug adverse events, and side effects. semGP is a graph partitioning method developed with the aim of identifying patterns in the knowledge graph; these patterns include clusters of patients that similarly react to lung cancer treatments, relations between drugs that allow for the explanation of drug adverse effects; and patterns between drugs and side effects that enable the discovery of potential new side effects and toxicity of a drug.

Next follows an overview of the Lung cancer pilot pipeline:


GO, DO, MESH, DRUG-Bank
STRUCTURED DATA HARVESTER
MONGO-DB:
HIERACHICAL INFO (IS-A, relation between concepts)
Drug-drug interactions
SEMANTIC INDEXING (UMLS)
PubMed
LITTERATURE HARVESTER (MESH terms for search)
LITERATURE ANALYSIS
NEO4J
Disease specific literature
Concepts related with articles
Disease specific predications: concept 2 concept relations of different types
KGCat [Privacy Aware]
Onco-App
Electronic Health Records
Phone-Calls
NATURAL LANGUAGE PROCESSING
NEO4J Entities
LcKG (Lung Cancer knowledge Graoh
SPARQL Query
MULDER++ [Privacy Aware]
Answer
SemGP
Patterns

MULDER++ is a query engine able to execute SPARQL queries over knowledge graphs accessible via SPARQL endpoints. MULDER++ implements the privacy-aware techniques defined by Endris et. al[2]. These techniques empower MULDER++ query planning strategies to generate query execution plans that respect query processing regulations. The novelty of this component is: (1) different granularities of data privacy and access control regulations, e.g. at the level of entities or attributes; (2) formal representation of the data privacy and access control regulations as privacy rules using a formal language, and (3) the generation of query plans to access data sources or the knowledge graph, that executes the privacy rules and enforces that data is collected, managed, and analysed respecting these policies. The privacy rules represent data access and management required to execute any query or knowledge management task; they include: Read (R), Merge (M), Storage (S), and Distribution (D), enabling the description of privacy characteristics, e.g. authority for processing the data. Thus, the privacy-aware techniques meet the requirements of Zeng et al[3]. and respect the DAC and RBAC models where data providers restrict data access to certain parties.

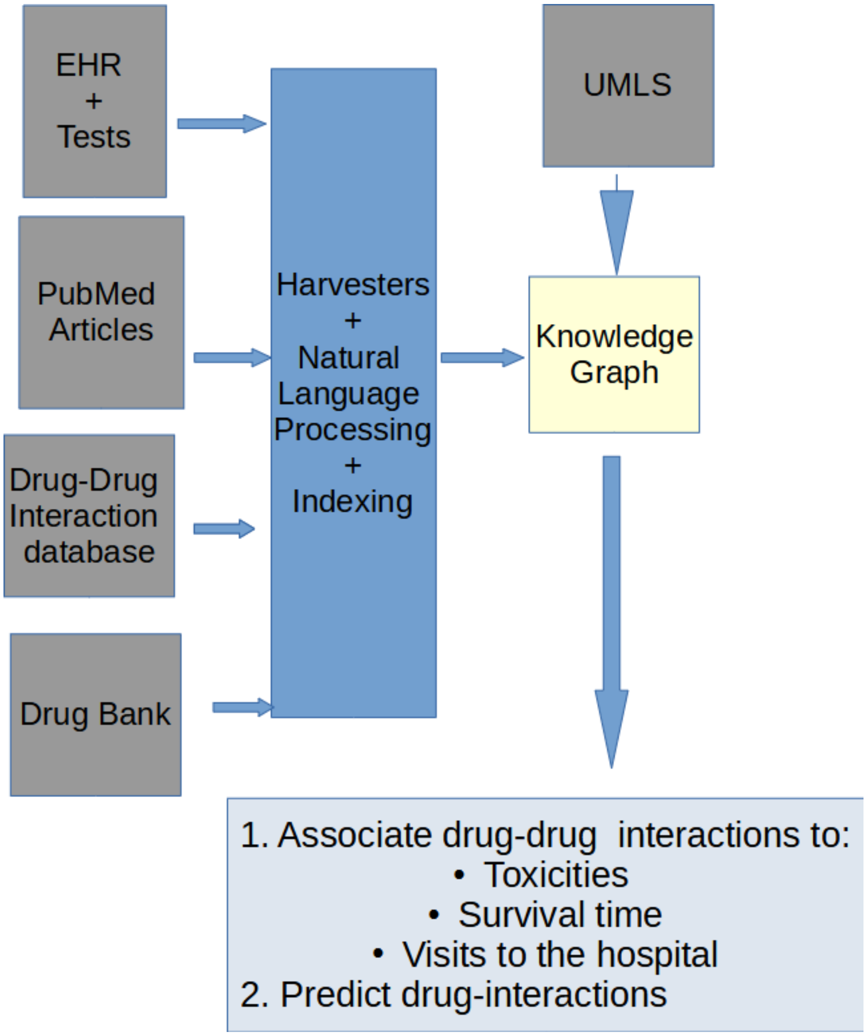
There are multiple pattern discovery/prediction algorithms applied on different places in the Lung cancer pilot:

[2] Kemele M. Endris, Zuhair Almhithawi, Ioanna Lytra, Maria-Esther Vidal, Sören Auer: BOUNCER: Privacy-Aware Query Processing over Federations of RDF Datasets. DEXA (1) 2018: 69-84

[3] Zeng, Q., Zhao, M., Liu, P., Yadav, P., Calo, S., & Lobo, J. (2015). Enforcement of Autonomous Authorizations in Collaborative Distributed Query Evaluation. IEEE Trans. Knowl. Data Eng

1. Pattern discovery on open source data:
    a. Metamap (https://metamap.nlm.nih.gov/) performs named entity recognition returning the named entities and a confidence.
    b. Semrep (https://semrep.nlm.nih.gov/) performs relation extraction. The result is stored in the knowledge graph.
2. Pattern discovery on health records and phone logs:
    a. C-LiKEs performs natural language processing to extract important entities and the result is stored in the knowledge graph.
3. Pattern discovery on the knowledge graph, where the data are stored as RDF triplets
    a. semGPs resort to community detection algorithms like semEP (https://github.com/gpalma/semep) and Metis (http://glaros.dtc.umn.edu/gkhome/metis/metis/overview), and semantic similarity measures, for partitioning the knowledge graph into subgraphs that represent meaningful patterns. The identified patterns are represented in the knowledge graph and correspond to actionable knowledge required for the support of precision medicine.

Next follows a diagram of the pattern discovery/prediction algorithms:


EHR
+
Tests
PubMed Articles
Drug-Drug Interaction database
Drug Bank
Harvesters
+
Natural Language Processing
+
Indexing
UMLS
Knowledge Graph
1. Associate drug-drug interactions to:
• Toxicities
• Survival time
• Visits to the hospital
2. Predict drug-interactions

2.4.2.5 Evaluation

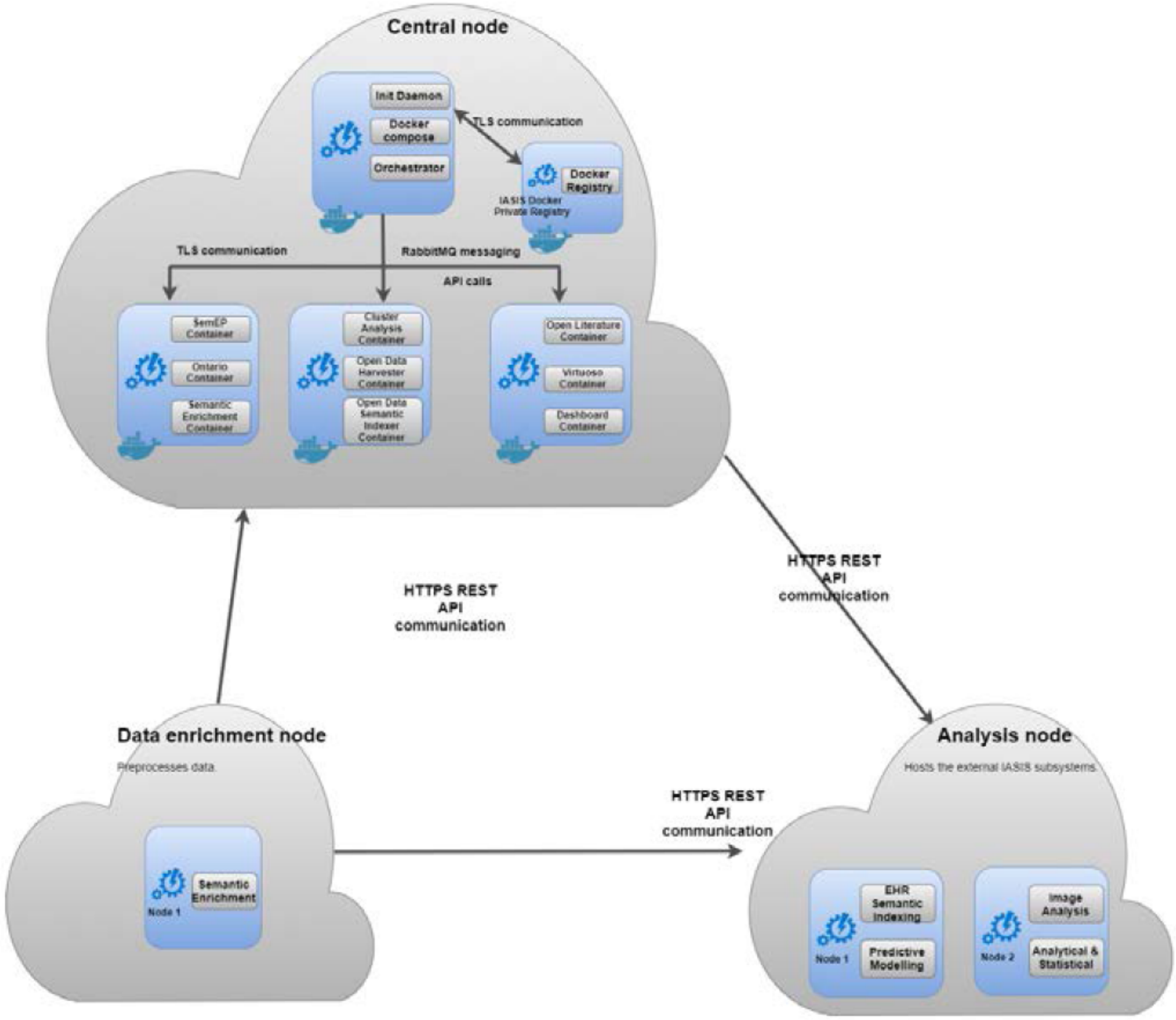
Description of Evaluation Setup (also mention if data is retrospective/prospective): The evaluation is carried out as retrospective study and evaluated within a 5 fold cross validation. In addition to that we explore how well a human doctor can solve the task of predicting one of the endpoints in the next 90 days. Then we compare this to our method and examine if a doctor can achieve better results including the prediction of our model.
Size of Data: Data includes more than 1400 patients with overall more than 100,000 data points, and then split into 70% training, 15% development and 15% test data.
Evaluation Method (e.g. ROC, F1): The model is evaluated using ROC and AUC-PC.
Results: Transplant Loss: ROC: 0.93, Rejection: 0.84, Infection: 0.80
Comparison to other related work (in terms of results): A comparison to other related work is difficult, as results might differ if cohort and data varies.

2.5 Security and privacy of data access and processing

The Lung Cancer pilot consists of three main physical nodes, each with a specific task.
The Analysis node is responsible to store the anonymized clinical data and make a first processing in order to reach to valuable conclusions and provide processed information to the other two nodes. The clinical data are provided by the hospital after a special permission was signed for usage in the context of the BigMedilytics project.

The Data enrichment node is responsible to retrieve the processed data and enhance them, using various sources, with semantic information. The semantically annotated data is transferred to the central node for usage from the node's components.

The central node is the node that facilitates most of the BigMedilytics components. It coordinates the off line and on line processes of the system, in order to enrich the semantically annotated data provided by the data enrichment node and to present the derived information to the users of the system.



2.5.1 Access Control

2.5.1.1 Authentication

Remote or physical access to the servers storing the data is provided only to authorised personnel. Regarding the application itself, there are three main roles, the clinicians, the experts and the administrators.
All user registrations are done from the administrator only after invitation of the new users and acceptance from their part. Valid email and password are required for the login in the application.

2.5.1.2 Authorization

All communication between the publicly available APIs of the platform and the components is done through secure protocols TLS/SSL. Additional to HTTPS, the different web services of the platform require also a token-based user and role authentication. This is accomplished with the help of the JWT framework which allows access to them only for the registered users of the system based on their role.

2.5.2 Data Protection

2.5.2.1 Data at rest

Users and roles are also defined on a database level, where different users have different privileges on the various database tables of the BigMedilytics database.
Remote or physical access to the servers storing the data is provided only to authorised personnel.

No data is saved on external devices

2.5.2.2 Data in transit

All data transfers are done through TLS/SSL and token based credentials.

Additionally at the network level, firewalls employ rules that allow the traffic flow only through specific ports and domains to the services and databases of the system.

2.5.3 Auditory and logs

2.5.3.1 System Auditory

There will be logs related only to the user evaluation and the measurement of KPIs

2.5.3.2 Services Auditory

Only standard system logs produced by the open source tools

2.5.4 Privacy measurements

2.5.4.1 Data Privacy Impact Assessment (DPIA)

2.5.4.2 Legal/Ethical process

2.5.4.3 Processes for complying with the current legislation

Different knowledge extraction methods are used to transform data represented in different formats into datasets annotated with biomedical ontologies. Clinical data is collected from the HUMPH; the data sets includes: a) Clinical notes of more than 1.000 lung cancer patients described in terms of 400 variables. Anonymisation techniques are performed by HUPHM in order to remove confidential attributes, e.g. national identifier, name, address, and profession. Moreover, these data can only be read and analysed in Spain by UPM partners and its use is regulated by the Spanish Law of Personal Data Access [4](Leyes 15/1999 and 41/2002). With the aim of extracting knowledge from the anonymised clinical records, the EHR text analysis and image analysis components are executed at the UPM servers. The results of these components correspond to records annotated with terms from medical vocabularies, e.g. UMLS. At UPM, new identifiers are generated to identify each record which logically represents a patient in terms of prescribed treatments, duration of the treatments, molecular biomarkers, evolution of the lung cancer tumor, and ECOG performance scores. These records can only be accessed by a Web REST Service and the permission to the execution of this service is only granted to LUH, the BigMedilytics (Lung Cancer Pilot) partner in charge of developing the techniques for creating the BigMedilytics knowledge graph.

Publications from PubMed are annotated with UMLS CUIs for the techniques developed by NCSR; also, relations among the extracted CUIs are recognised and described using the semantics relations of UMLS. The results of this extraction process are shared with LUH via secure access in an ownCloud server. LUH also performs tasks for recognising and linking relations and entities; these tasks are conducted over open data from scientific databases like DrugBank and SIDER. Different licenses regulate the access of these data sources, i.e. all the data management tasks are regulated by licenses, e.g. Attribution CC BY, Creative Commons, Creative Commons Attribution-NoDerivs License, CC BY-NC, Creative Commons Attribution 4.0 Unported License. They are respected during the data-driven process implemented in the BigMedilytics framework, i.e. data ingestion, knowledge graph creation and exploration, and knowledge discovery. The flow of information is as follows: LUH receives data from all the partners following diverse transfer methods, i.e. private clouds, secure FTP, and encoded Web REST APIs. The collected data is stored in a server of LUH that cannot be accessed from outside the internal network. After data is collected several processes are conducted with the

[4] https://www.boe.es/buscar/act.php?id=BOE-A-2002-22188

aim of creating the BigMedilytics knowledge graph, as well as supporting its exploration and knowledge discovery. Data access regulations and privacy-aware policies are described for classes, properties, and relations in the BigMedilytics knowledge graph in terms of classes and properties. These descriptions are used to guide all the data management processes and to ensure that data access regulations and privacy-aware policies are enforced whenever the BigMedilytics knowledge graph is accessed.

2.6 Trustworthy AI

Lawful: EHR generated at HUMPH in Spain, are only read and processed by UPM in Spain also. Access and management is according to the Spanish law. UPM provides only to LUH (in Germany) access to processed information from EHR through a Web REST service. The service uses authentication encoded in Base64 following the TLS protocol. Once the data are downloaded to LUH there is no public access to them.

Ethical: The patients that contribute their EHR for research reasons, sign a consent form. Also the hospital anonymizes the EHR by removing personal attributes, e.g. Id number, name, address, and profession

Robust: The layered design of the software contributes in the robustness, usability, performance, security, robustness, maintainability, availability, scalability & fault tolerance

2.6.1 technology/user adoption and establishing trust

2.6.2 ethical principles

- respect for human authority
- prevention of harm
- fairness
- explicability

2.6.3 key requirements

- Human agency and oversight
- Technical Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

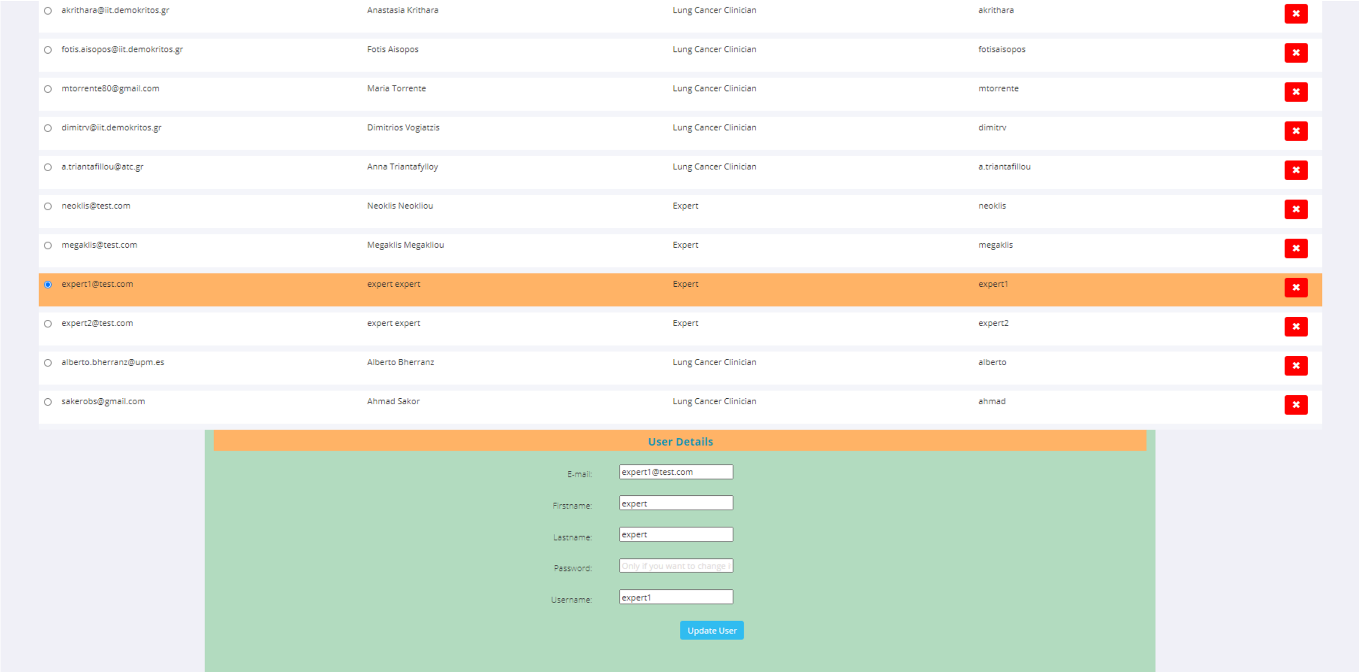
2.7 System-Interaction

2.7.1 Human-Machine Interface / GUI

As mentioned there three roles in the pilot application, administrator, clinician and expert.

The Experts can interact with clinicians through the chat interface of the application.
The Clinicians can search for descriptive analytics data, search the knowledge graph, ask in physical language the system and chat with experts.
The Administrators can manage the different users of the system.

Some screenshots depicting the aforementioned functionalities are provided below.

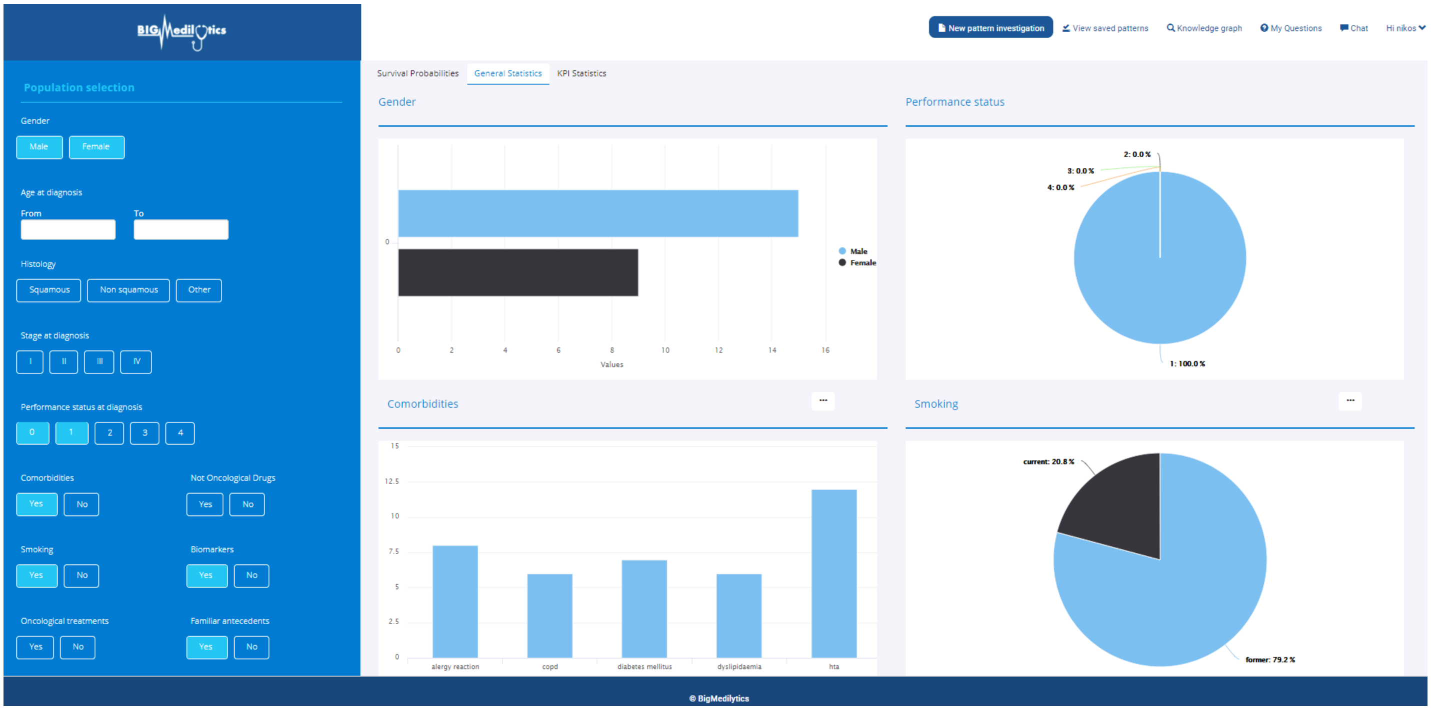


○ akrithara@iit.demokritos.gr	Anastasia Krithara	Lung Cancer Clinician	akrithara	✖
○ fotis.aisopos@iit.demokritos.gr	Fotis Aisopos	Lung Cancer Clinician	fotisaisopos	✖
○ mtorrente80@gmail.com	Maria Torrente	Lung Cancer Clinician	mtorrente	✖
○ dimitrv@iit.demokritos.gr	Dimitrios Vogiatzis	Lung Cancer Clinician	dimitrv	✖
○ a.triantafillou@atc.gr	Anna Triantafylloy	Lung Cancer Clinician	a.triantafillou	✖
○ neoklis@test.com	Neoklis Neokliou	Expert	neoklis	✖
○ megaklis@test.com	Megaklis Megakliou	Expert	megaklis	✖
● expert1@test.com	expert expert	Expert	expert1	✖
○ expert2@test.com	expert expert	Expert	expert2	✖
○ alberto.bherranz@upm.es	Alberto Bherranz	Lung Cancer Clinician	alberto	✖
○ sakerobs@gmail.com	Ahmad Sakor	Lung Cancer Clinician	ahmad	✖

User Details

E-mail: expert1@test.com
Firstname: expert
Lastname: expert
Password: Only if you want to change i
Username: expert1

Administrator - User management



Population selection

Gender: [✓ Male] [✓ Female]

Age at diagnosis: From ____ To ____

Histology: Squamous / Non squamous / Other

Stage at diagnosis: I / II / III / IV

Performance status at diagnosis: [✓ 0] [✓ 1] 2 3 4

Comorbidities: [✓ Yes] No

Not Oncological Drugs: Yes No

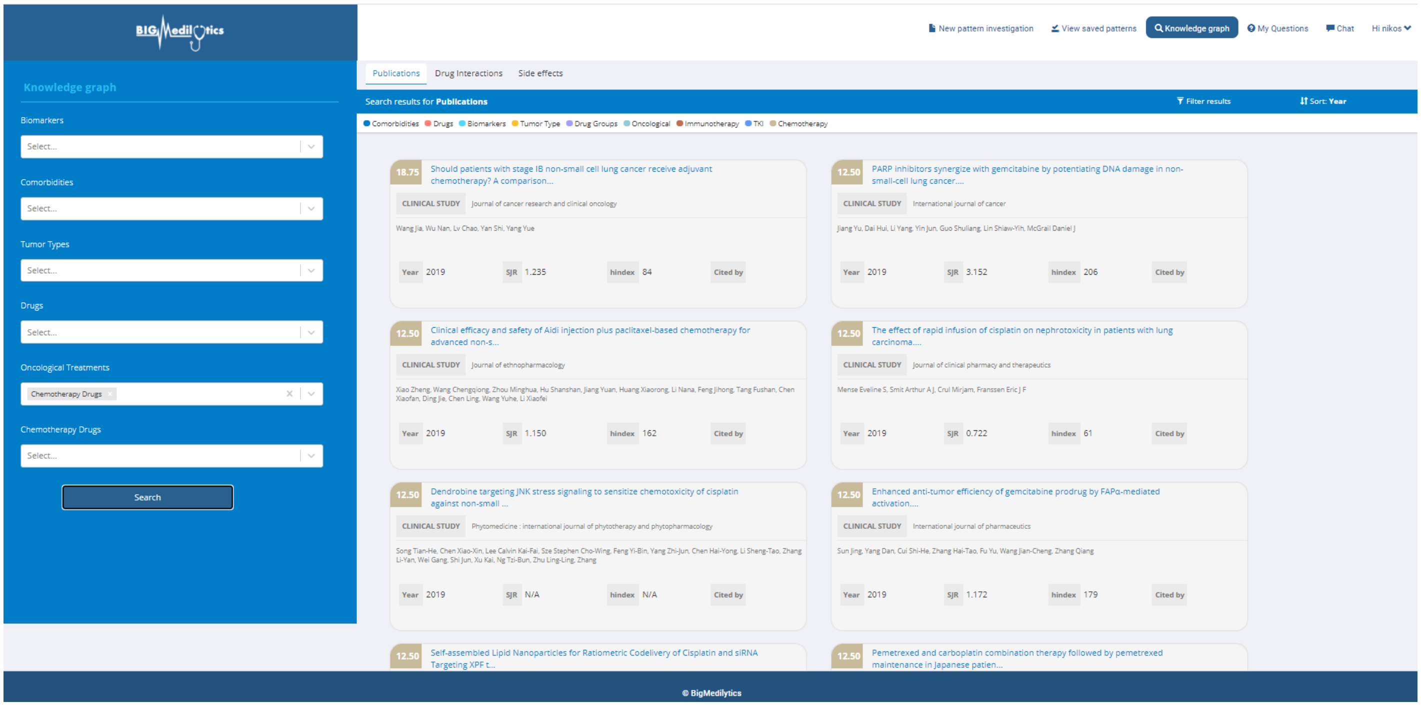
Smoking: [✓ Yes] No

Biomarkers: [✓ Yes] No

Oncological treatments: Yes No

Familiar antecedents: [✓ Yes] No

Clinician - Search based on population



Knowledge graph

Biomarkers: Select...

Comorbidities: Select...

Tumor Types: Select...

Drugs: Select...

Oncological Treatments: Chemotherapy Drugs

Chemotherapy Drugs: Select...

Search results for **Publications**

● Comorbidities ● Drugs ● Biomarkers ● Tumor Type ● Drug Groups ● Oncological ● Immunotherapy ● TKI ● Chemotherapy

18.75 Should patients with stage IB non-small cell lung cancer receive adjuvant chemotherapy? A comparison...
CLINICAL STUDY Journal of cancer research and clinical oncology
Wang Jia, Wu Nan, Lv Chao, Yan Shi, Yang Yue
Year 2019 SJR 1.235 hindex 84 Cited by

12.50 PARP inhibitors synergize with gemcitabine by potentiating DNA damage in non-small-cell lung cancer....
CLINICAL STUDY International journal of cancer
Jiang Yu, Dai Hui, Li Yang, Yin Jun, Guo Shuliang, Lin Shiaw-Yih, McGrail Daniel J
Year 2019 SJR 3.152 hindex 206 Cited by

12.50 Clinical efficacy and safety of Aidi injection plus paclitaxel-based chemotherapy for advanced non-s...
CLINICAL STUDY Journal of ethnopharmacology
Xiao Zheng, Wang Chengqiong, Zhou Minghua, Hu Shanshan, Jiang Yuan, Huang Xiaorong, Li Nana, Feng Jihong, Tang Fushan, Chen Xiaofan, Ding Jie, Chen Ling, Wang Yuhe, Li Xiaofei
Year 2019 SJR 1.150 hindex 162 Cited by

12.50 The effect of rapid infusion of cisplatin on nephrotoxicity in patients with lung carcinoma....
CLINICAL STUDY Journal of clinical pharmacy and therapeutics
Mense Eveline S, Smit Arthur A J, Crul Mirjam, Franssen Eric J F
Year 2019 SJR 0.722 hindex 61 Cited by

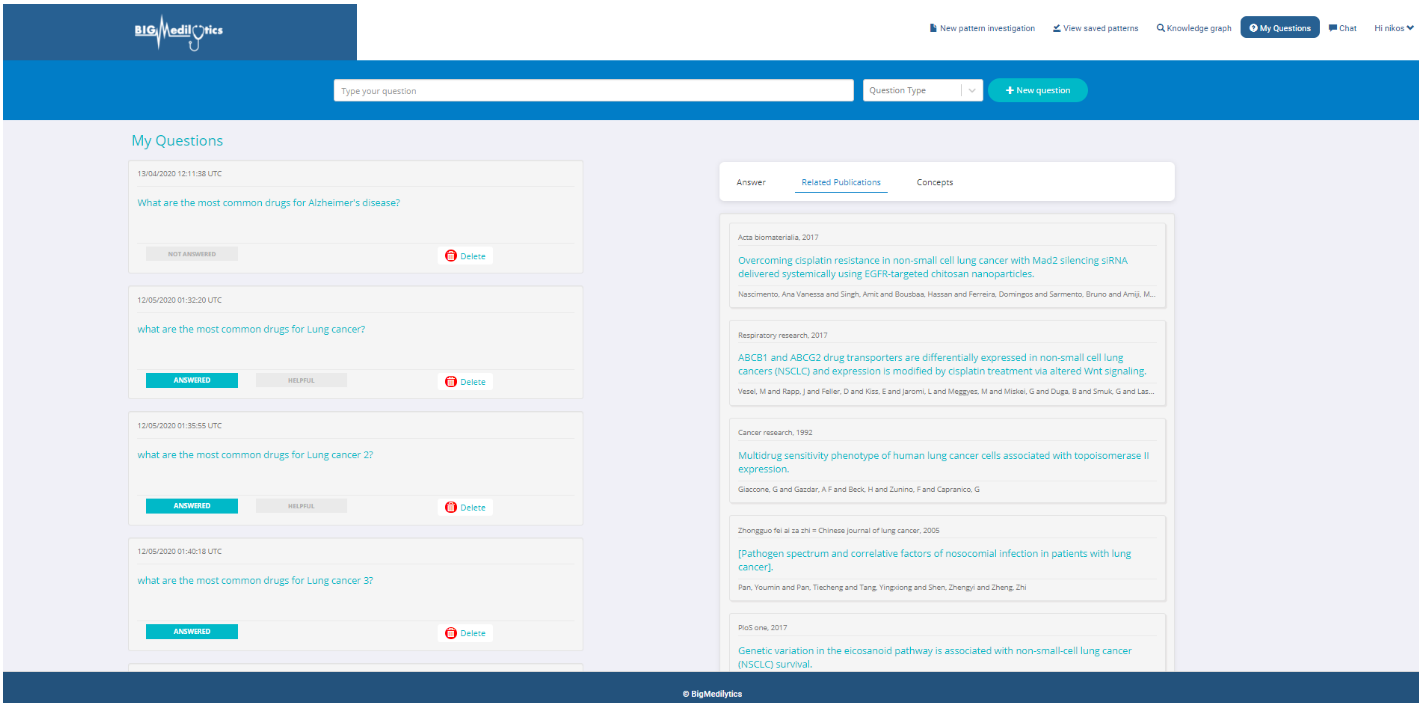
12.50 Dendrobine targeting JNK stress signaling to sensitize chemotoxicity of cisplatin against non-small ...
CLINICAL STUDY Phytomedicine : international journal of phytotherapy and phytopharmacology
Song Tian-He, Chen Xiao-Xin, Lee Calvin Kai-Fai, Sze Stephen Cho-Wing, Feng Yi-Bin, Yang Zhi-Jun, Chen Hai-Yong, Li Sheng-Tao, Zhang Li-Yan, Wei Gang, Shi Jun, Xu Kai, Ng Tzi-Bun, Zhu Ling-Ling, Zhang
Year 2019 SJR N/A hindex N/A Cited by

12.50 Enhanced anti-tumor efficiency of gemcitabine prodrug by FAPα-mediated activation....
CLINICAL STUDY International journal of pharmaceutics
Sun Jing, Yang Dan, Cui Shi-He, Zhang Hai-Tao, Fu Yu, Wang Jian-Cheng, Zhang Qiang
Year 2019 SJR 1.172 hindex 179 Cited by

12.50 Self-assembled Lipid Nanoparticles for Ratiometric Codelivery of Cisplatin and siRNA Targeting XPF t...

12.50 Pemetrexed and carboplatin combination therapy followed by pemetrexed maintenance in Japanese patien...

Clinician - Search in knowledge graph



My Questions

13/04/2020 12:11:38 UTC
What are the most common drugs for Alzheimer's disease?
NOT ANSWERED — Delete

12/05/2020 01:32:20 UTC
what are the most common drugs for Lung cancer?
ANSWERED — HELPFUL — Delete

12/05/2020 01:35:55 UTC
what are the most common drugs for Lung cancer 2?
ANSWERED — HELPFUL — Delete

12/05/2020 01:40:18 UTC
what are the most common drugs for Lung cancer 3?
ANSWERED — Delete

Acta biomaterialia, 2017
Overcoming cisplatin resistance in non-small cell lung cancer with Mad2 silencing siRNA delivered systemically using EGFR-targeted chitosan nanoparticles.
Nascimento, Ana Vanessa and Singh, Amit and Bousbaa, Hassan and Ferreira, Domingos and Sarmento, Bruno and Amiji, M...

Respiratory research, 2017
ABCB1 and ABCG2 drug transporters are differentially expressed in non-small cell lung cancers (NSCLC) and expression is modified by cisplatin treatment via altered Wnt signaling.
Vesel, M and Rapp, J and Feller, D and Kiss, E and Jaromi, L and Meggyes, M and Miskei, G and Duga, B and Smuk, G and Las...

Cancer research, 1992
Multidrug sensitivity phenotype of human lung cancer cells associated with topoisomerase II expression.
Giaccone, G and Gazdar, A F and Beck, H and Zunino, F and Capranico, G

Zhongguo fei ai za zhi = Chinese journal of lung cancer, 2005
[Pathogen spectrum and correlative factors of nosocomial infection in patients with lung cancer].
Pan, Youmin and Pan, Tiecheng and Tang, Yingxiong and Shen, Zhengyi and Zheng, Zhi

PloS one, 2017
Genetic variation in the eicosanoid pathway is associated with non-small-cell lung cancer (NSCLC) survival.

Clinician - Querying / Answering

2.7.2 Education

The user interface of pilot was designed by ATC in cooperation with the HUMPH hospital. Therefore, the oncologists are intimately knowledgeable of the user interface functionality.

3. Learnings

3.1 Challenges & Barriers

- Architecture

The software that implements the Lung Cancer platform should cover the following aspects: Usability, Performance, Security, Robustness, Maintainability, Availability, Scalability & Fault Tolerance. We have addressed all of the above in order to create a robust and fully flexible system with high response times and with all-time availability for its users. In addition, the platform secures all the interactions and the data that its components exchange, it is easy to maintain and covers all the needs of the project.

- Processing of large structured / unstructured data sources

The main data sources are open source publications and electronic health records (EHR) of patients. Both are unstructured and to be further analyzed, we need to extract some structure. The format of the EHR was proprietary to the hospital involved, and in particular they contained structured and unstructured parts in Spanish. The open data sources (typically from PubMed) were in English. The first major challenge is the data selection, and data quality. This challenge is manifested in the EHR as selection of the patients that have performed all the medical tests that are relevant, have no missing values, and are relatively recent because of the focus on immunotherapy which is a newer approach. Then a combination of Natural Language Processing techniques were used to extract structured information.

Regarding the open data sources, industry standard tools were used to extract structured information. The result is a graph of diseases, treatments, patients and relations, but a big issue is the quality of the resulting graph.

- Natural Language Processing

The natural language processing should be adapted to the native language of the hospital, and to the form of the EHR.

- Prediction Algorithms

- Security and privacy of data access and processing

- Trustworthy AI

The informatio

- System-Interaction

3.2 Lessons Learned

- Architecture & System Integration

The high-level software architecture of the Lung-Cancer prototype follows a layered architecture pattern and is consistent with modern distributed systems' architectural principles. Thus the Lung-Cancer platform follows a layered architecture pattern, thus the system's modules are divided into multiple distinct layers, each one responsible for a specific set of functionalities. By convention, the layers interact with each other in a top-down manner, with each layer being able to access all layers below. A lower layer should never interact with layers above. This convention helps to avoid circular dependencies between layers. There are also two vertical cross-cutting layers - the *Access Control* layer and the *Support* layer - that affects all other layers of the Pilot..

- Processing of large structured / unstructured data sources

The data quality in the open graph has been handled in two ways. The first way is a scoring function of the data which reflects the impact of the journal from which the information is extracted, as well as the recency of the publication.   The second way is through graph theoretic measures that aim to remove edges that are essentially noise.

- Integrations of structured and unstructured information in a knowledge graph

Data extracted from diverse biomedical data sources is integrated and represented into a knowledge graph. Knowledge graphs comprise entities and their relations in a structured fashion. Ontologies and controlled vocabularies are utilised to describe the meaning of relations, as well as for annotating entities in a uniform way in the knowledge graph. The Lung Caner data sources provide relevant data for the knowledge graph; however, each source presents the data in different formats, e.g., (un)structured, or may suffer from quality issues, e.g. missing or duplicate values. These issues impact on the tasks of knowledge graph creation, and have to be addressed in order to avoid that the same issues are transferred to the knowledge graph. In this document, the techniques followed for pre-processing, curating, and integrating the data are presented. First, the data sources are described, as well as the pipelines followed to pre-process the data to reduce redundancy, ambiguities, and entity mismatching. Next, the unified schema that describes the integrated data in the knowledge graph is defined; it includes classes and predicates that enable the description of biomedical concepts in the knowledge graph. The Resource Description Framework (RDF), RDF Schema, and OWL are utilised as the data models to express the Lung Cancer unified schema. Additionally, the pipeline followed to transform the input data into the knowledge graph is depicted in this document, as well as the integration rules used in this pipeline.

- Natural Language Processing

- Prediction Algorithms

Decision trees utilise simple decision rules, learned from the data features, predict the probable outcomes. They are easy to understand and interpret, allowing for explainable AI and visualization. They also require little data preparation.

- Security and privacy of data access and processing

Clinical data collected from HUMPH, consent forms signed by the patients;

HUMPH anonymizes to remove personal attributes, e.g. national identifier, name, address, and profession. These data are only read and analysed in Spain by the (UPM); access & management regulated by the Spanish Law

EHR text analysis components are executed at the UPM servers to extract knowledge from the anonymised clinical records. The results of these components correspond to records annotated with terms from medical vocabularies, e.g. UMLS. At UPM, new identifiers are generated to identify each record, which logically represents a patient in terms of prescribed treatments, duration of the treatments, services, mutations, evolution of the lung cancer tumor, and ECOG performance scores.

These records can only be accessed via a Web REST Service. The permission to execute this service is only granted to LUH The access of the Web REST Service requires the authentication of a username and password, which are encoded in Base64 following the TLS protocol, i.e., the communication established to download the data is encrypted, respecting a security protocol. Data downloaded using the API is stored in a server at LUH; this server is not accessible via the Internet.Once the data is collected at LUH, the following steps are conducted:

1. Preprocessing: The collected data is processed to extract aggregated values, e.g. lists representing the evolution of patient conditions, tumours, and mutations, toxicities, and services; the task of entity recognition and linking is also conducted over textual attributes like the description of the hospital services conducted to the patients
2. Secure Storing: The results of the data extraction and processing process are stored and encrypted in the LUH servers; these servers can only be accessed in the Intranet of the Scientific Data Management group at LUH
3. Data Integration and Semantic Enrichment: aggregated data used by the Semantic Enrichment process to populate the BigMedilytics LC knowledge graph. Mapping rules areutilised for guiding this process and describe in a declarative manner the whole process of integration performed to create the knowledge graph. In the knowledge graph, the aggregated data is only integrated with public and open data that includes drugs, mutations, publications, or side effects. No further attributes are included or integrated into the KG
4. KG: can be accessed via the APIs; only data collected from open data sources are available through these APIs

- Trustworthy AI

Components of Trustworthy AI:

<u>Lawful:</u> EHR generated at HUMPH in Spain, are only read and processed by UPM in Spain also. Access and management is according to the Spanish law. UPM provides only to LUH (in Germany) access to processed information from EHR through a Web REST service. The service uses authentication encoded in Base64 following the TLS protocol.Once the data are downloaded to LUH there is no public access to them.

<u>Ethical:</u> The patients that contribute their EHR for research reasons, sign a consent form. Also the hospital anonymizes the EHR by removing personal attributes, e.g. Id number, name, address, and profession

<u>Robust:</u> The layered design of the software contributes in the robustness, usability, performance, security, robustness, maintainability, availability, scalability & fault tolerance

Requirement

Human agency and oversight: The pilot provides information about length of stay in the hospital, medical exams undertaken before being diagnosed with lung cancer, toxicities because of drug interactions and related bibliographic references. This information allows oncologists to plan medical interventions.

Technical robustness & safety: The pilot is used periodically for longer term planning of management of patients. Thus a temporary deactivation of the pilot is not critical. Besides the software architecture is such to provide a robust system.

Privacy and data governance: No access to individual patient records is provided by the pilot. The pilot provides access only to aggregated patient information.

Transparency: The system does not provide suggestions to the oncologist, instead the oncologist will receive information about different parameters that affect the well being of the patients and then makes decisions.

Diversity: The data that have been analyzed concern Spanish patients.

Societal and environmental well-being: The pilot aims to reduce the cost of health care for lung-cancer patients and will at the same time improve their well being.

Accountability: The way to approach accountability is by tracing the process of data analytics from raw data to the final information that is acquired through the dashboard. Thus, ultimately the results of the pilot rely on information acquired from EHR and from publications. Naturally, the individual EHR are not readily available for security reasons, but if need be they can be acquired.

- System-Interaction

3.3 Main (quantifiable) achievements

- A Lung-Cancer Pilot with a Dashboard, which is accessible through any web browser
- The dashboard allows to filter through various aspects of the cancer stage, treatments line existing comorbidities, etc. and obtain survival curves. Then the focus shifts on information about Hospital stay, Identification of people at risk of developing lung cancer, and on Toxicities of drugs. Furthermore, the pilot can provide related references to PubMed ordered by a scoring function
- The participating hospital has conducted a series of interventions based on the findings of the pilot.

4. Output

4.1 Papers

4.2 Open Source & Resources (refer to ELG)

4.3 Demos