Deliverable 1.3

Updated prototypes of specific components for all BigMedilytics pilots (software)

Big Data for Medical Analytics

Project Coordinator	Supriyo Chatterjea, Philips Electronics Nederland B.V.		
Start date Project	January 1st 2018	Duration	38 months
Version	1.0		
Status	Final		
Date of issue	31/05/2019		
Dissemination level	Public		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780495.

The content of this document reflects only the author's view. The European Commission is not responsible for any use that may be made of the information.

Authors' data

Author	Beneficiary	e-mail
Roland Roller	DFKI, roland.roller@dfki.de	
Anne Schwerk	DFKI, anne.schwerk@dfki.de	
Final editor's address	Supriyo Chatterjea Philips Electronics B.V. High Tech Campus 34 5656AE Eindhoven / Netherlands	

Management Summary

The goal of work package 1 (WP1) is to oversee the transfer of mature Big Data technologies into the BigMedilytics pilots. The final outcome of WP1 is the Big Data Healthcare Analytics Blueprint, the BigMedilytics-BigMatrix, a mapping between the requirements and the technical components. The matrix will be multidimensional, taking into account aspects of technologies, of pilots/businesses, and of communities, as well as aspects of specific data sources. The first deliverable D1.1 addressed the technical requirements collected from all BigMedilytics pilots. D1.2 presented a more detailed analysis of the different pilots according to the WP1 tasks. In particular it presented a first overview about the different software prototypes, shows the different components, data and challenges to overcome.

Deliverable D1.3 is an update of D1.2. We take the previous input into account and try to gather more detailed information in order to understand similarities and differences across pilots easier. In this way we prepare a solid baseline towards the BigMedilytics-BigMatrix. Most updates are highlighted in blue.

Table of Contents

1. Introduction

1.1. Purpose of the document

The goal of work package 1 (WP1) is to oversee the transfer of mature Big Data technologies into the BigMedilytics use cases (hereafter “pilots”). The transfer will happen in three cycles: (1) initial prototypes, based on pilot requirements; (2) updated prototypes, based on pilot internal validation; (3) final implementations, based on pilot external validation. Each cycle will be described within the WP1 deliverable. The final outcome of WP1 is the Big Data Healthcare Analytics Blueprint, the BigMedilytics-BigMatrix, a mapping between the requirements and the technical components. The matrix will be multidimensional, taking into account aspects of technologies, of pilots/businesses, and of communities, as well as aspects of specific data sources. Moreover, the matrix will also make a distinction between data sources that have different velocities, e.g. mobile devices vs. sensor streams for telemedicine pilots or real-time location data vs. electronic medical records (EMRs) for hospital workflow focused pilots in the Industrialization of Healthcare theme.

Deliverable D1.2 collects information according to the different WP1 tasks from all BigMedilytics. In this way we shed light on all relevant components of each pilot in order to prepare a good comparison between all of them.

The current deliverable D1.3 addresses an update of the specific components of the initial software prototypes.

1.2. Related documents

Related documents: D1.1, D1.2, D2.1, D3.1, D4.1:
Similarly, to D1.3, also D1.1, D1.2, D2.1, D3.1, D4.1 provide a general overview of the different pilots.
However, D1.3 is an updated version of D1.2 and provides a more detailed overview on the different prototypes of each pilot. D1.1 instead addresses technical requirements from a Big Data perspective. D2.1, D3.1 and D4.1 focus on the requirements from an application perspective.

2. Overview of specific components

The following document provides an overview of the different prototypes developed within the different pilots of BigMedilytics. The overview provides a first analysis according to six different tasks: deep learning for Multilingual natural language processing (NLP) and image analytics (IA) (T1.1.), prediction algorithms (T1.2), complex real-time event detection (T1.3), processing of large structured / unstructured data sources (T1.4), multi-velocity processing of heterogeneous data streams (T1.5) and security and privacy of data access and processing (T1.7). The goal of this document is a better understanding and comparison between the pilot.

2.1. Task 1.1: Deep learning for multilingual NLP and IA

Task 1.1 involves the application of NLP and IA components within the following WPs and pilots: WP2 (comorbidities, kidney disease, COPD/asthma), WP3 (prostate, lung, breast cancers), and WP4 (radiology workflow). In the following, an overview of the different pilots in terms of NLP and IA is given. Results are presented in the tables below according to various aspects.

From all pilots listed here, only pilot 8 and 12 focus on IA, all others work on NLP methods. All NLP-related pilots focus on a variety of different subtasks, such as named entity (or concept) detection, relation (event) detection and negation detection. However, overall a variety of languages are covered, i.e. there is no overlap in terms of languages used.

In brief, the NLP-related pilots aim to automatically extract text data from electronic health records (EHRs) in order to enhance their prediction models. This is a time-efficient and cost-effective data retrieval method to leverage relevant information from patient data and ultimately to further reduce costs.

Table 1: Overview of participants of Task 1.1

Partner	Involved tasks	Language	How will NLP/IA support your pilot?	How will NLP help you to reduce costs?
Pilot 2: Kidney Disease				
DFKI	NLP	German	In EHRs, text is an additional source of information, which often includes other information that is not directly expressed in structured data. Thus, it is important to extract relevant information from text in order to get a better understanding of patients.	We plan to combine structured and unstructured information in our prediction models. As text includes additional information we assume, that models will provide better results using text. Better results will also mean that patient outcomes can be improved and hospitalizations reduced. The reduction of hospitalizations reduces costs.
Pilot 6: Prostate Cancer				
PHI	NLP	Swedish, Dutch, English	In comparison to structured data, text can store additional information in EHRs. For this reason, we try to integrate information from text in our prediction models. Manual extraction of data is too labour-intensive. NLP makes our pilot possible.	If including extracted information to our prediction models leads to further and meaningful improvements, this will help our pilot to reduce costs. The NLP pipeline helps to automatically extract patient data from radiology and pathology reports for use on a dashboard. The alternative would be to do this by hand (for example by a nurse) which would take a lot of time and increase cost. Now it only has to be checked, not manually extracted.
Pilot 7: Lung Cancer				
UPM	NLP	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	The retrieval of the EHR data along with the used in another sources could allow to have an improvement of the KPIs, which is translated in a reduction of the economic costs associated to the hospital.

			and richness of patient data to improve the knowledge that we have about the patients and their associated processes in the hospital.	
NCSR-D	NLP	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.
LUH	Knowledge Management	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.
Pilot 8: Breast Cancer				
IBM	IA	Medical imaging	Medical imaging in a non-intrusive method to get valuable personalized information about the patient condition. Specifically, in breast cancer, there are several modalities: magnetic resonance, mammography, and ultrasound to capture the patient condition, and performing image analytics on this data enables us giving a personalized and more effective treatment.	By using image analytics to predict response to neoadjuvant chemotherapy (NACT) treatment, we can influence the right treatment for the patient. This can save the costs of ineffective treatments as well as prolong patient's quality life making them productive and contributors to the EU economy for longer periods.
Pilot 12: Radiology Workflows				
CON	IA and in parts NLP	German	Image processing is crucial to perform the image pattern comparison, driving the image/case search during radiological diagnosis. DL is used to train it on the diseases relevant for the pilot, and to comparison measures based on imaging data reflecting the disease specific appearance. NLP is important to use clinical routine data that comes with free text reports on the findings in images.	The tool will save time, it will increase quality, completeness and confidence of reports. The DL based similarity is used to perform search in large case data bases. The search result helps radiologists to obtain relevant information faster than alternative searches (e.g., books, internet). Furthermore the search result and the summary of its findings aid completeness and confidence in reported findings.

Table 2: NLP components of pilots

Which NLP tasks do you address?	**Method**	**Software frameworks**	**Vocabularies/ corpus used**	**Describe your method in a few sentence.**	**Describe your corpus/training data?**
Pilot 2: Kidney Disease					
Named Entity Recognition	Bi-LSTM	tensorflow	corpus of German nephology reports (clinical notes & discharge summaries)	Character-level bidirectional LSTM which reads in training data and is trained for each single concept.	Fine grained annotation of 12 different concepts in German nephrology reports which include discharge summaries and clinical notes.
Relation Extraction	CNN	tensorflow	corpus of German nephology reports (clinical notes & discharge summaries)	Convolutional neural network for binary relation extraction, which uses as input information beside words and related concepts, also the distance between the two concepts.	Annotated relations between concepts in German nephrology reports (see NER).
Negation Detection	rule-based	NegEx	small manually annotated corpus of negations in German	Simple negation detection of medical conditions	no training required

			clinical reports and a modified set of negation triggers		
Pilot 6: Prostate Cancer					
NER for disorders, findings, and anatomy in prostate biopsy pathology reports	regular expression matching	none	a few hundred pathology reports from ~10 sites in US, Netherlands, and Sweden		
item (vial) detection	regular expression matching in combination with rules	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		
measurement detection	regular expression matching	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		
negation detection	Negex-like, with some improvements for our specific task	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		
Pilot 7: Lung Cancer					
Named Entity Recognition	Deterministic rule-based system	Apache UIMA	UMLS repository	Rule-based system that deterministically classifies tokens in previously defined entities in UMLS repository. It relates a token with a concept and its semantics.	The Unified Medical Language System (UMLS) 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	Proprietary	Metamap	Applied on	Named entity	Pretrained

recognition			PubMed, PubMed Central	recognition in biomedical text, it recognizes UMLS concepts in text	(https://metamap.nlm.nih.gov/)
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/
Pilot 12: Radiology Workflows					
Named entity recognition	However this is not focus of the development in this project, we are using existing technology in the company	tensorflow		Free text reports are parsed, words are mapped to a terminology and the terminology is used to extract location-finding pairs from the structured report	
Concept normalization		tensorflow	Radlex		

Table 3: Image Analytics components of pilots

Do you require regulatory approval?	Which Image Processing tasks do you address?	Do you use public data repositories? If yes, which?	Describe your method in a few sentence.	Which training technique do you use?	Which pathologies are covered?	On which level does classification happen (volume, slice/img, or px-level)?	What is the level of detail of your GT-annotations (volume, slice/img, or px-level)?
Pilot 8: Breast Cancer							
No as we are doing a retrospective study	Multi-modal classification of imaging and clinical data including longitudinal mammography (MG) images, ultrasound (US) images and magnetic resonance (MR) images	Yes, ISP1 and UCSF NACT datasets from https://wiki.cancerimagingarchive.net	We'll use several methods and then ensemble them altogether. The methods include (1) deep neural networks to analyze images, (2) traditional feature extraction from images, (3) machine learning methods such as xgboost.	Train using convolutional neural networks (CNN)	Breast cancer	On volume for MRI and on slice/image for MG and US	Annotation is per image level
Pilot 12: Radiology Workflows							
Yes, it needs CE certification	comparison of image content in medical imaging data (e.g. computed tomograph	yes, for research we use available repositories e.g., challenges	The technique compares image content of a query ROI (region of interest) with a large number of image segment in the data base (15Bn image segments, ~7000+	we use mainly weakly supervised from radiological routine information,	lung diseases	on pixel level, and in a subsequent step, ranking is performed on a volume level.	volume level, and regions of interest for evaluation

	y volumes)		volumes). It identifies the closest matches and retrieves the corresponding cases.	i.e., images together with reports			
	search based on a marked region of interest, to find image segments across a large data base that carry similar patterns, and rank these cases according to similarity		During training we learn an image similarity function and index the imaging data to provide fast retrieval		lung diseases		

2.2. Task 1.2: Prediction algorithms

Task 1.2 involves the integration of state-of-the-art machine learning (ML) algorithms for prediction scenarios within the following pilots: WP2 (Comorbidities (pilot 1), kidney disease (pilot 2), gestational diabetes (pilot 3), COPD/ asthma (pilot 4), heart failure (HF; pilot 5)), WP3 (prostate cancer (pilot 6), lung cancer (pilot 7), breast cancer (pilot 8)), and WP4 (hyper-acute workflows (pilot 9/10), asset management (pilot 11)).

Table 4. prediction problem landscape

Pilot name	Prediction problem categories			
	Medical event prediction	Comorbidity analysis	Treatment risk analysis	Other
Comorbidities (pilot 1)		1.1 Comorbidities aggrupation 1.2 Relevant features extraction from HER 1.3 Hospitalization & Mortality risk prediction model		-
Kidney disease (pilot 2)	2.1 Infection Detection 2.2 Detection of possible re-hospitalizations			2.3 Adherence Monitoring
Gestational diabetes (pilot 3)				3.1. Monitoring and categorizing of glucose levels
COPD and Asthma (pilot4)	4.1. Predict acute exacerbations of COPD	-	-	
HF (pilot 5)	5.1 Prediction of HF patient hospitalizations	5.2 Lasso Regression	-	-
prostate cancer (pilot 6)			- 6.1 Pre-surgical risk of post-surgical adverse prostate cancer pathology (i.e., pathology Gleason>=7) - 6.2 Pre-surgical risk of post-surgical advanced extent of disease (i.e., pathology disease stage >=pT3a) - 6.3 Pre-surgical risk of the presence of tumor infiltrated	

			lymph nodes.	
Lung cancer (pilot 7)	7.1 Predict Long Term Survivors	7.2 Correlation between Commodities and Toxocities	7.3 Prediction of Drug Interactions	
Breast cancer (pilot 8)	8.1 Predict pathologic complete response (pCR) to neoadjuvant chemotherapy (NACT) treatment			8.2 Predict cohorts for clinical trials towards next generation therapies
Hyper-acute workflows (pilot 9/ 10/11)				Examples: 9.1, 10.1, 11.1 Bottleneck formation in workflow 9.2, 10.2, 11.2 Timing characteristics in workflow 11.3 Periodic Automatic Replenishment (PAR) level

Data and prediction algorithms

The data and prediction algorithm categories have been chosen according to all pilot inputs from the first version of this document. One row per prediction algorithm is used to specify the data which is utilized within the algorithm.

Table 5. data and prediction algorithm used in each prediction problem

Pilot	Prediction problem ID	Prediction algorithm						Data								
		Tree-based	SVM	Probabilistic	Clustering	Neural Nets	Other	Medical time-stamped				biomedical text	Demographics	Environmental	Insurance	Other
								Structured time series	Text	Image	Other					
1	1.1			x							x					
1	1.2						x				x		x	x		x
1	1.3	x	x			x	x				x		x	x		x
2	2.1	x				x		(x)	x				x			x
2	2.2	x				x		(x)	x		x		x			x
2	2.3						x				x					x
3	3.1						x	x					x			
4	4.1	x						x					x	x		
4	4.1					x		x					x	x		
5	5.1						X						X		X	
6	6.1, 6.2, 6.3	x								x			x			x
7	7.1, 7.2, 7.3	x	x		x			x	x		x		x			
8	8.1					X	X			X			X			X
8	8.2					X	X			X			X			X
9, 10, 11	9.1, 10.1, 11.1	x				x		x								x
9, 10, 11	9.2, 10.2, 11.2	x				x		x								x
9, 10, 11	11.3	x				x		x								x

Features

For each prediction algorithm a brief description of the algorithm and the features used within it is provided in Table 6.

Table 6. Features that are used for each algorithm

Prediction algorithm ID	Brief description of algorithm	Features used within this algorithm
1.1 Groups discovery	Statistical inference techniques	1. Patients ICD codes related with comorbidities 2. Risk Indicators 3. Patient sex – age range
1.2 Feature Selection	Ensembler or genetic evolutive algorithms	1. Patients EHR 4. Risk Indicators

1.3 Neural Net	MLP Regressor, decision trees	1. Patients EHR 5. Risk Indicators
2.1 & 2.2 Tree-Based	Decision Tree, Random Forest, Decision Tree Regressor	1. Demographics 2. Vital Parameters 3. Lab Values 4. Co-Morbitites 5. Medications 6. Clinical Notes (todo) 7. Discharge Summaries (todo)
2.1 & 2.2 Neural Net	MLP	
2.3 Rule-based	Rule-based ranking of patients with a low adherence based on expert knowledge	1. App data from patients about drug intake
3.1 Fuzzy Rules	A set of fuzzy rules will be trained from existing thresholds, and collected data.	1. Demographic information of the patients 2. Self-measured and automatically collected time series of the glucose levels
4.1 Tree-based	Random forest	1. Self-reported daily patient symptom scores ([1..4]) 2. Daily weather data at patient location per patient 3. Daily pollution data at patient location per patient 4. Self-reported irregular CAT score
4.1 Neural Net	Neural networks for sequence analysis	1. Self-reported daily patient symptom scores ([1..4]) 2. Daily weather data at patient location per patient 3. Daily pollution data at patient location per patient 4. Self-reported irregular CAT score
5.1 Logistic Regression	Lasso	1. Patient claims data 2. Patient characteristics
6.1, 6.2, 6.3	Random forest	Patient characteristics, clinical info (PSA, T stage from DRE), pathology info from biopsy (Gleason score, # of positive biopsies, etc), MRI information (PIRADS, ECE, bulging, tumor location, lymph node)
7.1	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.
7.2	Embeddings of patients, feature selection methods, correlation matrices	• Patient embeddings are computed for representing the patient features. • Correlation matrics/coeficients to compute the scores of the features
7.3	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 computed missing links.
8.1 Neural Net	CNN	1. MRI images 2. Longitudinal MG images 3. US images 4. Clinical data
8.1 Other	XGBoost	1. Features from CNN 2. Features from traditional ML on imaging 3. Clinical data
8.2 Neural Net	CNN	1. MRI images 2. Longitudinal MG images 3. US images 4. Clinical data
8.2 Other	XGBoost	1. Features from CNN 2. Features from traditional ML on imaging 3. Clinical data
9.1, 10.1, 11.1, 11.3 Workflow characterization	Random forest, Long-short term memory	1. Location information 2. No. of patients

2.2.1. Detailed description of each pilot:

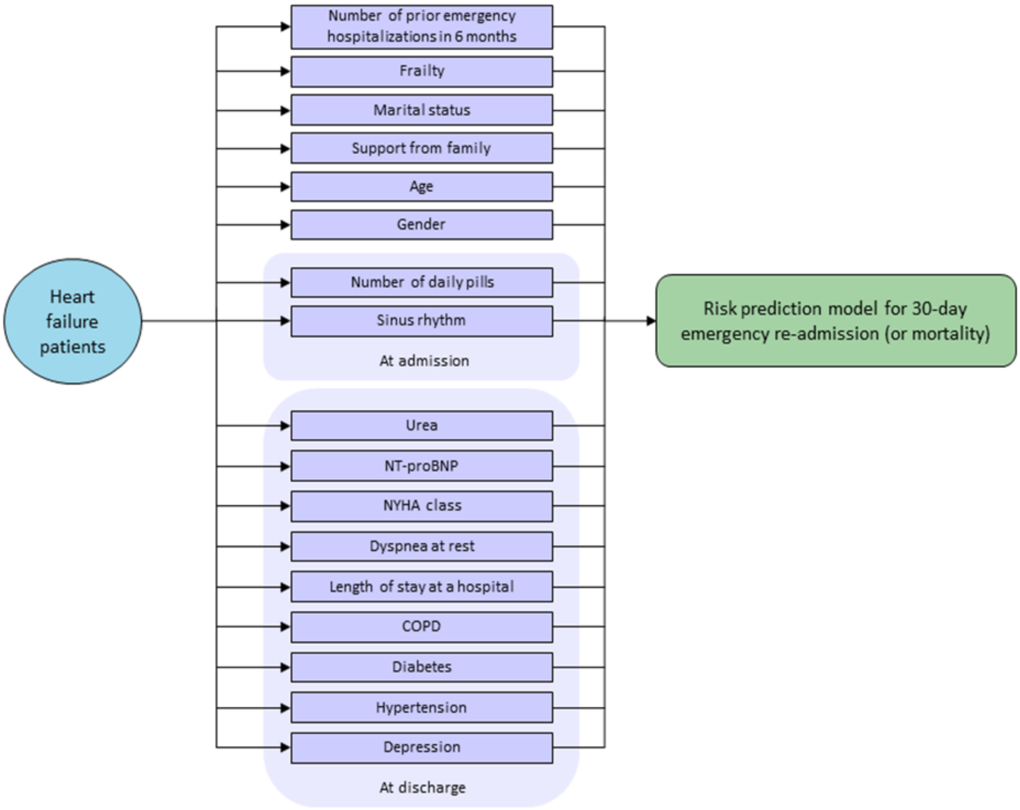
Pilot 1 - Comorbidities (WP2) **Partners: OPTI, ITI, PHI**

Description of the prediction problem:

This pilot uses as main data source around 5 million EHRs of the Valencian Region population over a timespan of approximately 7 years. Using this dataset, our goal is to understand how critical diseases influence each other and, then, to provide a more accurate risk for a specific patient. Therefore, the main prediction problem is to discover comorbidity clusters, i.e, patients that share similar diagnoses of relevant diseases such Diabetes or HF, and then analyse carefully the most influential variables that define such clusters. Then, we will use relevant metrics, such as the number of hospital readmissions, hospital emergencies or the number of visits to secondary care, to define predictive models about the healthcare assistance recommended. The resulting models will be tested in a specific health department of the Valencia Region to evaluate a potential improvement in the aforementioned metrics. To create such both clusterization and predictive models, we will use EHR information from several areas as described below:

- **Socio-demographic data**: General information about the patient (age, gender, residence, etc.) and health-oriented information such as smoking habits and physical activity
- **Healthcare metrics**: Per each patient, yearly visits to primary care, secondary care and hospital urgencies
- **Hospital discharges**: Per each hospital stay of the patient in a specific time range, this data source provides a set of diagnosis and procedures carried on the patient using ICD codes.
- **Previous diagnosis**: ICD codes related to diagnosis not specifically related to a hospitalization. This data source complements the previous one.
- **Treatment and prescriptions**: Provides information about drugs and pharmacy dispensation related to a specific patient treatment.
- **Clinical measurements**: Average values per each 6 months of analyses (such as a blood analysis) carried on the patient.

Additionally, this data will be used for analysis of existing and potential development of new risk prediction models for 30-day emergency readmission (or mortality) of (HF) patients. The next figure summarizes the use of the data for this task:



Considered machine learning solutions:

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

1. **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
2. **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 statistical inference techniques that help us to discover statistical differences between groups.

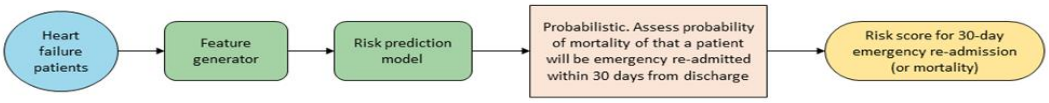
The next step requires the supervision of a human expert to look and understand for the ICD codes involved in each comorbidity group. This supervised step is required to check if the selected diseases for each group have a medical significance and they are accurate from a diagnosis point of view. This 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:

1. Pre-processing of all data types;

2. Aggregation and synchronization of multiple data sources;
3. Feature extraction and possibly feature selection
4. Risk prediction modelling which utilizes a probabilistic approach (see figure below for more details):


Heart failure patients
Feature generator
Risk prediction model
Probabilistic. Assess probability of mortality of that a patient will be emergency re-admitted within 30 days from discharge
Risk score for 30-day emergency re-admission (or mortality)

Summary:

- High-level aims: comorbidity cluster analysis, medical event prediction
- Data used: demographics, medical time-series
- Mentioned prediction algorithms: clustering methods, probabilistic methods, neural networks

Programing languages and/or proprietary software:

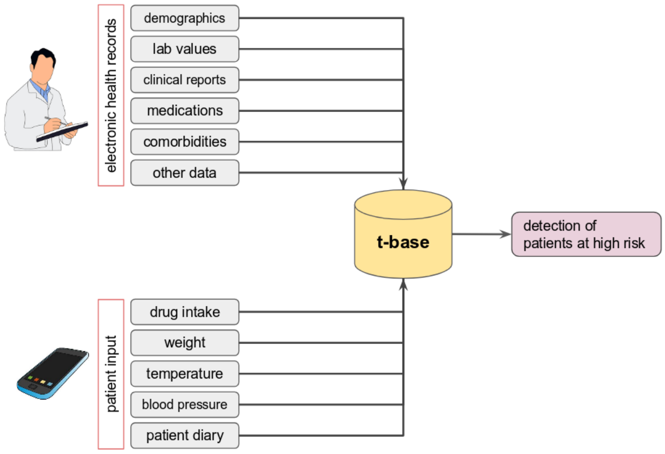
- Python
- R
- R-Studio
- Spark

Pilot 2 - Kidney disease (WP2) **Partners: DFKI**

Description of the prediction problem:

The baseline of the kidney pilot is a patient-centered smart electronic health-care service platform, which focuses on improving the safety of patients after kidney transplants. Particularly, the platform aims at improving the drug safety of patients, as well as the communication between patients-physicians and between physicians. In the course of this, the kidney pilot focuses on a reduction of unwanted re-hospitalizations, reduction of mortality, detection of possible acute kidney failures after transplantation and the extension of the graft survival (KPIs). Moreover, the pilot targets the support of adherence, which is a major reason for graft loss (if medications are not taken). A central element of the kidney pilot is a dashboard that visualizes critical patients according to various aspects. In order to address the KPIs we focus on three scenarios: a) monitoring adherence using information about drug intake, b) detecting risk factors for infections and acute kidney injury using historical data and c) patient monitoring and outlier detection focussing on patient input.

As input for our prediction models we consider two main data sources: data from EHRs and patient data inserted via an app. The EHR data is data from the transplant center and involves information, such as demographics, lab values or clinical reports, from transplanted kidney patients of the last 15 years. The patient input involves the drug intake which provides information about adherence, as well as weight, blood pressure etc.



Considered machine learning solutions:

The main foci of the pilot are time-series analyses and prediction models. Various methods will be examined for its efficiency in our pilot, such as random forest, LSTM and Convolutional LSTM. Moreover, beside the main risk prediction task we also target the explainability of our models. However, the ranking of 'critical' patients (according to e.g. adherence) will play a major role.

Summary:

- High-level aim: detecting critical patients according to adherence, risk for infections and acute kidney injury
- Data used: electronic health records from transplant center (demographics, lab values, medications, diagnoses, clinical notes, discharge summaries etc.), User input via APP (drug intake, weight, blood pressure, patient diary etc)
- Mentioned prediction algorithms: tree-based methods, neural networks

Programing languages and/or proprietary software:

- Python
- Java
- Flink

Pilot 3 – Diabetes (WP2) **Partners: HUA, NISS**

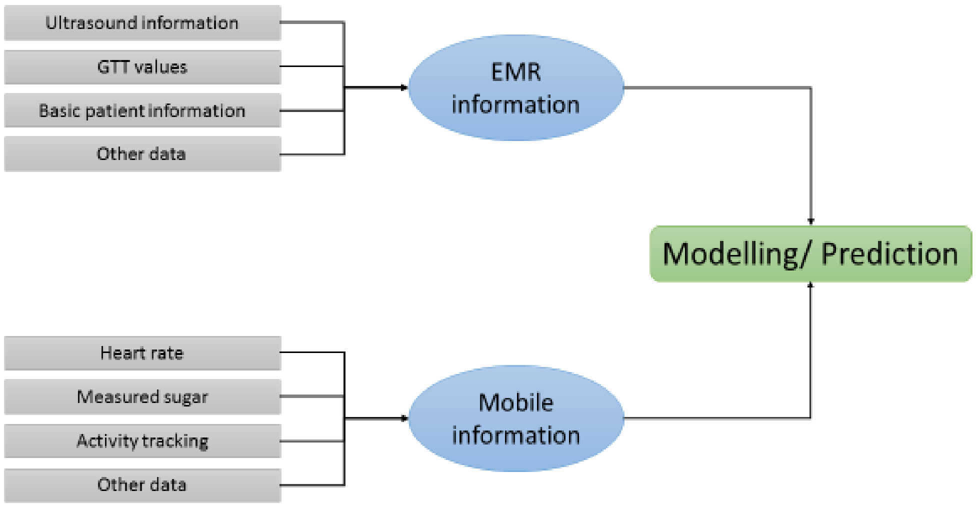
Description of the prediction problem:

This project will use data coming from two primary sources:

- EMR and retrospective studies. Collection of anonymized data and subsequent analysis of medical data received from the hospital, with a varied set of information, including ultrasound information, GTT values, basic information about the patient (BMI, patient age, gestational age), etc.
- Mobile data. A smartphone will collect the data from the sensors and send it to the cloud periodically. The data will contain information about the patient, such as heart rate, measured sugar, activity tracking, etc. It will also include self-reported nutrition information, where the user will report her/his intake, and it will be taken into account

Using this data, we will create a system to monitor the patients, allowing to react faster to emergencies and reduce hospital visits.

A graphic representation can be found in the diagram below.



In this diagram, other data includes the possibility to include new sources of data, given that technology develops in the span of the project, and new hardware is manufactured that may be useful for a better development of the final solution.

Considered machine learning solutions:

Please find below an explanation of the machine learning solutions that will be used. Please note that there will be a strong interaction between them, and the exact steps and parts within each part will be designed to solve the specific problem statement:

- **Data ingestion**. The data will arrive from the devices into the cloud
- **ETL**. Cleaning and transformation of incoming data. This step may involve creation of new features, normalization of the data, etc. An exploratory analysis will be done in order to decide which techniques are necessary for best results
- **Storage**. The data will be stored in the cloud. This will allow to create new models using the collected data, make sure that the currently deployed models are generating valuable results, etc.
- **Analytics**. The data will be fed into models that will be trained to generate results. Various analytical methods will be examined to select the most adequate approach to be used in the models to be developed. In addition, other processing steps (for example, dimensionality reduction) will be tested and used to refine the models. The initial selection of models to be tested and evaluated are listed below:
  - Recommender algorithms
  - Clustering
  - SVM
  - Random forest
  - Deep Learning models
  - Fuzzy rules

  These models are considered due to their great result in literature, but the exact model used may vary to accommodate the use case. Various tests will be carried out to select the most appropriate model in terms of accurate information for the purpose of this case. The final selected model will be implemented in the platform for further tests, validation and refinements.

  After initial analysis, we have decided to implement fuzzy rules as the analytics model for the initial version of the application. This type of model has several advantages. First, it's a model which is similar to what the medical team is currently using. This will easy the introduction into the hospital, and will increase the usage of the tool. Second, it is a model that aligns well with monitoring, as it allows flexible thresholds to personalize to each patient exact situation. As we collect more data during the pilot, we will keep testing different models, and we will decide whether new models achieve better performance and readability.
- **Visualization**. The visualization layer, will use the storage and analytics layer to generate insights. An example could be feeding a feedback loop, used to assess that the model deployed is giving good results. Further uses of the visualization layer may be implemented to adapt to the problem

Summary:

- High-level aim: medical event prediction
- Data used: medical time-series, demographics, images, real-time streaming
- Mentioned prediction algorithms: clustering methods, tree-based methods, support vector machines, neural networks

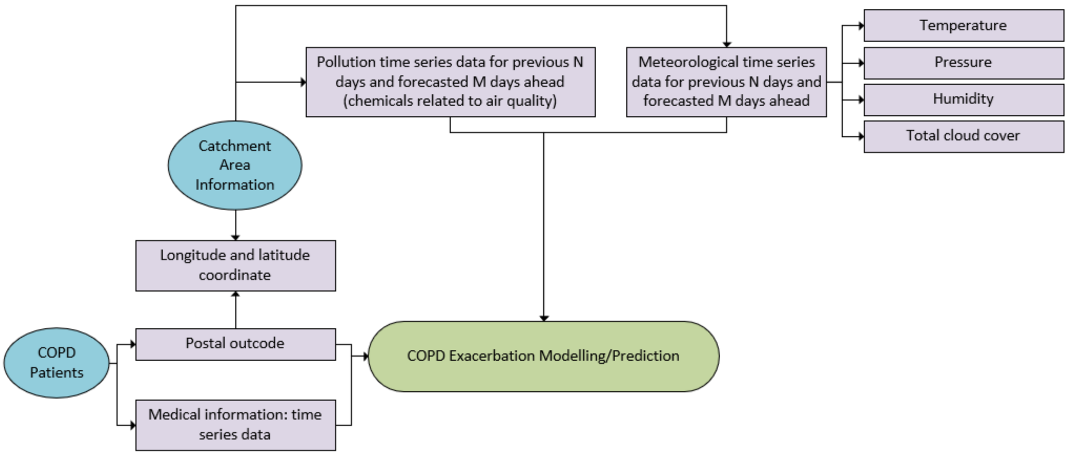
Programing languages and/or proprietary software:

- Python
- Scala/Java
- Flink
- Kafka

Pilot 4 – COPD and Asthma (WP2) **Partners: UNIS, MY**

Description of the prediction problem:

This pilot will utilise data captured using mobile and web enabled platforms MY COPD and MY Asthma (MyMhealth data). This is used to develop predictive models of acute exacerbations of COPD. These models will enable a move from a reactive to proactive approach to care. The pilot will utilise data captured on the platforms to create models using daily data on symptoms, treatment, environmental observation data including temperature, humidity, pollen counts and air pollution to create risk models which are individualised to the patient's own disease state and environment. For more details see figure below.


Temperature
Pressure
Humidity
Total cloud cover
Pollution time series data for previous N days and forecasted M days ahead (chemicals related to air quality)
Meteorological time series data for previous N days and forecasted M days ahead
Catchment Area Information
Longitude and latitude coordinate
COPD Patients
Postal outcode
Medical information: time series data
COPD Exacerbation Modelling/Prediction

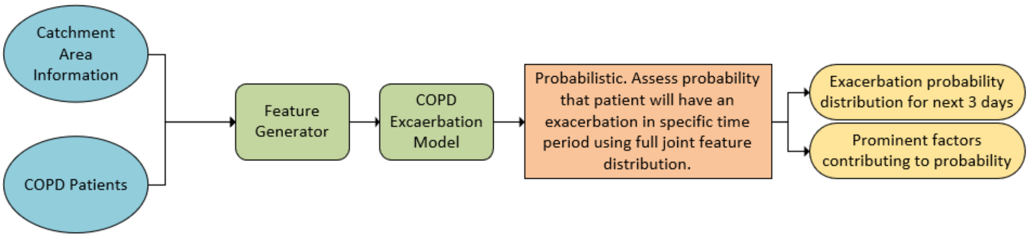
Considered machine learning solutions:

To this end the algorithm requires:

1. Pre-processing of all data types;
2. Aggregation of MyMhealth mobile/web generated patient data
3. Information about the environment that the patient lives in, i.e., weather and pollution levels
4. Feature extraction
5. COPD exacerbation modelling which we base on the prediction of the probability of an exacerbation event

We are investigating the following prediction algorithms:

1. Random forest
2. AdaBoost with Decision Trees
3. Hidden Markov models and variants
4. Neural networks for sequence analysis


Catchment Area Information
COPD Patients
Feature Generator
COPD Excaerbation Model
Probabilistic. Assess probability that patient will have an exacerbation in specific time period using full joint feature distribution.
Exacerbation probability distribution for next 3 days
Prominent factors contributing to probability

Summary:

- High-level aim: medical event prediction
- Data used: medical time-series, environmental time series, demographics
- Mentioned prediction algorithms: tree-based methods, probabilistic methods, neural networks

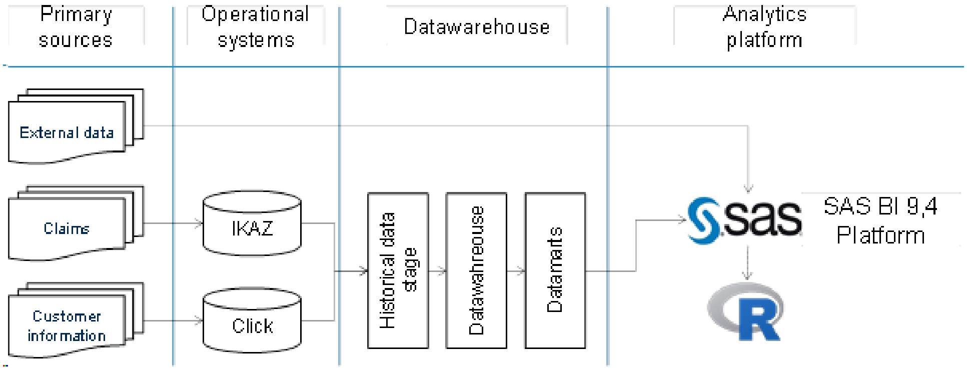
Programing languages and/or proprietary software:

- Python

Pilot 5 – Heart Failure (WP2) **Partners: EMC**

Description of the prediction problem:

Within the HF pilot we are using databases containing a large number of HF patients and their many comorbidities. Our key KPIs is a reduction in the number of hospitalizations and we are planning on using machine learning approaches to identify those comorbidities that have the strongest correlation with number of hospitalizations. Based on identified comorbidities we will design an intervention that will be tested in a prospective study on patients that fit our inclusion criteria. For details on the architecture for data storage and analysis see the below figure:



Considered machine learning solutions:

We are planning on using different types of machine learning solutions to investigate which ones are the most informative for our problem. Some of the machine learning approaches we will apply are:

- random forest
- KAGGLE
- LASSO
- elastic net.

Summary:

- High-level aim: HF comorbidity analysis to reduce number of HF related hospitalizations
- Data used: Hospital patient databases, Health Insurance company database
- Mentioned prediction algorithms: tree-based methods, probabilistic methods

Programing languages and/or proprietary software:

- R
- SAS

Pilot 6 – Prostate Cancer (WP3) **Partners: PHI**

Description of the prediction problem:

Surgery is one of the main treatment options for prostate cancer today. There are multiple aspects that need to be considered when planning for the removal of the prostate. On the one hand, the oncological control of the tumor is most relevant to ensure as much as possible that all cancer has been removed and does not return during follow-up. On the other hand, the aggressive removal of all prostate structures including nerve bundles etc. will likely lead to poor functional outcomes like urinary incontinence of sexual dysfunctions. Consequently, the appropriate balance between oncological and functional surgery outcome is of utmost relevance to the patient. We aim to support the decision-making process of how the surgery should be performed in order to provide the most optimal balance between tumor control and urological function after treatment; for this we will provide multiple risk models based on the integration of heterogeneous data sources like demographics, laboratory, imaging, histology and ultimately genomics.

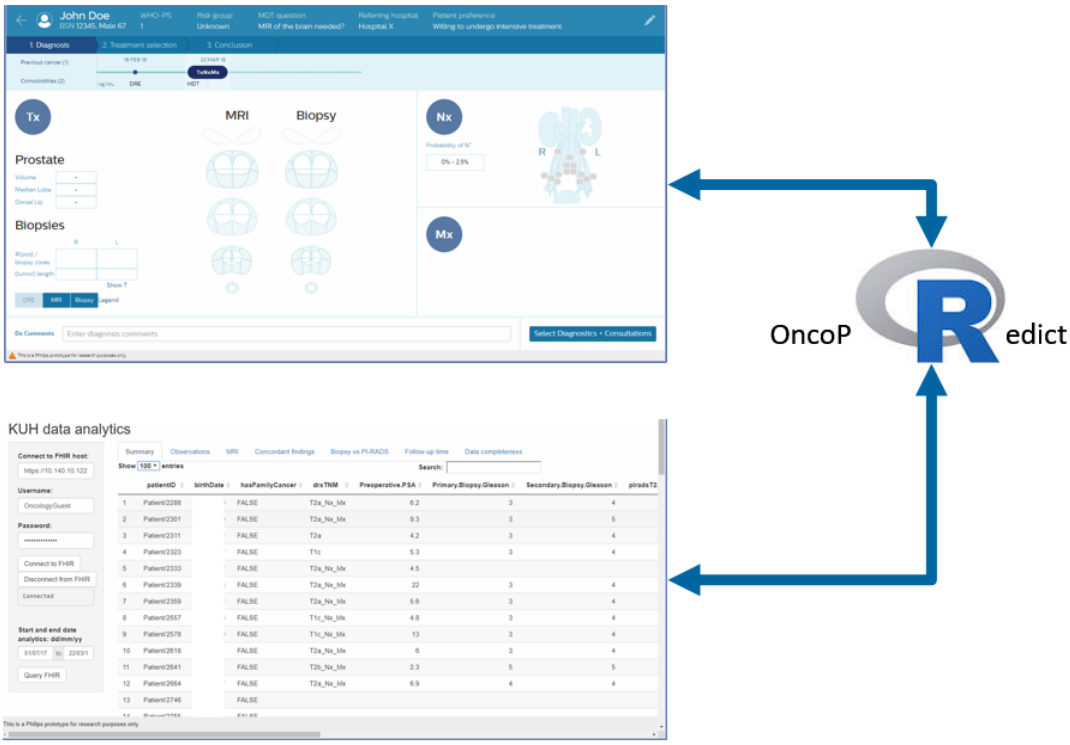
Considered machine learning solutions:

The current predicted risks are detailed below, while more risk models may follow:

- Pre-surgical risk of post-surgical adverse prostate cancer pathology (i.e., pathology Gleason>=7)
- Pre-surgical risk of post-surgical advanced extent of disease (i.e., pathology disease stage >=pT3a)
- Pre-surgical risk of the presence of tumor infiltrated lymph nodes.

Models that predict risk of urinary incontinence and sexual dysfunction will follow. The output of these models is the risk to experience the relevant adverse event.

We have implemented an online learning framework to update an initial risk model with prospectively collected heterogeneous patient data. A random forest classifier was used as the prediction model in all learning strategies. Feature selection was performed based on the impact of each feature on the internally computed accuracy of the model (accuracy of model after random selection of feature at each decision tree). Note that although the RF classifier has a built-in feature selection (by prioritization), feature selection still improves the model slightly. The learning framework is connected to a clinical data dashboard which contains data elements from various medical sources in a structured way.



	patientID	birthDate	hasFamilyCancer	drxTNM	Preoperative.PSA	Primary.Biopsy.Gleason	Secondary.Biopsy.Gleason	piradsT2
1	Patient/2288		FALSE	T2a_Nx_Mx	6.2	3	4	
2	Patient/2301		FALSE	T2a_Nx_Mx	9.3	3	5	
3	Patient/2311		FALSE	T2a	4.2	3	4	
4	Patient/2323		FALSE	T1c	5.3	3	4	
5	Patient/2333		FALSE	T2a_Nx_Mx	4.5			
6	Patient/2339		FALSE	T2a_Nx_Mx	22	3	4	
7	Patient/2359		FALSE	T2a_Nx_Mx	5.6	3	4	
8	Patient/2557		FALSE	T1c_Nx_Mx	4.8	3	4	
9	Patient/2578		FALSE	T1c_Nx_Mx	13	3	4	
10	Patient/2618		FALSE	T2a_Nx_Mx	6	3	4	
11	Patient/2641		FALSE	T2b_Nx_Mx	2.3	5	5	
12	Patient/2664		FALSE	T2a_Nx_Mx	6.9	4	4	
13	Patient/2746		FALSE					

The structured data is used to execute the initially implemented risk model(s). Structured data is defined as the input variables that are available in a structured format, i.e., not in form of a variable within a medical report but in the form of a field in a database. So, the value of the variable does not need to be extracted from a report. Any prospectively collected patient data is used to update the initial model within the implemented learning framework. Over time, the initial risk model will adapt to the characteristics of the local patient population.

Summary:

Support of the surgery strategy decision making

Implementation and presentation of risk models to balance oncological vs function control

Implementation of an online learning framework to prospectively model heterogeneous sources of data against patient relevant outcomes

- High-level aim: medical risk prediction
- Data used: medical time-series
- Mentioned prediction algorithms: tree-based methods

Programing languages and/or proprietary software:

Proprietary software

- miProstate (clinical data dashboard)
- OncoPredict (prediction learning framework)[1]

[1] The clinical dashboard miProstate is an HTML5 implementation with a FHIR database. The OncoPredict is a client-server application running from the R console and is implemented as a RESTful API such that the clinical dashboard can

Pilot 7 – Lung Cancer (WP3) **Partners: NCSR-D, A**

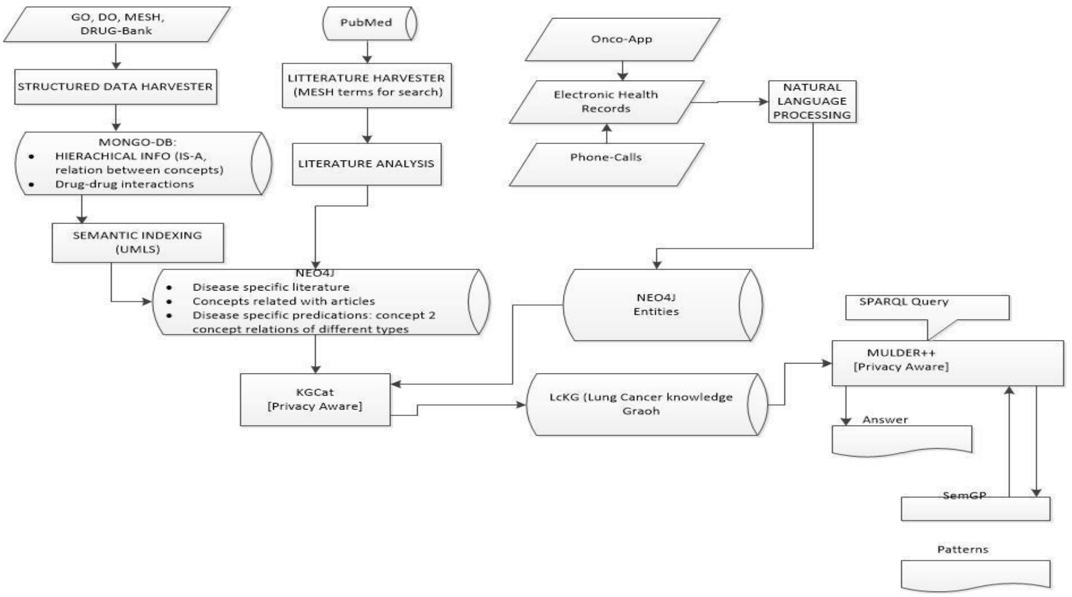
Description of the prediction problem:

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

Considered machine learning solutions:

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

1. Pattern discovery on open source data:
   a. Metamap (https://metamap.nlm.nih.gov/) performs named entity recognition

connect to the OncoPredict server. The OncoPredict server also has a direct connection to the FHIR database such that statistics analysis can directly be performed. The OncoPredict server also runs a Shiny dashboard for interactive statistical analysis of the FHIR data.

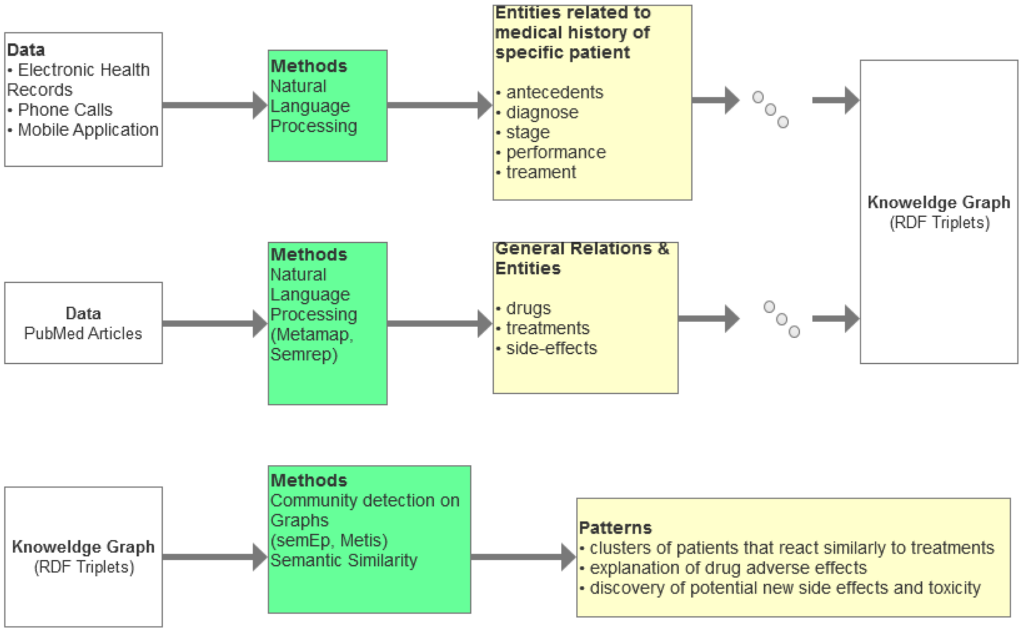
The logo of R is included to let the reader know that the wonderful world of R is being used, and the R becomes part of the OncoPredict world.

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:



Data
• Electronic Health Records
• Phone Calls
• Mobile Application
Methods
Natural Language Processing
Entities related to medical history of specific patient
• antecedents
• diagnose
• stage
• performance
• treament
Knoweldge Graph
(RDF Triplets)
Data
PubMed Articles
Methods
Natural Language Processing (Metamap, Semrep)
General Relations & Entities
• drugs
• treatments
• side-effects
Knoweldge Graph
(RDF Triplets)
Methods
Community detection on Graphs
(semEp, Metis)
Semantic Similarity
Patterns
• clusters of patients that react similarly to treatments
• explanation of drug adverse effects
• discovery of potential new side effects and toxicity

Summary:

- High-level aim: treatment effectiveness analysis, drug adverse effects prediction through link prediction in the knowledge graph.
- Data used: medical time-series, text documents
- Prediction algorithms used: natural language processing, community detection algorithms, semantic similarity measures

Programing languages and/or proprietary software:

Programming languages

- Python
- Java
- GCC version 5.x or higher
- Scala version 2.11.X

Software and frameworks

- Federated SPARQL Engine MUDLER++
- Virtuozzo: https://searchservervirtualization.techtarget.com/definition/Virtuozzo

- Docker containers, Linux, MongoDB, Neo4j
- Metamap for named entity recognition in bio-medical articles https://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html
- Semrep for relation extraction in bio-medical articles https://semrep.nlm.nih.gov/
- C-Likes (a software tool developed by UPM to extract information from Electronic Health Records)

Pilot 8 – Breast Cancer (WP3) **Partners: IBM**

Description of the prediction problem:

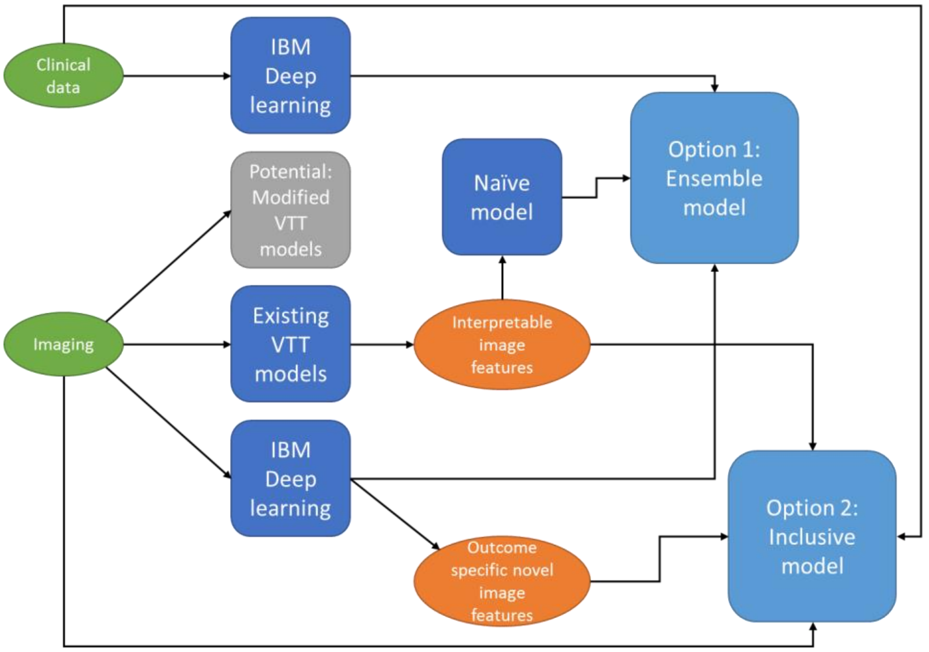
The pilot will analyse mammograms, ultrasound and MRI images along with structured clinical data and information extracted from pathology reports to automatically predict patient response to breast cancer treatments, specifically neoadjuvant treatments. The data was collected over the last several years in the databases of CUR, and will be made available offline to the processing collaborators IBM and VTT, through a VPN that allows access to a local server at CUR.

In summary, given images and clinical information, the models will retroactively predict the probability of success for each patient had they received neoadjuvant treatment. These models will allow evaluating the ability to make personalized treatment decisions rather than following global population guidelines and will allow assessing the economic effect of such protocols.

Considered machine learning solutions:

VTT will apply existing models that are immediately interpretable to the medical images.

IBM will apply deep-learning to the imaging data, extracting the most relevant features. Concurrently, a model using clinical features will be applied, and finally a model combining all available data will be applied (potentially also including the outputs from the VTT models)


Clinical data
IBM Deep learning
Potential: Modified VTT models
Naïve model
Option 1: Ensemble model
Imaging
Existing VTT models
Interpretable image features
IBM Deep learning
Outcome specific novel image features
Option 2: Inclusive model

Summary:

- High-level aim: treatment effectiveness analysis
- Data used: MRI images, longitudinal MG images, US images, clinical data
- Mentioned prediction algorithms: deep neural networks, traditional image analysis approaches, machine learning algorithms

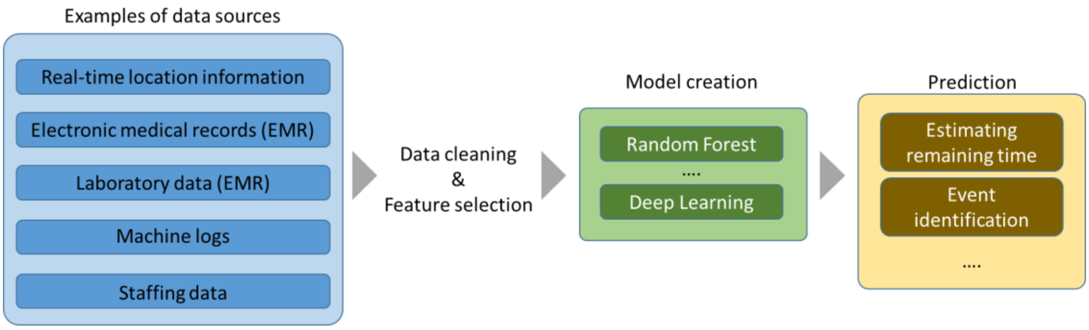
Programing languages and/or proprietary software:

- Python
- Java
- SPARK
- Docker

Pilot 9, 10 & 11 – Hyper-acute workflows, asset management (WP4) **Partners: PHI, TUE**

Description of the prediction problem:

This pilot focuses on monitoring and characterizing workflows within a hospital using multiple data streams available within a hospital. A workflow typically consists of all the processes that get triggered when a particular patient arrives at the Emergency Department. For example, when a stroke patient arrives at the emergency department of a hospital, the patient needs to go through triage, have a CT scan performed and blood tests taken. Once the appropriate tests have been performed, relevant care providers (e.g. neurologist, radiologist, etc.) analyse the available data and decide on the correct form of treatment. Multiple data streams are used to make various predictions about the care pathway, such as automatic prediction of where a patient is within a particular care pathway or how long a particular part of the care pathway will take to complete.



Considered machine learning solutions:

Data will arrive from multiple sources, e.g. real-time location information, Electronic Medical Records, Laboratory data, Machine logs, Staffing data, etc. Individual data sources will be cleaned and transformed. The next step will be to create a pipeline that integrates the data at the right level of granularity. The integrated data will be fed into feature selection algorithms and subsequently be used to develop predictive models. This will involve techniques such as random forest, deep learning (e.g. LSTM). Tests will be performed to select the most appropriate model for eventual deployment in the pilot.

Summary:

- High-level aim: Characterize events (e.g. detect, predict times) using multiple data sources available within the hospital
- Prediction algorithms: Random forest, SVM, LSTM

Programing languages and/or proprietary software:

- Python
- Scala/ Java

2.3. Task 1.3: Complex real-time event detection

2.3.1. Landscape of problems and approaches:

Problem landscape.

Almost in any process there is a need to be notified in some specific situations (so called situations of interest). It can be any situation that might request an action, like alerting a patient to take a medicine. Such a situation can be defined in a complex way, e.g. based on the combination of events from various streams. In the following we guide you through the different cases that might be relevant for your pilot. Table 7 allows to specify the types of notifications per pilot.

Type of notifications

There are three types of notifications

- Information sharing
  - Information are "just" shared – no action needed
- Warning
  - Awareness about a potentially "alarming" situation is created – no immediate action required, but it can follow
- Alarm
  - Critical information is shared - need for an immediate reaction
  - It can be divided into:
    - Automated reactions (the notification system can change the state of the patient related system, e.g. patient monitoring system)
    - Manual reaction (user can get "only" an alarm)

Table 7. Types of notifications and alerts to be issued

Pilot name	Need for notification services			
	Notification	**Warning**	**Alarm (automated / manual reaction)**	**Other**
Gestational diabetes (pilot 3)	Traffic light system – (red if above given value, green if normal, and orange if below). Risk level based on historical evolution (to medical team)	If red notification issued two times in one day. If the risk level increases too much.	If red notification more than 2 times in one day, issue an alarm. If the risk level arrives to dangerous levels.	-
COPD/Asthma (pilot 4)	Prediction of patient symptoms is performed. A 4-tiered score shows the severity of the COPD symptoms.	If the predicted patient symptoms are in the two top severe states (red and black), with a probability lower than the alarm threshold (currently set as 0.4)	If the predicted patient symptoms are in the two top severe states (red and black), with a probability higher than the alarm threshold (currently set as 0.4)	
Pilot 6	Likelihood score, and if available combined with a recommendation from the EAU or NCCN guidelines			
Lung Cancer (pilot 7)				No notifications or alerts are issued at this point. The clinicians access the Pilot application to observe statistical analysis, and various patterns around drug effect and patients.
Pilot 9		Excessive time for particular workflow step	Time exceeded for particular workflow step	
Pilot 11		Assets below required level	Asset out of bounds	
Pilot 12	Scores of top-ranked findings in search result-set related to the query case			

Type of situations of interest

There are three main types of situations of interest (to react on):

a) Simple
- Based on the value of parameters, e.g. if a parameter exceeds a value

b) Trends

- Based on the trends in the value of parameters, e.g. if a parameter increases/decreases fast
- It can be divided into:
  i. Time-window based (considering trends in measured values based on a period of time)
  ii. Frequency based (considering trends in several consecutive repetitions of measurements)

c) Complex

- Based on some complex combination of the value of parameters
- It can be contextualised with:
  iii. the values of several parameters (multi parameters)
  iv. past data (historical context)
  v. usually it is a comparison with the average value from the past, e.g. if the current value of a parameter is within the + - 10% range of average from a past period

Table 8. Types of complex events to react on

Pilot name	Type of situations of interest			
	Simple	**Trends (time-window / frequency based)**	**Complex (multiparameter / historical context)**	**Other**
Gestational diabetes (pilot3)	If the value of glucose sugar is above the predefined threshold	If the value of glucose sugar is increasing continuously in several consecutive measurements (or during one day) Glucose level are high in a time window for a patient with tendency to low levels	If the value of glucose sugar is above a predefined threshold and the value of blood pressure is high (increased) Glucose level show patterns shared with previous patients that developed issues	-
	When user takes 5th measurement of glucose today -> display motivational message	Each day at 9 am, 12 pm, 3 pm, 5 pm and 7 pm -> issue notification (reminder) to take glucose measurement	Doctors are alerted when they go to portal about patients whose glucose readings are above normal values (if around 30% of total glucose readings are above treshold) -> Medical staff can then issue notification with which they call patients to come to check-up	
COPD/Asthma (pilot 4)	-	-	If the predicted patient symptoms are in the two top severe states (red and black), with a probability higher than the alarm threshold	
Pilot 6	If the value of the likelihood is above the threshold defined by the EAU or NCCN guidelines			
Pilot 9, 10	If time exceeds a particular threshold		If patient enters a particular location and lab results are completed	
Pilot 11	If no. of assets fall below a particular level		If asset enters an area before being disinfected	
Pilot 12				If no similar case is found during the search the system does not contribute

Type of event processing

According to the literature[2], there are two main types of event processing, please indicate those in Table 9 according to the following scheme:

- Filter – selection of some of the values
- Transform – complex transformation, as presented in Figure 1

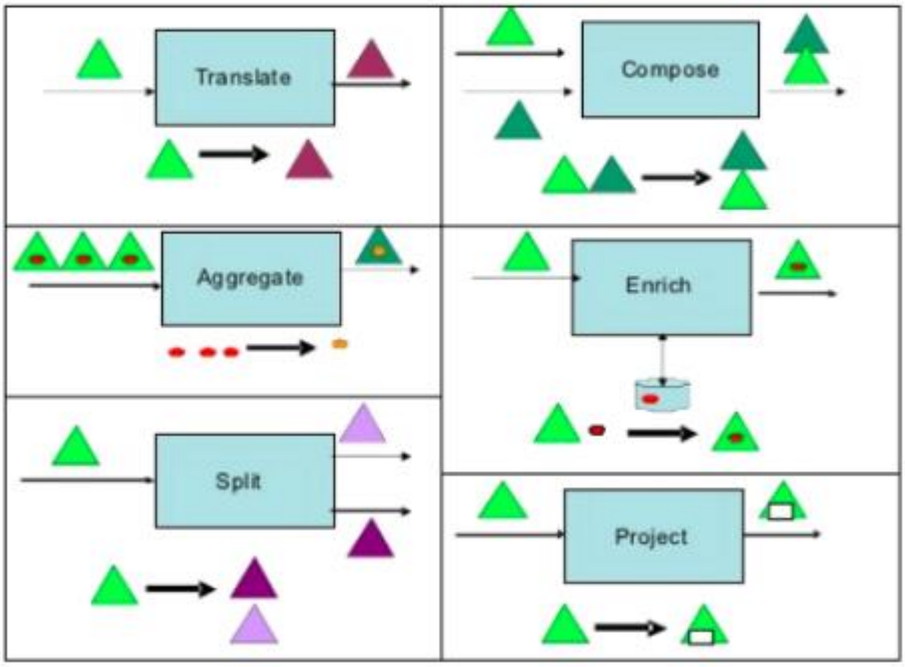


Table 9. Types of event processing actions

Pilot name			
	Filter	Transform	Other
Gestational diabetes (pilot3)	Only the values in a range are valid	The average value of several measurements should be sent to e.g. medical doctor	Enrich – Add model predictions information, together with historical data.
COPD/Asthma (pilot 4)	-	The input patient symptoms are enriched with the weather, pollution and pollen data and then projected into the future using the learned models.	-
Pilot 6	Filter patients of which a required value is unavailable for calculating the likelihood	Percentage calculation and normalization	
Pilot 9, 10	Tag spends more time in a location than a specified threshold	Reconstruct new event trace by aggregating RTLS and EMR data streams	-
Pilot 11	Tag spends more time in a location than a specified threshold	-	-

[2] https://euc-word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en%2DUS&rs=en%2DUS&WOPISrc=https%3A%2F%2Fdocs.philips.com%2Fpersonal%2Fsupriyo_chatterjea_philips_com%2F_vti_bin%2Fwopi.ashx%2Ffiles%2F45bd3c0f33854c5492f75bd60ea5c780&wdEnableRoaming=1&mscc=0&wdODB=1&hid=58B4D59E-303E-0000-6324-032F587C7DF1&wdorigin=ExternalSite&wdOrigin=DocLib&wdHostClickTime=1556004514351&jsApi=1&NewSession=1&corrid=15f42e0e-1b88-441a-a47d-a6497b069e3c&usid=15f42e0e-1b88-441a-a47d-a6497b069e3c&instantedit=1&wopicomplete=1&wdRedirectionReason=Unified_SingleFlushFallback

Event sources

Provide the source of events that can be used for defining situation of interest in Table 10.

Table 10. Event source during complex-event processing

Pilot name	Stream name	Contents of stream	Stream velocity	Description of the stream
Gestational diabetes (pilot 3)	Glucose monitoring	Glucose levels (variable-value)	4 times per day	The patients will use a glucometer connected to a mobile app to report their blood levels. The mobile app will send the values form the stream into the computing infrastructure.
Gestational diabetes (pilot 3)	Food monitoring	Food intake (variable-value)	Every time the patient eats	The mobile app also allows the sending of food data. The patient needs to input manually this data into the app, so its availability will be much sparser.
COPD/Asthma (pilot 4)	Patient self-recorded symptoms	Symptom Score	Every time the patient inputs their symptom score; between daily and weekly.	This is a single score taking the value 1, 2, 3, 4 in order of the severity of the COPD symptoms. It is retrieved from the patient inputting their symptom score via the MyCOPD app. This is the primary stream.
COPD/Asthma (pilot 4)	Patient self-recorded detailed evaluation score	CAT Score	Every time the patient completes the questionnaire; between weekly and monthly.	This is a single score taking any integer value between 0 and 40, and representing the self-assessment of the patient's recent symptoms and wellbeing. It is retrieved from the patient completing the questionnaire via the MyCOPD app.
COPD/Asthma (pilot 4)	Weather and pollution data	weather, atmospheric pollution and pollen exposures	Variable (e.g. 6-hourly, daily)	For each patient symptom score or CAT score event, corresponding weather, atmospheric pollution and pollen exposures will be extracted from the data provided by Copernicus Atmosphere Monitoring Service and Met Office. These will be retrieved at patient's geospatial location, using the event time stamp.
Pilot 6	Prostate surgery decision making	Treatment decision	Weekly	Every patient is discussed during the tumour board meeting for optimal treatment decision making
Pilot 9, 10, 11	RTLS data stream	Location and timestamp	Up to every 1.5s	Location data of a tag is reported every 1.5 seconds when the tag is in motion. The update frequency goes down if there's less movement.

Event processing engine

Table 11. Event processing engine

Pilot name	Engine	Comment
Pilot 6	NA	No complex engine is used to store the data
Pilot 9, 10, 11	Drools	Used to execute rules

[1] https://www.slideshare.net/opher.etzion/vldb-2010-event-processing-tutorial-5207547

2.4. Task 1.4: Processing of large structured / unstructured data sources

This section outlines how the pilots handle their data in terms of acquisition, cleaning, integration and analysis and processing methods.

Pilot 1: Comorbidities (WP2)	**Partners: OPTI, ITI, PHI**
Description of the prediction problem	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 e.g. previously unknown links between disease groups.
Data acquisition	The pilot will analyse Electronic Health Records from the Valencian Region Community. These data will be provided in bulk from the current databases per year (initially, 2015). 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 an Ethics Committee and the data exportation process is granted. Then, data is securely transferred to the Incliva premises. 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 upload into a PostgreSql database. For each patient the following set of information is provided: • Personal information • Family background • Diagnosis • Primary and Secondary care visits • Hospital discharges • Hospital urgencies • Health habits **What technologies (open source/ propriety) were used for the big data acquisition?** EHRs are stored in the Valencia Region Health Systems using relational databases. They provide us with a dump of the relevant tables of such system. This data then was uploaded to a PostgreSQL database **Why those technologies were selected?** As they original data source uses a relational database, we consider that to replicate that approach is the most convenient choice. Additionally, relational databases are suitable for performing analytical tasks. **What were the alternatives technologies and why they were not selected?** We consider the use of non-relational databases base on the "big table" paradigm such as Hive or Cassandra. However, in this stage we consider that there was no need as the amount of data was not huge enough.

Data cleaning	Next, we describe the data procedures carried 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 (from 30 to 35 years) and small towns are aggregated into wider geographic areas. Prescription data is analyzed and a diagnosis for such treatment is associated to the patient. This transformation helps further analysis. **What technologies (open source/ propriety) were used for the data cleaning?** We run specific scripts using R and Python to perform the aforementioned transformations. **Why those technologies were selected?** Data scientists involved in the pilot are proficient in such languages and the languages also provide ready-to-use libraries for several cleaning tasks. **What were the alternatives technologies and why they were not selected?** We do not consider alternatives, but for instance using an ETL framework will be an over-engineering approach for the tasks at hand.
Data integration and analysis	Each partner will define its own data integration procedure from the original data sources to carry on its specific analysis. Overall, each partner will query the original database to generate several datasets with the relevant information for analysis. For instance, ITI will store such datasets in a distributed Hive database, to improve query performance, and then upload in memory to a Spark cluster. In this memory cluster, specific ML algorithms for clusterization (as described in T1.2) will be executed. The output of the analysis will be stored in the Incliva premises. **What technologies (open source/ propriety) were used for the big data integration?** Selected technologies are Hadoop HDFS for integrating and storing the cleaned data and Apache Spark for analysis. Specifically, we use Spark data structures, as dataframes, to prepare the data for further analysis. Hive **Why those technologies were selected?** We selected these technologies because of their distributed processing nature. Using Spark data integration and analysis process could use several machines. The main benefit is the use of the combined amount of RAM memory, allowing bigger datasets and avoiding the continuous storage of intermediate results in physical hard drives. Hadoop HDFS stores transparently the files in several machines in a replicated way. This approach guarantees a high availability of data and resilience. As an additional advantage, Spark node could connect directly to several HDFS nodes for faster retrieval of the data. **What were the alternatives technologies and why they were not selected?** Though we use Hadoop HDFS for storage purposes, we discarded the use of Hadoop for data processing. The main drawback is that Hadoop is slower than Spark as it heavily writes intermediate results to disk. We also consider the Pandas framework for data structures, but our initial tests using Spark Dataframes showed better results. Cassandra and Elastic were also considered.
Considered machine learning solutions	Predictive models will be built using algorithms such as random forest, SVM and deep learning. These models will be used to not just identify particular events within the care pathway but will also be used to predict the remaining time for certain paths of the care pathway. **What technologies (open source/ propriety) were used for the big data analysis?** According to our requirements, selected machine learning technologies are Scikit, Scipy, SparkML and Pytorch. These technologies are used to discover groups of diseases. We consider this technology stack will support future requirements of the pilot. **Why those technologies were selected?** Currently, Scipy and Scikit are the “de-facto” standard in the ML community for implementing algorithms or prepare data for analysis. Regarding, SparkML and

	Pytorch, we selected them because their support to distributed computation. Scikit and Scipy complement these two frameworks, because not all the statistical functions and machine learning algorithms are currently supported by SparkML and Pytorch. **What were the alternatives technologies and why they were not selected?** Tensorflow is also a quite popular ML framework, but it requires low-level programming knowledge and, as a consequence, models implementation requires more effort. Keras is a wide popular framework for dealing with this issue providing a user-friendly layer on top of Tensorflow. However, There is a trend in the ML community to move towards Pytorch and this framework has increased its popularity and capabilities.

Pilot 2: Kidney disease (WP2)	**Partners: DFKI**
Description of the prediction problem	The ultimate aim of this pilot is to provide a data analytics system based on data by Charité Berlin, that enables to predict the health state of patients in the immediate/short-term as well as in the long-term. We believe that by achieving such ambitious goals it would be possible (a) to reduce the number of visits of a patient to the hospital, (b) to reduce the number of (re)hospitalizations, (c) to reduce the number of kidneys rejections, (d) to increase the lifespan of a transplanted kidney, (e) to reduce costs due to health treatments, and (f) to provide treatment options based on the effectiveness of the treatment, the side effects of the treatment, the impact of the treatment in a patient's QoL, and the cost of the treatment. In other words, by achieving such ambitious goals, it would be possible to increase the patients' QoL.
Data acquisition	The data is acquired at the Charité hospital and the AOK insurance company. At the Charité, the data assumes the form of EHRs (with structured, semi-structured, and unstructured formats depending on the attribute/field of the data) that are stored in the TBase relational database. At the AOK insurance company, the data has a proprietary insurance company form and is stored in a RDBMS. **What technologies (open source/ propriety) were used for the big data acquisition?** For the first stage of the use case we focus on the usage of TBase. TBase is a relational data base, based on MSSQL. To explore and extract the data, we use a Python environment to connect with the database and apply queries to create a target cohort. **Why those technologies were selected?** At this stage of the project this solution is sufficient. Also, for data science propose, in the phase of exploration, Python provide many tools that make the task more easily. **What were the alternatives technologies and why they were not selected?** There are more complex methods for acquisition, for example the creation of a distributed system, with Apache Flink or Spark, to start collecting the data and the process of exploration and analysis. However, for the first stage and test, such an architecture was not necessary. Regardless, it could be an option for next stages of the projects. With a distributed system, we will be more capable to work with more complex structures, getting data from different sources, applying complex event processing for monitoring and analysing the data with ML models.

Data cleaning	This project will devote considerable attention to data cleansing since the quality of data has a huge impact on the accuracy of the prediction models. As so, we aim to give special attention to the detection of anomalies and outliers, the identification of duplicates (if any), and the identification of extreme bias. We will also take into consideration the case of missing values since, in the medical domain, no one can just simply disregard records – patients' datapoints – just because they have one missing value. Besides the aforementioned tasks and others to improve data quality, we will also transform the data (e.g. rescaling of data, standardization of data, normalization of data, or binarization of data) to turn it amenable to best expose the structure of the underlying problem according to the selected ML algorithm, and to avoid biasing the predictions as much as possible. Within the stage of data cleansing, issues such as data anonymization of any private and sensitive information will also be taken into account. **What technologies (open source/ propriety) were used for the data cleaning?** Referring to the section above regarding data acquisition, the selected environment is Python working with Anaconda and including libraries as Numpy, Pandas and Matplotlib for visualization. **Why those technologies were selected?** This environment was chosen for data cleaning because the facilities that it provides when you are working in different machines. Simple installing Anaconda, you can start working with python and writing Jupyter Notebooks through the Browser. Also, for data science propose, in the phase of exploration, Python provide many tools that make the task more easily. **What were the alternatives technologies and why they were not selected?** R environment could also be chosen for the exploration and cleaning, but for propose like applying more complex Machine Learning algorithm in future stages, Python result to be a good choice, for the many libraries available that we can get and the large community.
Data integration and analysis	This project will not focus on data integration due to the restrictions that the partners have to follow, which are mainly imposed by privacy issues. Nevertheless, it will try to integrate the results of the two data analysis pipelines, that will be implemented, into an interactive visualization dashboard. The data analysis pipeline encompasses several modules that will support (a) connecting to a data source (TBase relational database) at the Charité, (b) performing the cleansing of data, including its pre-processing/transformations, (c) performing complex-event processing (CEP), (d) performing NLP, (e) running a set of prediction models over the transformed dataset, and (f) storing and documenting the achieved results. We also aim to implement data lineage, which is of utmost importance to ensure that traceability, reproducibility, explainability, and transparency will be available to allow decision makers (e.g. doctors) to trust in the system. **What technologies (open source/ propriety) were used for the big data integration?** At this point we use retrospective data from the patient database of the hospital. We simply select relevant patients and data to create a cohort to train our models on. **Why those technologies were selected?** N/A **What were the alternatives technologies and why they were not selected?** N/A

Considered machine learning solutions	We aim to apply supervised and unsupervised learning methods to perform predictions as well as a hybrid approach that uses both methods. As fundamental algorithms, we envision to use linear and logistic regression algorithms, clustering algorithms (e.g. KNN), and decision trees algorithms (e.g. Random Forests). On a later stage of the project, we aim to extend the aforementioned array of (supervised) learning algorithms with unsupervised learning algorithms, that would allow to explore a wider range of features including those we know too little about in terms of counterpart relevance, in order to reveal patterns and insights that are hidden in the data. Depending on the success of applying unsupervised learning methods, we aim to extend the implemented methods with hybrid prediction models that use unsupervised learning methods combined with supervised ones. **What technologies (open source/ propriety) were used for the big data analysis?** Mainly our methods are implemented using python in combination with existing libraries such as scikit learn and tensorflow **Why those technologies were selected?** Python is quite convenient to apply machine learning and provides a large range of useful libraries **What were the alternatives technologies and why they were not selected?** N/A

Pilot 3: Diabetes (WP2)	**Partners: HUA, NISS**
Description of the prediction problem	Gestational diabetes can generate long-term health problems for both the baby and the mother. Midwives are overwhelmed with work, and there are no standards for monitoring patients across hospitals, or country. The objective of the project is to improve the efficiency and frequency of blood sugar level monitoring by using both data from patients and information and experience coming from the midwives. This will allow healthcare professionals to focus on patients who are in greater risk, and reduce the visits to the hospital with patients who are having a non-problematic pregnancy.
Data acquisition	The pilot main stream of data is the sugar levels of the patients that will be measured with a glucometer, and send to analyse automatically via a mobile app. The mobile app also generates more data, namely the food intake of the patient. This data needs to be manually introduced by the patients, and therefore is less reliable. Finally, there is some information coming from the medical team, like abdominal circumference, ultrasound information, and/or the measures for the sugar levels that generated the gestational diabetes diagnose. **What technologies (open source/ propriety) were used for the big data acquisition?** Bluetooth as a standard communication protocol and Java/Kotlin application to accept data (Since data comes in bytestreams it needed to be filtered and adjusted to our data-model). **Why those technologies were selected?** Reliability, security and modularity. Bluetooth is safe and reliable, while providing maximum flexibility because we could use any sensor which has Bluetooth (which most digital sensors do have) and application could communicate with it, get the data from it, and whole system would function without problem. Java and Kotlin are both open-source, and they provide for a stable and well-tested environment. **What were the alternatives technologies and why they were not selected?** There are almost no alternatives to Bluetooth communication, since it's standardized and widely used.

Data cleaning	The data will be collected via applications specifically developed by the project. Therefore, the cleaning of the data has been reduced, since the data is structured from the beginning. Even with that consideration, there will be a Java application that will ensure that all the data is correctly formatted, and that there are no issues with it. Regarding missing data, the predictive system will be missing-values tolerant, since the patient can forget to take measurements of her blood sugar at some paint. This won't affect the system at any point, and it will continue monitoring the patient. **What technologies (open source/ propriety) were used for the data cleaning?** Java. **Why those technologies were selected?** Java is a programming language widely used in many different industries. It has proved to be robust, and well suited for working in real time environments. In this pilot, data is mostly clean from the start, due to the end to end nature of the project. Therefore, the organization of the data, and small cleaning tasks can be done efficiently with Java. **What were the alternatives technologies and why they were not selected?** Python was considered for the cleaning, but was decided against due to the small cleaning needed, and the increased robustness offered by java.
Data integration and analysis	The data will be collected on a continuous basis via the applications specifically developed by the project. Therefore, the data is integrated in the system from the beginning and handled following the GDPR rules Each new measurement will trigger a function that prepares the data necessary to apply the model (by querying the necessary history, and giving the format that the model will use), and will anonymize the data. This prepared data will then be sent for the predictions, and the results will be showed to the medical staff so that they can react accordingly. **What technologies (open source/ propriety) were used for the big data integration?** Java, Kafka, Flink, python, PostgreSQL. **Why those technologies were selected?** They are open source, highly reliable, regularly maintained, and offer the features that we require for the project. They interact with each other in a simple way, making the combination really well suited for the development of a new infrastructure **What were the alternatives technologies and why they were not selected?** R instead of Python. It was decided against due to availability of libraries, and familiarity with the language. Other SQL relational databases instead of PostgreSQL. PostgreSQL was decided due to familiarity. RabbitMQ instead of Kafka. Kafka was chosen since it is durable message broker where applications can process and re-process streamed data on disk, it has simpler routing approach, more light-weight for our solution, and it provides horizontal scaling - meaning many machines can communicate to it, easily and safely, while RabbitMQ is designed for vertical scaling which is not needed in our pilot.

Considered machine learning solutions	We are considering two main methods to have a standard and uniform approach for all monitoring needs with a particular focus on certain thresholds:: • Use classification models (random forest, xgboost, SVM, deep learning, etc.) to assign to each patient a risk, and act accordingly. However, the main disadvantage of this method is that it requires the labelling from an expert team, and therefore it is more difficult to obtain the required data to follow up. On the same logic, we could use regression models to assign a scaled score to each patient, and use it to calculate the risk level of each patient. • Use the already existing information at the hospital to create a fuzzy rules system. This will allow for much further interpretation for the doctors, and give a wide range of options for them, while keeping it simple for the patients (by simplifying the thresholds). The other advantage of this approach is that it does not need each data point to be labelled individually, since it can be labelled using the rules that are on place on hospitals. **What technologies (open source/ propriety) were used for the big data analysis?** Fuzzy rules - Python **Why those technologies were selected?** We decided to use fuzzy rules as the first analytics method. It is intuitively more similar to the system currently being used in hospitals (hard rules), and therefore it will be easier for medical teams to understand and adapt to it. We used a python implementation to train the model. **What were the alternatives technologies and why they were not selected?** Other analytics methods – Python. They were not selected for the current version of the pilot. It was decided that fuzzy rules would be the initial model, therefore the rest of the models were postponed, but not discarded.

Pilot 4: COPD and Asthma (WP2)	**Partners: UNIS, MYM**
Description of the prediction problem	The aim is to predict the risk of a patient experiencing a COPD exacerbation event within a specified timeframe, based on the known influences of the event: a medical time series from the MyCOPD app, meteorological data and atmospheric pollution exposures at a patient's given location.
Data acquisition	The medical time series data will be provided by MYM. This includes information about the symptom score and CAT score (patient inputs), as well as demographic features for each patient. The environmental data will be derived at patient's given location and timestamp. This data will be retrieved from the Copernicus Atmosphere Monitoring Service and UK Met Office. Furthermore, Socio-economic data may be retrieved from the Office for National Statistics UK. **What technologies (open source/ propriety) were used for the big data acquisition?** Apache Spark is considered for data acquisition. This will include data quality checks and transformation rules. Mongo DB is considered for storage of data and is converted into target format. Each Application is isolated functionally so that we can mix and match the technologies easily (support extensibility). **Why those technologies were selected?** Being the application part of data analytics platform, Spark is the major player in data ingestion and distributed processing (both batch and real-time). Mongo DB has been considered as stable distributed data storage for big data. **What were the alternatives technologies and why they were not selected?** Hadoop HDFS supports storage of files in Parquet and Avro formats but it needs infrastructure setup and maintenance. Kafka will be considered for real-time data ingestion, if needed in the future. Apache NiFi was considered as it has good processing capabilities with ease of pipeline development but it will be complicated for the long run as it needs to be connected to other softwares like Apache Spark, Flink or Storm for efficient processing.

Data cleaning	We will require that any patient used in the production of the prediction data-driven model(s) has the following requirements: 1. The patient has an associated location 2. They have at least one symptom score measurements. As the timestamps in the medical time series are daily timestamps, we must aggregate some of the meteorological data, as the data is usually given at smaller intervals, e.g. 6-hour intervals. We also expect to eliminate anomalies by utilising known correlation between the symptom score and CAT score. **What technologies (open source/ propriety) were used for the data cleaning?** Apache Spark combination, Drools with Mongo DB, Python **Why those technologies were selected?** Ease of rule definition and distributed processing and persistence of valid and invalid records for further processing of analytics. As for Python: open source; contains efficient tools to perform data cleaning activities. **What were the alternatives technologies and why they were not selected?** Other technologies such as R and MATLAB were also considered. MATLAB is propriety software so we elected to avoid it. R doesn't integrate well with our other uses of Python within the project.
Data integration and analysis	The data sources will be integrated at patient level. Each entry in the demographics file has an associated anonymised patient ID, as well as every point in the time series. This will allow us to construct a consistent record of medical events for each patient. The environmental information will be extracted based upon the patient's postal outcode of their home address, as well as the relevant timestamps associated with the respective measurements in the time series. **What technologies (open source/ propriety) were used for the big data integration?** Apache Spark, Mongo DB, Python **Why those technologies were selected?** There are many Python modules available to process meteorological and pollution data available that easily integrate with the other tools we have developed. While Apache Spark and Mongo DB are used for big data processing and storage. **What were the alternatives technologies and why they were not selected?** Python integrates well with the other parts of the service, allowing for smoother integration activities later.
Considered machine learning solutions	We are investigating the following prediction algorithms: 1. Random forest 2. AdaBoost with Decision Trees 3. Hidden Markov models and variants 4. Neural networks for sequence analysis The decision tree-based algorithms above can both predict the likelihood of exacerbations, as well as produce a classification label if we so choose. **What technologies (open source/ propriety) were used for the big data analysis?** Python: TensorFlow, scikit-learn modules **Why those technologies were selected?** Versatile and powerful libraries which are capable of handling big data. **What were the alternatives technologies and why they were not selected?** Python integrates well with the other parts of the service, allowing for smoother integration activities later.

Pilot 5: Heart Failure (WP2)	**Partners: EMC**
Description of the prediction problem	There are currently 15 million patients living with HF in Europe, and over 23 million worldwide. Annual mortality risk in HF patients is 10-20%, higher than that due to various cancers. Because OMT is not adequately applied, HF patients can be characterized as 'revolving door patients', with frequent (re)hospitalizations, resulting in high costs for society. OMT is currently not applied in 70% of HF patients. In particular, patients with multiple comorbidities are at risk, including patients with diabetes, hypertension, vascular disease, COPD and other yet unknown comorbidities. Within the HF Pilot, we are using several databases containing large numbers of HF patients and their comorbidities. These databases are both Hospital and Health Insurance based systems. Description of prediction problems: 1. Our goal is to achieve a reduction in the number of hospitalizations and we are planning on using machine learning approaches to identify those comorbidities that have the strongest correlation with number of hospitalizations by exploring the data in the above mentioned databases. These identified comorbidities will be used to guide an intervention that is designed to reduce the number of hospitalizations. 2. We will develop a system called Multi Party Computation (MPC) that allows for the secure combined analysis of datasets residing in the hospital databases and health insurance databases. The key feature of this approach is that it allows for a secure combined analysis of two (or more) datasets that either owner cannot share with the other.
Data acquisition	We will use readily available data in existing databases at Achmea and Erasmus MC. Due to confidentiality, the MPC will work with generated datasets from both Achmea as EMC. In parallel to the proof-of-concept demonstration, the next steps towards operationalization of the MPC analysis will be investigated. **What technologies (open source/ propriety) were used for the big data acquisition?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already acquired before the start of the project and therefor no data acquisition was needed. **Why those technologies were selected?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already acquired before the start of the project and therefor no data acquisition was needed. **What were the alternatives technologies and why they were not selected?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already acquired before the start of the project and therefor no data acquisition was needed.
Data cleaning	The patient data will be cleaned with regards to our inclusion criteria for chronic HF. We will make sure the data is complete and normalized per feature For the MPC setting, the data cleaning will happen locally at each party (Achmea and EMC). The MPC solution will deal with data integration and apply a regression analysis on the patients that are available in both datasets. This will be done in another MPC solution, so that it remains hidden which patients are in the intersection of both datasets. **What technologies (open source/ propriety) were used for the data cleaning?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already cleaned before the start of the project and therefor no data cleaning was needed. **Why those technologies were selected?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already cleaned before the start of the project and therefor no data cleaning was needed. **What were the alternatives technologies and why they were not selected?** The data used in our pilot resides in databases present at EMC and Achmea. This data was already cleaned before the start of the project and therefor no data cleaning was needed.

Data integration and analysis	The data will be integrated on the patient level. We are making use of MPC to allow for secure data integration and analysis. This approach allows us to securely and in an anonymized fashion combine and integrate databases from different owners who do not wish to share their database with each other. The MPC solution makes sure that the data input of both EMC as Achmea remains private. A trusted third party (TTP) is added to the MPC scheme in order to speed up the MPC computations. Note that this TTP does not learn anything about the private data inputs of the other parties, but only assists in the computations. The data integration will be performed by a hidden set intersection, where no one learns the identity of the shared data records (patients) and yet a regression analysis can be performed on the intersection of the data. The outcome of the MPC protocol are the weights for the LASSO regression coefficients, so that all involved parties learns the relation between various life style parameters and the number of hospitalizations. **What technologies (open source/ propriety) were used for the big data integration?** MPC is used to integrate the data input from EMC and Achmea. **Why those technologies were selected?** MPC using Secure Lasso Regression was chosen as this combination was found to be most efficient with a large number of features. **What were the alternatives technologies and why they were not selected?** Linear regression and Ridge regression were also tested but Secure lasso regression was most efficient with large numbers of features.
Considered machine learning solutions	**What technologies (open source/ propriety) were used for the big data analysis?** LASSO regression was used to predict hospital admission out of > 4.000 features, based on an individual level. **Why those technologies were selected?** LASSO regression was used to predict hospital admission because it was the better technology for our data. **What were the alternatives technologies and why they were not selected?** Alternative technologies tested are random forest, KAGGLE and elastic net but they were less optimal as compared to Lasso regression for our data.

Pilot 6: Prostate Cancer (WP3)	**Partners: PHI**
Description of the prediction problem	The pilot captures prostate cancer related data that will be used to: 1. Derive primary treatment decisions in a multidisciplinary setting, 2. Derive treatment and VBHC related quality outcome measures, 3. Create decision models to improve functional outcome predictions after primary intervention, 4. Apply health economic modelling to test cost-effectiveness of the implemented big data technologies.
Data acquisition	The various data sources will be integrated at the patient level in the tumor dashboard. Data that is analysed by the R software is anonymized and provided with a database unique id when connecting and accumulating data from the FHIR database. **What technologies (open source/ propriety) were used for the big data acquisition?** R , Shiny **Why those technologies were selected?** R is open-source and freely available. Shiny is also open source, freely available and was selected for its ease of use to create dashboards to support clinicians in data mining **What were the alternatives technologies and why they were not selected?** Alternative would be to use python. R is preferred for the availability of statistical packages used to obtain insights from the collected data

Data cleaning	Data blending will be applied to the collected data to aggregate, filter or merge data. Feature engineering operations will be performed to create context relevant features (e.g. PSA density form PSA and prostate volume). Furthermore, feature values are validated to clinical context (e.g. PIRADS scores are always between 1 and 5, discrepancy between PIRADS scoring and biopsy outcome, etc) Data cleansing to automatically identify duplicates, outliers and anomalies. Feature value distribution analysis to identify features with extreme bias between treatment groups. Automated feature selection to select the optimal set of features for each prediction model. **What technologies (open source/ propriety) were used for the data cleaning?** The in-house developed shiny dashboard has a multiple visualisation to bring attention to incomplete, incorrect, inaccurate or irrelevant parts of the data. An example of such a visualisation is an interactive barplot showing counts of collected patients and related variables. Missing values e.g. are not counted such that the user knows how many and which patients need additional info. **Why those technologies were selected?** The visuals need to be simple and easy to understand for clinicians. **What were the alternatives technologies and why they were not selected?** Excel, not a practical tool for maintenance. SPSS, not easy to use for clinicians.
Data integration and analysis	An in-house developed R package is used to connect to the FHIR database of the tumor dashboard. The tumor dashboard is connected to 3rd party data source (EPD, PACS) and collects the relevant data into the FHIR database. Analysis is performed in the R environment on anonymized data. **What technologies (open source/ propriety) were used for the big data integration?** Patient data is stored in the FHIR database of the application. We have an in-house developed R package that is able to connect to the FHIR backend directly (with user authentication and HTTPS protocol). **Why those technologies were selected?** Ease of use. **What were the alternatives technologies and why they were not selected?** All data from different sources is collected by the tumor board application and stored into the FHIR database. No alternatives available.
Considered machine learning solutions	RStudio with R as the modelling framework **What technologies (open source/ propriety) were used for the big data analysis?** R with available statistical packages **Why those technologies were selected?** R has very well validated statistical packages and is intensively used by the healthcare community **What were the alternatives technologies and why they were not selected?** SPSS, Excel. Not the preferred utilities.

Pilot 7: Lung Cancer (WP3)	**Partners: NCSR-D**
Description of the prediction problem	The aim of the Lung cancer pilot is to improve the management of patients with lung cancer in order to improve their satisfaction, but also save substantial costs to the health budget. To accomplice that, we will discover correlations between the medical history of a patient, and the reasons for teleconsultation, as well as correlations between teleconsultation reasons and the diagnosis at the hospital. Finally all these pieces of information will be related to the adverse effects of drugs. The final goal is to reduce the duration of hospitalisation.

Data 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. The second source of raw data are the EHR of patients, which become available once the hospital releases them. **What technologies (open source/ propriety) were used for the big data acquisition?** Data harvesters have been built to access the Application Programming Interface (API) of PubMed/PubMed Central through REST. **Why those technologies were selected?** This is the access technology supported by PubMed. **What were the alternatives technologies and why they were not selected?** N/A
Data cleaning	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. Once EHR data is integrated in the knowledge graph, restrictions represented in the knowledge graph ontology are used to validate the completeness and consistency of the integrated data. Data that does not respect the restrictions is manually checked with the team of knowledge engineers and clinicians. **What technologies (open source/ propriety) were used for the data cleaning?** Design patterns from ontologies, SPARQL queries, and constraint validation **Why those technologies were selected?** The data integrated in the knowledge graph is described using ontologies whose properties are formally stated in design patterns. The satisfaction or not of the design patterns enables to detect inconsistencies and missing values in the data integrated in the knowledge graph by evaluating queries that express the design patterns **What were the alternatives technologies and why they were not selected?** Formalisms like Shape Constraint Language (SHACL) is commonly used to represent constraints and evaluate their satisfaction. Although SHACL could be used for expressing design patterns, the implementation of these design patterns in SPARQL enables for a more efficient execution as well as for scaling up to large knowledge graphs.
Data integration and analysis	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. **What technologies (open source/ propriety) were used for the big data integration?** We represented all the information, irrespective of its origin (I.e. be it open data, or information from the Electronic health records) as RDF triplets that constitute a big knowledge graph, thus representing the semantics. **Why those technologies were selected?** The semantic representation offers many advantages, such as the ability to represent complex relations (as they can be found between drugs, or between drugs and people). Moreover, complex queries can be formed in the SPARQL language on the knowledge graph. Finally, since the knowledge graph is based on logic, it can support reasoning. **What were the alternatives technologies and why they were not selected?** The closest alternative to the knowledge graph is to use a graph database, but then it poorer representation than description logics upon which the knowledge graph is based. Moreover, a graph database would not have a reasoning engine.

Considered machine learning solutions	At the level of raw data, NLP techniques are applied to extract named entities and relations. In the knowledge graph, clustering algorithms are implemented to discover correlations among the entities that may be patients, drugs, treatments etc. Also, the federated SPARQL engine MUDLER++, which scales well to big data, can be used to query the knowledge graph. **What technologies (open source/ propriety) were used for the big data analysis?** For mining open unstructured data we have used the SemRep and MetaMap to extract named entities (such are references to drugs, symptoms, patients etc.), and relations between them. **Why those technologies were selected?** They are industry standards for retrieving named entities and relations from biomedical texts. **What were the alternatives technologies and why they were not selected?** There are tools for named entity recognition based on Java (Stanford library) or Python (NLTK, and SpaCy) but there are not tuned to biomedical texts and thus have a very low accuracy.

Pilot 8: Breast Cancer (WP3)	**Partners: IBM, CUR, VTT**
Description of the prediction problem	The pilot aims to develop a radiomics system that uses deep learning algorithms on multi-modal big data in order to improve patient outcomes and reduce costs. The pilot will analyse mammograms, ultrasound and MRI images along with structured clinical data to automatically predict patient response to breast cancer treatments, specifically neoadjuvant treatments. Additionally, we'll try to predict cohorts for clinical trials towards next generation therapies. Within the pilot study, various models for the different tasks will be tested and scored. Then, we will ensemble the various models to further improve the prediction.
Data acquisition	Image data will be provided in bulk, with some structured clinical data embedded in the DICOM image header. Additional clinical information will be extracted from the EHR systems in CUR and restructured into patient level flat files. This will include the index date of treatment initialization, some summary clinical and demographic features prior to diagnosis, tumour properties, NACT treatment indications, surgery properties, and the outcome of the treatment. **What technologies (open source/ propriety) were used for the big data acquisition?** Pilot clinical data is extracted from Curie multi-purpose SQL repository called BioMedics. This repository includes aggregated information coming from all different sources (software used by doctors, medical records, handwritten events, etc). The extracted data is anonymized and stored in a CSV file. Pilot MG, US and MRI imaging data is extracted from Curie PACS system as DICOM files and then go through an anonymization process that is different per modality. The identifying DICOM tags are erased or replaced with non-identifiable content. Additionally, US images may also have some identifiable information at the edge of the figure, so Curie developed a procedure to replace that information in the image with non-identifiable patch. Then, the anonymized images are stored in the filesystem in an agreed per-patient structure. **Why those technologies were selected?** We wanted to prepare the data for analysis as fast as possible and concentrate on extracting the data and anonymize it for pilot usage. Thus, we chose to represent the curated anonymised clinical data in CSV file and the anonymised images in special structure on the file system. **What were the alternatives technologies and why they were not selected?** We considered creating a NO SQL repository based on some open source software e.g. Cassandra for the pilot data. We didn't choose that option because the focus of our research is the analytics algorithms and not data management. We thought that for the amount of data we have, it may be sufficient to have a simpler infrastructure for the data that is based on CSV and filesystem structure.

Data cleaning	We will exclude from the data cases with multifocal tumours, bilateral cancer, skin invasive cancer or inflammatory tumours. We will also exclude patients who have relapsed from a previous disease. Cleaning and filtering of image data will include removal of images with low resolution, images with obstructions, and may include removal of images with previous surgical indications. Clinical data will be examined to identify extreme outliers, and to identify features with extreme bias between treatment groups. Patients who are extreme outliers will be excluded from further analysis, and features that show extreme bias will be further analysed to check if the bias can be corrected or whether analysis must be stratified by such features. The clinical outcomes will be analysed for inconsistencies between the various data sources. The association between features and outcomes will be analysed to see the baseline predictive power available in the data. **What technologies (open source/ propriety) were used for the data cleaning?** The data cleaning is done by specific propriety code written for the purpose. We may also use Python plotting packages to visualize the features bias and outliers. **Why those technologies were selected?** As the data cleaning is something specific for our data and the specific task we need to perform in our pilot, we are obliged to create this specific code ourselves. **What were the alternatives technologies and why they were not selected?** There were no real alternatives as the data cleaning is specific to our data and pilot task.
Data integration and analysis	The various data sources will be integrated at the patient level. Each image file will be identified by an anonymized patient id, and this ID will be used to identify the rest of the clinical and outcome features. The analysis itself will create algorithms based on all the inputs in the relevant stratified data or using the reweighted data in order to provide results that are not biased by treatment assignment. The algorithms may be polyglot namely written in different programming languages (Python, java, C) and use different deep learning or other frameworks (Tensorflow, Pytorch). **What technologies (open source/ propriety) were used for the big data integration?** We will use the IBM Biomedical Framework (BMF) that creates configurable reusable pipelines and exposes them as REST microservices in Docker containers. The framework also enables transparent run of pipelines on a SPARK cluster where workers run in parallel and each worker runs on a separate GPU. It is doing that by automatic translation from a descriptive pipeline flow to efficient SPARK application that can perform multi-modal analytics and utilize analytics modules written in various programming languages and deep learning frameworks. The data model in BMF uses the HL7 FHIR object model. **Why those technologies were selected?** The proposed technologies are scalable and can easily grow from one node to as much as needed without changing the algorithm code. The architecture enables scale in all axis: x-axis (containers), y-axis (function partition), z-axis (data partition). By using the Apache SPARK open source, we get a fault-tolerant, distributed backend for robustly analyzing large datasets in a scale-out cluster. The use of Docker containers enables easy deployment in heterogenous environment. **What were the alternatives technologies and why they were not selected?** There is no alternative that provides all the required above but there are alternatives to some aspects. For inference distributed models, we could use tensorflow distributed model or pytorch distributed model, but then we would be locked to one specific deep learning framework.
Considered machine learning	We will utilize DNN algorithms as well as traditional image processing methods. The deep learning algorithms will be based on various convolutional neural network (CNN) and use architectures such as customized Inception ResNet V2 and modified U-Net architecture. For longitudinal image analytics we are using algorithms based on LSTM architecture. Traditional methods include algorithms such as Expectation Maximization (EM) for image segmentation. This can be used to extract numerical measures of the lesion in

	respect to the surroundings and location of the cancer tissue. All the features extracted from the various algorithms will then be ensembled using ML Models such as XGBoost. **What technologies (open source/ propriety) were used for the big data analysis?** We leverage the state-of-the-art deep learning frameworks such as tensorflow and pytorch and the state-of-the-art architectures such as ResNet, U-Net, LSTM. We also leverage the state-of-the-art ML algorithms such as Logistic Regression, Random Forest and XGBoost. **Why those technologies were selected?** We selected the state-of-the-art technologies suitable for our pilot algorithms in order to get the best possible accuracy and algorithmic performance. As the project progress, we'll continue to investigate new coming state-of the-art frameworks and networks. **What were the alternatives technologies and why they were not selected?** There are other alternatives for deep learning frameworks e.g. Caffe and other architectures but our experience shows that so far, the selected technologies give the best results.

Pilot 9, 10 & 11: Hyper-acute workflows, asset management (WP4)	
Description of the prediction problem	This pilot focuses on monitoring and characterizing workflows within a hospital using multiple data streams available within a hospital. A workflow typically consists of all the processes that get triggered when a particular patient arrives at the Emergency Department. For example when a stroke patient arrives at the emergency department of a hospital, the patient needs to go through triage, have a CT scan performed and blood tests taken. Once the appropriate tests have been performed, relevant care providers (e.g. neurologist, radiologist, etc.) analyse the available data and decide on the correct form of treatment. Multiple data streams are used to make various predictions about the care pathway, such as automatic prediction of where a patient is within a particular care pathway or how long a particular part of the care pathway will take to complete.
Data acquisition	Data will arrive from multiple sources, e.g. real-time location information, EMRs, Laboratory data, Machine logs, Staffing data, etc. **What technologies (open source/ propriety) were used for the big data acquisition?** Proprietary RF technology used to gather data from all tagged entities. Java was used to parse the incoming data. **Why those technologies were selected?** Proprietary RF technology considered most reliable and energy efficient and meets application requirements. Java considered robust and suitable for real-time applications. **What were the alternatives technologies and why they were not selected?** N/A
Data cleaning	Individual data sources will be cleaned and transformed. The precise technique used for cleaning will depend on the specific data source. For example, interpolation and noise reduction techniques will be applied to real-time location data to ensure missing or erroneous data points are adequately addressed. Various data imputation strategies will be tested (e.g. using decision trees, Bayesian networks, etc.) to address missing data in electronic medical records. **What technologies (open source/ propriety) were used for the data cleaning?** Java **Why those technologies were selected?** Robust and suitable for real-time applications. **What were the alternatives technologies and why they were not selected?** Others not considered due to familiarity with Java.

Data integration and analysis	Data will be transformed and different data streams will be combined in order to derive higher level semantics that describe particular events of interest within the target care pathway. Once higher level events have been detected, feature selection techniques will be used to ensure relevant features are used to train, test and validate the models. **What technologies (open source/ propriety) were used for the big data integration?** Python, MySQL **Why those technologies were selected?** Open source, highly reliable and regularly maintained. Also used widely in the industry. All required libraries were easily available. **What were the alternatives technologies and why they were not selected?** R. Not chosen due to lack of familiarity and availability of libraries.
Considered machine learning solutions	Predictive models will be built using techniques such as random forest, SVM, deep learning (e.g. LSTM). These models will be used to not just identify particular events within the care pathway but will also be used to predict the remaining time for certain paths of the care pathway. **What technologies (open source/ propriety) were used for the big data analysis?** Python, Spark **Why those technologies were selected?** Familiarity and availability of libraries. **What were the alternatives technologies and why they were not selected?** R. Unfamiliarity and lack of libraries.

Pilot 12: Radiology Workflows (WP4)	**Partners: CON**
Description of the prediction problem	This pilot focusses on providing radiologists with relevant information during the reading of cases. During typical assessment of radiological imaging data the radiologist, parses the image, reports on findings, and in cases of difficult cases, consults a range of sources, to identify the finding, verify suspected findings, or to put the finding in the context of the disease. The prototype supports this by enabling radiologists to trigger search by marking a region of interest in the imaging data. The software then compares the marked patterns with a large data base of cases, rankes cases and shows the most similar cases, together with a summary and scoring of findings, and additional information such as differential diagnosis guidance, or direct links into curated literature optimized for supporting radiologists.
Data acquisition	Data is acquired during clinical routine with heavily regulated image acquisition technology of different vendors (medical products). The data then typically resides in highly secured hospital data systems called picture archiving and communication systems (PACS) and internal data bases. **What technologies (open source/ propriety) were used for the big data acquisition?** The data forming the corpus is provided in bulk, after undergoing anonymization following standard (DICOM), and is transferred to the search data base. These techniques are proprietary technology under the control of the hospital. Standard used involve DICOM image standards and transfer protocols. **Why those technologies were selected?** Hospital policy requires that the extraction and anonymization is under the control of the hospital. DICOM standards and protocols were chosen, because they are the standard used in all hospitals, and provide unified storage and transport protocols for the data. **What were the alternatives technologies and why they were not selected?** Alternative technologies might involve standard image formats such as jpg, they were not used because they are practically non-existent in the radiology environment, that relies fully on the DICOM standard. Bulk transfer for the search data base was chosen to optimize effort needed.

Data cleaning	Data cleaning consists primarily of selecting a sub-set of imaging data relevant for diagnosis, and for being included in the search corpus. Data such as navigator scans are excluded. Most of this is possible based on technical parameters. **What technologies (open source/ propriety) were used for the data cleaning?** Data cleaning was performed before transfer, using technical parameters of image acquisition such as resolution, or field of acquisition. This is exploiting DICOM standard encoded information. Sanity checks of region imaged were performed using small vignet visualizations. Proprietary software was used. **Why those technologies were selected?** They are feasible for a large-scale data set, and necessary information is included in the data, therefore this cleaning is sufficient for a first pass. Fine-tuned passes can be performed based on image recognition algorithms. **What were the alternatives technologies and why they were not selected?** Alternatives would be manual checks of the data. However, this is not feasible given the data set size. Available software for individual checks, did not allow for fast parsing of large-numbers of images.
Data integration and analysis	Data integration encompasses the transfer of the data into an internal data base, indexing of linked imaging and textual information, and preparation of search. Analysis is part of the indexing, and extracts searchable features that capture relevant information in the imaging data linked to disease and relevant radiological terms. **What technologies (open source/ propriety) were used for the big data integration?** For analysis proprietary algorithms were used, including image parcellation algorithms, deep learning-based algorithms that extract features from imaging data, and segmentation algorithms that identify anatomical regions. **Why those technologies were selected?** These techniques are proprietary algorithms available at the partner, and are optimized for this purpose. Aside from being optimized to link image information and search relevant categories, these algorithms are optimized for speed at search time. **What were the alternatives technologies and why they were not selected?** Alternative approaches involve standard feature extraction algorithms such as for instance GLCM features. These were found to be not sufficiently specific.
Considered machine learning solutions	Machine learning solutions are used to process imaging data, and conduct a search. **What technologies (open source/ propriety) were used for the big data analysis?** The technology is based on Julia framework, and the algorithms are proprietary. **Why those technologies were selected?** They were optimized for the purpose of the prototype, no existing sufficiently accurate algorithms were identified. **What were the alternatives technologies and why they were not selected?** Alternative technologies involve standard feature extractors such as GLCM, or comparable texture features. During evaluation, these did not exhibit sufficient specificity during search.

2.5. Task 1.5: Multi-velocity processing of heterogeneous data streams

This section provides an initial overview on how pilots handle multi-velocity of heterogeneous data streams for pilots diabetes (pilot 3), lung cancer (pilot 7) and hyper-acute workflows, asset management (pilot 9, 10 & 11).

Pilot 3: Diabetes			
Stream name	**Contents of stream**	**Stream velocity**	**Description of the stream**
Background data	Medical data (variable-value)	Initial patient input which can be	The doctors will collect this data into the system. When a doctor registers a

		updated in additional visits to the doctor	new patient into the pilot, there are different measurements that can be introduced (BMI, age, ethnicity, results of initial glucose test,smoking status, employment status, etc)
Antenatal ultrasounds and other specific data	Medical data (variable-value)	Periodic visits to the hospital during pregnancy	Realization of ultrasounds (e.g. macrosomia predicted, polyhydramnios), gestational weight gain, need for insulin, etc
Abdominal circumference	Medical data (variable-value)	Initial input, then measured every few weeks	During the pregnancy, the doctors monitor the abdominal circumference, as well as the centile.
Glucose monitoring	Glucose levels (variable-value)	From 4 to 7 times per day	Glucose monitoring constitutes the main stream of the project. The patients will use a glucometer connected to a mobile app to report their blood levels. The mobile app will send the values form the stream into the computing infrastructure.
Food monitoring and activity data	Food intake and patient activity (exercise)(variable-value)	Every time the patient eats regular food and drinks and when increased exercise or moderate activity is recommended	The mobile app also allows the sending of exercise and food data. The patient needs to input manually this data into the app, so its availability will be much sparser.

Combination of different streams

There are two main objectives to combine different streams in this pilot:
The models that we will use for monitoring the patients benefit greatly from the additional information generated by each one of the different streams. By combining them, the models will have access to a richer set of information, helping them to improve their recommendations to the patients.
The doctors will have a visualization screen that they can use to follow the evolution of the patients. This screen will present the data in a way that is useful to the doctors, as well as present the results of the models. This will allow doctors to take the most informed decision, both using raw data and the model results.

Challenges combining different streams

The data that we are working with, specially the background data is extremely sensitive. We needed to design a system that kept all the confidential information inside the hospital infrastructure, and that allowed communication with more powerful resources. As an additional difficulty, those resources are located in a different country, outside the hospital infrastructure. The design addresses this difficulty, making sure that it ensures users privacy.

Solution

We decided to deploy an Apache Kafka broker in a server inside the hospital infrastructure. This broker has a wide range of tasks to perform, namely:
All the messages generated go through the broker infrastructure. When a doctor introduces a new patient, it goes through the broker before going into the database. The broker and the doctor’s screen are hosted in the server inside the hospital infrastructure, and stay secured. The mobile app sends data (glucose and food) to the broker. The broker cleans and organizes it, and then redirects it to the database (for storage), and to the outside server (to apply the model).
The data that is sent to the outside server is completely anonymized beforehand. Before the data leaves the hospital infrastructure, we make sure that it cannot be traced back to any individual person. Then, the model is applied and the results are sent back to the broker. The broker distributes them to the database. From there, the doctors can consult the data through

their interface.
The intermediate broker allows us to ensure data privacy in a scalable fashion.

Streams mixed	Technology used	Purpose of the process	Result (name if is a stream)
Background data Glucose monitoring Food monitoring Antenatal ultrasounds Abdominal circumference	Java, Kafka, Flink	Prepare the data before sending it to the analytics model.	Stream (Analytics input).
Analytics input Analytics output	Java, Kafka, Flink	Store the results of the analytics, and show them to the medical team.	Database storage Visualization layer.

Stream combination	Difficulties	Solutions
Analytics input	Data is being collected at different speeds, and with different frequencies. Organizing the data to serve as an input for the analytics model.	Ordering the data by date, as our model needs the temporal information. Order by date, inform to the model which variable is being sent.

Pilot 7: Lung Cancer		
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 in 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 the 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 un-structured 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).

Pilots 9, 10 & 11: Hyper-acute Workflows, Asset Management		
Stream name	**Contents of stream**	**Stream velocity**
Real-Time Locating System (RTLS) data stream	Data packets contain information about the location of tags, button presses, motion flags, battery level, tag type, etc.	Every tag can transmit data up to once every 1.5 seconds. A typical large scale deployment can generate around 500 data points per second.
Other Hospital IT systems, e.g. EMR, Laboratory data, Staffing information	Contains medical information, information about number and type of personnel present.	Updates occur in the order of minutes/hours.

Combination of different streams

Different data streams are integrated in order to characterize the performance of the workflows. Integrated data streams are plugged into models which help predict the estimated waiting times/deduce position of a patient within a particular care pathway.

Challenges combining different streams

Every data stream can have missing/noisy data at any point of time. This can make it difficult to combine data for deriving higher-level context information. Moreover, certain data streams can also have erroneous data.

Solution

Data imputation techniques are used to fill up missing data prior to data integration. In addition to that filtering techniques are used to remove noise in the data streams.

Streams mixed	Technology used	Purpose of the process	Result (name if is a stream)
RTLS data Data from Hospital IT system (e.g. EMR, Laboratory, Staffing Records)	Java & Python	Create integrated data stream before it is sent for storage and analysis.	Integrated/cleaned RTLS and EMR data stream.

Stream combination	Difficulties	Solutions
RTLS+EMR analytics input	Data is collected at different speeds and quality. Missing/delayed data on all data streams can make data integration a challenge.	Use data imputation methods or use predictive analytics (e.g. using a Bayesian network approach for estimation) to handle missing/delayed data.

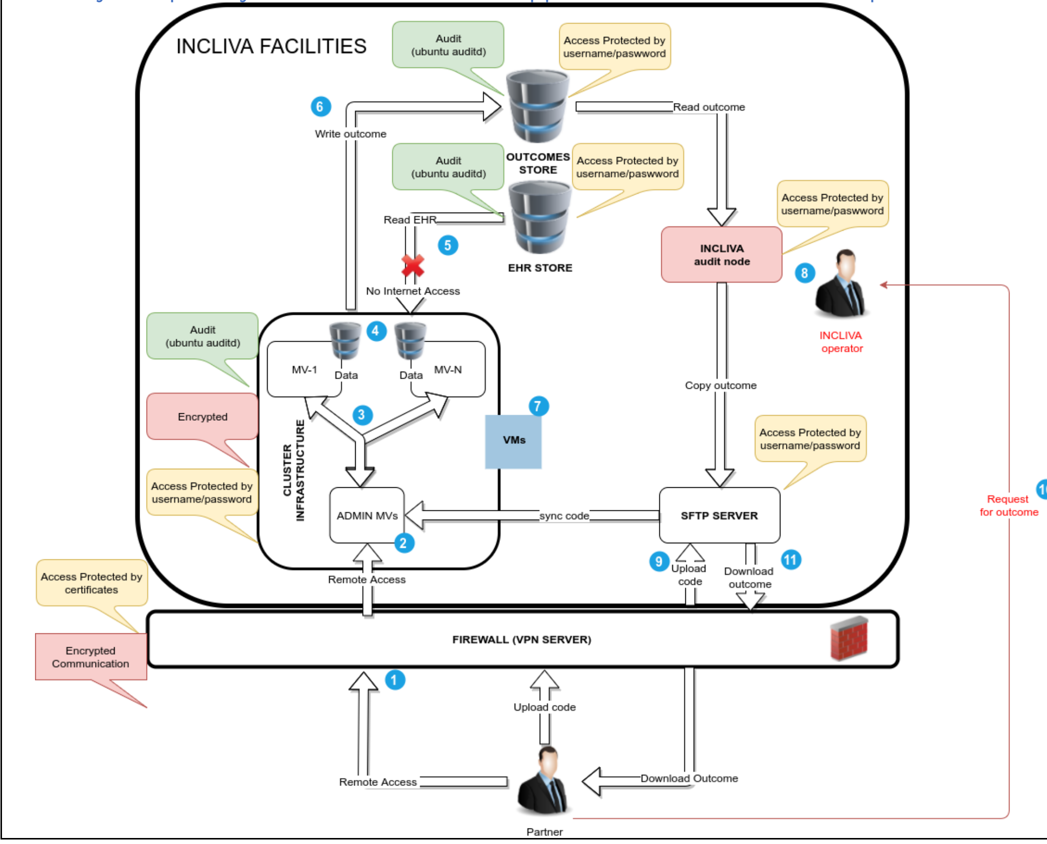
2.6. Task 1.7: Security and privacy of data access and processing

Task 1.7. overviews different approaches for privacy- and security-preserving data access, processing and access control with support for auditing.

Pilot 1 - Comorbidities (WP2)	Partners: OPTI, ITI, I

General Description

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. The following figure 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.

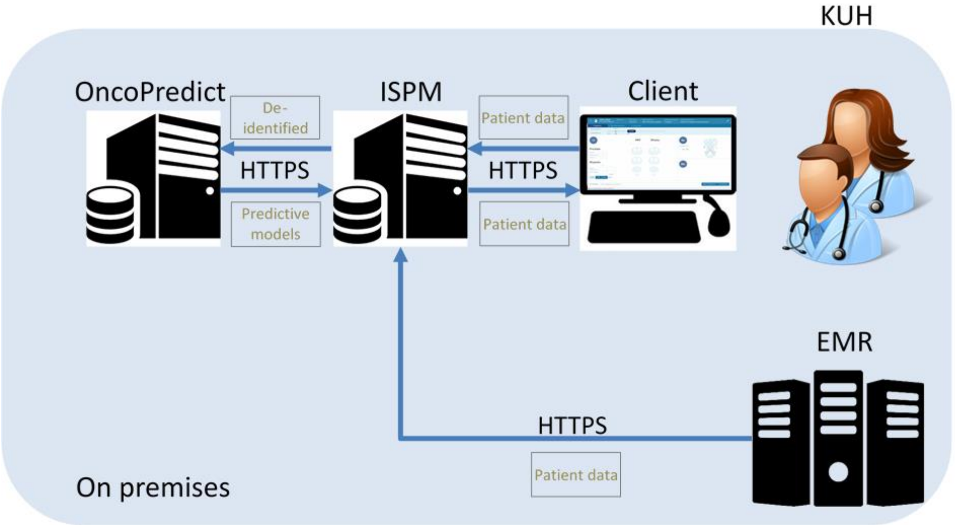


	As the figure shows, all accesses and processes that involves sensitive data are always performed within the Incliva facilities. Every access from a remote computer to the infrastructure, will be done using a Virtual Private Network (VPN). This mechanism guarantees the encryption of the communications 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 summarize 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.	
Access Control		
	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.
	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 to 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 to 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 communicate to the partners by means of an encrypted e-mail. *Access to infrastructure resources for a specific partner* • Each partner will be granted to access to the admin node using a remote

		desktop console provided by the hypervisor of the cluster (Proxmox). To access to the admin node, the end user must be authenticated into the system before hand, 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.
Data protection		
	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 round 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 for that the services provided by the Unix operative system. Data will not be stored in any external media such as pendrives, optical media or portable hard drives under any circumstances.
	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 are encrypted using properly 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.
Audit/Log measurements		
	Application/Services logs	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.
	System logs	The main tools for auditing the system will be the Linux auditd 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

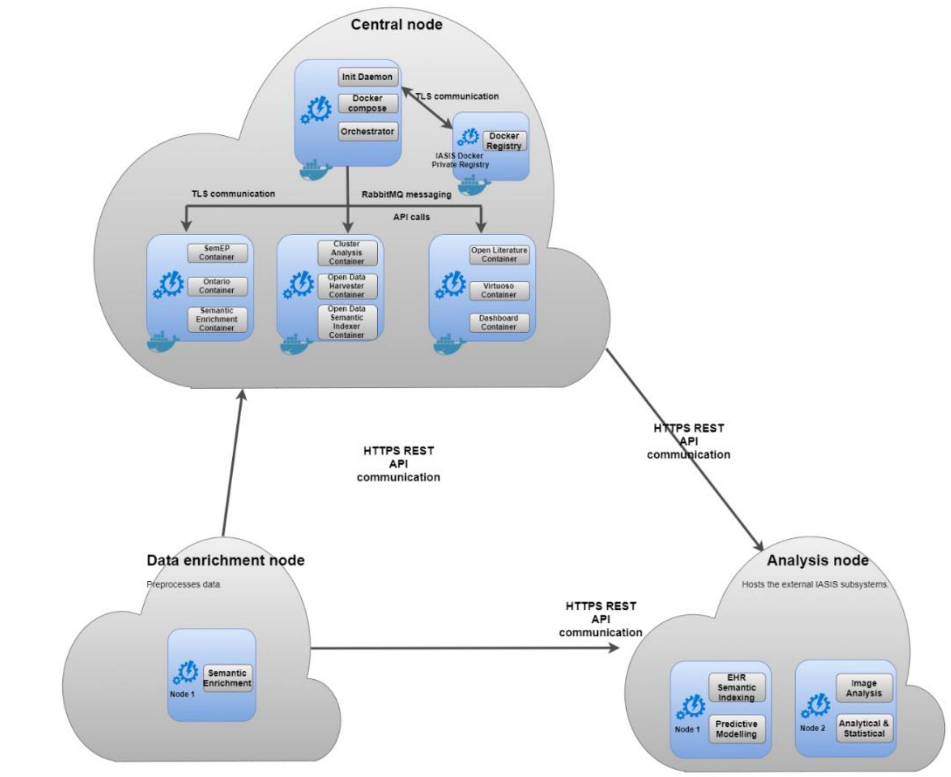
Pilot 5 – Heart Failure (WP2)		**Partners: EMC**
General Description		
		There are currently 15 million HF patients in Europe. The annual risk of death for these patients is 10% to 20%. In particular patients with multiple comorbidities do not currently receive the optimum medical treatment, which leads to many hospital admissions. Achmea, Erasmus MC and TNO work together in this pilot to offer personalized medical treatments based on secure data analyzes. Multi-party computation (MPC) is used for this, where analyzes are carried out on the combination of data from Achmea and Erasmus MC without harming the privacy of individuals. The MPC solution consists of two parts: 1. **Hidden set intersection**. The Achmea and Ergo datasets both contain of a group of patients and attribute values are known to either Achmea or Ergo. For the analysis following this step, it is important to use the overlap of patients in the two groups. That is, the group of patients that appear in both the Achmea and Ergo data. This is also called the intersection of the two datasets. The intersection itself remains concealed during the hidden set intersection; only the encrypted set intersection is computed. The only thing that becomes known during the hidden set intersection is the size of the intersection, so the number of patients that occurs in both datasets (in the intersection). The result of the hidden set intersection is an encrypted dataset of the patients that appear in the data from both Achmea and Ergo. Encrypted means that nothing is known about these patients or their data. 2. **Secure LASSO regression**. The encrypted data of patients in the intersection set is used to find a statistical relationship between influencing factors (such as: number of cigarettes per day, weight, medication compliance, etc.) and the number of hospitalization days. During the secure LASSO regression, no information is revealed other than the outcome of the regression. The outcome of the secure LASSO regression is the regression coefficients that indicate the linear relationship between the influence factors and the number of hospitalization days. The results of the secure LASSO regression can be used to advice (new) HF patients on the influence of their behaviour (smoking, exercising, etc) on the expected number of hospitalization days. Instead of the current qualitative advice, doctors can give a quantitative advice on lifestyle and medicine compliance.
Access Control		
	Authorization	Not applicable for the MPC proof of concept.
	Authentication	The MPC computation uses for the communication TLS/SSL. The SSL certificates are generated and distributed to all parties beforehand. Each party can check the certificates of the other parties to authenticate. Additionally, only specific IP addresses are allowed to join the MPC protocol.
Data protection		
	Data at rest	During the MPC computation, the involved data will remain private to each partner. The data and outcome of computations that are performed are encrypted by means of Shamir secret sharing and Paillier additive homomorphic encryption. The (pseudo) randomness is implemented by PEP 506. That means that nothing can be derived from the (secret shared) data of one party.
	Data in transit	The data is transit is encrypted as described above. In addition to that, the encrypted data exchanges are sent using TLS/SSL.

Audit/Log measurements	
Application/Services logs	For the MPC proof of concept demonstration, there will be no application or services logs in place.
System logs	For the MPC proof of concept demonstration, there will be no system logs in place.

Pilot 6 – Prostate Cancer (WP3)	**Partners: PHI**
General Description	
	ISPM (the Philips tumour board application) is an official released product for clinical use, where all the required security measures are implemented. ISPM is installed on premise, on a windows server VM administrated and controlled by the hospital IT. ISPM consists of angular/typescript frontend and a java micro services backend, running on local web server. Data is stored in local FHIR database. To install/update software, the VM can be accessed using a secure VPN connection set up by hospital IT. ISPM runs completely locally inside the hospital firewall. OncoPredict is a separate research prototype, also installed on the VM. Is retrieves de-identified data from ISPM, to execute statistical analysis.  KUH OncoPredict De-identified HTTPS Predictive models ISPM Patient data HTTPS Patient data Client EMR HTTPS Patient data On premises
Access Control	
Authorization	A 2-factor authentication using a token procedure is performed the gain access to the ISPM. VPN is used to remotely connect to the VM running ISPM and OncoPredict using Pulse Secure delivered by KUH; thereafter a Windows account login (security login).
Authentication	ISPM has multiple roles defined to gain access to the system. For each role, there are a number of writing and reading regulations. Typical roles are: • Tumour board chair • Radiologist • Nurse OncoPredict enables data analytics and can get access to anonymized data using a

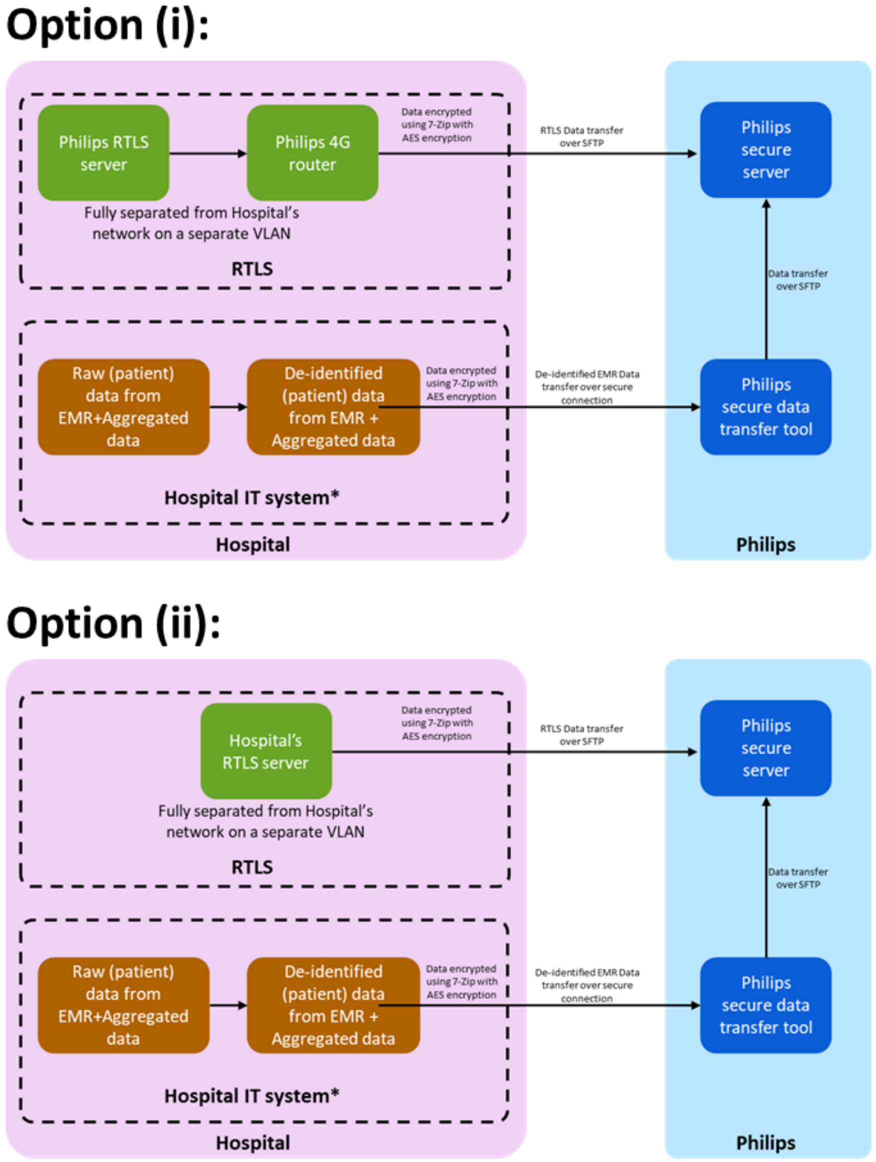
		2-way authentication with read only access.
Data protection		
	Data at rest	HTTP TLS encryption is used for the web interface of ISPM. Login to intranet of KUH via VPN. RDP connection to VM. There is no encryption for the HD of the server. Hospital IT says: “There is no encryption for the HD of the server. Encryption is usually mostly used on discs that run the risk of getting in the wrong hands, so at our hospital they are rolling out BitLocker for laptops and stationary PC’s as they can be stolen or lost but servers are protected by perimeter security so discs are not encrypted.”
	Data in transit	No data is transferred outside the hospital’s intranet. All statistical analysis is performed on premises. At the end of the project all data will be deleted and the VMs are shut down and erased.
Audit/Log measurements		
	Application/Services logs	VPN logging, Windows logging by hospital IT.
	System logs	ISPM internal logging

Pilot 7 – Lung Cancer (WP3)	**Partners: NCSR-D, ATC**
General Description	
	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 sematic 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.

	 Central node Init Daemon Docker compose Orchestrator TLS communication Docker Registry IASIS Docker Private Registry TLS communication RabbitMQ messaging API calls SemEP Container Ontario Container Semantic Enrichment Container Cluster Analysis Container Open Data Harvester Container Open Data Semantic Indexer Container Open Literature Container Virtuoso Container Dashboard Container HTTPS REST API communication HTTPS REST API communication Data enrichment node Preprocesses data Semantic Enrichment Node 1 HTTPS REST API communication Analysis node Hosts the external IASIS subsystems EHR Semantic Indexing Predictive Modelling Node 1 Image Analysis Analytical & Statistical Node 2	
Access Control		
	Authorization	Remote or physical access to the servers storing the data is provided only to authorised personnel based on their role.
	Authentication	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.
Data protection		
	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
	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.
Audit/Log measurements		

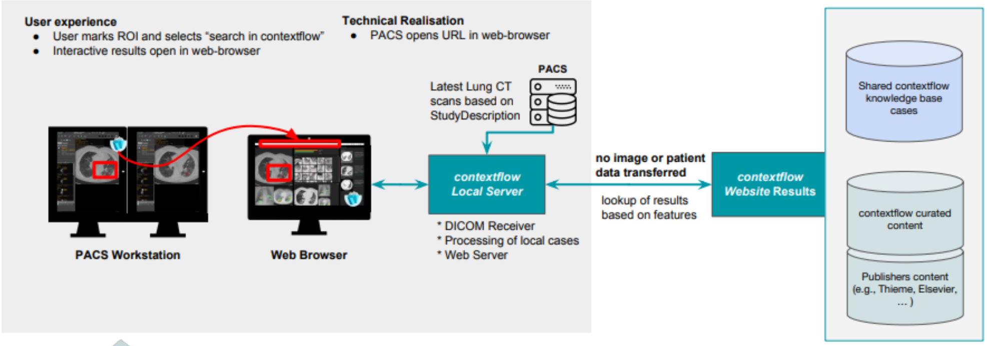
	Application/Services logs	There will be logs related only to the user evaluation and the measurement of KPIs
	System logs	Only standard system logs produced by the open source tools

Pilot 9, 10 & 11 – Hyper-acute workflows, asset management (WP4)	**Partners: PHI, TUE**
General Description	
	Pilots 9 and 10 have two sources of data streams: Real-time Locating System (RTLS) and Electronic Medical Record (EMR). Pilot 11 only has a single RTLS data stream. RTLS data is collected locally at a server situated at the hospital. The server can be connected to Philips in either one of the following two ways: (i) over a 4G connection or (ii) the server is connected to the internet over a separate VPN at the hospital. All data is encrypted prior to transmission over the internet. All data is transmitted to Philips over a secure connection. EMR data is de-identified at the hospital and relevant fields are encrypted and manually uploaded to a secure Philips server over a secure connection. This general procedure is shown in the figure below:

		 Option (i): Philips RTLS server Philips 4G router Data encrypted using 7-Zip with AES encryption RTLS Data transfer over SFTP Philips secure server Fully separated from Hospital's network on a separate VLAN RTLS Data transfer over SFTP Raw (patient) data from EMR+Aggregated data De-identified (patient) data from EMR + Aggregated data Data encrypted using 7-Zip with AES encryption De-identified EMR Data transfer over secure connection Philips secure data transfer tool Hospital IT system* Hospital Philips Option (ii): Hospital's RTLS server Data encrypted using 7-Zip with AES encryption RTLS Data transfer over SFTP Philips secure server Fully separated from Hospital's network on a separate VLAN RTLS Data transfer over SFTP Raw (patient) data from EMR+Aggregated data De-identified (patient) data from EMR + Aggregated data Data encrypted using 7-Zip with AES encryption De-identified EMR Data transfer over secure connection Philips secure data transfer tool Hospital IT system* Hospital Philips
Access Control		
	Authorization	Access to RTLS data stream will be limited only to specific Philips personnel working on the pilot. Access to EMR data stream will be limited to specific hospital staff members working on the pilot.
	Authentication	(i) Local access with a standard Windows account authentication using a strong password with more than 8 characters which locks after a certain number of wrong attempts or (ii) remote access using a Philips-compliant remote access service.
Data protection		
	Data at rest	Data will be encrypted before transfer using 7-zip with AES encryption
	Da	TLS encryption

Audit/Log measurements		
	Application/Services logs	All activities actions on the RTLS server, secure server for data storage and secure data transfer server are logged for auditing purposes.
	System logs	VPN/Windows logging

<table>
<tr><th><strong>Pilot 12: Radiology Workflows (WP4)</strong></th><th><strong>Partners: CON</strong></th></tr>
<tr><td colspan="2"><em>General Description</em></td></tr>
<tr><td colspan="2">The main aim of this pilot is to reduce the time of diagnosis, and to increase quality of diagnosis in radiology departments by providing an efficient search engine for radiological data.
Radiologists need fastest access to information and documental evidence to back-up their initial interpretation of the images before the diagnosis. Access to external resources occurs in a 20% of the cases, consuming a significant amount of time. In many cases, radiologists need to ask their colleagues, or search for reference literature or web resources, which is time consuming and error prone. The pilot intends to use a big data approach taking into consideration a wealth of knowledge to help radiologist to take informed decisions and backing up their conclusions.
The figure below shows the way the solution is devised.


User experience
• User marks ROI and selects "search in contextflow"
• Interactive results open in web-browser
Technical Realisation
• PACS opens URL in web-browser
PACS
Latest Lung CT scans based on StudyDescription
PACS Workstation
Web Browser
contextflow Local Server
* DICOM Receiver
* Processing of local cases
* Web Server
no image or patient data transferred
lookup of results based on features
contextflow Website Results
Shared contextflow knowledge base cases
contextflow curated content
Publishers content (e.g., Thieme, Elsevier, …)

The pilot will deploy a local server at the user premises (typically the hospitals) and open a connection to a centralized server where the search of the similar results will be performed. It is important to point out that no image or patient data will be transferred outside the hospital. The main results obtained during the development of this pilot will be integrated with the tools used by the Vienna and Puerta de Hierro Hospitals. Therefore, it will inherit all the security measurements already in place.</td></tr>
<tr><td colspan="2"><em>Access Control</em></td></tr>
<tr><td>Authorization / Authentication</td><td>It can be divided the type of access management control applied in these pilots into two different categories: those components which access to sensitive data (radiological images) and the components without.
With regards the components with access to sensitive data (within the grey square), it is worth highlighting that this access will be always executed in the hospital facilities and therefore into a controlled environment. As mentioned above, these components will be fully integrated with the tools already in place and used by the health institutions as a brand-new functionality, hence the control access (Authentication and Authorization) will rely on the already used by these institutions.
Regarding the other components deployed outside of the hospital facilities is worth mentioning that they will not access to the sensitive information hence although access control is applied, they don't need to be so strict as the components described above.</td></tr>
<tr><td colspan="2"><em>Data protection</em></td></tr>
<tr><td>Data at rest</td><td>The sensitive data involved on the development of this pilot will always stay within the hospital facilities in a controlled environment. However, in any case the data store will be encrypted to guarantee that they will only be accessible with the appropriate authentication and authorization permissions.</td></tr>
<tr><td>Data in transit</td><td>All the communication between the different components involved in the pilot will rely on SSL/TLS protocol for the communications, hence the data will be encrypted, and therefore protected, in all the communications</td></tr>
</table>

Audit/Log measurements		
	Application/Services logs	As mentioned initially this pilot will be integrated with an already full functional application within the health institutions hence for the components with access to the sensitive data, the auditory of the access to the services will rely on the services already in place. However, for the components which are outside of the hospital premises, although they don't access directly to the sensitive data, will be monitored and audit in order a whole track of the operations realized will be controlled
	System logs	