Pilot 1: Comorbidities

1. Key Information

1.1 Involved Partners

- INCLIVA
- Philips
- Optimedics
- ITI
- ATOS
- Technische Universiteit Eindhoven

1.2 Involved Countries

- Spain, Netherlands, Germany

1.3 Keywords

- Risk Prediction
- Comorbidities aggrupation
- Prediction of hospitalization & mortality

1.4 Task Description

This pilot primarily addresses the long-term treatment of chronic disease patients and aims to develop a risk prediction model to reduce costs by directing patients to primary or secondary care where emergency care and hospitalization are not required. Using a Big Data approach, the disease trajectories and care pathways of a large patient population are characterized over an extended time period. Thus, this task has the potential to unravel the pattern of a disease as well as previously unknown links among disease groups.

2. Building Blocks

2.1 Architecture

2.1.1 System Architecture

The computing cluster has been deployed in a secure computational infrastructure within the INCLIVA's facilities. Every remote access to this cluster from the pilot's partners is monitored

and encrypted, requiring a two-step authentication and providing only a graphical frontend that is configured to avoid the possibility of extracting data from it. Through this frontend, partners will be allowed to interact with the cluster in order to perform the different proposed analysis.

Figure 1 summarizes the architectural layers and the specific technologies deployed in the INCLIVA's infrastructure in order to support the Comorbidities pilot. The technological stack is mainly inspired by the Cloudera distribution for Hadoop, a widely popular stack in the Big Data Analytics domain. This stack enforces distributed data processing, in order to maximize the use of computing resources and, consequently, improving the accuracy of the resultant models. Our architecture is made up of six layers according to the provided functionality. Next, we explain the involved technologies in each layer:

- **Persistence**: Original data from the EHR repository, specifically a PostgreSQL database, is imported into the deployed infrastructure after performing an ETL process. The resulting CSV files with relevant information from health records, are distributed into several persistence nodes using HDFS. This step guarantees the distributed processing of the data and, indirectly, reduces the amount of information to be analysed.
- **Resource Manager:** Yarn supports the management of the Hadoop infrastructure services: HDFS file system, computing resources and data storage. Currently the infrastructure is made up of a master node and three additional processing nodes
- **Data Access**: Two access modes to the Hadoop File System (HDFS) data store are available in our current architecture. First, Hive enables to define SQL queries directly over the HDFS data and to perform exploratory analyses, such as aggregated queries or the computation of statistical indicators. Secondly, Apache Spark enables in-memory processing of the data using both SQL (SparkQL) or a map-reduce paradigm. Using these technologies, the data scientist transforms the data into the most suitable structure for further analysis or performs feature-engineering tasks.
- **Analytics**: This layer introduces two frameworks to perform the analytical tasks over the data, mostly to create analytical models. Firstly, Scikit-learn provides a wide array of algorithms, specifically for clustering and classification purposes. Secondly, Spark-ML provides a similar role with a smaller subset of algorithms available. However, Spark-ML algorithms are implemented in a distributed way and could benefit from the in-memory capabilities provided by the deployed Spark cluster. Since using Spark-ML provides faster training times, this will be the preferred choice when possible.
- **User Interface**: Due to security limitations, EHR data cannot be exported outside the INCLIVA's infrastructure. Consequently, the role of the user interface layer is to provide a controlled interaction environment with the data and the infrastructure. The architecture offers two web-based interfaces for achieving this goal. On the one hand, Hue provides monitoring capabilities over the infrastructure and provides a SQL-like editor to execute queries and create subsets of the data. On the other hand, Jupyter provides a coding editor to create analytical scripts (notebooks) using the aforementioned frameworks and technologies.
- **Models**: The result of the analytics process is a model for classifying patients according to a specific set of features. These models are accessed using a REST API to send data

and receive a risk classification. This API enables the use of the models by external applications. In the context of this pilot, models will be consumed by a mobile application deployed in primary care. Additionally, models will be used for developing graphical charts using libraries such as Matplotlib or Bokeh.

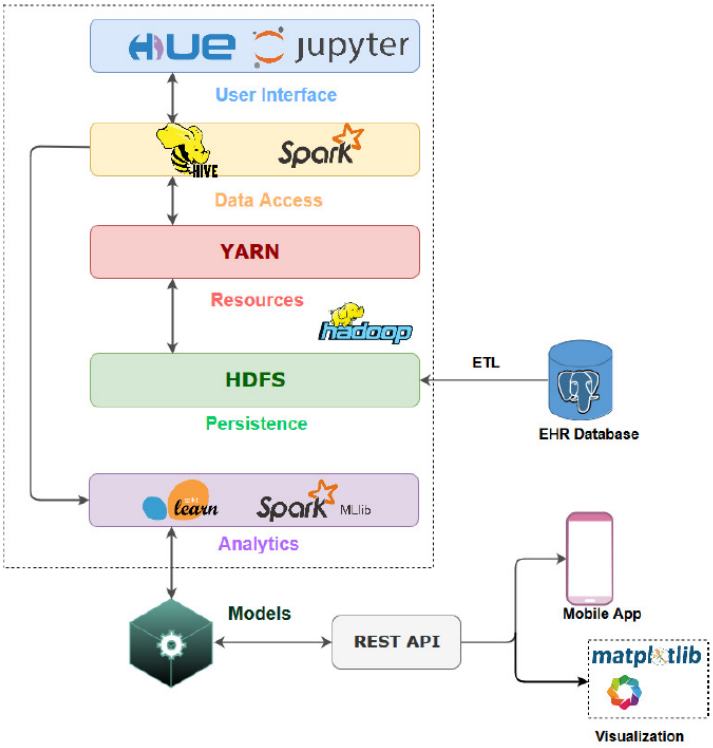

Figure 1 Architectural layers and the specific technologies deployed in the INCLIVA's infrastructure

2.1.2 Data Flow & Interoperability of services

The data flow following this architecture (Figure 1) could be summarized as follows:

- First, the original data is extracted and processed to generate a set of files suitable for analysis. This set of files is stored into the HDFS file system.
- Next, the data scientist executes the required queries, with Hive or SparkQL to generate a dataset with the relevant features.
- These datasets are executed in combination with a script to train a classification model. If Sparks supports the required algorithm, this script is launched in the distributed environment.
- Finally, models are stored in an external server to be accessed through the rest API.

2.1.3 Necessary Hardware

The hardware which will be necessary for the implementation of the project requires a distributed Cloudera cluster.
This cluster has been created by using 2 servers with Ubuntu Server 18.04; CPU: Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz with 12 cores; RAM: 1536 GB RAM DDR2 ECC (24 modules, 64 GB, 2666 MHz) and HDD: 5.24TB SSD RAID5 & 32.74TB HDD SATA RAID5.

2.1.4 Software Components

Please see 2.1.1 description and figure 1.

2.3 Data Processing

2.3.1 Processing of large structured / unstructured data sources

2.3.1.1 Data Sources

The following table describes the anonymized data used for comorbidities' estimation risk, the electronic health record (EHR) from period 2012-2016 and individuals from the Region of Valencia.

Data Source	**Description**	**Acquisition**	**Characteristic (Size, Patients, Years, Origin/Region)**
EHR	Data was requested to the Health Care Authorities (Generalitat Valenciana) after obtaining Ethical Committee approval and classification by the Spanish Drugs and Medical Devices Agency (AEMPS) (Minister of Health). They provide a large set of data files which contains all the information collected in the EHR (Primary Care, Secondary Care, Treatments, Prescriptions, Hospital Admissions, Procedures, Mortality, Demographics) of the period from 2012 to 2016 (approximately 100 data files separated by year and source). After a process of pseudo-anonymization where identifiers such as patient's name, number of Health Care Card or treatments identification number were codified, avoiding any possibility to go back to the original data for the INCLIVA team, data is delivered to INCLIVA by using encrypted storage devices.	Data gathered from the Health Care Authorities (Generalitat Valenciana)	> 4 million subjects from 2012-2016, Valencia region

	Once delivered, files were stored on a specific server inside the INCLIVA's Data Centre where isolation from other potential networks and other security measurements were taken to provide a high secure environment. This server will act as a data storage server and security back-ups to ensure the availability of the data in case of unexpected errors. It should be pointed out that the partners of this pilot considered the European Regulation 2016/679.		

Pilot	**Multiple sources**	**Integration to data warehouse**	**Data access**	**Data stored in cloud**	**Multi-party architecture**	**Secure environment**	**Transform raw / unstructured data**
1	yes	no	Processed on secure infrastructure/secure server at INCLIVA and accessed remotely by other pilot partners and processed in virtual machine	no	no	yes	no

2.3.1.2 De-Identification and anonymisation

First a pseudo-anonymization was carried out. Identifiers such as patient's name, number of Health Care Card or treatment identification number were codified, avoiding any possibility to go back to the original data for the INCLIVA team. Data is delivered to INCLIVA by using encrypted storage devices.

A second de-identification process has been performed to increase the security and privacy of the data and it does not affect the final risk prediction. This process is an iterative three-phases process (see figure 2):

1. In the first phase, the variables were evaluated to define which ones needed to be anonymised and how.
2. In the second phase, the different anonymisation techniques are applied to complete the process.

3. The third phase consists of a risk assessment to evaluate if the anonymisation previously done satisfies a chosen security threshold, based on the very low probability of an individual to be identified. If this threshold is not reached, then we will go back to the first phase in order to improve the anonymisation until an acceptable risk level is obtained.

Mainly, three different techniques are used to anonymise the chosen variables, following the process proposed by the Spanish Data Protection Agency. However, not all of them are going to be applied to each variable. These techniques are briefly described in the following lines.

- **Elimination/Reduction of variables**: Removing non-useful information that has a direct impact on reducing the re-identification risk. For this reason, all the non-useful variables to achieve the goals of the project are removed from the datasets.
- **Disturbance of data**: Depending on the type of the variables, different data disturbance techniques are applied. These techniques are based on increasing cardinality inside analyzed groups with the objective of decreasing the identification probability of an individual inside a specific group (e.g. the micro-aggregation of numeric data such as clinical measurements or the decrement of detail in data by putting it into ranges).
- **Encryption of data**: In case of key identifiers a combination of hashing and encryption algorithms are used.

These algorithms applied on a single data generates a unique key that can be used to represent this data. These hash functions are unidirectional so there should not be a way back to the original data from the output hash code, but since the mere application of a hashing algorithm does not guarantee the irreversibility of the process, it is used in conjunction with an encryption algorithm.

An encryption algorithm encrypts the information using a key that can also be used to decrypt it. As there is no interest in reverting this process, data is encrypted by using a randomly generated key that is then destroyed. Encryption algorithms are easily combined with hashing algorithms and together, they provide an optimum level of security.

Figure 2 shows the different paths that the data/variables could follow in the anonymisation techniques.

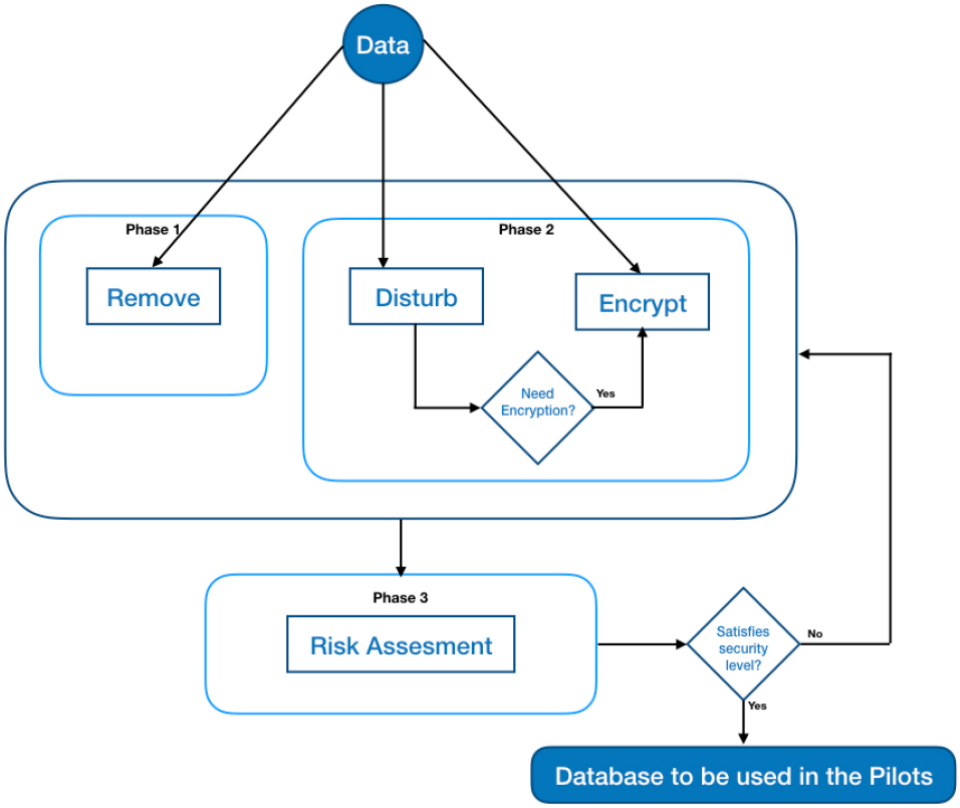

Figure 2 De-identification and anonymisation process

2.3.1.3 Acquisition

The pilot analyses EHRs from the Valencian Region Community. These data will be provided in bulk from the current databases per year (initially, 2012). In order to access this data, INCLIVA fulfils a data requirement with the set of required variables and the purpose of the project. This data requirement is reviewed by the Hospital Ethics Committee which is highly skilled and used to these matters and the data exportation process is granted. Then, data is securely transferred to the INCLIVA premises which are not accessible from outside. Personal IDs are anonymized to assure privacy and confidentiality. Provided data is mainly structured in tabular format and no natural language is included in the current data set. All this data is uploaded into a PostgreSQL database. For each patient the following set of information is provided:

- Personal information (age, gender, comorbidities and some general clinical features).
- Family background
- Diagnosis
- Primary and Secondary care visits
- Hospital discharges (yes or no, and date).
- Hospital urgencies (yes or no, date)
- Health habits (smoker, non-smoker, ex-smoker)

EHRs are stored in the Valencia Region Health Systems using relational databases. They provide us with a dump of the relevant tables of such systems. This data then was uploaded to a PostgreSQL database.

2.3.1.4 Cleansing

The following data procedures have been carried out on the data:

- ICD codes are translated to the version 10 due to data is encoded with the ICD 9 version.
- An additional anonymization process is applied over the patient ids. This process guarantees that partners analyzing the data have not access to original ids.
- Sensitive information to identify a patient is translated to a specific group. For instance, age is provided in ranges (25+, 40+, 55+, 70+, 85+) and small towns are aggregated into wider geographic areas. The greater the aggregated data, the more general the risk prediction will be.

Prescription data is analyzed and a diagnosis for such treatment is associated with the patient. This transformation helps further analysis. We run specific scripts using R and Python to perform the aforementioned transformations.

2.3.1.5 Data Integration

Each partner defines its own data integration procedure from the original data sources to carry on its specific analysis. Overall, each partner queries the original database to generate several datasets with the relevant information for analysis. For instance, ITI stores such datasets in a distributed Hive database, to improve query performance, and then upload it in memory to a Spark cluster. In this memory cluster, specific ML algorithms for clusterization will be executed. The output of the analysis will be stored in the INCLIVA premises. Selected technologies are Hadoop HDFS for integrating and storing the cleaned data and Apache Spark for analysis. Specifically, we use Spark data structures, as data frames, to prepare the data for further analysis.

2.3.1 Multi-velocity processing of heterogeneous data streams

Does not apply

2.3.5 Complex real-time event detection

Does not apply.

2.3.5.1 Notifications

2.3.5.2 Situations of Interest

2.3.5.3 Event Processing

2.3.5.4 Event Sources

2.3.5.5 Evaluation

2.4 AI Components

2.4.1 Deep learning for multilingual NLP and image analytics

2.4.1.1 Natural Language Processing

Does not apply.

2.4.1.1.1 Evaluation

Does not apply.

2.4.1.2 Image Processing

Does not apply.

2.4.1.2.1 Evaluation

Does not apply.

2.4.2 Prediction Algorithms

2.4.2.1 Task

1. Comorbidities aggrupation

This pilot is implemented in two phases, and aims to obtain the relevant comorbidity clusters to improve personalized medical care. The first one aims to discover comorbidity groups that produce an increment in exitus (death) and hospitalization (KPIs) in the population of such groups, with respect to the general population groups classified by sex and age ranges (25+, 40+, 55+, 70+, 85+). On the other hand, the second phase studies with more detail the most relevant comorbidity population groups previously detected. This second phase analyses several clinical features and provides KPI predictions for a specific patient and, then, compares it with the population of his/her comorbidity group.
The discovery of comorbidity groups, during the first phase, uses statistical inference as a technique. Given a comorbidity group of two (or more) diseases, KPIs are calculated if an

additional comorbidity disease is included. The overall approach is to create a novel comorbidity group only when a statistical significance is detected, in at least one of the KPIs.

2. Missing data imputation.

A model has been developed to perform the missing data imputation in a multidimensional way.

3. Feature selection

A logistic regression has been applied to filter the most relevant clinical features for each comorbidity group.

4. Hospitalization & Mortality risk prediction model

The initially discovered comorbidity groups are utilised to create specific models for each one. These models use, as additional information, a set of the most relevant clinical features for each KPI, according to the diseases involved in the group, as explained in 2.
These models, also offer us predictions about the KPI risk per patient considering using its clinical features as an explanation.
The development of the second phase uses a model-training approach for each comorbidity group. First, a model has been developed to perform the missing data imputation in a multidimensional way. Then, a logistic regression has been applied to filter the most relevant clinical features for each comorbidity group. Lastly, a decision tree has been used to obtain the KPIs prediction for a patient and the relevance weight of each feature to obtain the final prediction.
As the created decision tree model is specific for each comorbidity group, it could be useful for the physicians to check or review the most relevant clinical features. On the other hand, as interpretable models have been used to obtain the specific KPI prediction for a patient, it is possible to know the specific state of a patient in relation with the group. This fact could help to know what actions or analysis are required to improve a KPI in the context of his/her population group. Finally, the error estimation for each model provides a confidence estimation of the prediction to the physicians.

2.4.2.2 Data, Data Modelling

For the implementation of the two phases, a dataset formed by approximately four million patient records from the Valencian Region has been used. The first phase has been developed considering 19 specific diseases to obtain different comorbidity groups up to four diseases. The result is around 5000 relevant groups in terms of the death and hospitalization KPIs. Then, the specific models for the previously generated comorbidity groups with a significant population will be created.

2.4.2.3 Features

The models use retrospective EHR data. Particularly the pilot uses for the model the following data:

- Socio-demographics (e.g. smoking)

- Lab values
- Diagnoses
- Medications
- Stays at hospital

Task	**Brief description of algorithm**	**Features used within this algorithm**
1. Groups discovery	Statistical inference techniques. Binomial distribution and confidence intervals to find out relevant specific groups from their corresponding general groups	1. Patients ICD codes related with comorbidities 2. Risk Indicators 3. Patient sex – age range
2. Feature Selection	Logistic regression over standardized clinical features for the corresponding group	1. Patients EHR 4. Risk Indicators
3. Risk Prediction	Decision trees finding out the explicability of the model	1. Patients EHR 5. Risk Indicators

At the end, the model uses 8 features.

2.4.2.4 Model

Before the application of the ML solutions we carry out two processes to prepare the data:

- Anonymization: as we are dealing with sensitive data, we will apply an anonymization methodology to avoid the risk of de-identification. This anonymization process is a previous step before sharing the data from the INCLIVA (data owner) to the rest of the pilot partners
- ETL and data discovery: as we receive data as database dumps from their original sources, a process of transformation and cleaning is required to prepare the data for analysis.

After data is ready for analysis, the second step is to know how different diseases are grouped, defining a set of higher risk comorbidities groups. The goal of this step is to understand how different diseases relate to each other. To analyse these diseases, we use a directed tree: the n-level of the tree represents a group of n comorbidities that provide a higher mortality or hospitalization risk than in isolation. New levels in the comorbidities tree are defined only if there is a statistical difference between the adding a new comorbidity to the group defined in the previous level. These groups are calculated by using a binomial t-student statistical test that helps us to discover statistical differences between groups.

The next step requires the supervision of a human expert to look and understand the ICD codes involved in each comorbidity group. This supervised step is required to check if the selected diseases for each group have a medical significance and they are accurate from a diagnosis point of view. This step can help to establish the right confidence level during the statistical inference process.

Finally, from each group we will obtain the most significant features that best explains the target. Then, we foresee to obtain models that help us to predict the pursued target for each patient according to such relevant features.

Additionally, the risk prediction models for 30-day emergency readmission (or mortality) of (HF) patients, follows a quite similar approach summarized as:

- Pre-processing of all data types;
- Aggregation and synchronization of multiple data sources;
- Feature extraction and possibly feature selection
- Risk prediction modelling which utilizes a probabilistic approach (see figure 3 for more details)

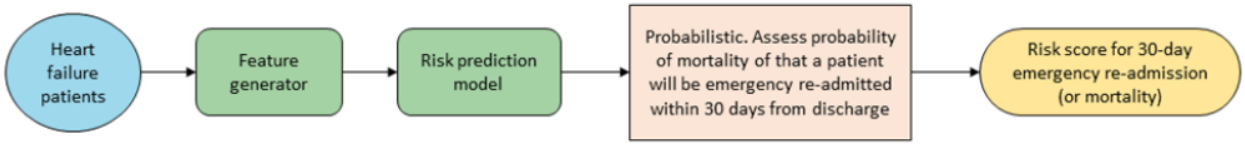

Figure 3 Risk prediction modelling

2.4.2.5 Evaluation

The models have been automatically validated under a mathematical point of view by using the cross-validation technique that allows us to measure the error of available data not seen during the training process and obtaining the different confidence intervals for those errors. From a non-mathematical point of view, the results of each model and the final decision tree should be supervised by a clinician.

2.4.3 Data Analytics

The main goal of the data analysis to fulfil in this pilot is to find **how comorbidities increase the mortality and hospitalization risk of a given patient**. This analysis is carried out in two stages:

1. In the first one, only relevant diagnoses (according to statistical tests) associated with comorbidities are considered in order to find groups or clusters of high-risk patients. To fulfil this stage, we perform some aggregation queries over the data to create relevant

comorbidities clusters. This cluster will be reviewed with the help of domain experts to check their relevance.
2. In a second stage, we will include not only the previous information but also additional features provided in the EHRs dataset such as smoking habits, TAS, TAD, glucose, HbA1c, creatinine, LDL-cholesterol, HDL-cholesterol. As the amount of data increases dramatically in this stage, we will use ML algorithms to discover clusters that not only involve comorbidities but also clinical features.

The result of both stages will be a classification model that, given the set of diagnosis/features associated with a patient, provides a risk level for the corresponding comorbidity cluster taking as information the clinical features, which will be assigned according to a centered 30-days window. This model will be deployed in primary care as a guideline for involved physicians to redirect a patient to secondary care or to take a similar action (hospitalization, drug prescription etc.)

In order to build such model, we define an approach made up of three main steps:

1. **Feature representation from original data**: This step takes the original EHR data and generates a tabular representation suitable for analysis. In the resulting matrix, each row represents a patient and each column defines a feature related to the patient (for a total of 8 clinical features). Since the pilot involves around 5 million of EHR, the process of generating this matrix requires a considerable amount of computing and memory resources. To accomplish this task, we will implement an ETL-like process using Spark and its distributed processing capabilities. The result of this stage is a sparse high-dimensional matrix with the features to be considered for analysis.
2. **Missing data imputation**: A model has been developed to perform the missing data imputation in a multidimensional way.
3. **Feature selection**: A logistic regression has been applied to filter the most relevant clinical features for each comorbidity group.
4. **Cluster modelling**: A decision tree has been used to obtain the KPIs prediction for a patient and the relevance weight of each feature to obtain the final prediction. As the created decision tree model is specific for each comorbidity group, it could be useful for the physicians to check or review the most relevant clinical features. On the other hand, as interpretable models have been used to obtain the specific KPI prediction for a patient, it is possible to know the specific state of a patient in relation with the group. This fact could help to know what actions or analysis are required to improve a KPI in the context of his/her population group. Finally, the error estimation for each model provides a confidence estimation of the prediction to the physicians

Clustering analytics of the stored data tested:

- The parameters of the KPIs from the most prevalent chronic conditions (diabetes, hypertension, coronary heart disease, heart failure, stroke, peripheral vascular disease, chronic kidney disease, depression, arthritis, atrial fibrillation, COPD) in two different

years were calculated obtaining the following parameters: mortality, age of dead, days with sickness leave, number visits to specialist and to Emergency room, days in Critical Care Unit and hospitalization, cost of hospitalization and cost of emergency room (INCLIVA).

- Predictive model of hospital re-admission in patients hospitalized for Acute Heart Failure (PHI, INCLIVA)

Predictive model of hospital re-admission in patients hospitalized for several comorbidities (PHI, INCLIVA) has been created. We used part of the aggregated data to generate simulated data reflecting the HF hospitalized population in Valencia. The aim was to validate the OPERA model, an existing prediction model for 30 day unplanned readmission or mortality in HF patients [14]. This model was developed on UK population data and had an original area under the ROC curve of 70%. As shown in Figure 3, we used distributions fitted on the model input variables to simulate the Valencian data. In the Valencian data, we identified 53789 positive events (HF patients having an unplanned readmission or dying within 30 days after discharge) and 281081 negative events (patients not having any of the aforementioned events within 30 days after discharge). Then, the OPERA model was tested for the newly generated data set. This process was repeated 10 times, aiming to suppress estimation variability due to randomness. The obtained area under the ROC curve was similar to the original reported on the development dataset with very low variance. This is an indication that the model can be a good fit on this population and should be tested further on real data.

2.5 Security and privacy of data access and processing

The aim of this section is to provide the security procedures and measures adopted to guarantee the security of the data involved in the Comorbidities pilot. This data is, specifically, a set of EHRs (EHRs) previously de-identified, from which different partners will gain insights by using analytical tools, mainly scripts and binaries. Figure 4 shows the high-level security architecture diagram, and next we detail the security and privacy measurements to be applied in the context of this pilot.

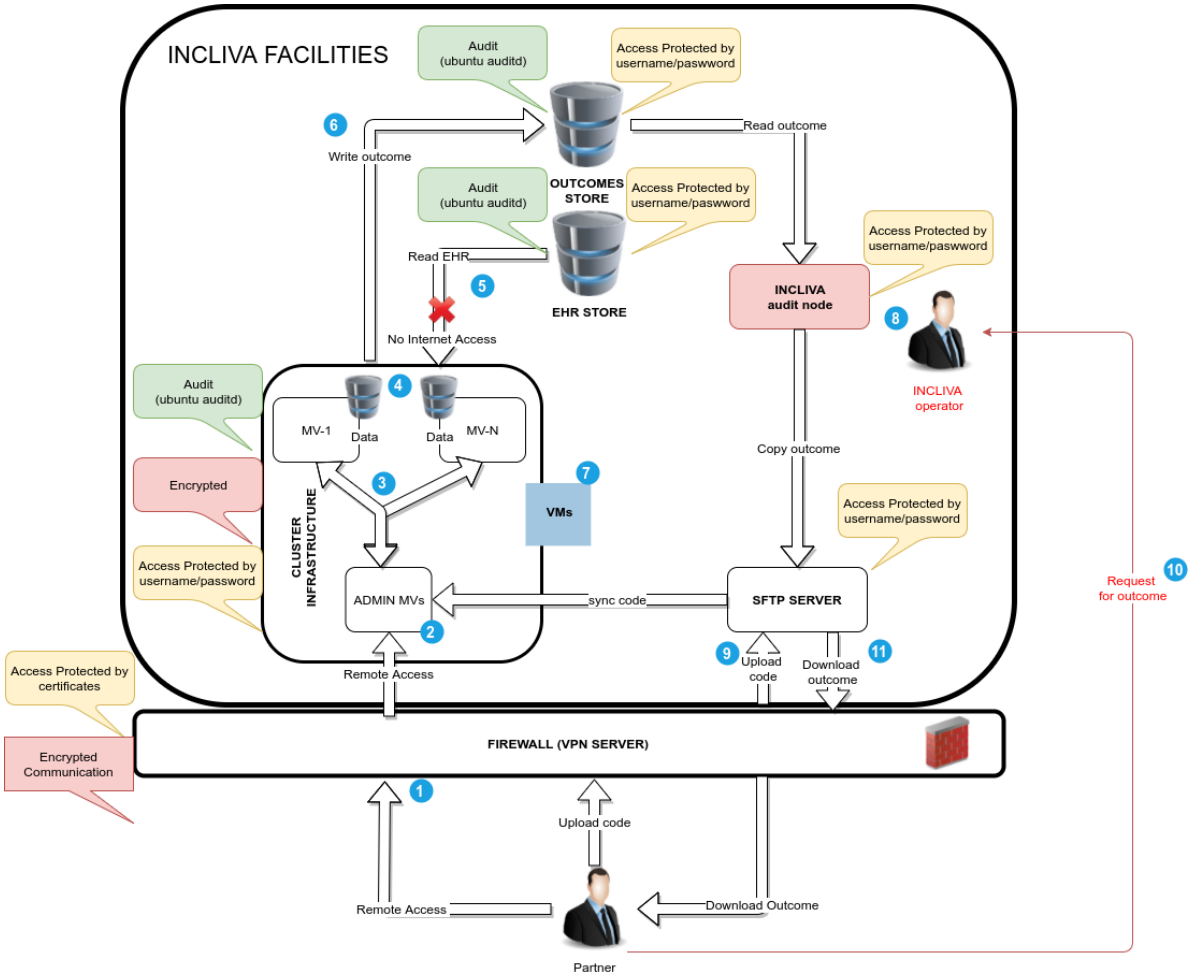

Figure 4 Security access data architecture

As the figure shows, all accesses and processes that involve sensitive data are always performed within the INCLIVA facilities. Every access from a remote computer to the infrastructure, will be done using a Virtual Private Network (VPN). This mechanism guarantees the encryption of the communications between the INCLIVA infrastructure and the remote computer and, in addition, it provides an authentication procedure. Inside this VPN, the access to the resources is controlled and audited according to the specific user credentials and the procedures described in the following section. The main components within the infrastructure are summarized as follows:

- **Cluster Infrastructure:** a set of VMs provided to facilitate the execution of big data analytics tasks. All data processing will be carried inside the INCLIVA infrastructure, namely a cluster (2). There will be a different cluster for each partner. Each partner will be granted with an admin node and, optionally, one or more processing nodes (3). From the technical point of view, a node is a virtual machine with a set of assigned computer resources (memory, disk and processors). The default operating system is Ubuntu 16.0.4 LTS. For each processing node, a mounted volume (4) with additional storage space will be assigned to store the processed data. These volumes support the storage of the processed data.

- **EHR database:** EHRs with sensitive and personal data are stored in a database isolated from the rest of the clusters. Accesses to this data will be monitorised in detail and only will be possible when access to external networks is disabled.
- **SFTP server**: the SFTP server will be used to upload the required source code and binaries for the analytical tasks.
- **Outcomes storage:** The outcomes from the analyses carried out in the pilot will be stored within this component. The download of the outcomes using the SFTP server will be controlled by INCLIVA and only possible when specifically granted.

2.5.1 Access Control

2.5.1.1 Authorization

Two access modes or roles are available to the partners:

- Setup mode: with root access to the virtual machines, including access to external networks, to install the required software and to configure the environment during a predefined time slot. The setup mode allows each partner to setup all the processing software (scripts, frameworks, editors) to support their specific tasks in the pilot. In the setup mode, connection to the database/data will never be granted under any circumstances.
- Development mode: the partners will have access to the system without the root user permissions, the access to the database/data will be available, but no access to external networks is granted.

Once the system is configured, the root access will be revoked. In case a partner requires a new software or root access to change the environment configuration, it should communicate to INCLIVA the setup procedure and they will carry it out. During the period where the partners will have root access, they won't be able to access any sensitive data.

2.5.1.2 Authentication

The authentication mechanism used on this pilot relies on two different services: external access to the infrastructure and access to the infrastructure resources. The detailed description of both of them is as follows:

External access to the INCLIVA Infrastructure

- Each partner will be granted a unique user id and password to access the VPN. It will be each partner responsibility to ensure the proper use of this user/password in their organization. The system will include a policy when creating/updating the end user passwords in order to assure that it will be secure (at least 8 characters, at least one capital character, at least a symbol, etc)
- In order to enable a two-factor authentication, INCLIVA will generate a client certificate with an encryption tool, such as OpenSSH, that will be required to access the VPN in combination with the previously provided user/password.
- Private certificate keys will be securely stored by INCLIVA.
- Password and certificate key will be periodically renewed by INCLIVA and communicated to the partners by means of an encrypted e-mail.

Access to infrastructure resources for a specific partner

- Each partner will be granted access to the admin node using a remote desktop console provided by the hypervisor of the cluster (Proxmox). To access the admin node, the end user must be authenticated into the system beforehand, using the mechanism provided by the operative system. From this admin node, the partner could connect to the rest of

the processing nodes via ssh or similar tools. Both admin and processing nodes assigned to a partner will not be accessible under any circumstance to the rest of partners. Only INCLIVA will have root access for setup tasks, i.e. in order to change nodes from development mode to internet mode. Partners should encrypt their working folders to ensure no information leakage.

- At the setup mode, root access and connection to external networks will be available to each partner. Each partner will be able to setup all the processing software required (scripts, frameworks, editors) to support their specific tasks in the project.
- In the setup mode, connection to the **EHR repository will never be granted** under any circumstances.

2.5.2 Data Protection

2.5.2.1 Data at rest

The data to be stored is a set of EHRs with clinical information. Previously to be stored into the infrastructure, two de-identification processes will be applied to this data. The first one, carried out by the IT Department of the data owner, will codify all specific personal ids. The second one, carried out by INCLIVA, will include additional rounds of de-identification, together with elimination of all data that potentially can identify a patient. All data stored within any of the hardware components of the infrastructure (mainly, hard disks) will be encrypted with a 256 bit AES encryption algorithm, using the services provided by the Unix operating system. Data will not be stored in any external media such as pendrives, optical media or portable hard drives under any circumstances.

2.5.2.2 Data in transit

All external connections will be established by means of a VPN connection (1) using the Fortigate software deployed in the INCLIVA firewall. The firewall will ensure that all communication to the infrastructure is encrypted using proper standards (SSL/TLS). Transfer of output data from the infrastructure will be made using a SFTP connection once a previous VPN connection has been established.
Regarding the internal communications within the VPN, they will be encrypted as far as performance requirements or the deployed technology allows that. Potential non-encrypted communications among nodes inside the VPN, will be communicated to INCLIVA for granting them and implementing the suitable monitoring measures.

2.5.3 Auditory and logs

2.5.3.1 System Auditory

VPN logs: both positive and negative (non-granted) accesses (timestamp and user login) to the VPN will be audited and securely stored by INCLIVA.
SFTP logs: both positive and negative (non-granted) accesses (timestamp and user login) to the SFTP server will be audited and securely stored by INCLIVA. Every file transaction (upload/download, timestamp, filename, file size) performed will be audited and securely stored by INCLIVA.
EHR Database server logs: both positive and negative (non-granted) accesses to the database server (timestamp, login, client IP) will be audited and securely stored by INCLIVA.

Every query performed (timestamp, login, SQL command or equivalent) will be audited and securely stored by INCLIVA.

2.5.3.2 Services Auditory

The main tools for auditing the system will be the Linux audited tool and the system logs available at /var/log. During development mode, INCLIVA will store these logs in an external node to support traceability in case of a security incident. Specific events to audit are:

- In setup mode, software and libraries installed into the master and processing nodes (maybe a ls of the whole system before moving to the development mode)
- In setup mode, external URL/IPs accessed and files transferred
- Commands ran by the partners in both master and processing nodes
- User accesses to the different nodes.
- Files created in development mode

2.5.4 Privacy measurements

2.5.4.1 Data Privacy Impact Assessment (DPIA)

Prior authorization and ethics committee. Data protection by design and by default.
INCLIVA designs the processing conditions through internal procedures. Prior to the collection, the data to be collected are analysed according to the purpose of the research, and whether they should be anonymous or personal data. This design is integrated into the documentation to be submitted to the ethics committee and should include:

- Purpose of the processing.
- Data subjects concerned.
- Data categories.
- Data retention period.
- If an authorized access to a third party is provided.
- A model of transparency and consent in data protection, if necessary.

Collection of data.
The collection of data takes place in two different ways depending on whether it is anonymised or personal data.
In the case of anonymized data, it is obtained from the public health service by means of a binding agreement that states:

- The data may only be used for the specific research purpose and within the framework of the project for which they were requested.
- The data will be processed in INCLIVA means (HPC).
- In case of personal data, the collection will be done in the terms described to ethics committee in above.

Warranties.
The data processing is designed in accordance with the provisions of national laws (Organic Law 15/1999, of 13 December, on the Protection of Personal Data and Act 41/2002, of 14 November, which regulates the autonomy of the patient and the rights and obligations regarding clinical information and documentation. This implies:

- Guarantee the rights of patients in the collection of data (see (i) above).
- Adopt the security measures provided for in the regulations when personal data are processed (Royal Decree 1720/2007, of 21 December, approving the Regulations for the

implementation of Organic Law 15/1999, of 13 December, on the protection of personal data).
- Taking reasonable security measures when processing anonymized data. These measures provide for the risks of access by unauthorized third parties, ensure non-re-identification, and the availability and resilience of the information systems.

2.5.4.2 Legal/Ethical process

The process follows the following rules:
- Data storage is only allowed at authorized terminals. Appropriate security measures are defined for this purpose.
- Access to the data is only allowed to authorized researchers and support staff. These staff must be trained and know their security and privacy obligations.
- The conservation period must be defined by the research project, or by the pilot. This usually involves conservation limited to the duration of the Grant Agreement. Any subsequent use or storage terms must be provided for in the applicable regulations.
- Secure deletion procedures must be used to destroy data. The terminals or other means containing data to be deleted can't be reused. If the means containing data are removed, it must be subjected to a secure destruction process. If the destruction is carried out by a service provider, the duty of confidentiality and the certification of the destruction shall be ensured by contract.

2.5.4.3 Processes for complying with the current legislation

INCLIVA complies with the following laws:

a. EUROPEAN UNION:
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

b. NATIONAL LAWS. (we reproduce only those provisions that are specific in front of GDPR):
- Organic Law 15/1999, of 13 December, on the Protection of Personal Data:
  Article 4. Quality of the data
  5. Personal data shall be erased when they have ceased to be necessary or relevant for the purpose for which they were obtained or recorded.
  On a regular basis, the procedure shall be determined by which, exceptionally, it is decided to keep the entire set of particular data, in accordance with the specific legislation, because of their historical, statistical or scientific value.
  Article 7. Data with special protection
  3. Personal data which refer to racial origin, health or sex life may be collected, processed and assigned only when, for reasons of general interest, this is so provided for by law or the data subject has given his explicit consent.
- Act 41/2002, of 14 November, which regulates the autonomy of the patient and the rights and obligations regarding clinical information and documentation.

Article 16. Uses of medical records.
3. Access to medical records for judicial, epidemiological, public health, research or teaching purposes is governed by the provisions of Organic Law 15/1999, of 13 December, on the Protection of Personal Data, and by Act 14/1986, of 25 April, General Health, and other applicable regulations in each case. Access to the medical record for these purposes requires that the patient's personal identification data, separated from those of a clinical-care-related nature, be preserved so that, as a general rule, anonymity is guaranteed, unless the patient has given his or her consent not to separate them.
Exceptions are made for cases of investigation by the judicial authority in which it is considered essential to unify the identifying data with the clinical assistance data, in which the judges and courts will be in charge of the corresponding process. Access to medical record data and documents is strictly limited to the specific purposes of each case.
When this is necessary for the prevention of a serious risk or danger to the health of the population, the health administrations referred to in Act 33/2011, General Public Health, may access the identification data of patients for epidemiological or public health protection reasons. In any event, access shall be granted by a health professional subject to professional secrecy or by another person subject to an equivalent obligation of secrecy, subject to a statement of reasons by the administration requesting access to the data.

2.6 Trustworthy AI

2.6.1 technology/user adoption and establishing trust

In order to establish trust in the context of technology adoption, and due to the lack of complete information related to this new technology, an end-user must make an “act of trusting” when committing to the new technology. This is “trust”.

This study uses the three components of Trustworthy AI (a lawful, ethical and robust system) to meet the characteristics of a new technology to the trusting behaviour of the potential technology end-user (in this case, only clinicians). This trusting behaviour is then linked to the technology adoption decision.

The research models were obtained complying with all applicable laws and regulations, ensuring adherence to ethical principles and values, related with anonymization, aggregation, de-identification of individuals. And the app only obtains the final predictions of the models and not the proper models which are in secured internal severs. Moreover, the models only contain aggregated data and mathematical formulas. This should be clarified for end-users in order to make them aware of these fair working conditions, both from a technical and social perspective, trying to avoid any unintentional harm.

2.6.2 ethical principles

- Respect for human authority: Predictions are only a clinician´s decision help and never replace them.
- Prevention of harm: Models only try to extract the relevant information present in the dataset. This is done from an aggregated and mathematical point of view to achieve to objectives of the pilot. These objectives, which are the prediction of KPIs and the considered clinical feature selection for a given comorbidity group, best explain the pilot KPIs.
- Fairness: Models only consider as relevant information the age range, gender, comorbidity groups and clinical features.
- Explicability: The final prediction models will be obtained by using decision trees that allow a good explanation of the results for users who are not specialised in the mathematical interpretation of those models.

2.6.3 key requirements

- Human agency and oversight: The pilot app and dashboard allow end-users to make informed decisions and foster their fundamental rights, always with a human-in-command approach.

- Technical Robustness and safety: AI systems need to be resilient and secure: the anonymised dataset and the processing scripts are duly secured in the internal server. Models try to learn functions which relate an objective with several explanatory variables. They are accurate, reliable and reproducible, avoiding unintentional harm can be minimized and prevented.

- Privacy and data governance: Full respect for privacy and data protection, adequate data governance mechanisms are ensured, considering the quality and integrity of the data, and ensuring legitimised access to data.

- Transparency: End-users will be aware that they are interacting with an AI system, and will be informed of the system's capabilities and limitations with a clear message.

- Diversity, non-discrimination and fairness: The pilot app and dashboard will be accessible to all clinicians, regardless of any disability, and involve relevant stakeholders throughout their entire life cycle.

- Societal and environmental well-being: The pilot app and dashboard will benefit all human beings, including future generations. They will be sustainable and environmentally friendly at the same level of a mobile phone or a computer.

- Accountability: The whole system has been validated at different levels: the usability of the interfaces, operative testing, and models cross-validation to measure the error in the prediction of the different models.

2.7 System-Interaction

2.7.1 Human-Machine Interface / GUI

A dataset formed by approximately four million patient records from the Valencian Region has been used.
Patients are not identified as the dataset has been anonymized and the only feature required for a patient consists of the corresponding gender, age range, diseases selection and the quantized values for the corresponding clinical features in the corresponding comorbidity group. The app will only be available for clinicians.
A first phase has been developed considering 19 specific diseases to obtain different comorbidity groups up to four diseases. And a second phase has consisted in, the creation of specific models for the previously generated comorbidity groups with a significant population.
The results of both phases have been deployed using a mobile application and a web dashboard. Such applications are briefly described:
Figure 5 shows several examples of the developed mobile application. As shown in the different pictures, the first step (A) consists on the gender selection. Then, the physician selects the age range (B) and a set of diseases that define a comorbidity group (C). Finally, a new window will show us the results for the two KPIs (D) according to our previous study. These values are compared with the corresponding KPIs values for the general population, but independently of the diseases considered in the study. A prototype of the deployed mobile app is available in the following URL: http://app.bigmedilytics.eu:443/ .

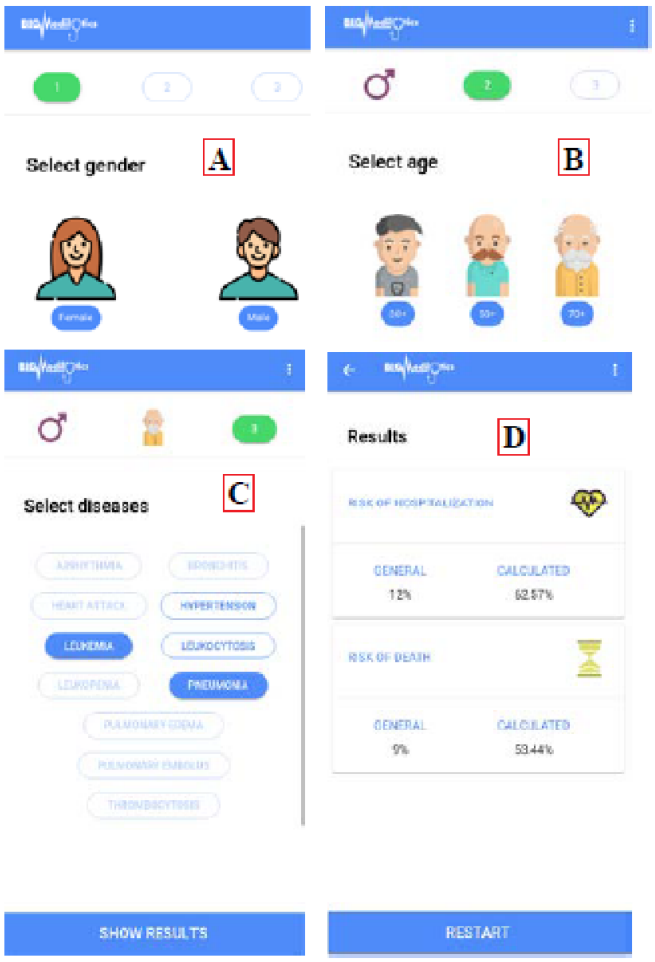

Figure 5 Mobile app interface

In order to provide an easy and intuitive way of exploring the great volume of data analysed within this pilot, a reporting web dashboard has been designed and developed. The main goal of the dashboard is to explore how the hospitalization and exitus KPIs change across different comorbidity groups or clusters. The dashboard is composed by one panel for each KPI. Each panel shows different charts displaying information related to the selected comorbidity group. Figure 6 shows the different data charts displayed for the exitus KPI page.

At the top of the page, two bar charts display information about the gender and age distributions of the specific comorbidity cluster that the user has selected, in terms of the exitus rate and the number of patients affected. A more general overview of the data can be found in the scatter plot shown in Figure 6. This chart displays the number of patients and KPI rate for the comorbidity cluster of the selected diseases, as well as for all the clusters that contain them. The user can also filter by age group and/or gender, zoom in/out, hover over the elements to see the information of a cluster, and download the image.

The current prototype for the dashboard application is available in the following link: http://app.bigmedilytics.eu:8090/ .

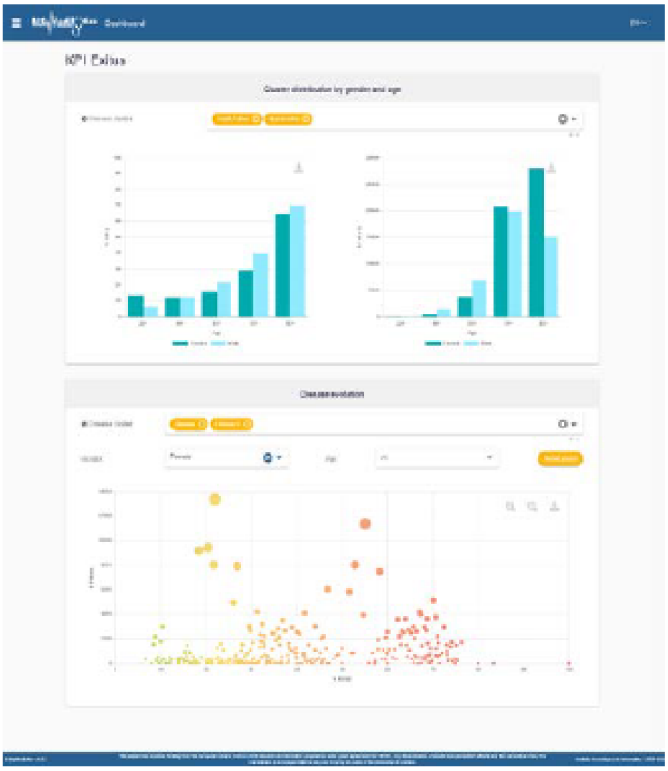

Figure 6 Dashboard

2.7.2 Education

Initially, the idea was to the show the app and the dashboard to clinicians in a short workshop. The workshop will consist in a 2 parts session:

- The introduction of the problem.
- Practical demonstration on the use of both GUIs interfaces.

3. Learnings

3.1 Challenges & Barriers

- Architecture: Creation of a dynamic cluster to launch the different ETLs from a large size dataset. Use of a distributed database (Hive).

- Processing of large structured / unstructured data sources: Applying functional programming by using Spark over a cluster, because of the stringent need to stick to the framework. Creation of the different comorbidity clusters by using a statistical method, in this case hierarchical combinatorial search.

- Multi-velocity processing of heterogeneous data streams: N/A

- Complex real-time event detection: N/A

- Natural Language Processing: N/A

- Image Processing: N/A

- Prediction Algorithms: Quantisation of data and missing values imputation need an up-to-date specific expertise, medical and mathematical respectively. The clinical feature selection for each relevant comorbidity group needs both medical and mathematical supervision. Obtention of an easily explainable model to make the risk prediction.

- Security and privacy of data access and processing: Ensure anonymisation of data. Perform the corresponding aggregation and quantisation of features.

- Trustworthy AI: Clinicians must validate the results in order to make this technology friendlier and must be informed of the system capabilities and limitations with a clear message.

- System-Interaction: Discovering the appropriate interface that reduces the interaction during the input of data to obtain the corresponding risk prediction. The best final results configuration has to be clearly understandable by clinicians.

3.2 Lessons Learned

- Architecture: Creation of clusters allowing the increase/decrease of computation power in transparent way without affecting the scripts.

- Processing of large structured / unstructured data sources: The functional programming is a relevant framework to be used in a distributed cluster. Use of the hierarchical combinatorial search in order to create different comorbidity clusters by using a statistical method.

- Multi-velocity processing of heterogeneous data streams: N/A

- Complex real-time event detection: N/A

- Natural Language Processing: N/A

- Image Processing: N/A

- Prediction Algorithms: Quantisation of data. Missing values imputation and clinical feature selection for each relevant comorbidity groups. Obtention of an easily explainable model to make the risk prediction. Extract the maximum available information from the different models.

- Security and privacy of data access and processing: The problematic behind the sensitive data and how to protect patients' private data.

- Trustworthy AI: At the end of the day, the AI gives a recommendation. However, a human supervision is required before taking the final decision behind a prediction.

- System-Interaction: How to create the appropriate visual metaphors in order to make easy the introduction of data and interpretation of results.

3.3 Main (quantifiable) achievements

App:

This app allows clinicians to know 2 different risk predictions (mortality and hospitalisation) for a given comorbidity group and even for an individual.
It also allows the discovery of the most relevant clinical features (initially selected in the study) for each comorbidity group and their relevance. (see 2.7)

Dashboard:

This dashboard is a descriptive tool which allows the discovery of the relationship between the comorbidities aggregation complexity and their corresponding computed risk. In this dashboard, the most probable complex comorbidities in relation to other less comorbidities are shown. (see 2.7).

4. Output

4.1 Papers

Writing: "Towards a Personalized Primary Health Care by Using a Hierarchical Divisive Approach for Comorbidity Clusters Establishment and Conditioned Group Risks Prediction".

4.2 Open Source & Resources (refer to ELG)

N/A

4.3 Demos (warning: for internal use only)

App: http://app.bigmedilytics.eu:443/ .
Dashboard: http://app.bigmedilytics.eu:8090/