Deliverable 1.4

Final implementations 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	44 months
Version	1.0		
Status	Final		
Date of issue	30/06/2020		
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	
Salvador Vigo Marmol	DFKI, salvador.vigo_marmol@dfki.de	
Alireza Rezaei Mahdiraji	DFKI, alireza.rezaei_mahdiraji@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 and D1.3 presented a more detailed analysis of the different pilots according to the WP1 tasks, including a first overview of the different software prototypes, components, data, and challenges. The current document presents the final implementations of the different developments in the pilots. In particular, this deliverable is a first step towards the BigMedilytics blueprint by consolidating and compressing information as they will be presented in the blueprint.

Table of Contents

1. Introduction 5

1.1. Purpose of the document 5

1.2. Related documents 5

2. Overview of specific components 6

2.1. Short Project Introduction 6

2.1.1. Pilot 1 - Comorbidities 6

2.1.2. Pilot 2 - Kidney disease 7

2.1.3. Pilot 3 - Diabetes 8

2.1.4. Pilot 4 - COPD/Asthma 9

2.1.5. Pilot 5 - Heart failure 10

2.1.6. Pilot 6 - Prostate cancer 12

2.1.7. Pilot 7 - Lung cancer 13

2.1.8. Pilot 8 - Breast cancer 13

2.1.9. Pilot 9 & 10 - Hyper-Acute workflows: Stroke and Sepsis management & Pilot 11 Asset Management 15

2.1.10. Pilot 12 - Radiology workflows 16

2.2. Multi-velocity processing of heterogeneous data streams 17

2.3. Processing of large structured / unstructured data sources 17

2.4. Complex real-time event detection 19

2.5. Deep learning for multilingual NLP and image analytics 20

2.6. Prediction algorithms 21

2.7. Security and privacy of data access and processing 24

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 is realised in three cycles: (1) initial prototypes, based on pilot requirements, (2) updated prototypes, based on pilot internal validation, and (3) final implementations, based on pilot external validation. The WP1 deliverables describe each of the cycles. The final outcome of WP1 is the Big Data Healthcare Analytics Blueprint, the BigMedilytics-BigMatrix, that is 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 current deliverable D1.4 presents the final implementations of all BigMedilytics pilots. While the previous WP1 deliverables were relatively long, this deliverable uses the previous input, as well as the collected information from personal interviews with technical partners from all pilots. We shortened much information from the previous deliverable D1.3 to make it easily accessible for the reader. This condensed form is also necessary to create the BigMedilytics Blueprint at the end of the project.

1.2. Related documents

Related documents: D1.1, D1.2, D2.1, D2.2, D2.3, D3.1, D4.1

2. Overview of specific components

This document presents information collected from the previous deliverables, as well as different interviews with people working on the technical aspects within the pilots. All the information has been "compressed" to access and understand the information in an easier way. This compression is a necessary step towards the BigMedilytics Blueprint and the BigMedilytics-BigMatrix which will combine high-level, as well as, detailed information. In particular, the high-level information is necessary to perform a better comparison among the manifold pilot solutions.

The current document is structured as follows: First, we present a high-level architecture of each system, including a brief description. Then, we will present different relevant aspects of the pilots according to various Big Data topics, which represent the different WP1 tasks. Please note that the order of the tasks will be different as originally stated in the BigMedilytics proposal. This re-ordering has been already presented at the last review meeting in Luxembourg in autumn 2019, in order to present a more logical way of a data processing pipeline. Moreover, note that this is a static text document which makes it more difficult to navigate in details various aspects of different pilots. Therefore, the BigMedilytics-BigMatrix will be a digital solution to overcome this disadvantage.

In this document, we will indicate the contributions of each pilot to each task in a table similar to this one below. If a pilot contributes to a given task, the pilot will be marked with a red dot. Please note that by putting the content of each pilot together we found that the contributions of various aspects should be sorted differently. The reason for that will be explained in each section and these situations will be highlighted with a yellow dot.

Table 1. Overview of all BigMedilytics pilots across the three themes.

Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidities (1)	Kidney Disease (2)	Diabetes (3)	COPD/ Asthma (4)	Heart Failure (5)	Prostate Cancer (6)	Lung Cancer (7)	Breast Cancer (8)	Stroke (9)	Sepsis (10)	Asset Management (11)	Radiology Workflow (12)

2.1. Short Project Introduction

In order to get a better understanding of each single pilot, this section provides a short high-level presentation of each single BigMedilytics pilot. Table 1 shows all the pilots of the project categorized in 3 groups. Each pilot will be presented with some technical keywords, a description and at least one schematic overview.

2.1.1. Pilot 1 - Comorbidities

Keywords:

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

Description:
This pilot primarily addresses the long-term treatment of chronic disease patients and aims to develop a risk prediction model to reduce costs by directing patients to primary or secondary care where emergency care and hospitalization are not required. Using a Big Data approach, the disease trajectories and care pathways of a large patient population are characterized over an extended time period. Thus, this

task has the potential to unravel the pattern of a disease as well as, e.g., previously unknown links among disease groups.

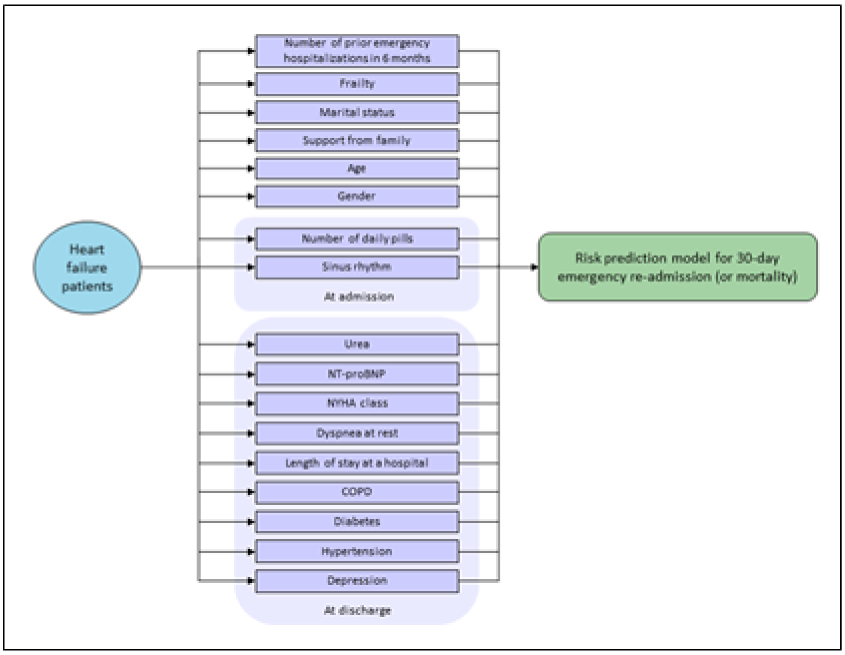


Figure 1: Different information which are fed into the risk prediction model of Pilot 1.

Figure 1 shows a tentative example of the large variety of different information which are then fed into the risk prediction model to identify: exitus and hospitalization due to comorbidities.

2.1.2. Pilot 2 - Kidney disease

Keywords:

- Adherence monitoring
- Risk prediction
- Prediction of rejection, transplant loss, infections
- Complex event processing

Description:

The aim of this pilot is to support clinical staff in the kidney transplant center which treats patients before and after a kidney transplant. Main data sources are the TBase system, an electronic health record system, containing patients of the last 20 years, and a patient app which provided vital parameters and daily drug intake of the patients. Using this data, the pilot addresses the following tasks: a) adherence monitoring, b) risk prediction and c) complex event processing. All tasks aim to detect critical patients to reduce complications, (therefore, reduce hospitalizations), reduce costs, and improve the quality of life.

Figure 2 presents the high-level architecture of the pilot. Outpatient data, such as vitals, can be sent via app to the hospital and is then stored in the EMR system TBase, which includes many other patient relevant information. The development of the different AI models is carried out at the hospital infrastructure and interacts with the EMR system. The information from the EMR, the patient app, and the results of the AI models are then visualized in the dashboard.

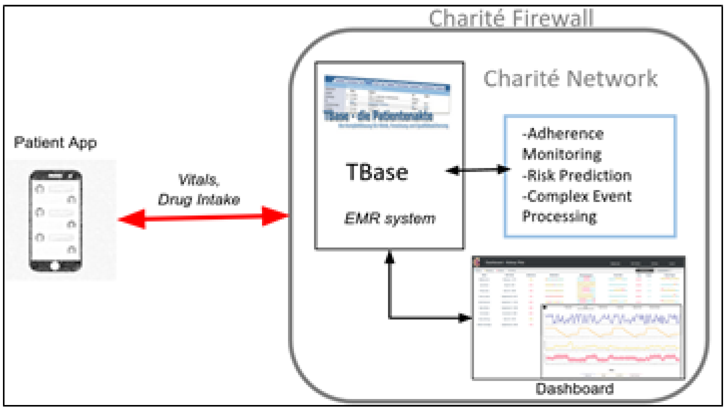


Figure 2. The high-level architecture of Pilot 2.

2.1.3. Pilot 3 - Diabetes

<u>Keywords:</u>

- Complex event processing
- Detection of problematic patients
- Blood sugar monitoring
- Glucose level monitoring

<u>Description:</u>
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 countries. The objective of the pilot is to improve the efficiency and frequency of blood sugar level monitoring by using data from patients, experience coming from the midwives, and medical guidelines. This will allow healthcare professionals to focus on the patients who are in greater risk, and reduce the visits to the hospital of patients who are having a non-problematic pregnancy.

Figure 3 describes diabetes patients using an application to record, track and manage their glucose levels. The data can be sent via app to the hospital server, and is shown to medical staff on the medical portal. In this use-case medical staff (doctor) uses the portal to create new patients, review their data, and directly communicate with patients by sending notifications to their app to call them for checkup. Only after being enrolled by medical staff, patients can log into the app and start using it.

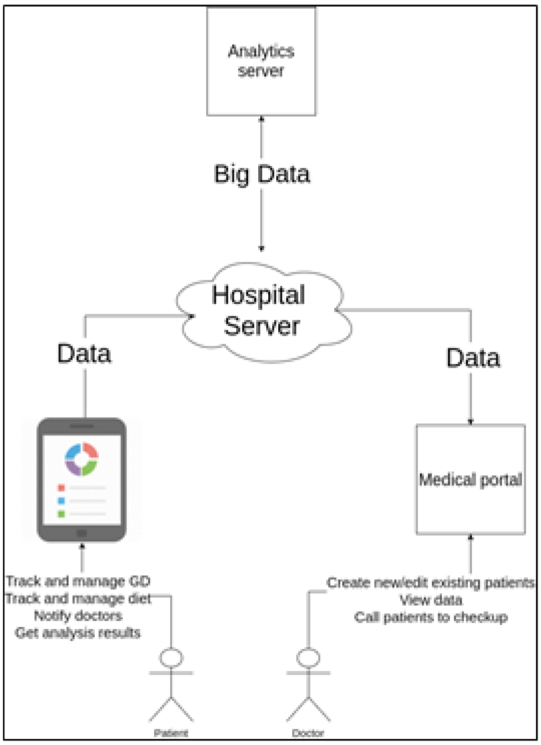


Figure 3. Overview of architecture of Pilot 3.

2.1.4. Pilot 4 - COPD/Asthma

Keywords:

- Risk prediction
- Predict acute exacerbations of COPD

Description:

This pilot utilises data captured using mobile and web enabled platforms MY COPD and MY Asthma (MyMhealth data), which 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 utilises data captured on the platforms to create models using daily data on symptoms, treatment, and environmental observation data including temperature, humidity, pollen counts and air pollution to create risk models which are personalized to the patient's own disease state and environment.

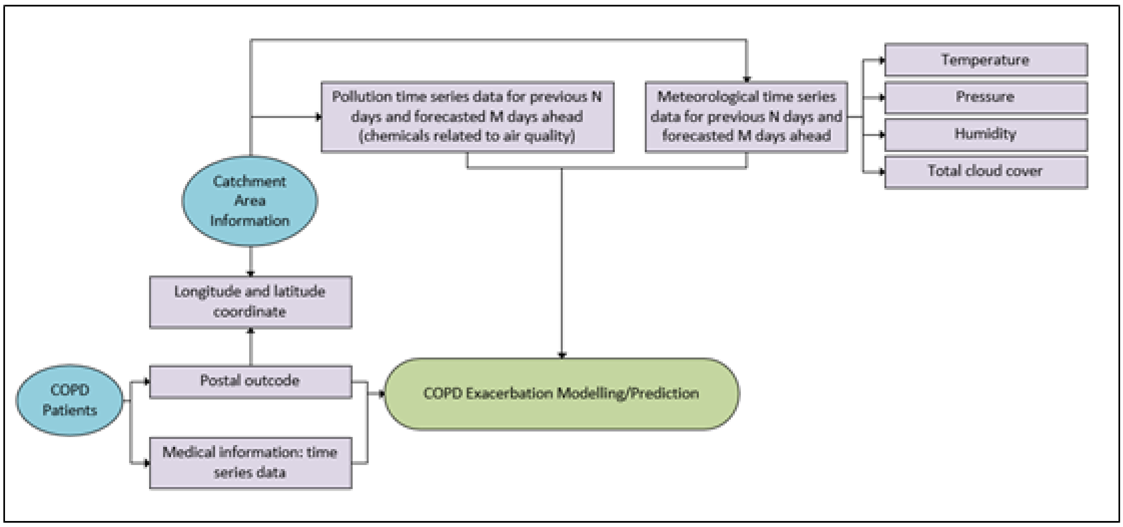

Pollution time series data for previous N days and forecasted M days ahead (chemicals related to air quality)
Meteorological time series data for previous N days and forecasted M days ahead
Temperature
Pressure
Humidity
Total cloud cover
Catchment Area Information
Longitude and latitude coordinate
COPD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

Figure 4a. Overview of Pilot 4.

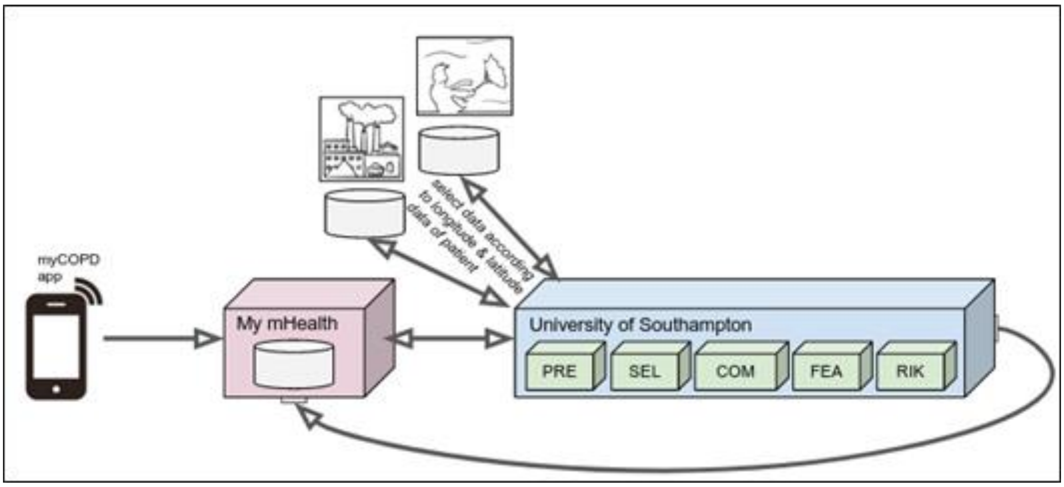

myCOPD app
My mHealth
select data according to longitude & latitude data of patient
University of Southampton
PRE
SEL
COM
FEA
RIK

Figure 4b. Overview of Pilot 4.

Figure 4a and Figure 4b show an overview of Pilot 4. Users of the myCOPD app provide data (demographics, self-reported symptoms and CAT) which is then sent to My mHealth. FIrst, this data is preprocessed (PRE) which includes the identification of longitude and latitude position of a patient according to the given postcode. Note only the first part of the postcode is available. Using this longitude and latitude data, weather and pollution data according to those areas can be selected from the corresponding databases (SEL). Next data from all three sources is combined (COM) and relevant features extracted (FEA). Finally those features are fed into the risk prediction models (RIK). The output score is then stored back at My mHealth under the profile of the patient.

2.1.5. Pilot 5 - Heart failure

Keywords:

- Multi-Party Computation
- Encryption
- Secure lasso regression
- Prediction model

Description:

The goal of this pilot is to identify comorbidities that have the strongest correlation with the number of hospitalizations in context of heart failure. These comorbidities are then used to guide an intervention that is designed to reduce the number of hospitalizations. A Multi Party Computation (MPC) system is developed 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 a secure analysis of two datasets that the owners cannot share with the others.

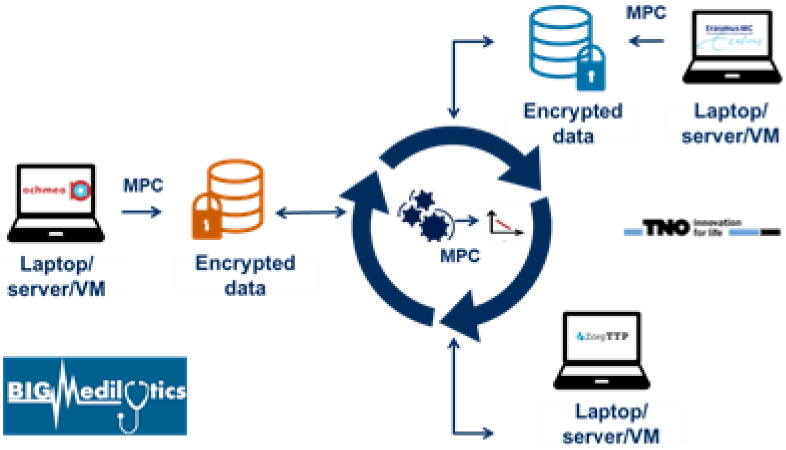


Figure 5a. Overview of Pilot 5.

Figure 5a presents an overview of pilot 5 which is described in the following:

1. The data is collected and is loaded into the MPC application that is installed at a laptop, server or VM at both Achmea and Erasmus MC.
2. The MPC tool encrypts all data. Each party uses its own public key for encryption and keeps the private key secret.
3. The MPC tool of the different parties (Achmea, Erasmus MC and ZorgTTP) exchange encrypted data in such a way that the prediction model can be computed without revealing anything else about their data.
4. One or more of the parties receive the computed prediction model.
5. Before, during or after the computation process, no input data of one of the parties is received by any other party (only in encrypted form, which cannot be decrypted by the receiving party).

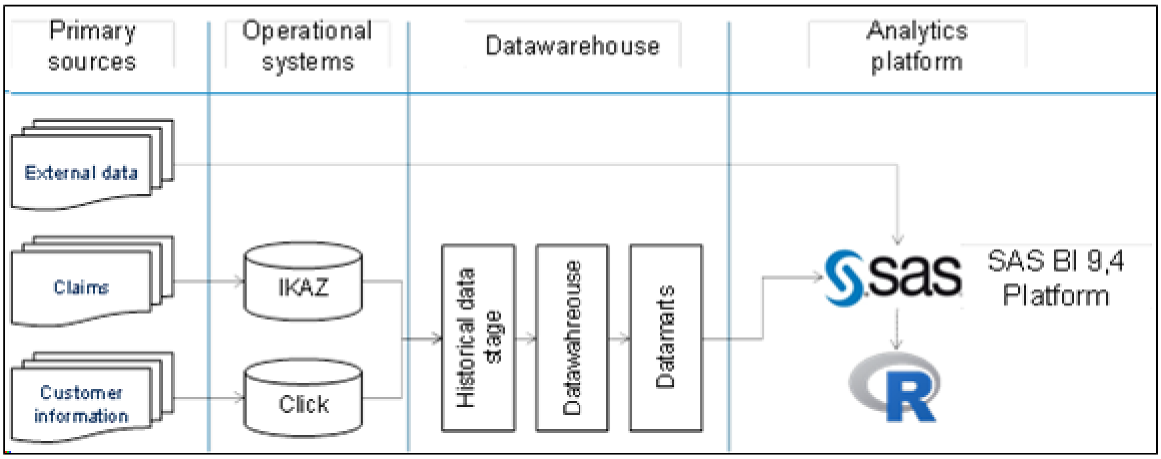


Figure 5b. Overview of Pilot 5.

2.1.6. Pilot 6 - Prostate cancer

Keywords:

- Clinical decision support
- Medical risk prediction
- Treatment risk prediction
- Support of the surgery strategy decision making process

Description:
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 the 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, e.g. nerve bundles 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. This pilot aims 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. To this end, multiple risk models are provided based on the integration of heterogeneous data sources like demographics, laboratory, imaging, and histology.

ISPM (the Philips tumour board application) is an officially released product for clinical use, where all the required security measures are implemented (Figure 7). ISPM is installed on premise, on a windows server VM administered and controlled by the hospital IT. ISPM consists of angular/typescript frontend and a java micro services backend, running on a local web server. Data is stored in the 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 developed in BigMediyltics, also installed on the VM. It retrieves de-identified data from ISPM, to execute statistical analysis. The results of the analysis are then visually processed and presented to the medics.

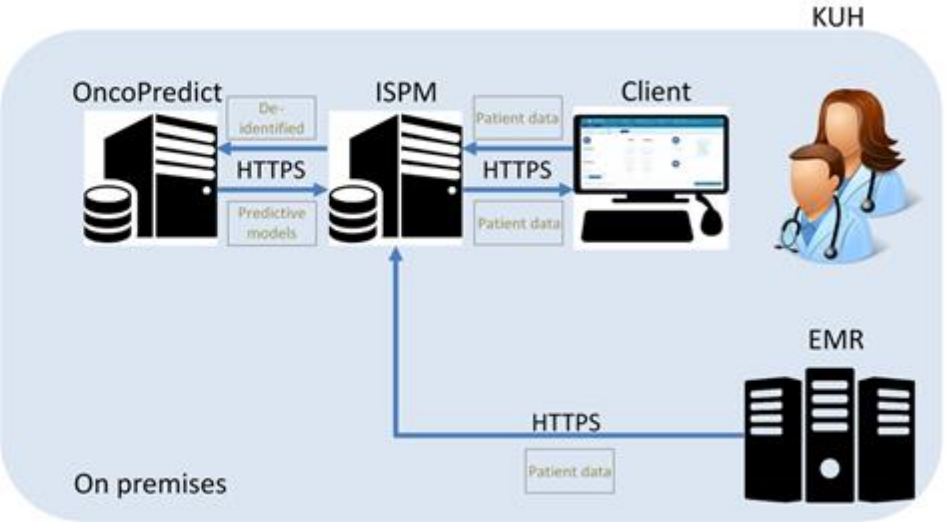


Figure 6. ISPM product of Pilot 6.

2.1.7. Pilot 7 - Lung cancer

Keywords:

- Treatment effectiveness analysis
- Drug adverse prediction

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

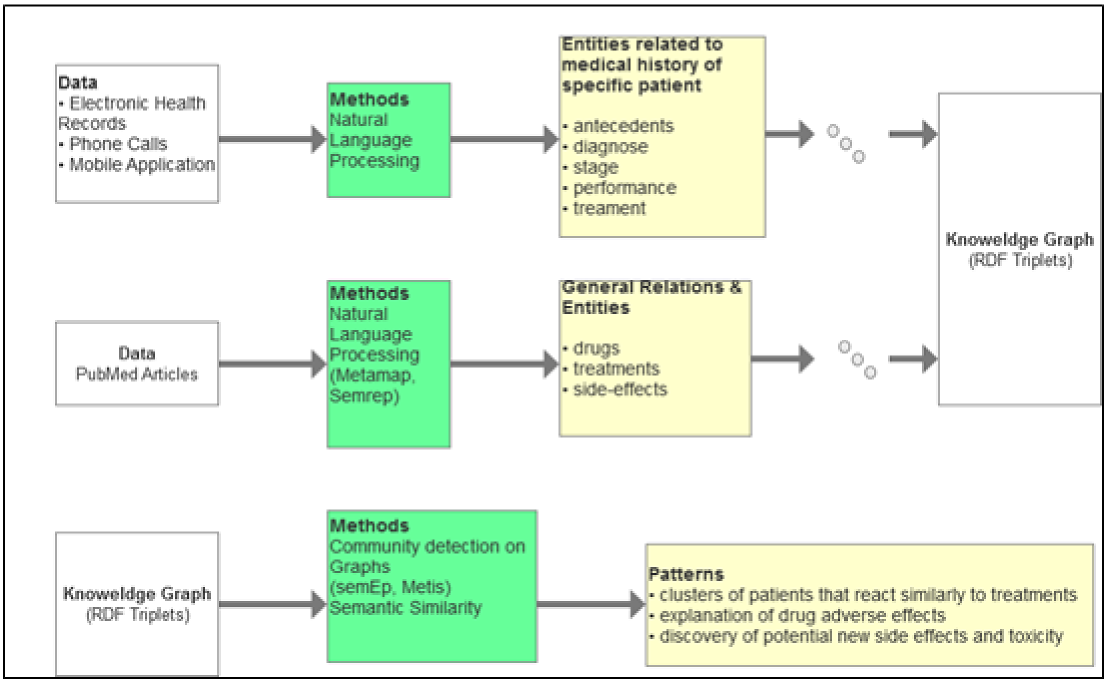

Data
• Electronic Health Records
• Phone Calls
• Mobile Application
Methods
Natural Language Processing
Entities related to medical history of specific patient
• antecedents
• diagnose
• stage
• performance
• treament
Knoweldge Graph
(RDF Triplets)
Data
PubMed Articles
Methods
Natural Language Processing (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

Figure 7. Different data sources and processing steps of Pilot 7.

Figure 7 presents the different data sources and processing steps in the lung cancer pilot. Data from electronic health records and publications as preprocessed applying NLP methods, such as named entity recognition and relation extraction and then normalized to a structured vocabulary. Extracted information are then stored as triplets in a knowledge graph. On this graph further methods are applied to detect patterns, e.g., among potential drug-drug interactions, toxicities, and visits to the patients to the hospital.

2.1.8. Pilot 8 - Breast cancer

Keywords:

- Clinical decision support
- Predict patient response to breast cancer treatments
- Predict treatment effectiveness

Description:
The pilot is a retrospective study that analyses 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 in IBM and VTT, through a VPN to access a local server at CUR. In summary, given images and clinical information, the models predict the probability of success for each patient who received neoadjuvant treatment. These models allow for evaluating the ability to make personalized treatment decisions rather than following global population guidelines and allow for assessing the economic effect of such protocols.

The overall pilot architecture is depicted in Figure 8a and 8b. To comply with regulations as GDPR, we use a model-to-data paradigm where all the data remains at Institut Curie infrastructure. All computations are applied on a strong GPU enabled server that resides in Curie, and various docker containers and pipelines of analytics models are transferred to the server and executed there. The overall flow is as follows: the anonymized imaging and clinical data are transferred from Curie expert repositories to the pilot server hosted within the institute infrastructure. Training and inference pipelines utilize the data and produce analytics results. The analytics results are stored in a repository and an application is used to visualize those analytics results.

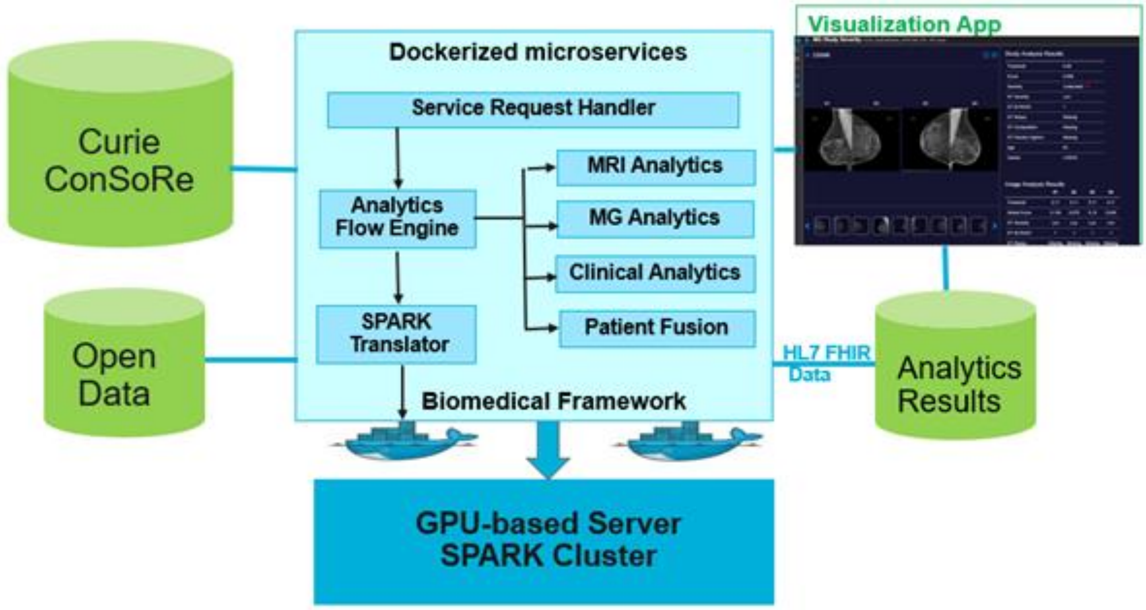


Figure 8a. Architecture of Pilot 8.

The pilot includes multiple pipelines. For example, one of the pipelines published in SPIE Medical Imaging 2020 predicts pathologic complete response via the pipeline depicted in Figure 10. Main data source used for this pipeline is the data of Curie, which includes the clinical data (structured information) and image data (MG: mammogram). We use XGBoost to create a model out of the structured clinical data. For the imaging data, we use a pre-trained deep learning (DL) model to get the tumor location and then compute radiomics texture features (Gabor, LBP, GLCM, wavelets) within tumor area and in the peritumoral area. The output of both models is then stacked into an ensemble model and a final prediction is made.

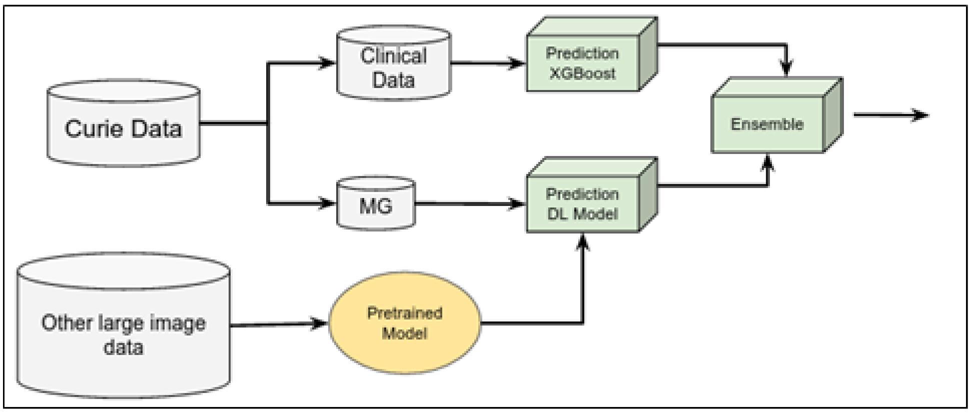


Figure 8b. Example Pipeline of Pilot 8.

2.1.9. Pilot 9 & 10 - Hyper-Acute workflows: Stroke and Sepsis management & Pilot 11 Asset Management

Keywords:

- Workflow optimization
- Bottleneck detection
- Data quality

Description:
The pilots focus 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) 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. Pilot 9, 10 and 11 have strong similarities, however while pilot 9 and 10 use staff/patient tracking, pilot 11 targets assets.

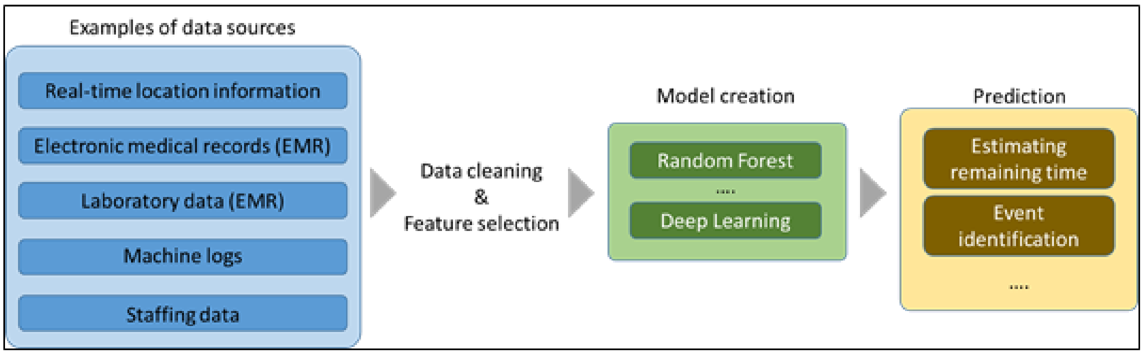


Figure 9. Overview of Pilot 9, 10 & 11.

As Figure 11 shows, data arrives from multiple sources, e.g., real-time location information, electronic medical records, laboratory data, machine logs, staffing data. Individual data sources will be cleaned and transformed, which is necessary to provide reliable results. Then a pipeline is created 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. Tests will be performed to select the most appropriate model for eventual deployment in the pilot.

2.1.10. Pilot 12 - Radiology workflows

Keywords:

- Clinical decision support
- Help to find the correct finding
- Finding similar cases
- Similarity search

Description:

This pilot focuses 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 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 database of cases, ranks those cases and shows the most similar ones. Furthermore, it provides a summary and scoring of the findings, and additional information are presented for supporting radiologists.

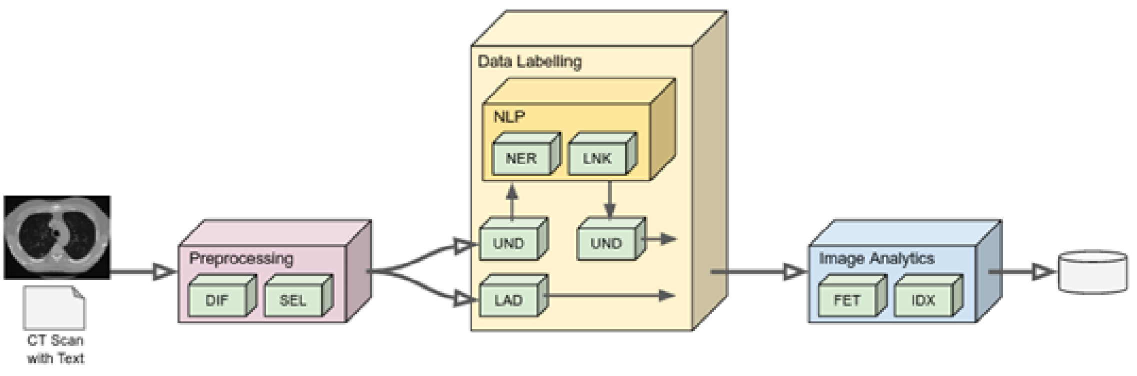


Figure 10a. Preprocessing Overview of Pilot 12.

Figure 10a shows the pipeline to generate the index: Preprocessing includes de-identification (DIF) and subset selection (SEL), then in data labelling a small set of labelled data (LAD) is manually generated. The rest instead is unlabelled data (UND), which is processed with methods of natural language processing (NL), namely named entity recognition (NER) and entity linking (LNK) in order to generate weakly labelled data (WLD). LAD and WLD are then further processed in image analytics. First features are extracted (FET) and then data is indexed (IDX) for fast access.

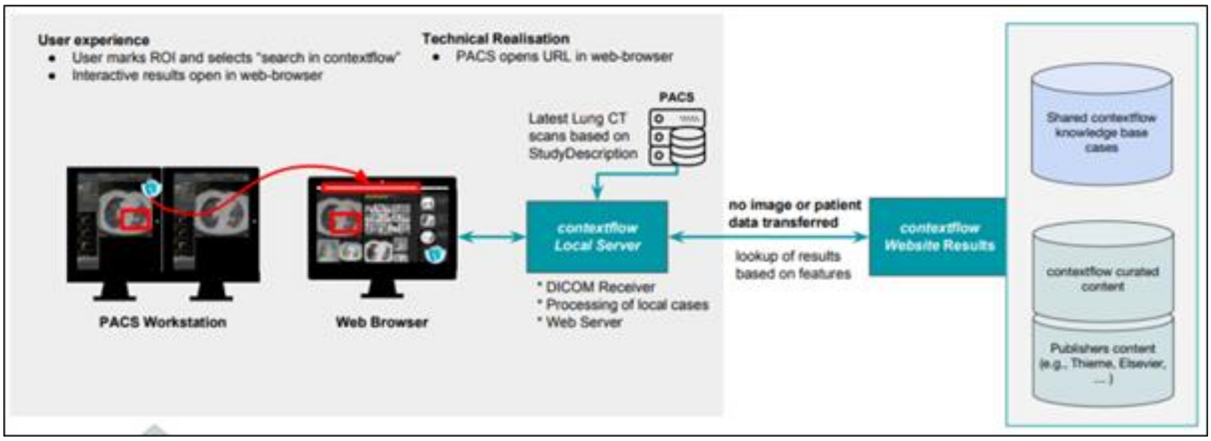
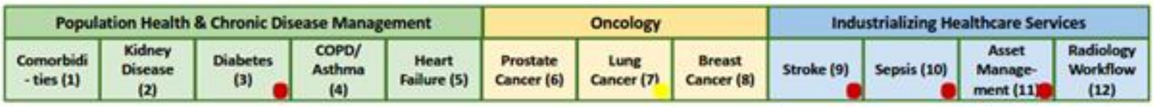

User experience
• User marks ROI and selects "search in contextflow"
• Interactive results open in web-browser
Technical Realisation
• PACS opens URL in web-browser
PACS
Latest Lung CT scans based on StudyDescription
PACS Workstation
Web Browser
contextflow Local Server
• DICOM Receiver
• Processing 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, ...)

Figure 10b. Pilot 12 in use.

Figure 10b shows the pilot's use case in the clinic: A region of interest is selected, features extracted and then sent to the Contextflow search engine/index. Note, not image or patient data is transferred - only features.

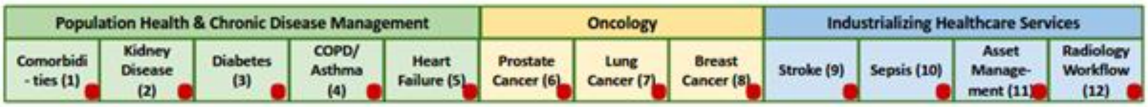
2.2. Multi-velocity processing of heterogeneous data streams



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi - ties (1)	Kidney Disease (2)	Diabetes (3) ●	COPD/ Asthma (4)	Heart Failure (5)	Prostate Cancer (6)	Lung Cancer (7) 🟡	Breast Cancer (8)	Stroke (9) ●	Sepsis (10) ●	Asset Manage- ment (11) ●	Radiology Workflow (12)

Multi-velocity processing of heterogeneous data streams, and is relevant for four different pilots, namely, gestational diabetes (pilot 3), hyper-acute workflows (pilot 9&10), and asset management (pilot 11). Pilot 7 (lung cancer) has been removed from the original list, as data presented in the previous deliverables did not include multi-velocity data. However, the presentation of the data streams along with their frequency is merged into Table 2 in the next section "Processing of large structured/unstructured data sources".

2.3. Processing of large structured / unstructured data sources



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi - ties (1) ●	Kidney Disease (2) ●	Diabetes (3) ●	COPD/ Asthma (4) ●	Heart Failure (5) ●	Prostate Cancer (6) ●	Lung Cancer (7) ●	Breast Cancer (8) ●	Stroke (9) ●	Sepsis (10) ●	Asset Manage- ment (11) ●	Radiology Workflow (12) ●

The consortium includes a large variety of different data sources and datasets which are used for the different pilots. This section presents a structured overview about these difference sources including additional pilot-specific information.

Table 2. Data variety & speed

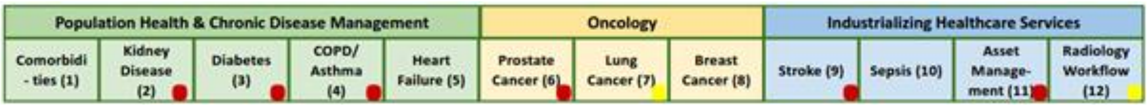
DATA	Pilot 1	Pilot 2	Pilot3	Pilot 4	Pilot 5	Pilot 6	Pilot 7	Pilot 8	Pilot 9/10	Pilot 11	Pilot 12
EMR	retrospective, 5mil	each visit	each visit		each visit	each visit		retrospective			retrospective
Socio-demographi	smoking, activities	smoking, drinking	x	smoking, rehabilitatio	smoking, drinking	x		x			

				n							
structured	lab, diagnoses, medications, hospital stay	lab, diagnoses, medications, hospital stay			lab, diagnoses, medications, hospital stay	x		treatment response, diagnoses, pathological complete response	lab (time-stamps)		
image						Histology		MRI, Mommo-grams, Ultrasound			CT
text		GER				SWE	SPA	FRE			GER
APP		once a day	each time								
food intake			x								
medication intake		x									
vitals		x									
Exercise			x								
gestational weight			x								
CAT scores				variable/monthly							
Symptom scores				every day							
location data				once							
Devices											
Glucometer			3-4x/day								
Other											
Pollution				daily							
Weather				daily							
Insurance					retrospective						
Publications							ENG				
Knowledge Bases							DrugBank, PubMed,				

							SIDER, UMLS, DBpedia				
phone logs							x				
RTLS									1.5 sec	1.5 sec	
Machine Logs									x	x	
Staffing									X		

Table 2 shows the full variety of all different data sources used. While some data sources have a high velocity (e.g. RTLS), most resources are rather infrequently updated (e.g. EMR is updated each time a patient comes around for a visit).

2.4. Complex real-time event detection



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi - ties (1)	Kidney Disease (2) 🔴	Diabetes (3) 🔴	COPD/ Asthma (4) 🔴	Heart Failure (5)	Prostate Cancer (6) 🔴	Lung Cancer (7) 🟡	Breast Cancer (8)	Stroke (9) 🔴	Sepsis (10)	Asset Manage- ment (11) 🔴	Radiology Workflow (12) 🟡

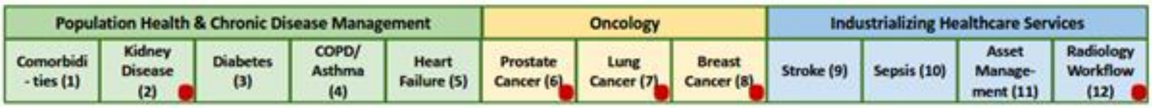
Complex Event Processing (CEP) for event detection has been a crucial part for many pilots in the project. Now more than ever the telemedicine approach for patient monitoring is a key resource to improve the functional service of the healthcare sector. Chronic disease patients are attached to a full lifetime of medic therapy and they need periodic supervision from specialist physicians. This monitoring process poses some problems for both patient and healthcare resources, e.g., the expenses of traveling to hospitals, waiting lists, and risk of contagion inside care facilities. The following table describes the main insights and specifications from the pilots that apply CEP categorized in three sections, namely, notifications, event type, and actions. For instance, the pilots have been forced to adapt from simple (e.g., parameter limits) to complex situations (events along time), to produce notifications and warnings (awareness) or alarms (critical) situations identification. Some specific actions in the workflow have been made in the process. Table 3 summarizes notifications, event type, and actions related to CEP in pilots dealing with CEP.

Table 3. Structured Overview about CEP

Pilot	Notifications	Event type	Actions
2	· Traffic Light system · Risk level based on thresholds related with blood pressure, heart rate, body weight and temperature	· Simple: warnings/alarms based on single thresholds · Complex: Simple events along period evolution. Weekly measurements	· Patient status (warnings, alarms) output triggers back into the main hospital system. · Dashboard for medics interface display
3	· Traffic Light system · Risk level based on historical evolution	· Threshold: glucose levels · Time complexity: 1 day	· Filter: values in range · Transform: sink average measurements to clinician · Other: Enrich model prediction info with historical records
4	· Risk level of worsening of COPD symptoms	· Predicted patient symptoms within severe states	· Transform: input patient symptoms are enriched with the weather, pollution and pollen

			data
6	· Likelihood score · If available combined with recommendation from the EAU or NCCN	· Simple event above likelihood score	· Filter: Noisy data · Transform: Percentage calculation and normalization
9, 10	· Time exceeded for particular workflow step	· Particular time thresholds proposed	· Filter: Tags depend on the location · Transform:Reconstruct new event trace by aggregating RTLS and EMR data streams
11	· Assets below required level · Assets out of bounds	· No. of assets below a threshold · Assets in disinfected areas	· Filter: Tags depend on the location

2.5. Deep learning for multilingual NLP and image analytics



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi-ties (1)	Kidney Disease (2) ●	Diabetes (3)	COPD/ Asthma (4)	Heart Failure (5)	Prostate Cancer (6) ●	Lung Cancer (7) ●	Breast Cancer (8) ●	Stroke (9)	Sepsis (10)	Asset Manage-ment (11)	Radiology Workflow (12) ●

Natural language processing (NLP) and image analytics is applied in various pilots. In most cases NLP is a way to extract additional structured information to support the main task. However, for the image analytics pilots instead, this technology describes the main contribution. Pilot 8 uses image processing to find the best treatment. In this case the image analytics part is closely interconnected to the main prediction task. Conversely, Pilot 12 is not involved in the prediction task, instead the image analytics is used to find similar image material to support the radiologist. The following tables present the core NLP and image analytics components. Table 4 shows a short description of each pilot dealing with NLP and image analytics. Table 5 lists NLP techniques and implementations used in BigMedilytics pilots. Table 6. shows an overview of core image analytics (sub-)tasks.

Table 4. Short description of different pilots regarding natural language processing and image analytics

Pilot 2	**Pilot 6**	**Pilot 7**	**Pilot 8**	**Pilot 12**
Information is extracted from clinical notes to access information better and to support the risk prediction model with additional information.	Extracted information from clinical text documents are used to support the main task, the risk prediction. These extracted information serve, together with various other information, as input for the prediction models.	NLP is one of the core aspects in pilot 7. Multiple text sources in English and Spanish exist which need to be processed so that structured information can be extracted. This data is then used to detect possible patterns.	Image processing is the main task of pilot 8. This data is used to find the most suitable treatment. To support this task also text data is used to a small extent, extracting e.g. tumor size or properties from notes.	Main focus lays on finding similar images in order to provide appropriate findings, however also text is used for semi-supervision, but it plays a minor role in the pilot.

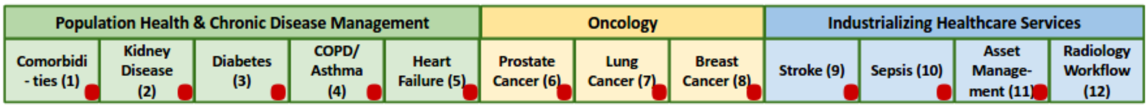
Table 5. NLP techniques and implementations used in BigMedilytics pilots

	Pilot 2	Pilot 6	Pilot 7	Pilot 8	Pilot 12
Language	German	Swedish	Spanish, English	French	German
Part-of-Speech Tagging			MLP (OpenNLP)		
Named Entity Recognition	Bi-LSTM (Flair)	Regular Expression (custom)	Rule-based (Stanford Core NLP, MetaMap)	custom implementation	custom
Relation Extraction	CNN (Flair)		(SemRep)		
Negation Detection	Rule-based (NegEx)	Rule-based (NegEx)			
Normalization			Rule-based (MetaMap)		custom

Table 6. Overview about core image analytics (sub-)tasks

	Pilot 8	Pilot 12
Image Data	Mammograms, MRI	CT
Feature Extraction		Moving CNN over image to reduce and compress information for feature (vector) extraction
Classification	Using CNN to process longitudinal image data and to train model	
Similarity Search		Using custom search engine to find similar images based on given feature vectors

2.6. Prediction algorithms



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi-ties (1) ●	Kidney Disease (2) ●	Diabetes (3) ●	COPD/ Asthma (4) ●	Heart Failure (5) ●	Prostate Cancer (6) ●	Lung Cancer (7) ●	Breast Cancer (8) ●	Stroke (9) ●	Sepsis (10) ●	Asset Manage-ment (11) ●	Radiology Workflow (12)

The following tables present the risk prediction within the different pilots. Table 7 presents the different risk prediction tasks which have been addressed and Table 8 presents the selected methods to solve the problem. As the table shows, most pilots rely on multiple solutions. Moreover, tree-based methods such as decision trees are the most popular method of choice, followed by a large variety of different neural approaches.

Table 7. Overview about BigMedilytics prediction tasks

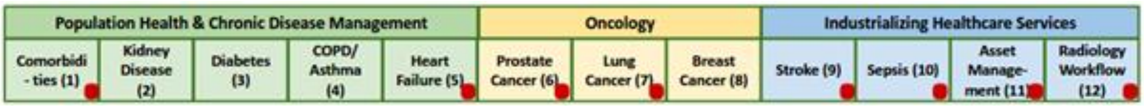
Pilot	Prediction problem categories			
	Medical event prediction	**Comorbidity analysis**	**Treatment risk analysis**	**Other**
1		• Comorbidities aggrupation • Relevant features extraction from HER • Hospitalization & Mortality risk prediction model		
2	• Risk Prediction related to: Infection, Rejection, Transplant Loss			• Adherence Monitoring
3				• Monitoring and categorizing of glucose levels
4	• Predict acute exacerbations of COPD			
5	• Prediction of heart failure patient hospitalizations	• Secure Lasso Regression		
6			• 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.	
7	• Predict Long Term Survivors	• Correlation between Commodities and Toxicities	• Prediction of Drug-Drug Interactions	
8	• Predict pathologic complete response (pCR) to neoadjuvant chemotherapy treatment (NACT)			• Predict cohorts for clinical trials towards next generation therapies
9				• Bottleneck
10				

11				formation in workflow • Timing characteristics in workflow • Periodic Automatic Replenishment (PAR) level

Table 8. Prediction Models - High Level Machine Learning Solutions

Pilot	**Machine Learning Solutions**				
	Tree	**SVM**	**Probabilistic**	**Neural Net**	**Other**
1	Decision Tree		Statistical inference techniques	MLP Regressor	
2	Gradient Boosted Regression Trees			Custom neural architecture consistent of pooling and fully connected layers	Rule-based Complex Event Processing
3					Rule-based (fuzzy rules) Complex Event Processing
4	Random Forest			Neural networks for sequence analysis	
5					Secure lasso regression
6	Random forest				
7	x	x		x	
8	XGBoost, Random Forest			Convolutional Neural Network (see also Image Processing)	Logistic Regression, Texture Features
9	Random forest			Long-short term memory	
10	Random forest			Long-short term memory	
11	Random forest			Long-short term memory	

2.7. Security and privacy of data access and processing



Population Health & Chronic Disease Management					Oncology			Industrializing Healthcare Services			
Comorbidi - ties (1) ●	Kidney Disease (2)	Diabetes (3)	COPD/ Asthma (4)	Heart Failure (5) ●	Prostate Cancer (6) ●	Lung Cancer (7) ●	Breast Cancer (8)	Stroke (9) ●	Sepsis (10) ●	Asset Manage- ment (11) ●	Radiology Workflow (12) ●

The aim of this section is to provide an overview of the security procedures and measures adopted by the pilots to guarantee the security of the data involved in each of the pilots among the project. The heterogeneity of the clinical partners requirements in each specific use case has represented a major challenge in most of the cases. At the same time the GDPR legislation took place in 2018, adding more intensity to the data management process. The following Table 9 provides a structured overview about the technical security and privacy solutions of the different BigMedilytics pilots to implement material GDPR fulfilling conditions. Table 9 provides a structured overview about the technical security and privacy solutions of the different BigMedilytics pilots to implement material GDPR enforcement conditions.

Table 9. BigMedilytics Security and Privacy Overview

Pilot	Access Control	Data Protection	Audit/Log measurement
1	**Authorization:** Two acces mode . Setup mode: Root access to virtual machines . Development mode: access without root user permission **Authentication:** . External access to infrastructure: Each partner has a unique user id and password to access to VPN . Access to infr. resources: Each partner will be granted to access to the admin node using a remote desktop console provided by the hypervisor of the cluster	**Data at rest:** Sensitive data remains on premise. Data is encrypted with a 256 bit AES encryption algorithm. **Data in transit:** External connections are established by VPN connection. The communication to the infrastructure is encrypted using SSL/TLS. Transfer of output data via SFTP once VPN connection is established. Internal communication is encrypted.	**App/Serv. logs:** . VPN logs: positive and negative accesses to VPN is audited and securely stored . SFTP logs: both positive and negative accesses to SFTP will be audited and securely stored . EHR Database server logs: Securely stored by Incliva **System logs:** Linux auditd tool for auditing the system. Specific events to audit are: . Software and libraries installed into master and processing nodes . External URL/IPs accessed and files transferred . User accesses to different nodes . Files created in development mode
3	**Authorization:** Three access modes: - Developer: Root access to infrastructure - Medical team: Access through the medical portal - Patient: Access to own data through app **Authentication** - All types of access are secured using user and password.	**Data at rest:** Data remains on premise. The database is encrypted **Data in transit:** External connections are established by a VPN. Communications are encrypted using SSL/TLS. Messages are pseudonymized before encryption.	**App/Serv. Logs:** - VPN logs - System logs

	- To connect as developer or to medical premises, an extra authorization (VPN) is needed		
5	**Authorization:** Access to the servers of Achmea and Erasmus MC with fake data is provided to specific authorized persons with a private ssh key. In some cases, the IP address of the device is also checked. **Authentication:** The MPC computation used 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	**Data at rest:** During computation, involved data remains private to each partner. The outcome is encrypted by means of Shamir secret sharing and Paillier additive homomorphic encryption. The (pseudo) randomness is implemented by PEP 506. **Data in transit:** The data is encrypted and exchanged using TLS/SSL. Additionally, the data itself is also encrypted with secret sharing or Paillier encryption.	**App/Serv. logs:** For MPC proof of concept demonstration, there are internal app loggings in place. This enables experts to investigate any errors that may have occurred. **System logs:** Only standard Linux and Docker logging.
6	**Authorization:**Multiple roles defined to gain access to the system. Typical roles are: . Tumour board chair . Radiologis . Nurse OncoPredict enables data analytics and can get access to anonymized data using a 2-factors authentication with read only access **Authentication:** A 2-factor authentication using a token procedure is performed to 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).	**Data at rest:** HTTP TLS encryption is used for the web interface. Login to the intranet 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, but servers are protected by perimeter security” Therefor 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.	**App/Serv. logs:** VPN logging, Windows logging by hospital IT. **System logs:** ISPM internal logging
7	**Authorization:** Remote or physical access to the servers storing the data is provided only to authorised personnel based on their	**Data at rest:** Users and roles are also defined on a database level, where different users have different privileges on the	**App/Serv. logs:** Only logs related to the user evaluation and the measurement of KPIs **System logs:**

	role **Authentication:** Communication between the publicly available APIs of the platform and the components is done through secure protocols TLS/SSL. Additionally 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.	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. At network level, firewalls employ rules that allow the traffic flow only through specific ports and domains.	Only standard system logs produced by open source tools
8	**Authorization:** Remote access to the servers storing the data is provided only to the people that work on BigMedilytics. **Authentication:** Communication is done via VPN with several passwords using secure protocols TLS/SSL	**Data at rest:** Users and roles with different privileges are defined. Access to data servers only to authorised personnel. No data is saved on external devices. **Data in transit:** The data doesn't leave Curie Institut, so there is no data in transit.	**App/Serv. logs:** IBM Websphere logs **System logs:** VPN/Linux logging
9,10& 11	**Authorization:** - Access to RTLS data stream is limited to specific persons working on the pilot - Access to EMR data stream will be limited to specific hospital staff members **Authentication:** - Local access with a standard Windows account authentication - Remote access using a Philips-compliant remote access service	**Data at rest:** Data will be encrypted before transfer using 7-zip with AES encryption **Data in transit:** TLS encryption	**App/Serv. 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
12	**Authorization/ Authentication:** - Components which access sensitive data: 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	**Data at rest:** Sensitive data remains on premise. Data store is encrypted to guarantee that it is accessible only with authentication and authorization permissions. **Data in transit:** Communication between components relies on SSL/TLS protocol, hence data is encrypted, and therefore protected	**App/Serv. logs** **System logs:** - Pilot integrated with an already full functional application within the health institutions. - The auditory of the access to the services relies on the services already in place - Components which are outside of the hospital premises, are monitored and audit in order to have a whole track of

	functionality, hence the control access (Authentication and Authorization) will rely on the already used by these institutions. . Components deployed outside of the hospital: 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.		the operations