



Deliverable 1.2

Initial 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 1 st 2018	Duration	38 months
Version	1.0		
Status	Final		
Date of issue	30/11/2018		
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 instead presents a more detailed analysis of the different pilots according to the WP1 tasks. In particular this deliverable presents a first overview about the different software prototypes, shows the different components, data and challenges to overcome.

Table of Contents

1.	Introduction	4
1.1.	Purpose of the document.....	4
1.2.	Related documents	4
2.	Overview of specific components	5
2.1.	Task 1.1: Deep learning for Multilingual NLP and image analytics.....	5
2.2.	Task 1.2: Prediction algorithms.....	8
2.3.	Task 1.3: Complex real-time event detection	21
2.4.	Task 1.4: Processing of large structured / unstructured data sources	22
2.5.	Task 1.5: Multi-velocity processing of heterogeneous data streams	29
2.6.	Task 1.7: Security and privacy of data access and processing	31

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 for hospital workflow focused pilots in the Industrialization of Healthcare theme.

Deliverable D1.2 collects information according to the different WP1 tasks from all 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.

1.2. Related documents

Related documents: D1.1, D2.1, D3.1, D4.1:

Similarly to D1.2, also D1.1, D2.1, D3.1, D4.1 provide a general overview of the different pilots. However, D1.2 provides an initial, more detailed overview on the different prototypes of each pilot on the level of the different subtasks, such as prediction models (T1.2) or the processing of large structured/unstructured data sources (T1.4) while D1.1 addresses technical requirements from a Big Data perspective. D2.1, D3.1 and D4.1 instead focus on the requirements from an application perspective.

2. Overview of specific components

The following sections provide an overview of the different prototypes developed within the different pilots of BigMedilotics. The overview provides a first analysis according to six different tasks: Deep learning for Multilingual NLP and image analytics (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 first comparison between the different pilot developments.

2.1. Task 1.1: Deep learning for Multilingual NLP and image analytics

Task 1.1 involves the application of natural language processing (NLP) and image analytics (IA) components within the following 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 12 focuses on image analytics, all others work on natural language processing 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 is 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 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 role of NLP/IA across relevant pilots

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 electronic health records, text is an additional source of information which often includes other information that is not directly expressed in the structured data. Thus, it is important to extract relevant information from text in order to get a better understanding of a patient.	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 electronic health records. 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 are written in natural language in the clinical texts. This	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

			unstructured information can be used to enrich structured data and thus significantly increase the amount and richness of patient data to improve predictive analysis.	a reduction of the economic costs associated to 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 from the UML semantic network	
Pilot 12: Radiology Workflows				
CON	Image Analytics and in parts NLP	German	Image processing is crucial to perform the image pattern comparison. DL is used to train it on the diseases relevant for the pilot. 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

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 nephrology 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 nephrology 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 clinical reports and a modified set of negation triggers	Simple negation detection of medical conditions	no training required
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	Regular expression are used to find the (relatively few) disorders and findings. Anatomies are found by a combination of regular expressions and rules to interpret the anatomical locations.	Disorders, findings, and anatomies were manually annotated in a few hundred pathology reports from sites in US, The Netherlands, and Sweden.
item (vial) detection	regular expression matching in combination	none	a few hundred pathology and radiology reports from	First item candidates are detected with regular expressions. These are then	Items were manually annotated in a few hundred pathology reports from sites in US, The Netherlands, and Sweden.

	with rules		~10 sites in US, Netherlands, and Sweden	interpreted (is the item Roman, Arabic, etc) and then, based on a set of rules, candidate items are selected or discarded.	
measurement detection	regular expression matching	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden	Regular expressions are used to find measurements.	Measurements were manually annotated in a few hundred pathology reports from sites in US, The 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	Negex-like method: first negation cues (a predefined set of words) are detected. Then the scope of the negation is determined. This depends on the negation cue (cue before the scope or after the scope), a scope length limit, and scope ending tags (periods, words like "but", etc).	Negations of findings and disorders were manually annotated in a few hundred pathology reports from sites in US, The 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
Date Recognition	Deterministic rule-based system	SUTime from Stanford Core NLP	Ad-hoc Dictionary	Rule-based system that recognizes dates in a deterministic way. It can also recognize expressions of time in natural language and relate these expressions to the date of the document to extract the exact date to which the expression of time refers.	Dictionary created by Stanford Core NLP and translated to deal with spanish narratives.
Event (Concept-Date Relation) Recognition	Cyclic Bidirectional Dependency Network	Stanford Core NLP	AnCor 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 recognition	Proprietary	Metamap	Applied on PubMed, PubMed Central	Named entity recognition in biomedical text, it recognizes UMLS concepts in text	Pretrained (https://metamap.nlm.nih.gov/)
Relation Extraction	Proprietary	Semrep	Applied on PubMed, PubMed Central	Relation extratction in biomedical text. Extracts semantic predications where the concepts are from the UMLS Methathesaurus and the relation from from the UML semantic network	Pretrained (https://semrep.nlm.nih.gov/)
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		

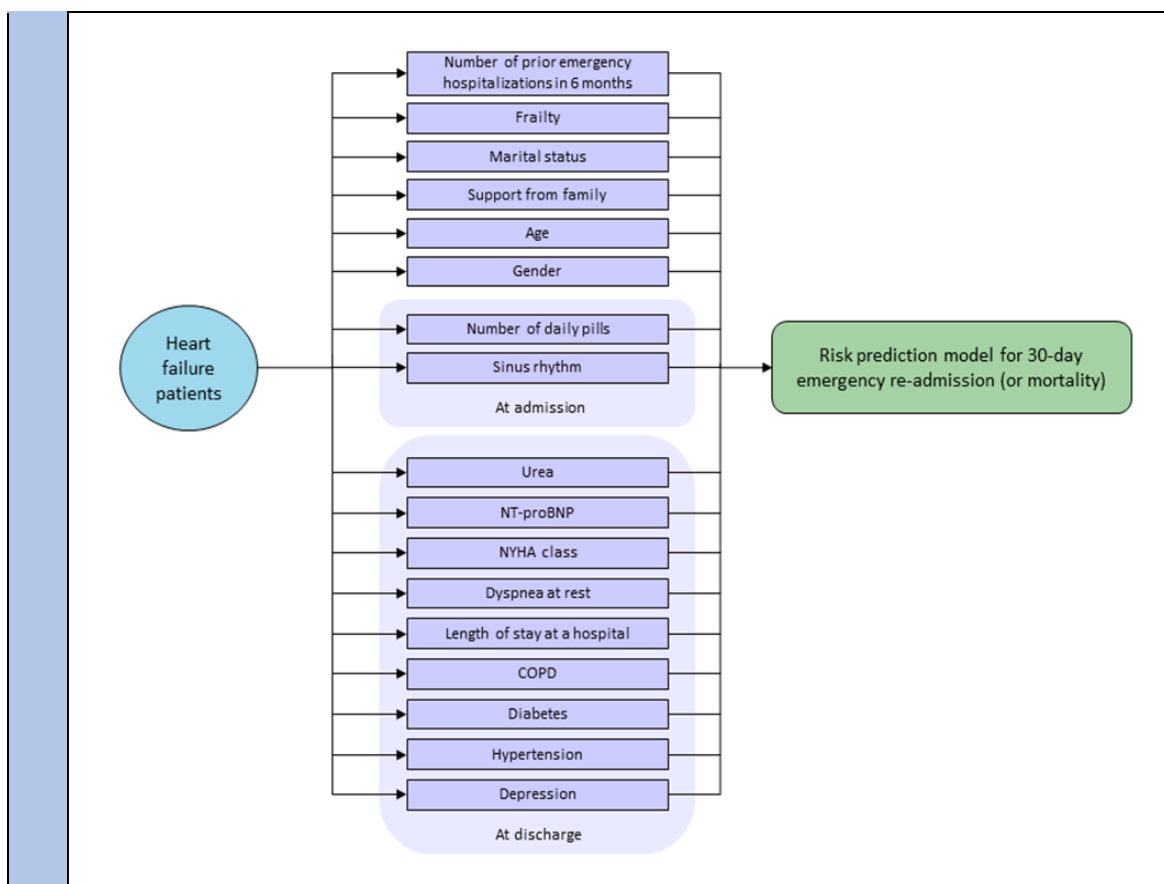
Table 3: Image Analytics components of pilots

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 12: Radiology Workflows							
Yes, it needs CE certification	comparison of image content	yes, for research we use available repositories e.g., challenge s	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+ volumes). It identifies the closest matches and retrieves the corresponding cases.	we use mainly weakly supervised from radiological routine information, i.e., images together with reports	lung diseases	on pixel level	volume level, and regions of interest for evaluation
	search based on a marked region of interest, to find image segments across a large data base that carry similar patterns		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, kidney disease, COPD+Asthma, diabetes, heart failure), WP3 (prostate, lung, breast cancers), and WP4 (hyper-acute workflows, asset management).

Pilot 1 - Comorbidities (WP2)	Partners: OPTI, ITI, PHI
<p><i>Description of the prediction problem:</i></p> <p>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 Heart Failure, 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:</p> <ul style="list-style-type: none"> • 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. <p>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 (heart failure) patients. The next figure summarizes the use of the data for this task:</p>	



Considered machine learning solutions:

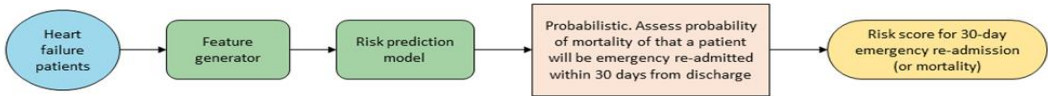
Before the application of the ML solutions we will carry two processes to prepare the data:

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.

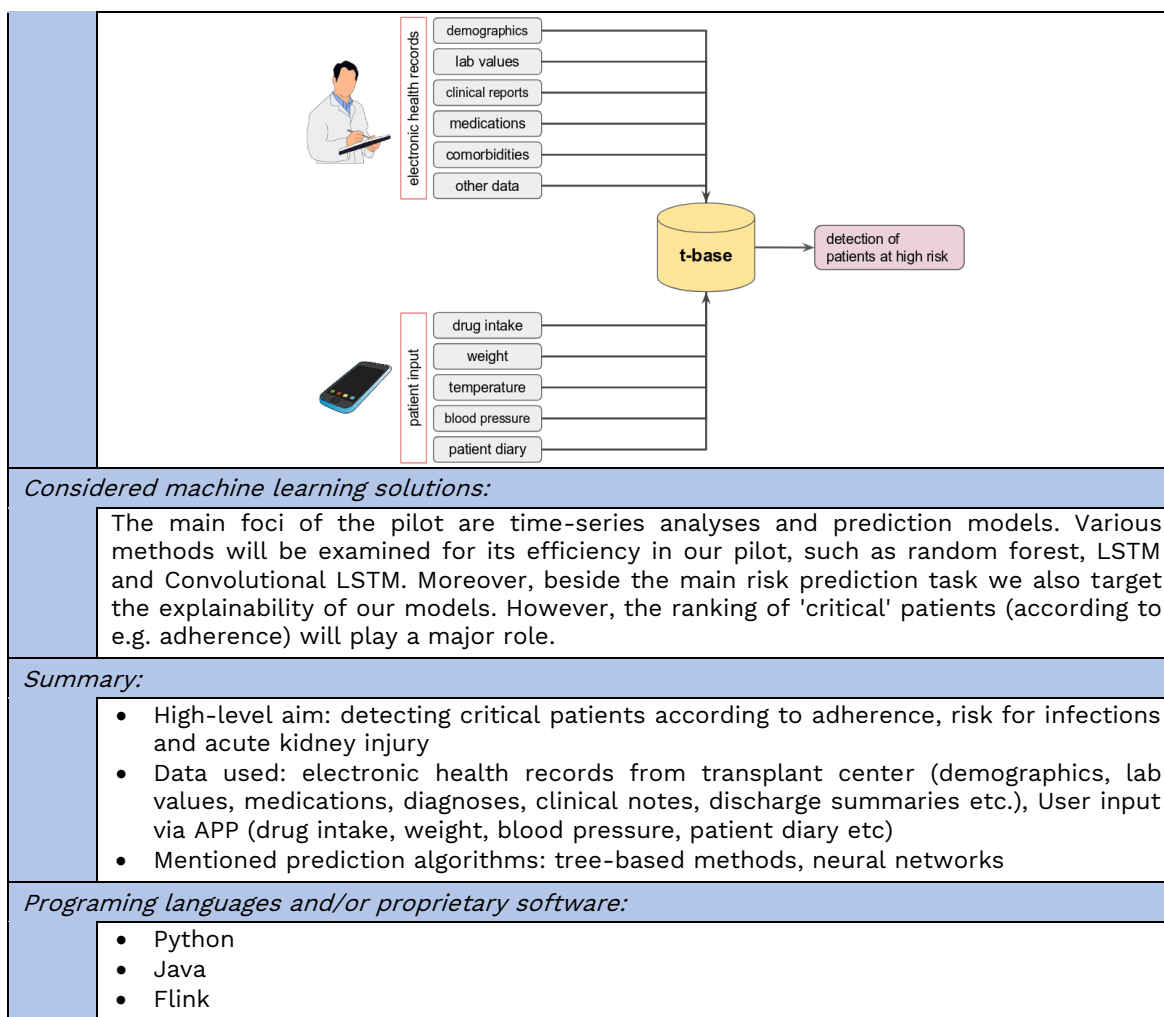
After the data is ready for analysis, the second step is to obtain how patients aggregate into clusters. For this task, several parameters influence the best solution: the data structure, how the data is distributed in a multidimensional space, how far a cluster must be from another one, the distance metric considered in the cluster learning process or the final number of clusters to obtain. As we do not know in advance the answer to these specific points, our overall approach will be to test different clusterization techniques in order to find the most suitable one. The proposed set of algorithms cover a wide range of scenarios and they are trusted techniques in the ML community. Current candidate solutions are:

- K-means.
- Spectral clustering.
- Gaussian mixture.
- Hierarchical clustering.
- Nearest Centroid.
- Self-organizative maps.

The next step requires the supervision of a human expert to look for ICD codes related with comorbidities. This supervised step is required to check if the selected features for each cluster have a meaning from the medical point of view. Taking into account this supervision, we highlight two potential ML solutions, Hierarchical Clustering and Self

	<p>Organizative Maps, as they provide better graphical representations of the clusters. The first one by using dendograms, trees that join individuals different hierarchical levels. The second one by using a topological 2D representation of a multidimensional space that consolidates the nearest individuals. Finally, from each cluster we will obtain the most relevant features, according to statistical indicators, and classify each patient to the “nearest” cluster according to such features.</p> <p>Additionally, the risk prediction models for 30-day emergency readmission (or mortality) of (heart failure) patients, follows a quite similar approach summarized as:</p> <ol style="list-style-type: none"> 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):  <pre> graph LR A([Heart failure patients]) --> B[Feature generator] B --> C[Risk prediction model] C --> D[Probabilistic. Assess probability of mortality of that a patient will be emergency re-admitted within 30 days from discharge] D --> E([Risk score for 30-day emergency re-admission (or mortality)]) </pre>
	<p>Summary:</p> <ul style="list-style-type: none"> • High-level aims: comorbidity cluster analysis, medical event prediction • Data used: demographics, medical time-series • Mentioned prediction algorithms: clustering methods, probabilistic methods, neural networks

Pilot 2 - Kidney disease (WP2)	Partners: <u>DFKI</u>
<p><i>Description of the prediction problem:</i></p> <p>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 physicians-physicians. In the course of this, the kidney pilot focuses on (KPIs) a reduction of unwanted re-hospitalizations, reduction of mortality, detection of possible acute kidney failures after transplantation and the extension of the graft survival. Moreover, the pilot targets the support of adherence which can be one of the major reasons for graft loss (if medications are not taken). Central element of the kidney pilot is a dashboard which 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.</p> <p>As input for our prediction models we consider two main data sources: data from electronic health records (EHR) 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.</p>	



Pilot 3 – Diabetes (WP2)

Partners: HUA, NISS

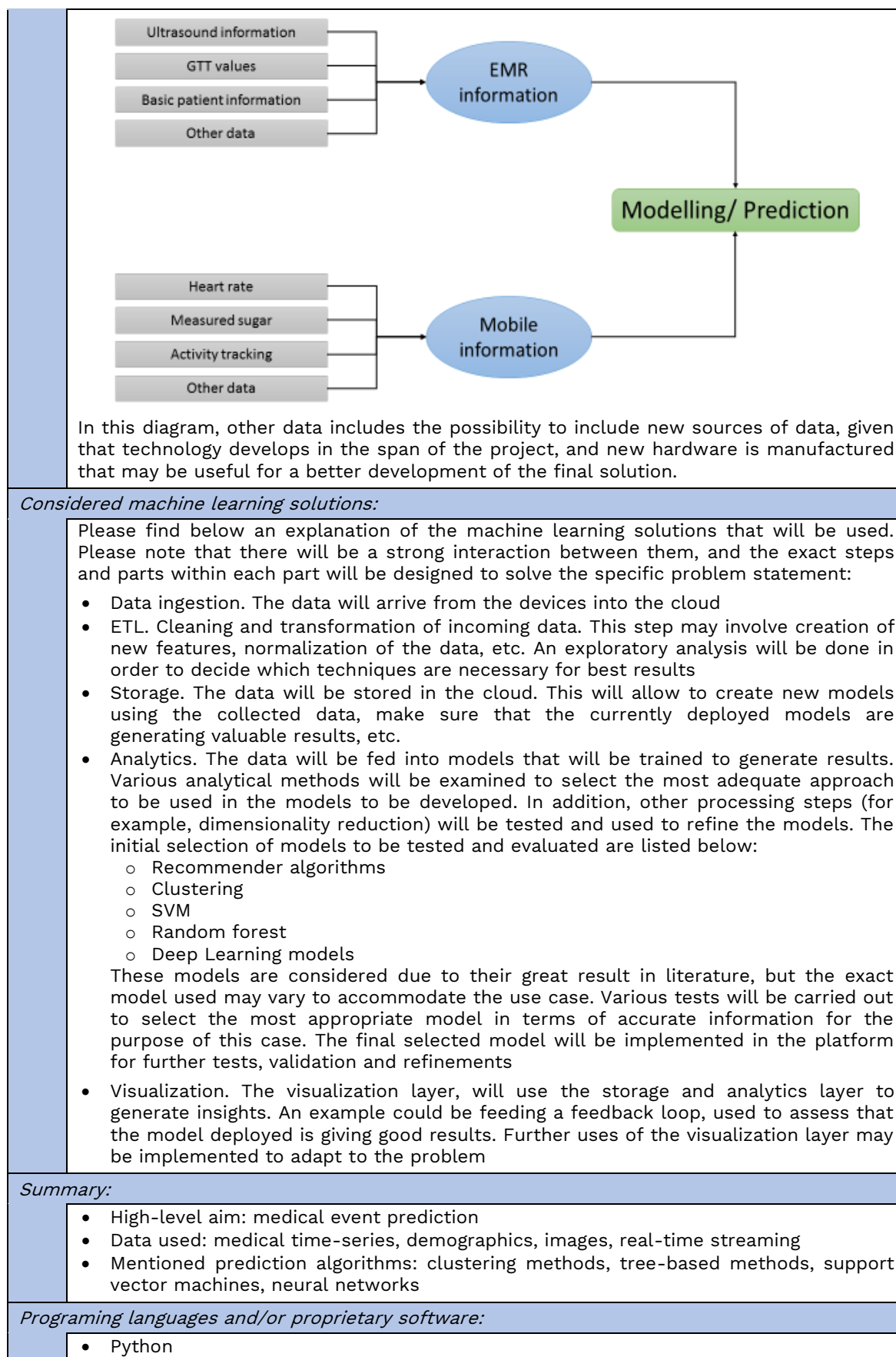
Description of the prediction problem:

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.



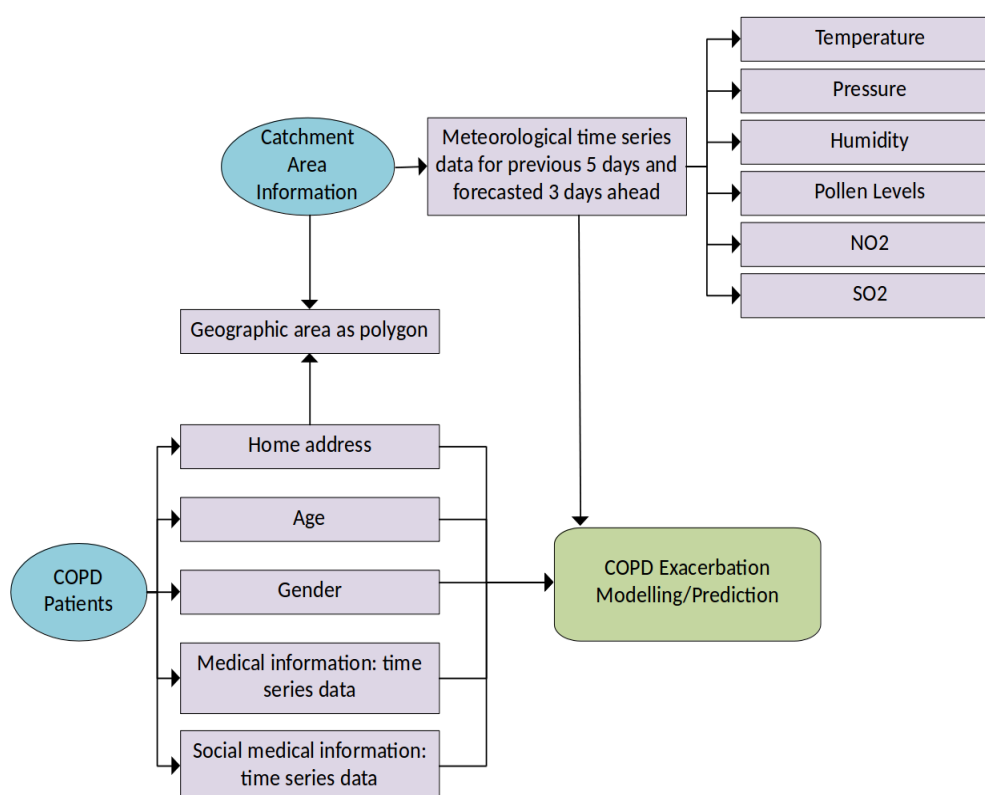
- Scala/Java
- Flink
- Kafka

Pilot 4 – COPD and Asthma (WP2)

Partners: UNIS, MYM

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.



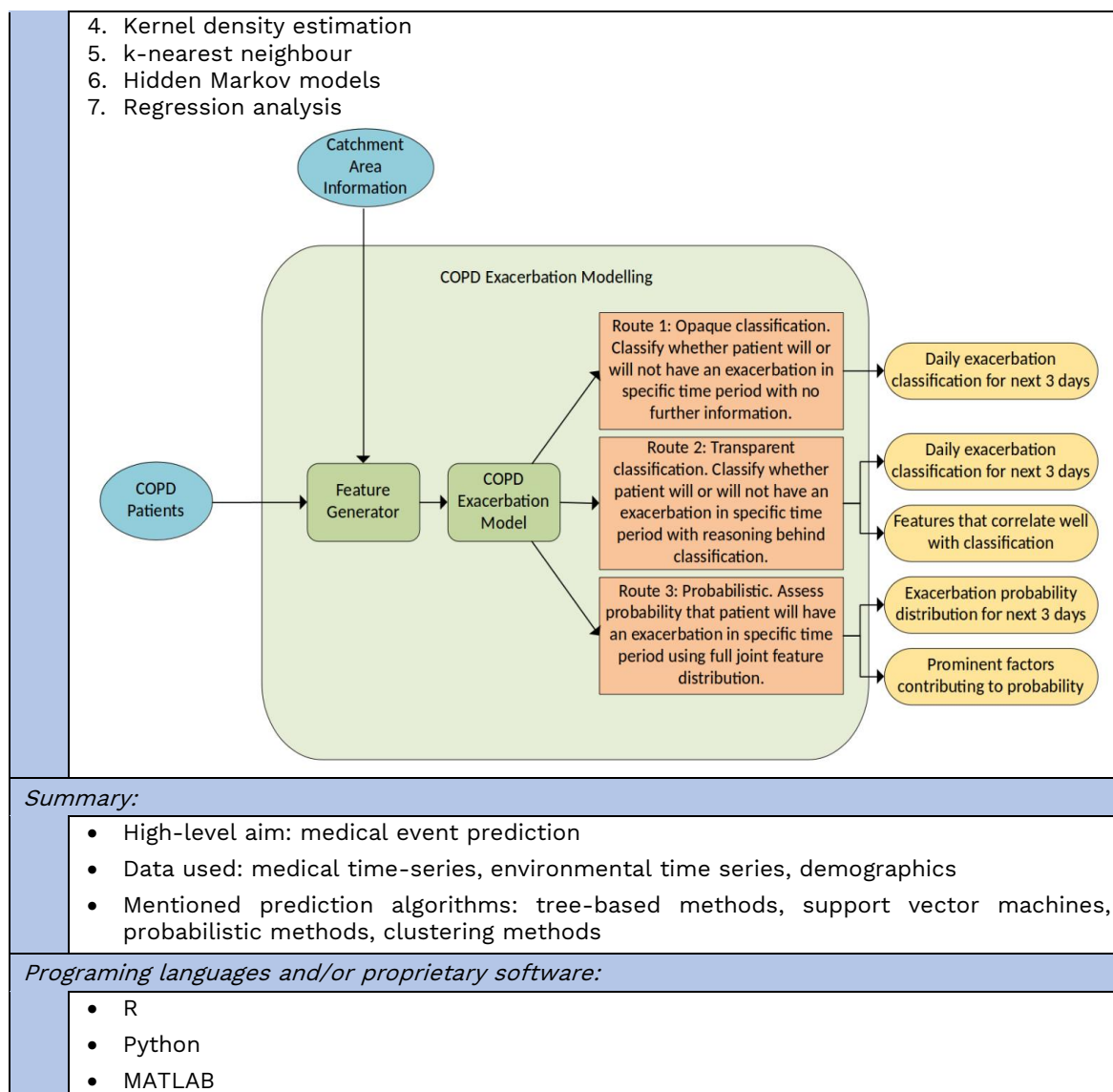
Considered machine learning solutions:

To this end the algorithm requires:

1. Pre-processing of all data types;
2. Aggregation of heterogeneous primary, secondary and MyMhealth mobile/web generated patient data;
3. Semantic harmonization;
4. Feature extraction and possibly feature selection
5. COPD exacerbation modelling, which can be based on one or multiple routes as follows (see figure below for more details):
 - a. Opaque classification
 - b. Transparent classification
 - c. Probabilistic

We expect to investigate the following prediction algorithms:

1. Random forest
2. Support vector machines
3. AdaBoost with Decision Trees



Pilot 5 – Heart Failure (WP2)

Partners: EMC

Description of the prediction problem:

Within the Heart Failure Pilot we are using databases containing a large number of Heart Failure patients and their many comorbidities. One of our 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.

Considered machine learning solutions:

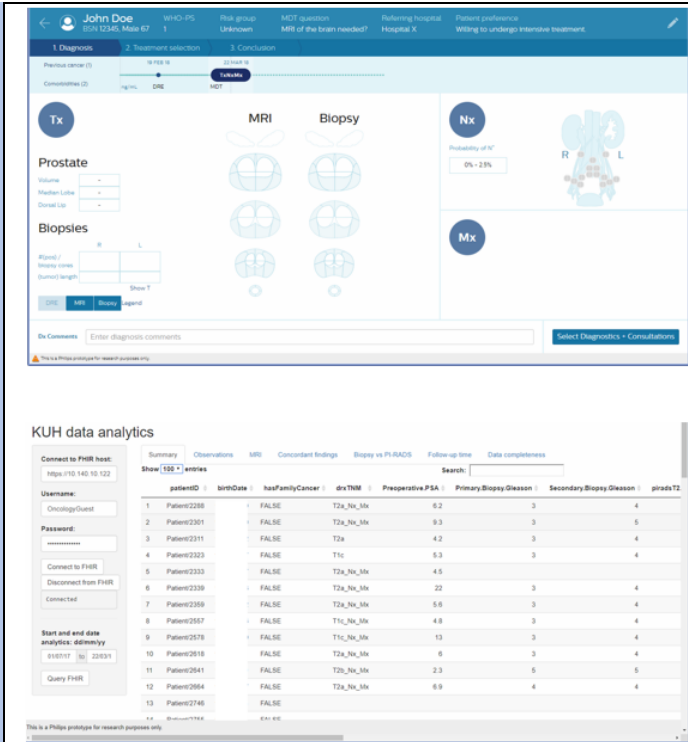
We are planning on using different types of machine learning solutions to investigate which ones are the most informative for our problem. Some of the machine learning approaches we will apply are:

- random forest
- KAGGLE
- LASSO
- elastic net

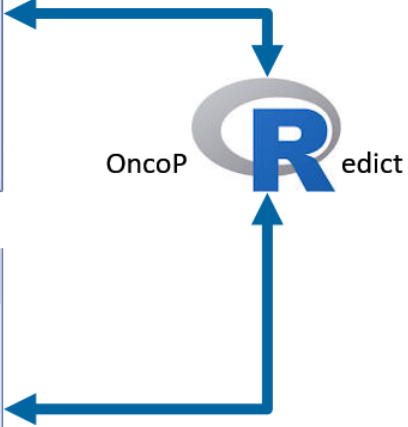
Summary:

	<ul style="list-style-type: none"> 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
<i>Programming languages and/or proprietary software:</i>	
	<ul style="list-style-type: none"> R SAS

Pilot 6 – Prostate Cancer (WP3)		Partners: <u>PHI</u>
<i>Description of the prediction problem:</i>		
	<p>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.</p>	
<i>Considered machine learning solutions:</i>		
	<p>The current predicted risks are detailed below, while more risk models may follow:</p> <ul style="list-style-type: none">• Pre-surgical risk of post-surgical adverse prostate cancer pathology (i.e., pathology Gleason\geq7)• Pre-surgical risk of post-surgical advanced extent of disease (i.e., pathology disease stage \geqpT3a)• Pre-surgical risk of the presence of tumor infiltrated lymph nodes. <p>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.</p>	



The top screenshot shows the 'miProstate' clinical dashboard for a patient named John Doe. It includes sections for Diagnosis, Treatment selection, and Conclusion, with visual aids for MRI and Biopsy. The bottom screenshot shows the 'KUH data analytics' interface, which includes a table of patient data with columns for patientID, birthDate, hasFamilyCancer, dxTNM, Preoperative.PSA, Primary.Biopsy.Gleason, Secondary.Biopsy.Gleason, and gradeT2.



The OncoPredict logo, featuring a stylized 'R' and the text 'OncoPredict', is shown with two blue arrows pointing from it to the clinical dashboard and data analytics interfaces, indicating its role in the system.

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

Programming languages and/or proprietary software:

Proprietary software: miProstate (clinical data dashboard); OncoPredict (prediction learning framework)¹

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

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

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 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.

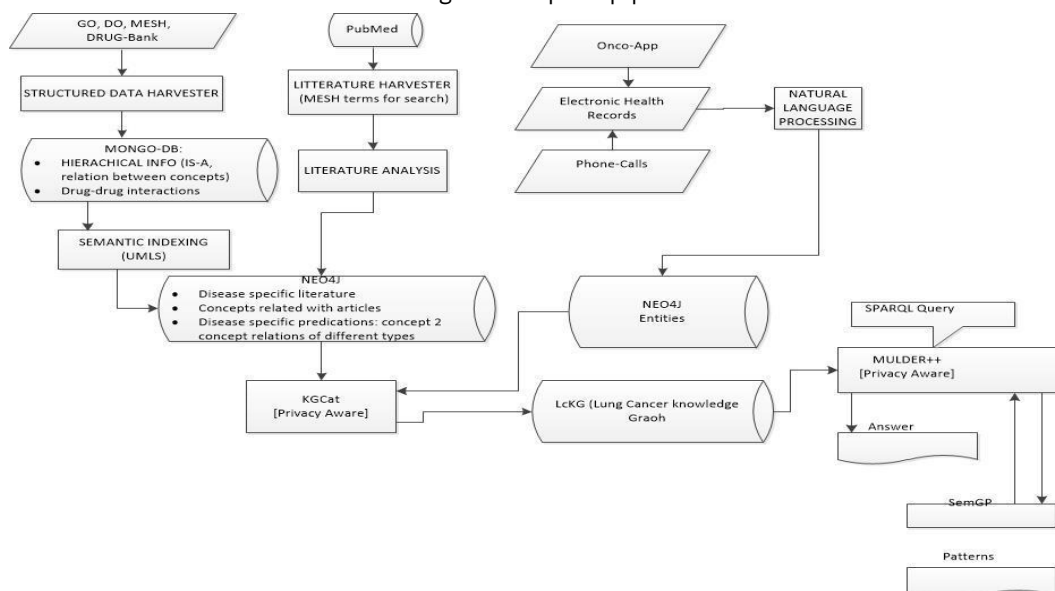
The logo of R is included to let the reader know that the wonderful world of R is being used, and the R becomes part of the OncoPredict world.

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.

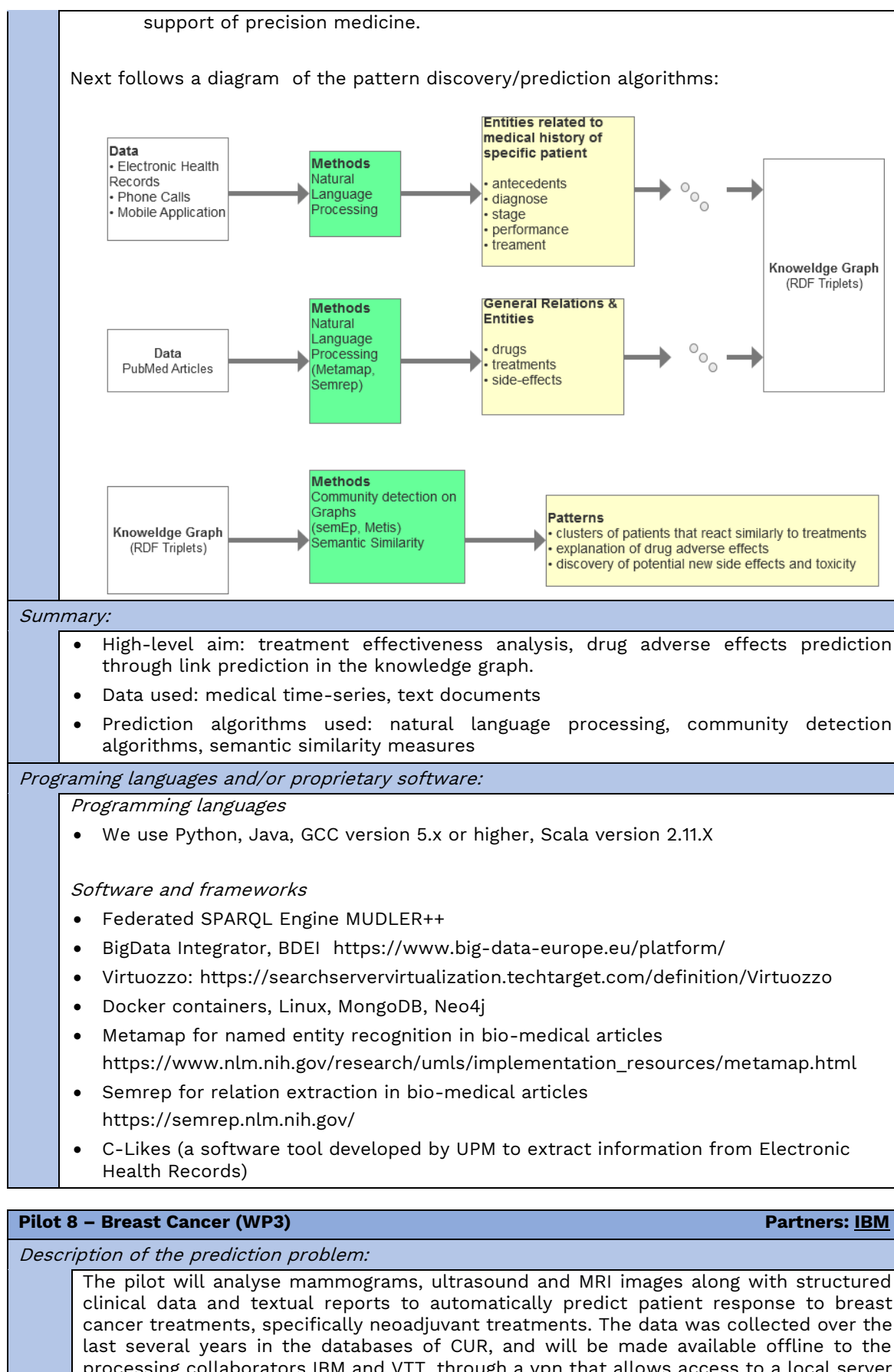
Next follows an overview of the Lung cancer pilot pipeline:



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 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

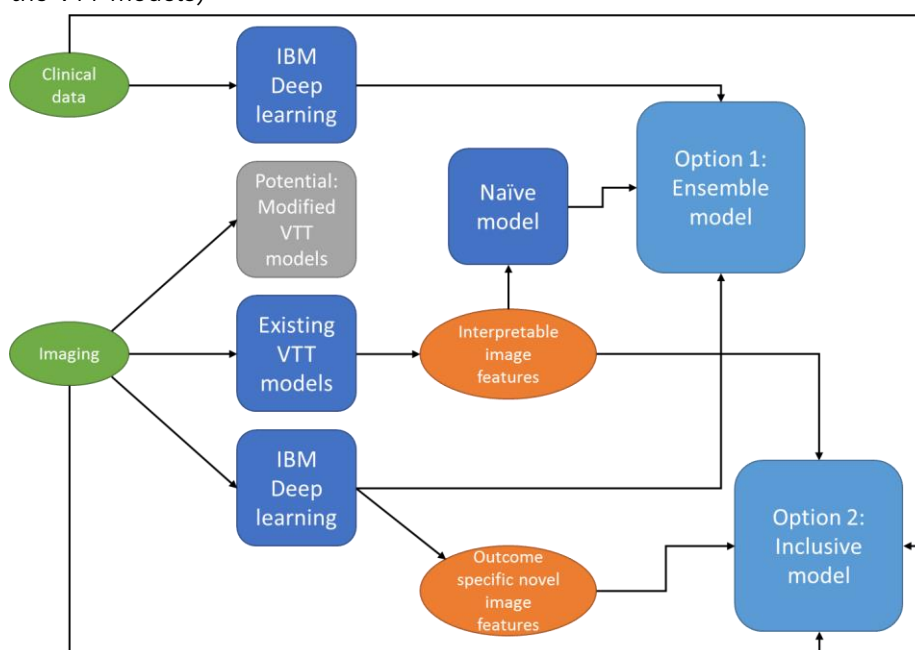


at CUR.

In summary, given images and clinical information, the models will retroactively predict the probability of success for each patient had they received each neoadjuvant treatment option. 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)



Summary:

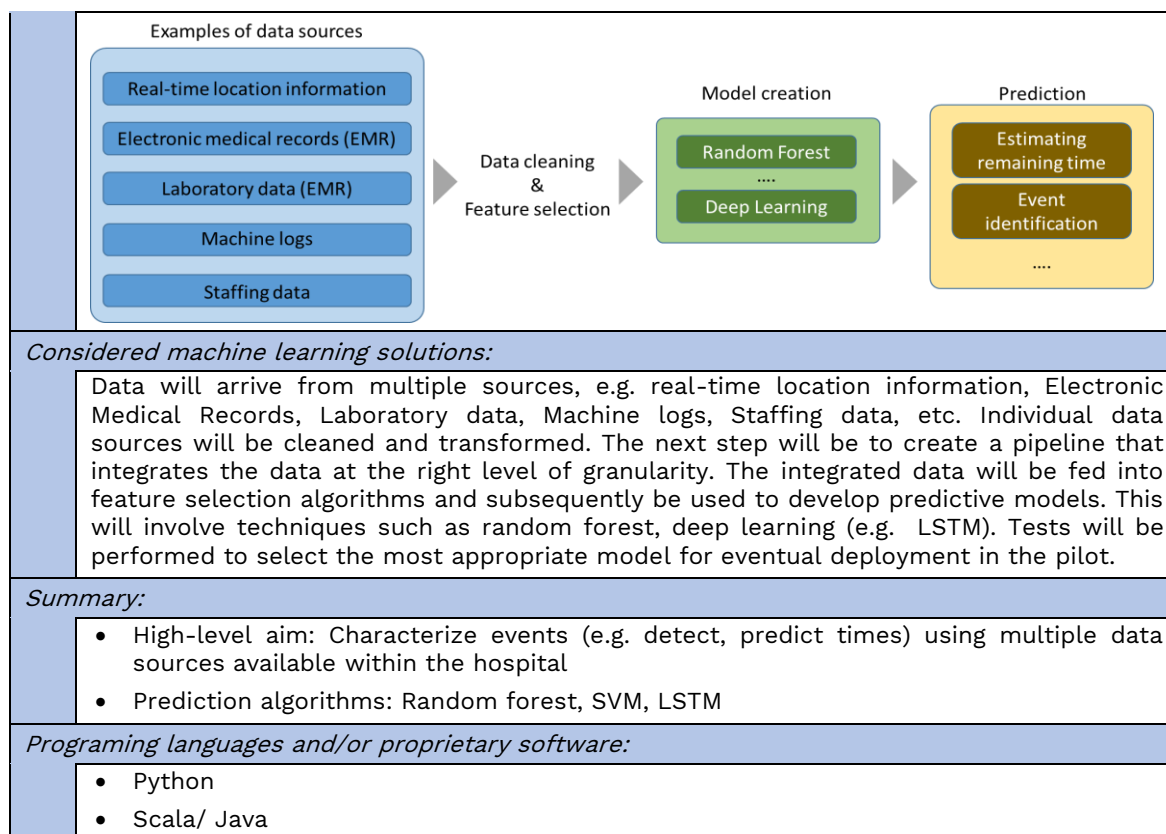
- High-level aim: treatment effectiveness analysis
- Data used: images, medical time-series, text documents
- Mentioned prediction algorithms: neural networks

Pilot 9 & 10 – 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.



2.3. Task 1.3: Complex real-time event detection

This task involves the integration of event detection components of health care related real-time data streams, which are encountered in pilot 3 (Diabetes), in pilot 4 (COPD/Asthma), in pilot 9 (hyper acute workflows), and in pilot 10 (asset management).

Concerning data streams, all pilots stream electronic health record data (EHR) that are transmitted through inputs of the doctors, including demographic information and laboratory values, such as glucose test results. Only pilot 4 (COPD/Asthma) uses symptoms filled in by the patients themselves and only this pilot integrates sensor data from environmental parameters, e.g. pollen and pollution information matched to the patient's location. Pilot 3 (diabetes) integrates glucose monitoring and food monitoring, entered by the patients, in addition to abdominal circumference measures entered by the doctors. Pilot 9 and pilot 10 use a real-time locating system (RTLS). The fastest velocity of the streams is 500 data points per seconds (RTLS, pilot 9/10), most other streams depend on patient entries and vary between 4 entries a day (glucose level, pilot 3) up to one entry per month (CAT score pilot 4). None of the pilots uses wearable data streams.

For pilot 3 and pilot 4, the specific situations of interest concern primarily symptom worsening/exacerbation (pilot 3 and pilot 4), based on glucose monitoring and COPD symptoms, respectively. Pilot 3 uses alerts in the form of a traffic light system, which alarms the physician and the patient of glucose levels and recommends actions to be taken. This pilot also needs to correlate events with historical data to assess unexpected results. Pilot 9/10 uses alerts when equipment enters unauthorized areas and it combines several streams into one complex event to raise alarms based on waiting time and particular stages of the care pathway.

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	<p>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:</p> <ul style="list-style-type: none"> • Personal information • Family background • Diagnosis • Primary and Secondary care visits • Hospital discharges • Hospital urgencies • Health habits 	
Data cleaning	<p>Next, we describe the data procedures carried on the data:</p> <ul style="list-style-type: none"> • 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. <p>Prescription data is analyzed and a diagnosis for such treatment is associated to the patient. This transformation helps further analysis.</p>	
Data integration and analysis	<p>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.</p>	

Pilot 2: Kidney disease (WP2)		Partners: <u>DFKI</u>
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 EHR (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.	
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.	
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.	
Considered machine learning solutions	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 natural language processing (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.	

Pilot 3: Diabetes (WP2)		Partners: <u>HUA, NISS</u>
Description of the prediction problem	<p>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.</p> <p>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.</p>	

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.
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 point. This won't affect the system at any point, and it will continue monitoring the patient.
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.
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: <ul style="list-style-type: none">• 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.

Pilot 4: COPD and Asthma (WP2)		Partners: <u>UNIS</u> , 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.	

Data cleaning	<p>We will require that any patient used in the production of the prediction data-driven model(s) has the following requirements:</p> <ol style="list-style-type: none"> 1. The patient has an associated location 2. They have at least one symptom score measurements. <p>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.</p> <p>We also expect to eliminate anomalies by utilising known correlation between the symptom score and CAT score.</p>
Data integration and analysis	<p>The data sources will be integrated at patients 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 following the patient's geo-location, as well as the relevant timestamps associated with the respective measurements in the time series.</p>

Pilot 5: Heart Failure (WP2)		Partners: EMC
Description of the prediction problem	<p>There are currently 15 million patients living with Heart failure (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 Heart Failure Pilot we are using several databases containing large number of Heart Failure patients and their comorbidities. These databases are both Hospital and Health Insurance based systems.</p> <p>Description of prediction problems:</p> <ol style="list-style-type: none"> 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	<p>We will use readily available data in existing databases at Achmea and Erasmus MC. Due to confidentiality, the Multi Party Computation 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.</p>	
Data cleaning	<p>The patient data will be cleaned with regards to our inclusion criteria for chronic heart failure. We will make sure the data is complete and normalized per feature</p> <p>For the MPC setting, the data cleaning will happen locally at each party (Achmea and EMC).</p> <p>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.</p>	

Data integration and analysis	<p>The data will be integrated on the patient level. We are making use of Multi Party Computation 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.</p> <p>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 learn the relation between various life style parameters and the number of hospitalizations.</p>
Considered machine learning solutions	<p>ad 1. We are applying Lasso regression for finding significant features. We will identify association rules for finding patterns in medication and other treatments. For this we will test 3 models: regression, Random Forrest and K-nearest neighbour.</p> <p>ad 2. The Multi Party Computation will focus on LASSO regression with a Stochastic Gradient Descent solution method for solving the underlying minimization problem.</p>

Pilot 6: Prostate Cancer (WP3)		Partners: PHI
Description of the prediction problem	<p>The pilot captures prostate cancer related data that will be used to:</p> <ol style="list-style-type: none"> 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	<p>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.</p>	
Data cleaning	<p>Data blending will be applied to the collected data to aggregate, filter or merge data.</p> <p>Feature engineering operations will be performed to create context relevant features (e.g. PSA density from 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)</p> <p>Data cleansing to automatically identify duplicates, outliers and anomalies.</p> <p>Feature value distribution analysis to identify features with extreme bias between treatment groups.</p> <p>Automated feature selection to select the optimal set of features for each prediction model.</p>	
Data integration and analysis	<p>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.</p>	

Considered machine learning solutions	RStudio with R as the modelling framework
---------------------------------------	---

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 accomplish 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.	
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.	
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.	
Considered machine learning solutions	At the level of raw data, natural language processing 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.	

Pilot 8: Breast Cancer (WP3)		Partners: <u>IBM</u> , <u>CUR</u> , <u>VTT</u>
Data acquisition	<p>The pilot will analyse mammograms, ultrasound and MRI images along with structured clinical data and textual reports to automatically predict patient response to breast cancer treatments, specifically neoadjuvant treatments.</p> <p>Image data will be provided in bulk, with some structured clinical data embedded in the DICOM image header.</p> <p>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, indications of complications and adverse events during treatment, and the outcome of the treatment.</p> <p>Another type of data will be extracted from the clinical reports. This may require NLP analysis of the clinical reports (which are mostly written in French)</p>	
Data cleaning	<p>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. This removal will be performed automatically.</p> <p>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.</p> <p>The clinical outcomes will be analysed for inconsistencies between the various data sources (HER vs. clinical report). The association between features and outcomes will be analysed to see the baseline predictive power available in the data</p>	
Data integration and analysis	<p>The various data sources will be integrated at the patient level.</p> <p>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.</p> <p>Data extracted from the image DICOM header will be added to the csv file containing the restructured clinical and outcome data and will be read by the analysis algorithm separately from the images.</p> <p>The analysis itself will create models 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.</p>	

Pilot 9, 10 & 11: Hyper-acute workflows, asset management (WP4)	
Description of the prediction problem	<p>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.</p>
Data acquisition	<p>Data will arrive from multiple sources, e.g. real-time location information, Electronic Medical Records, Laboratory data, Machine logs, Staffing data, etc.</p>
Data cleaning	<p>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.</p>

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.
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.

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 updated in additional visits to the doctor	The doctors will collect this data into the system. When a doctor registers a 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	The mobile app also allows the sending of exercise and food data. The patient needs to input manually this data into the app, so

		moderate activity is recommended	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.

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 is admitted to the hospital, or when there is an update to the EHR, and the hospital sends the new data to UPM.

Combination of different streams:

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

Challenges of different streams:

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

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 be queried 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.

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.

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

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

Techniques used for security preservation are de-identification of EMR (P9/10/11), RTLS (P9/10/11) and the use of HTTPS protocols (P9/10/11) and a single factor authentication (P9/10/11). Pilot 1 uses a VPN access for external users including a two-factor authentication, and access to the admin node for partners provided by the hypervisor of the cluster (Proxmox). 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.

Authorization in pilot 1 will be achieved through root access in the set-up mode and full-access in the development mode, yet without access to external networks. In the setup mode, connection to the EHR repository will never be granted under any circumstances.

Data protection of data at rest is achieved through de-identification of EMR and location data in pilot 9/10/11. Pilot 1 uses two de-identification processes, one is executed by the IT department of the data owner, and the other is executed by Incliva. Moreover, pilot 1 will also encrypt all data stored in the hardware with a 256 bit AES encryption algorithm and ensures that data will not be stored in external media. Concerning the data protection of data in transit, pilot 1 uses a VPN connection and Fortigate software in the firewall to ensure that all communications are encrypted with SSL/TLS standards. 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.

Pilot 1 will audit and store positive and negative (non-granted) accesses (timestamp and user login) to the VPN, to the SFTP server, and to the EHR Database server logs.

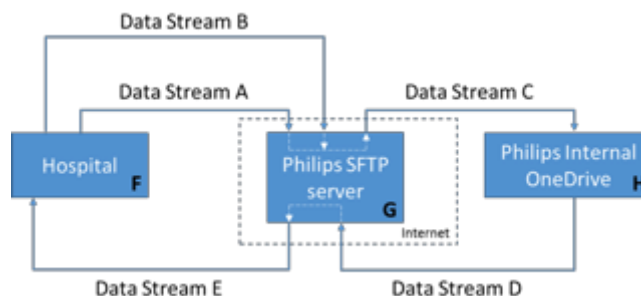


Figure 1: Security architecture of Pilots 9, 10 and 11

Multi-party computation techniques

Multi-party computation (MPC) enables a common secure analysis of datasets that cannot be openly shared. TNO advises to use Lasso regression, a secure regression variant that has an automatic feature selection and can be solved with Stochastic Gradient Descent algorithms.

The use case is built on fake data for a proof-of-concept. The data and outcome of computations that are performed are encrypted by means of Shamir secret sharing. The (pseudo) randomness is implemented by PEP 506. That means that nothing can be derived from the (secret shared) data of one party. Only if multiple involved parties would leak data, information may be retrieved by combining the shares.