

Grant Number:	720270	Grant Title:	Human Brain Project SGA1
Deliverable Title:	D8.6.4 (D48.4 D60) SP8 Medical Informatics Platform - Demo		
Contractual Number and type:	SGA1 D8.6.4 (D48.4, D60) - Demonstrator		
Dissemination Level:	PU (= PUBLIC)		
Version / Date:	ACCEPTED: 29 Oct 2018		
V1.0	31 Dec 2017	Delivered to EC	
V1.1	10.01.2018	Updated with François Junique's comments Purpose and the content of the document changed to MIP system validation plan Document name changed to "System Validation Plan at the end of SGA1"	
V1.2	04.05.2018	Updated with the answers to all comments, questions and suggestions from the EC Experts Review Report, end of January 2018 - the new Appendix IV: Answers to Experts Review Report Updated with an overview of the size and the content of each participating hospital's dataset, the names of the proposed clinical system validation scenarios and the type of the proposed studies (local hospital / cross-hospital) - the new Appendix V: Hospital Selection Criteria and Expected Dataset Size.	
V1.3	07.05.2018	Document updated with all 4 editorial suggestions, as specified in the EC Expert Review Report in Chapter 2.4 on page 7	
Abstract:	This document describes the final Medical Informatics Platform system validation plan for SGA1 M24.		
Keywords:	Medical Informatics Platform, SP8, plan, demo, SGA1		

Targeted users/readers	SP8 Project Management, EC PMO
Contributing Package(s):	Work- WP 8.6
Initially Planned Delivery Date:	SGA1 M21 / 31 Dec 2017
Authors:	Ferath KHERIF (CHUV, P27), Dušan MILOVANOVIĆ (CHUV, P27)
Compiling Editors:	
Contributors:	SP8 Team Hospitals: CHUV (CH), Lille (F), Brescia (I) Clinical Use Cases: Ferath KHERIF and Jean-François DÉMONET
SciTechCoord Review:	
Editorial Review:	EPFL (P1): Annemieke MICHELS, Guy WILLIS

Table of Contents

1. Purpose of The Document	5
2. Introduction.....	5
3. Strategic and Operational Objectives	6
4. Achievements	7
4.1 Positioning of MIP in Medical Informatics Solutions Eco-system	7
4.2 Clinical Benefits of the Medical Informatics Platform.....	7
4.3 High-level Medical Informatics Platform Architecture	8
5. Clinical, Neuroscientific Scope and Need for Future Developments.....	10
6. System Validation Scope	11
6.1 Validation Stakeholders	16
6.2 System Validation Schedule	17
6.3 Managing Validation Results	17
7. System Validation Test Cases	19
7.1 Medical Informatics Platform Deployment Validation Test Case	20
7.2 Clinical Validation Test Cases	22
7.2.1 Clinical Utility of Volume of Medial Temporal Lobe Sub-regions for AD Diagnostic.....	23
7.2.2 Clinical Utility of CSF Markers For Alzheimer's Disease	25
7.2.3 Differential Diagnostic: Fronto Temporal Dementia and Alzheimer's Disease	28
7.2.4 Biological Signature of Alzheimer's Disease Using Pathological Measurements.....	30
Appendix I: Overview of MIP Use Case Model	33
Software Installation	33
Data Factory	33
Web Applications	35
Data Mining	36
Data Analysis Accuracy Assessment	37
Clinical Validity	38
Clinical Utility	39
Appendix II: Medical Informatics Platform Component Model	41
Functional Architecture Overview	41
Data Capture Sub-system	41
Data Factory Sub-system	43
Feature Data Store Sub-system	58
Knowledge Extraction Sub-system.....	60
Web Sub-system	66
Deployment Architecture Overview	69
Microservice Architecture.....	69
Docker Images As Microservices.....	69
Automated Installation and Configuration of MIP Software	70
Medical Informatics Platform Software Installation Use Case Specification	70
Appendix III: Components: Old Name - New Name Mapping	73
Appendix IV: Answers to Experts Review Report	77
Appendix V: Hospital Selection Criteria and Expected Dataset Size	87
Appendix VI: Acronyms and Abbreviations	88
Appendix VII: References	90

List of Figures

Figure 1: Medical Informatics Platform High-level Architecture	10
Figure 2: MIP Software Installation Use Case	33
Figure 3: MIP Data Factory Use Cases	34
Figure 4: MIP Web Application Use Cases	36
Figure 5: MIP Data Mining Use Cases	37

Figure 6: Analytical Validity Use Case	38
Figure 7: Clinical Validity Use Case	39
Figure 8: Clinical Utility Use Case	40
Figure 9: Data Capture Sub-system.....	42
Figure 10: Data Folder Organisation for the De-identification Processing	42
Figure 11: Apache Airflow Concept.....	43
Figure 12: Data Factory Sub-system.....	44
Figure 13: Apache Airflow Dashboard.....	45
Figure 14: De-identified DICOM and EHR Data.....	45
Figure 15 - De-identified NIfTI and EHR Data	45
Figure 16: Reorganisation Pipeline	46
Figure 17: Neuromorphometric Processing.....	48
Figure 18: Apache Airflow Image Processing Pipeline Status	48
Figure 19: Brain Scan Pre-processing and Brain Feature Extraction Workflow	49
Figure 20: Original T1-weighted MRI scan (left), along with automatically extracted grey (middle) and white matter (right) tissue maps. The tissue maps encode the probability of each tissue type calculated using the given model and data	50
Figure 21: Grey and white matter from the original tissue atlases (left) along with registered versions (middle and right)	51
Figure 22: Automatically labelled image, showing most probable macro anatomy structure labels	51
Figure 23: Multi Parameter Mapping high-resolution quantitative MRI acquisition protocol	52
Figure 24: Voxel Based Quantification data analysis for studying microanatomy of the human brain <i>in vivo</i>	52
Figure 25: Brain Scan Metadata and EHR Data extraction pipelines.....	53
Figure 26: I2B2 tranSMART Foundation's research data warehouse for clinical, biomedical and pharmaceutical research.....	55
Figure 27: I2B2 Schema	55
Figure 28: Feature Data Transformation, Normalisation and Load Pipeline	56
Figure 29: MIPMap user interface	57
Figure 30: Feature Data Store Sub-system	59
Figure 31: Knowledge Extraction Sub-system.....	62
Figure 32: Algorithm Factory Communication Diagram	63
Figure 33: Distributed Query Engine Architecture Overview.....	64
Figure 34: Web Sub-system	67
Figure 35: List of MIP Docker Images.....	70

List of Tables

Table 1 - Medical Informatics Platform Deployment Use Cases	12
Table 2 - Medical Informatics Platform Web Applications and Data Analysis Use Cases	13
Table 3 - Medical Informatics Platform System Validation Stakeholders	16
Table 4 - Requirements and Validation Traceability Matrix	17
Table 5 - Use Case Specification: Medical Informatics Platform Software Installation	71
Table 6 - MIP Data Components	73
Table 7 - MIP Software Components.....	73
Table 8 - Selection Criteria for Hospitals and Clinical Use Scenario Demonstrations	87

1. Purpose of The Document

The main objective of this document is to provide a system validation plan for collecting objective evidence that the Medical Informatics Platform (MIP) fulfils its strategic and operational objectives and the needs of clinicians and researchers.

The document contains a plan for validation of the compliance of developed MIP functions to clinicians' needs and SP8 mission objectives. A system validation report is presented in Deliverable D8.6.3 as well as a presentation and evaluation of the following research-related SP8 tasks:

- Disease modelling in epidemiological and research cohorts
- Calibration model of healthy age-related brain changes
- Identification and use of a biological signature of disease in clinical trial data
- Robust image pre-processing and template creation adapted to clinical brain MRI scans
- Atlas of brain diseases
- Data analytics research artefacts, such as mathematical modelling, Bayesian inference and machine learning methods

End users from the three selected hospitals in Lausanne, Lille and Brescia will execute system validation actions specified in this document using a fully operational platform running in their hospital execution environments. The system validation test cases shall cover data analytics scenarios, using local patient datasets, as well as federated analytics using patient datasets from different hospitals.

The MIP system validation strategy and system validation plan specification are described in Chapters 6 and 7, respectively.

Chapters 3, 4 and 5 contain an overview of the Medical Informatics Platform strategic and operational objectives, achievements and the clinical scope of the supported use scenarios. The information in those sections of the document provides the functional context for MIP system validation.

More details about the MIP SGA1 use case model and the MIP end-to-end component model are provided in Appendix I: Overview of MIP Use Case Model and Appendix II: Medical Informatics Platform Component Model.

2. Introduction

As part of the Human Brain Project (HBP), the Medical Informatics Platform initiative is an innovative data analysis system that can be accessed by a wide public (clinicians, neuroscientists, epidemiologists), it can be used to analyse clinical and research data without moving the data from the hospital/Institute servers where they reside, and without infringing on patient privacy.

The MIP strategy is to use computational and machine learning approaches (from data pre-processing, brain feature extraction to data mining) and create a meeting place for neuroscience and IT for collaborative brain disease research, as well as benefitting clinicians, on a daily basis. Specifically, we designed the MIP to help clinicians with IT on site, aiming to adopt advanced analytics for diagnosis and research in clinics and to promote collaborative neuroscience research using hospital data.

3. Strategic and Operational Objectives

SP8 FPA Operational Objectives - what the SP wants to achieve over the lifetime of the HBP

FO1. Design, implement and operate a federated clinical infrastructure, comprising tools for harmonising heterogeneous clinical databases, data anonymisation, ontology-based query interfaces, federated search and distributed analysis of clinical data.

FO2. Establish agreements or memorandums of understanding (MoUs), in consultation with authorised representatives of involved HBP Partners, for access to hospital data, centralised large-scale clinical research databases and biobanks. Provide documentation, training and support to the users.

FO3. Develop generic tools for data curation, quality control and provenance. Develop, implement and deploy tools to extract brain morphology, genomic, proteomic behavioural and cognitive features from clinical and research databases.

FO4. Develop, implement and deploy mathematical methods for predicting multi-level features of diseases; develop tools for identification of homogeneous diseases using biological signatures; construct unified models of brain diseases.

FO5. Contribute data, novel disease classification for disease simulation and *in silico* experimentation.

SP8 SGA1 Objectives - what the SP wants to achieve by the end of SGA1 (from FPA)

8a. First version of Medical Informatics Platform; access for academic researchers, epidemiologists and clinicians

8b. Federation nodes in five hospital nodes for *in situ* querying of anonymised data

8c. Web-based services for neuro-epidemiological studies, interactive analysis and exploration of the biological signatures of Alzheimer's disease

8d. Initial publications demonstrating the value of the Platform

Fast-track objectives

To transition from the pilot developed in the RUP to a full production-ready proof of concept (POC) operational in real hospitals, SP8 will implement the Fast-track plan delivered and approved during the Review in October 2016.

The Fast-Track Plan has 4 objectives:

FT1. Complete the infrastructure including hardware and software supporting the clinical aims of SP8;

FT2. Rationalise the integration of services and products into the MIP- microservices architecture facilitating its deployment, maintenance and operation in hospitals;

FT3. Direct the deployment and hand-over to users (neuroscientists and clinicians) to ensure that the quality of the data, the services and the products meet the standard for accurate and validated results of the methods used from pre-processing.

FT4. Integrate methods for data analysis, data mining (WP8.3), disease models (WP8.4) and visualisation (WP8.5) towards testing hypotheses on neuropathology and structure-function relationship in diseased brain

From 01.08.2017 - following the recommendations of the June 2017 Review

To keep the tasks directed towards the 3rd main goal of the Fast-Track Plan focused, SP8 will strictly follow the four recommendations decided by the EC after the June 2017 review meeting in Lausanne:

[R1] Demonstrating and testing rigorously MIP-Local with all features functioning in a well posed coherent clinical protocol for disease signature discovery over one or two diseases

- [R2] Stopping the SP8 data mining work on new features, and focusing only on the implementation of the current advances into the MIP
- [R3] Showcasing a practical proof-of-principle of operational MIP-Federate that has ethical approval over hospitals using real life data
- [R4] Showcasing an interface for MIP-Local and MIP-Federate to be used by end-users and data providers, and how data-protection is being ensured

4. Achievements

4.1 Positioning of MIP in Medical Informatics Solutions Eco-system

With the introduction of electronic health records (EHR) and picture archiving and communication systems (PACS), clinical researchers got the means to access information belonging to groups of patients in their hospital, on condition that they have informed consent from each of the patients.

Due to the data protection regulations concerning patient information privacy and security, both EHR and PACS systems were designed for the collection of data from patients in one and only one hospital. Patient medical data remained scattered across a vast number of hospitals, clinics and private practices around the world, as was the case with paper-based medical records before the introduction of their electronic form.

The integration of dispersed EHRs and PACS systems is a big challenge today, not only because of patient data protection, but also due to incompatible ICT solutions. As a consequence, clinical researchers can only access data stored in systems belonging to their own hospitals.

Global leaders in medical informatics have been addressing this challenge by developing solutions for two distinct purposes:

- Content management and research data processing solutions (for example LORIS and CBrain)
- EHR systems for sharing patient data between clinicians (for example, Cerner and Epic Systems)
- Data catalogues (for example, EMIF and GAAIN)

None of the three distinct groups of solutions supports data analytics use cases.

The Medical Informatics Platform provides support, not only for clinical research data collection and storage, but also for data analysis across clinical and research datasets. It is a unique solution that adds value to patient data by analysing data inter-connectedness across massive data collections. It provides powerful tools to clinicians and researchers for descriptive, predictive and prescriptive data analytics with measurable reliability and accuracy achieved by validation of learned machine-learning models to estimate predictive model errors.

The Medical Informatics Platform is, therefore, designed with the objective to become a complete solution for descriptive and predictive disease diagnosis because it provides complete data analytics information, including the assessment of the accuracy of predictive errors ^{[27][28]}.

4.2 Clinical Benefits of the Medical Informatics Platform

The clinicians from the contributing hospitals, participating in the DGDS committee, selected the clinical use cases for the demonstrations. The goal of these demonstrations is to highlight how the MIP can provide the tools needed for a better understanding of the clinical population. The MIP will be demonstrated in the three selected hospitals - Lausanne, Lille, and Brescia.

Deployed locally at each node, the MIP Local provides the following benefits:

- Clinicians can explore their own variable dataset in a well-structured web interface through the Data Exploration web application component.
- Clinicians and researchers can run data-mining algorithms on their own datasets through the Model Builder and Experiment Builder & Model Diseases web applications.
- Clinicians can compare specific patients and their measured variables against the whole cohort stored locally in one execution environment, observe disease severity and therefore provide more accurate diagnosis and decide the optimal treatment.
- Clinicians can visualise and interpret the quantitative measurements of the brain MRIs after their pre-processing and morphometric features extractions, and further link these measures to diagnostic, behavioural and clinical measurements.
- Source brain imaging data is processed and features extracted with benchmarked industry tools and standards within the Data Factory sub-system. The results are, therefore, considered as relevant for publication in journals.
- Models, articles and data mining experiments may be shared between the users of the node.

Once federated, the data stored in the local hospital MIP deployments becomes accessible for multi-centre, multi-dataset studies. The benefits of the data federation are the following:

- Clinicians and researchers can explore and compare cases and measurements observed in their own clinic with the whole MIP data space
- Clinicians and researchers may analyse models, articles and data mining experimental results created and shared by other members of the MIP community

Specific example in Alzheimer's disease

Currently, Alzheimer's disease (AD) clinical diagnostic criteria rely on symptoms that do not precisely reveal the underlying AD biological processes. These criteria lack the ability to identify preclinical cases and objectively quantify the disease severity. A recent paper (Frisoni, *et al.* 2017) authored by World experts in dementia concluded that "the provision of high-quality care to patients is negatively affected because the informative value of biomarkers cannot be used with full reliability in clinical practice". The group proposed a new strategic five-phase roadmap to foster the clinical validation of biomarkers in Alzheimer's disease. The five phases include providing sufficient evidence of analytical validity (phase 1), evidence of clinical validity (phases 2 and 3) and clinical utility (phases 4 and 5). In the use cases described below, we followed the same strategy and the methodology proposed by Frisoni *et al.* 2017. We also argue that the implementation of this strategy requires standardisation of the methods used to extract each potential biomarkers and the use of algorithms able to combine multiple biomarkers. From a clinical perspective, the most important MIP application is to use routinely collected data at the hospitals for:

- Computing, testing and validating the biomarkers (MRI-derived, bio-specimen, etc.) proposed in the research with the clinical data
- Improving the classification of different dementia subtypes using differential patterns of cortical atrophy associated with cognitive decline
- Improving the classification of different dementia subtypes using neuropathological examination

4.3 High-level Medical Informatics Platform Architecture

The Medical Informatics Platform (<https://mip.humanbrainproject.eu>) is a distributed information system that:

- Collects de-identified health-related and privacy-sensitive patient data from hospital information systems (EHR Systems and PACS) and research datasets (ADNI, EDSD, PPMI, etc.) - **Data Capture** sub-system components
- Processes captured neuroimaging and other patient biomedical and demographic data to extract patient health-related features - **Data Factory** sub-system components
- Harmonises and normalises feature data types across the data sets captured from different hospitals and research databases - **Data Factory** sub-system components
- Provides permanent patient feature data storage at each participating hospital - **Feature Data Store** sub-system components
- Provides a set of pre-integrated statistical methods and predictive machine learning algorithms, including benchmarking and cross-validation of learned models, for patient data exploration, creation and execution of new experiments, and visualisation of the results:
 - a. Locally at a hospital level, including only that hospital's de-identified patient data and locally available research datasets - **local hospital Knowledge Extraction** sub-system components
 - b. Remotely from the central federation node, including de-identified patient data from one or more federated hospitals and any of the available research datasets - **centralised federal Knowledge Extraction** sub-system components
- Provides web applications for data extraction, building of statistical and machine learning models, designing new experiments, development of disease models, collaborative writing of articles and visualisation:
 - a. Locally at a hospital level, including only that hospital's de-identified patient data and locally available research datasets - **local hospital Web** sub-system components
 - b. Remotely from the central federation node, including de-identified patient data from one or more federated hospitals and any of the available research datasets - **centralised federal Web** sub-system components

The Medical Informatics Platform (MIP) includes functionalities for integration of new methods and algorithms, permanent storage for experiments and results, support for the collaborative writing of articles and access to 3rd party web applications for visual analytics.

Data Privacy Aspects

The MIP ecosystem provides distributed analytics, applying data mining and machine learning methods on patient cohorts from one or more participating hospitals. Distributed analytics are run from a Central Federation Node (Figure 1), from which the queries and machine learning algorithms are executed and orchestrated. The federated analytic results are visualised in the central federation node's web-based user interface, by means of aggregation, meta-analysis and cross-hospital validation. De-identified patient data never leaves local hospital's MIP execution environments.

The MIP has been designed to keep de-identified patient data in the execution environments of participating hospitals. The MIP is designed not to allow de-identified patient health feature data to leave hospital data centres. Patient population data, stored inside a hospital data centre's perimeters, is queried and analysed locally, using sets of locally deployed algorithms from the web-based user interface.

Optionally, clinicians and researchers, with the help of hospital data managers, can capture data from numerous open research cohort data sets of their interest in their local version of the Medical Informatics Platform. It is possible to capture and store open research cohort data sets on the Federation Node too.

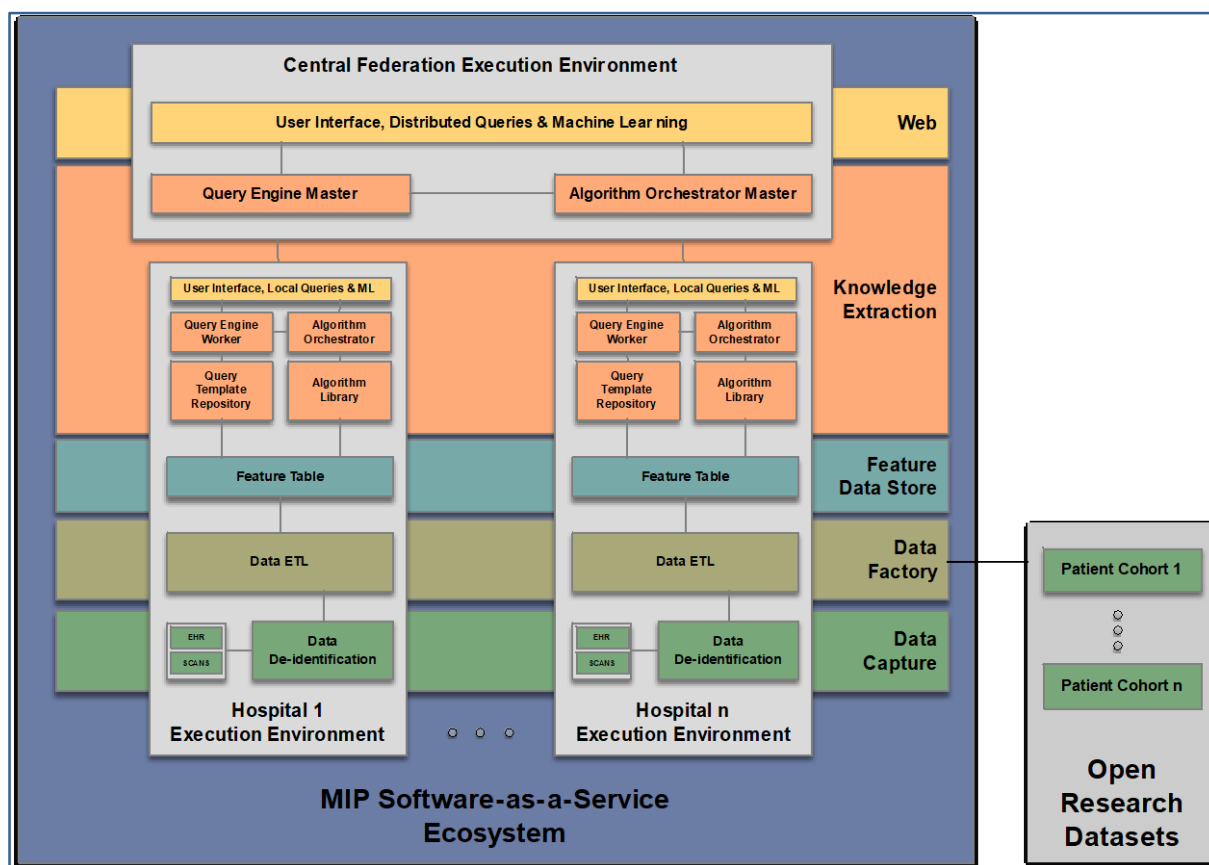


Figure 1: Medical Informatics Platform High-level Architecture

5. Clinical, Neuroscientific Scope and Need for Future Developments

Although SP8 aims for a broad coverage of human brain pathologies, the present use cases address Alzheimer's disease (AD) and related syndromes (defined by DSM V as 'Neurocognitive disorders'), as these brain pathologies represent an important global health challenge, acknowledged as such by the WHO in May 2017. AD affects about 47 million people worldwide and the affected population will probably double at least over the next 30 years. Currently we can estimate the direct costs for western countries as about one billion Euros per year/ million inhabitants. Most likely, milder and prodromal syndromes ("Mild Cognitive Impairment") are 2-3 times more prevalent than the full-blown dementia cases.

While the scientific advances in Alzheimer's disease are expanding continuously, an ever-growing gap is developing between primary evidence i.e. clinical data and the biology- or imaging-based research findings.

SP8 aims to bridge the gap between research biomarkers and real-world clinical data so that we can seamlessly integrate and compare at a statistical level clinically derived multimodal datasets to reference research-derived databases ("research gold standard").

The AD-MIP use cases address how to map real-world based evidence to "gold standard" datasets to statistically define expected discrepancies and identify the sources of variance that may not be captured by the current models. This endeavour is crucial for a better accounting of the associated causal neuro-pathologies and their complex interactions, as well as the relative contributions of genetic and environmental factors. Failure to address these complex phenomena likely accounts for the negative results of many costly AD drug trials. In most of the trials, a simplified model (the "amyloid cascade") was used, while, for a number of participants, the cognitive disorders were already too severe with widespread lesions, or had different neuropathology than mere accumulation of Abeta-amyloid peptide in brain tissue.

Here, we present use cases that allow end users of the currently functional local and Federated MIPs to explore the locally imported datasets and compare them to a reference dataset (Research data) and to other dataset in other hospitals, concerning a biologically relevant AD measure.

Given the constraints of the current concept relying exclusively on MRI data on neurodegeneration, we see a clear need for future developments in the following domains:

- Expansion of Data Factory capabilities towards other neuroimaging data - brain computer tomography (CT), ligand positron-emission-tomography (PET), single-positron-emission-computer-tomography (SPECT)
- Expansion of Data Factory capabilities for automated feature extraction in MRI brain data with lesions - e.g. after stroke

6. System Validation Scope

Validation is a transverse activity to every life cycle stage of the system. The Medical Informatics Platform has a fully agile incremental lifecycle - the functionalities have been incrementally developed, integrated, verified and validated.^[25]

The MIP mission objectives and compliance to user needs will be validated using a full final operational system in real hospital environments. Planned system validation process consists of validation actions and procedures in the context of the following operational scenarios:

- MIP deployment scenario (Chapter 7.1):
 - Preparation for the deployment
 - Software installation
 - Health-related feature data preparation
- Clinical scenarios (Chapter 7.2):
 - Measuring clinical utility of the hippocampal volume for diagnosing Alzheimer's disease (Chapter 7.2.1)
 - Measuring clinical utility of CSF markers for Alzheimer's disease (Chapter 7.2.2)
 - Differential diagnostic between frontotemporal dementia and Alzheimer's disease (Chapter 7.2.3)
 - Biological signature of Alzheimer's disease using pathological measures (Chapter 7.2.4)

MIP system validation process specified in this document is in close relation to the MIP mission analysis process, which established targeted Platform's operational capabilities at different stages of SP8 SGA1 project. (Chapter 3 - Strategic and Operational Objectives)

The objective of the MIP system validation is to prove satisfaction of the desired operational capabilities by showing through execution of operational scenarios that user needs are met.

Platform's operational capabilities and user needs are formally defined using use case modelling approach. MIP operational scenarios selected for MIP system validation include the execution of one or more MIP use cases listed in Table 1.

Appendix I: Overview of MIP Use Case Model contains a short conceptual description of all MIP use cases and their use case diagrams.

Table 1 - Medical Informatics Platform Deployment Use Cases

Medical Informatics Platform Use Case List			
ID	Name	Short Description	Relationship
Deployment Use Cases			
Software Installation			
UC_ITL_01	Software Installation	MIP execution environment configuration and software installation	
Data Capture / Data Factory			
UC_DFY_01	Data Preparation	Orchestration of source EHR and brain imaging data extraction, data transformation and data loading pipelines, including data quality assurance and data provenance storage	
UC_DFY_02	Patient's Feature Extraction from EHR, DICOM and NIFTI	Extraction of patient demographic, biological, genetic and cognitive data from HER and extraction of the metadata from patient's brain scan DICOM or NIFTI files	Included in UC_DFY_01
UC_DFY_03	Patient's Neuromorphometric Feature Extraction	Extraction of neuromorphometric data from patient brain scans	Included in UC_DFY_01
UC_DFY_04	Patient's Feature Extraction From Open Research Cohort Dataset	Extraction of patient feature data from open research cohort datasets	Included in UC_DFY_01
UC_DFY_05	Data Validation	Checking of pre-processed brain images for artefacts and quality metrics, check data for confound and biases, check metadata	Included in UC_DFY_01
UC_DFY_06	Data Harmonisation	Transformation of source patient biomedical and health-related features to harmonised data structure and data vocabulary	Extends UC_DFY_01
UC_DFY_07	Harmonised Data Loading	Loading of transformed source datasets to permanent harmonised feature data store for federated multi-centre multi-dataset analytics	Included in UC_DFY_05

Table 2 - Medical Informatics Platform Web Applications and Data Analysis Use Cases

Medical Informatics Platform Use Case List			
ID	Name	Short Description	Relationship
Clinical Study Use Cases			
Web Application			
UC_WEB_01	Data Exploration	Statistical exploration of patient feature data (i.e. variables)	
UC_WEB_02	Model Building	Configuration/design of statistical or predictive machine learning models	
UC_WEB_03	Model Validation	Validation of learned model against the test dataset. Calculation of the predictive error rate	
UC_WEB_04	Experiment Design	Selection of a statistical, feature extraction or machine learning method, the configuration of the method's parameters and the parameters for the trained model validation for supervised machine learning	
UC_WEB_05	Experiment Execution	Launching of the machine learning experiment. Displays experiment validation results as bar charts and confusion matrices	
UC_WEB_06	Article Writing	Writing scientific articles using the results of the executed experiments	
Data Mining			
UC_DTM_01	Test Correlation Between Health-relevant Features	Testing the correlation between two or more variables using a statistical or machine learning method	Included in UC_DTM_02 Included in UC_DTM_03
UC_DTM_02	Test Health-relevant Feature Outliers	Discovering outliers after testing the correlation between variables	
UC_DTM_03	Classify Disease	Using classification machine learning algorithms to create (learn), validate and/or apply the classifier	
UC_DTM_04	Predict Disease	Apply a learned classifier to predict pathology	
UC_DTM_05	Discover Health-relevant Feature Patterns	Discover patterns of correlated variables in a population	Included in UC_DTM_03 Included in UC_DTM_04
Data Analysis Accuracy Assessment			

Medical Informatics Platform Use Case List			
ID	Name	Short Description	Relationship
Clinical Study Use Cases			
UC_ACC_01	Measure Biomarker's Analytical Validity	Measure analytical validity of tests - assess the ability of the test to accurately detect and measure patient's health-related features of interest. Analytical validity measured using MIP is the probability that the test results in a dataset chosen for the study will be in the same expected range with the results of the same test under the same conditions in different control datasets, i.e. other research cohorts whose data are part of the MIP. Analytical validity is a measurement of the MIP data quality.	
UC_ACC_02	Measure Biomarker's Analytical Sensitivity	Measure the probability that a test will detect an analyte when it is present in a specimen	Included in UC_ACC_01
UC_ACC_03	Measure Biomarker's Analytical Specificity	Measure the probability that a test will be negative when an analyte is absent from a specimen	Included in UC_ACC_01
UC_ACC_04	Measure Biomarker's Reproducibility Under Different Conditions	Evaluating the results of the a test when it is performed under different conditions	Included in UC_ACC_01
UC_ACC_05	Measure Health-relevant Feature's Clinical Validity	Measure clinical validity of a biomarker or other health-relevant feature, i.e. to assess whether the biomarker or other tested health-relevant patient's feature is associated with a disease or outcome or the response to a treatment	
UC_ACC_06	Measure Health-relevant Feature's Clinical Sensitivity	Probability that the test is positive in people who have or will get the disease: $TPR = TP / P = TP / (TP + FN)$	Included in UC_ACC_05
UC_ACC_07	Measure Health-relevant Feature's Clinical Specificity	Probability that the test is negative in people who do not have or will not get the disease: $TNR = TN / N = TN / (TN + FP)$	Included in UC_ACC_05

Medical Informatics Platform Use Case List			
ID	Name	Short Description	Relationship
Clinical Study Use Cases			
UC_ACC_08	Measure Health-relevant Feature's Clinical Predictive Value	Positive Predictive Value (PPV) and Negative Predictive Value (NPV) results depend on feature's clinical sensitivity and specificity as well as on the prevalence of the disease in the population. $PPV = TP / (TP + FP)$ $NPV = TN / (TN + FN)$	Included in UC_ACC_05
Clinical Utility Assessment			
UC_CLU_01	Assess Health-relevant Feature's Clinical Utility	Three factors are generally considered when evaluating the clinical utility of a test: 1) Patient outcomes, 2) Diagnostic thinking, 3) Societal impacts	
UC_CLU_02	Measure Patient Outcomes	Do the results of the test ultimately lead to improvement of health outcomes (e.g. reduced mortality or morbidity) or other outcomes that are important to patients such as quality of life?	Included in UC_CLU_01
UC_CLU_03	Assess Diagnosis and Prognosis	Does the test confirm or change a diagnosis? Does it determine the aetiology for a condition or does it clarify the prognosis?	Included in UC_CLU_01
UC_CLU_04	Assess Societal Impact	Does the test identify high-risk race/ethnicities, and the impact on health systems and/or populations?	Included in UC_CLU_01

The Platform has been developed in a joint effort by pan-European cross-disciplinary research and development teams. Each of the SP8 teams was focusing on delivering functions and corresponding components in their specific area of expertise using different software technology stacks.

MIP software components are independently deployable, small, loosely coupled microservices, each one running a unique process and communicating through a well-defined lightweight mechanism. Updating a component or adding a new one does not require redeployments of the entire application. MIP microservice deployment architecture supports continuous integration and continuous deployment approach.

A detailed list of MIP software components, including their names, PLA component ID, WP/task number and the SP8 team responsible for the component development or integration is provided in Appendix III: Components: Old Name - New Name Mapping. An end-to-end functional overview of the Platform, describing logical component architecture, component roles and the interactions between them is provided in System Validation Strategy

The purpose of the Medical Informatics Platform validation process is to provide objective evidence that the integrated and verified platform fulfils its mission objectives and user needs in its intended operational environment.^[26]

Medical Informatics Platform system validation plan specifies user acceptance test activities that demonstrate how the Platform meets users' needs under their own local conditions in real hospital environments.

At different stages of the MIP SGA1 project, Medical Informatics Platform Deployment and Evaluation Agreements were negotiated and signed with the following seven hospitals:

- Lausanne University Hospital, Switzerland
- Regional University Hospital Centre in Lille, France
- IRCCS Centro San Giovanni di Dio - Fatebenefratelli in Brescia, Italy
- Metropolitan Hospital Niguarda in Milano, Italy
- University Clinic in Freiburg, Germany
- Sourasky Medical Centre, Tel Aviv, Israel
- Medical University Plovdiv, Bulgaria

The three hospitals in Lausanne, Lille and Brescia are selected for MIP system validation. The selection criteria were as follows:

- **Diversity** - hospitals that are in different countries in Europe to test the MIP in different environments, i.e. different healthcare systems, different exposure to risk factors, disease prevalence, etc.
- **Size** - hospitals that have a significant number of patients and large patient datasets
- **Clinical excellence** - the best national hospitals with expertise in clinical neuroscience and clinical care, willingness to share data with ethics consent procedures in place
- **Available resources** - hospitals that have personnel and IT equipment resources, including long-term commitment to maintain the Medical Informatics Platform infrastructure
- **Influence** - hospitals that will promote the Medical Informatics Platform through collaboration with other hospitals in the same region or country

6.1 Validation Stakeholders

Table 3 provides a list of the Medical Informatics Platform system validation stakeholders, their roles and responsibilities in the system validation / user acceptance test process.

Table 3 - Medical Informatics Platform System Validation Stakeholders

Role ID	Role	Responsibility
CLR	Clinician	Executes data selection activities in Platform deployment system validation test case. Executes clinical system validation test cases
RES	Researcher	Executes data selection activities in Platform deployment system validation test case. Executes clinical system validation test cases
CDM	Clinical Data Manager	Executes data selection activities in Platform deployment system validation test case
HES	Hospital Ethics Committee	Provides clearance for capturing patient data by the Platform

HIT	Hospital IT Engineer	Executes hospital data centre preparation activities in Platform deployment system validation test case
MIT	MIP Deployment Engineer	Executes installation and configuration activities in Platform deployment system validation test case
DGDS	MIP Data Governance and Data Selection Committee	Executes data selection and data harmonisation activities in Platform deployment system validation test case
SPR	SP8 Representative	Executes system validation project presentation, deployment and evaluation agreement signature and handover of the validated platform activities in Platform deployment system validation test case
HMG	Hospital Management	Executes deployment and evaluation agreement signature activity in Platform deployment system validation test case

6.2 System Validation Schedule

MIP system validation shall be executed by the users in the selected three hospitals:

- Lausanne University Hospital, Switzerland
- Regional University Hospital Centre in Lille, France
- IRCCS Centro San Giovanni di Dio - Fatebenefratelli in Brescia, Italy

Analysis of the system validation results, including the user acceptance testimonials and readiness for use assessment, will be delivered in Deliverable D8.6.3.

6.3 Managing Validation Results

The validation results and their formal user acceptance will be recorded in the validation report presented in Deliverable D8.6.3. Bidirectional traceability of the validated system functions and the validation actions will be maintained using the requirements and validation traceability matrix (Table 4).

Anomalies observed during the validation process will be analysed and resolved by executing corrective actions or improvements, using the MIP quality assurance process.^[25]

The performance of a validation action is compared with the expected result. The comparison enables the assessment of the validated item's acceptability.

Table 4 - Requirements and Validation Traceability Matrix

Objective /Use Case	Validation Action	Actor Role ID	Validation Technique	Validated Item Type	Validated Item ID	Expected Result	Obtained Result



7. System Validation Test Cases

This Chapter contains the specification of MIP system validation actions using the system deployment and clinical operational scenarios.

System deployment operational scenario will validate the procedure and “technical” use cases - software installation use case and data factory use cases for preparing patient data for analytics.

Clinical operation scenarios will validate all web application and data analytics use cases.

A validation action description contains the following information:

- **Objective / Use Case** - the item being validated (the FPA objective ID, use case name, other short description)
- **Validation Action** - validation action ID as per the system validation specification
- **Validation Technique** - applicable techniques for the planned MIP system validation are: demonstration, test, and inspection
- **Validated Item Type** - type of the item on which the validation process is performed, e.g. requirement, function, use case, procedure, sub-system, system component, document, presentation, agreement
- **Validated Item ID** - identification of the item on which the validation process is performed (use case ID or a name of the validated item)
- **Expected Result** - short description of the expected result of the validation action, i.e. validation criteria, from the system validation specification
- **Obtained Result** - short description of the result of the validation action, including the comparison with the expected result and any comment

The following three techniques are applicable for the MIP system validation process:

- **Inspection** - visual examination of a validated item, including peer reviews of process artefacts
- **Demonstration** - presenting correct operation of the validated item against operational observable characteristics with no measurements. It usually consists of a set of actions selected to show that the validated item’s response is compliant to the expected behaviour or to show that the users can perform their assigned tasks
- **Test** - quantitative verification of functional, measurable characteristics, such as operability, supportability or performance

7.1 Medical Informatics Platform Deployment Validation Test Case

Validation Objectives	<ol style="list-style-type: none"> 1) Hospital's data centre has a centralised platform for processing, storing and analysing de-identified and harmonised neuroimaging, neuropsychological, biological and demographic data of its patient population 2) Efficient, configurable and automated end-to-end software installation, unifying operation system configuration, middleware installation and microservice building minimises the IT efforts to keep the focus on using the MIP Platform for the scientific and clinical activities 3) Harmonisation of the full set of Medical Informatics Platform's patient biomedical and other health-related features enables large multi-centre, multi-datasource studies, increasing the accuracy of the analysis methods and probability for new scientific discoveries 4) Extraction and harmonisation of patient biomedical and other health-related features from the source patient data is a first step in the process of creating the data model for comprehensive molecular-level analysis of both individual patients and populations. Unification of biomedical and other health-related data provides the best opportunity to discover new biological signatures of diseases, improve taxonomy of diseases, develop preventive strategies, and improve medical treatment
Validation Actors	Neurologist (CLR), Neuroscientist (RES), Clinical Data Manager (CDM), Hospital Ethics Committee (HEC), Hospital IT Engineer (HIT), MIP Deployment Engineer (MIT), MIP Data Governance and Data Selection Committee (DGDS), SP8 Representative (SPR), Hospital Management (HMG)
Pre-conditions	Hospitals selected for the evaluation of Medical Informatics Platform (users acceptance test) agreed to participate in systems validation activities

Description of Validation Actions						
Action ID	Actor ID	What is validated	Item Type	Technique	Item ID or Name	Validation Criteria
A01	SPR	Project presentation	Process	Inspection	N/A	User assessment positive
A02	SPR HMG	Adapt and sign MIP Deployment and Evaluation Agreement	Documentation	Inspection	MIP Deployment and Evaluation Agreement	Signed MIP Deployment and Evaluation Agreement
A03	CDM DGDS	Gather meta-data, including data acquisition protocol identification	Process, Component	Inspection, Test	Data Element Specification	Metadata Registry, component of Web sub-system successfully updated with new data elements

Description of Validation Actions						
Action ID	Actor ID	What is validated	Item Type	Technique	Item ID or Name	Validation Criteria
A04	HIT MIT	Provide compliant machine and remote access	Functionality	Test	MIP server in hospital's execution environment	Remote SSH access through port 22 and HTTPS access through port 443 are successfully tested
A05	DGDS CDM	Data selection - variables of interest based on hypothesis/questions	Process, Documentation	Inspection	Data Element Specification	The list of variables selected for capturing by MIP is agreed with Clinical Data Manager
A06	CLR RES HEC	Checking whether ethics approval applies for using MIP	Process	Inspection	N/A	Hospital's Ethics Committee clearance applies to MIP Platform
A07	CLR RES CDM DGDS	Variable harmonisation and structuring	Use Case, Documentation	Demonstration, Inspection	UC_DFY_06 Online Data Integration Module Data Mapping and Transformation Specification	Online Data Integration Module is configured for automatic source data transformation to harmonised MIP data and loading to MIP CDE Database
A10	HIT MIT	Installation of MIP software package	Use case, Deployment components, Configuration script	Demonstration, Test	UC_ITL_01 Deployment components (Docker images), MIP Installation and Configuration Script	MIP software is installed on all servers with all processes up and running
A11	CDM HIT MIT DGDS	Installation/configuration/running of non-automated imaging capture	Use case, Sub-systems	Demonstration	UC_DFY_01 Data Factory sub-system Data Capture sub-system	Captured patient data is de-identified and stored in De-identified data storage in Data Factory sub-system
A12	MIT	Configuration/running of image pre-processing	Use case, Sub-system	Demonstration, Testing	UC_DFY_03 Data Factory sub-system	All the images are successfully processed with no error reported
A13	CDM HIT MIT	Installation/configuration/running of non-automated EHR data capture	Use case, Sub-systems	Demonstration	UC_DFY_01 Data Factory sub-system	Captured patient data is de-identified and stored in De-

Description of Validation Actions						
Action ID	Actor ID	What is validated	Item Type	Technique	Item ID or Name	Validation Criteria
	DGDS				Data Capture sub-system	identified data storage in Data Factory sub-system
A14	MIT	Configuration/running of data mapping	Use case, Sub-system	Demonstration, Testing	UC_DFY_02 UC_DFY_04 Data Factory sub-system	The data mapping is processed with no error reported
A15	RES CDM DGDS MIT	Data validation: <ul style="list-style-type: none"> Check pre-processed images for artefacts and quality metrics Check data for confound and biases Check meta-data 	Use case, Sub-systems	Demonstration, Testing	UC_DFY_05 Data Factory sub-system Web sub-system	Check for outliers using web applications then compare the results with high-quality open research dataset available in MIP
A16	SPR	Hand-over: <ul style="list-style-type: none"> Presentations Demo Training 	Process	Inspection	N/A	User assessment positive

Post-conditions	<ol style="list-style-type: none"> 1) MIP software is installed on all servers with all processes up and running 2) Harmonised patient biomedical and other health-related features are permanently stored in Feature Data Store sub-system's Feature Table for multi-centre, multi-dataset clinical studies
-----------------	--

7.2 Clinical Validation Test Cases

7.2.1 Clinical Utility of Volume of Medial Temporal Lobe Sub-regions for AD Diagnostic

Validation Objectives	<ol style="list-style-type: none"> 1) Measuring the clinical utility of the volume of Medial temporal lobe subregions for AD diagnostic. Hippocampal atrophy is a well-established biomarker for AD. However, there are very few studies on the clinical validity and generalisability of this biomarker using “real world patient data”. 2) Primary aim: measure the association between hippocampal atrophy and current clinical diagnostic using the data and the methods available in the MIP. 3) Secondary aim: measure the effect of confounding variables (age, gender, ...)
Validation Actors	Neurologist (CLR)
Pre-conditions	Data available and pre-processed in the MIP

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A01	CLR	Data Preparation: <ul style="list-style-type: none"> • get the summary statics on all the variables of interest (number of patients/ mean and variance) • get information about the acquisition protocol and pre-processing methods • filter to select the patients by setting inclusion and exclusion criteria 	Use Case Component	Demonstration	UC_WEB_01 UC_WEB_02 MIP-EE web App	Variables (hippocampal volume, diagnostic) are selected Population of interest (within one hospital and across) defined and described. Model of interest defined and built.
A02	CLR	Analytical Validity and data quality Test if the variables are accurate and sensitive enough with a valid range by	Use Case Components	Demonstration	UC_WEB_04 UC_WEB_05 MIP-EE web App	Use the MIP-interactive web-app and select ANOVA or linear regression to compare variables from the clinic to the variable from research data (e.g. ADNI).

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
		comparing the grey-matter volume atrophy from clinical scan to those from research scan.				Test the significance of the interaction
A03	CLR	Analytical Validity and data quality Test if variables are reproducible in different settings (different scanners, different environment, or cohorts)	Use case Components	Demonstration	UC_DTM_01 UC_DTM_02 MIP-interactive web-app	Use the MIP-interactive web-app and select ANOVA or linear regression to compare variables from the clinic to the variable from research data (e.g. ADNI). Test the significance of the interaction between scanners and disease diagnostics
A04	CLR	Clinical Validity: Test if the variables are associated with the disease diagnostic (e.g AD vs cognitively normal or with mild cognitive impairments) or disease outcome?	Use case Components	Demonstration	UC_DTM_01 UC_DTM_02 MIP-BSD webapp univariate linear regression and/or multivariate inference methods (e.g Anova, MLM)	significance test of the association between variable of interest and disease diagnostics
A05	CLR	Clinical Validity: Test the predictive value and Performance of the test sensitivity (positive and negative predictive values)	Use case Components	Demonstration	UC_DTM_03 UC_DTM_04 UC_ACC_01 to UC_ACC_08 MIP-BSD webapp predictive models and machine learning tools (e.g. naïve Bayes, knn, rule based, tree classification)	Train, test and validate the model against the selected cohort data. The MIP-BSD provides the information about the performance of the test: <ul style="list-style-type: none"> • Accuracy • Sensitivity • Specificity Benchmark the models obtained using different machine learning tools.
A06	CLR	Model validation across hospitals:	Use case Components	Demonstration	UC_DTM_03 UC_DTM_04	Compare the predicted label to the current diagnostic label

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
		Apply the selected model to the data of the other hospitals.			Use the MIP-BSD webapps to create a new model including the education variables. Use the MIP-BSD webapps to model comparisons Use the MIP-IA for further exploration	
A07	CLR	Publish results	Use case	Demonstration	UC_WEB_06 MIP-writing article webapp	save the results and output of the model (graph, table). Model is available for use by other users.
Post-conditions		1) Scientific results and validation of the MIP methods (pre-processing, data quality, machine learning performance) 2) User feedback reports: feedback from the clinical users on the UI (data exploration/selection, model building/testing and results interpretation) 3) Recommendation reports: recommendation from the SP8 team and the users				

7.2.2 Clinical Utility of CSF Markers For Alzheimer's Disease

Validation Objectives	1) Measuring the clinical utility of the cerebrospinal fluid (CSF) markers across different clinical centres. The aim is to measure the added value of the CSF markers 2) Primary aim: measure the association between cerebrospinal fluid markers (total Tau, phospho-Tau and AB42) and current clinical diagnostic using the data and the methods available in the MIP. 3) Secondary aim: measure the effect of confounding variables (age, gender ...)
Validation Actors	Neurologist (CLR)
Pre-conditions	Data available and pre-processed in the MIP

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A01	CLR	Data Preparation: <ul style="list-style-type: none"> get the summary statics on all the variables of interest (number of patients/ mean and variance) get information about the acquisition protocol and pre-processing methods filter to select the patients by setting inclusion and exclusion criteria 	Use Case	Demonstration	UC_WEB_01 UC_WEB_02 MIP-EE web App	Variables (cerebrospinal fluid markers, diagnostic) are selected Population of interest (within one hospitals and/or across hospitals) defined and described. Model of interest defined and built.
A02	CLR	Model building: create, compare and select the model that best discriminates the FTD and AD cases.	Use Case	Demonstration	UC_WEB_02 MIP-EE web App	Use the MIP-interactive web-app and select ANOVA or linear regression to compare variables from the clinic to the variable from the research data (e.g. ADNI). Test the significance of the interaction
A03	CLR	Analytical Validity and data quality: Test if variables are reproducible in different settings (different scanners, different environment, or cohorts)	Use case	Demonstration	UC_DTM_01 UC_DTM_03 MIP-interactive web-app	Use the MIP-interactive web-app and select ANOVA or linear regression to compare variables from the clinic to the variable from the research data (e.g. ADNI). Test the significance of the interaction between scanners and disease diagnostics
A04	CLR	Clinical Validity: Test if the variables are associated with the disease	Use case	Demonstration	UC_DTM_03 UC_DTM_04	Significance test of the association between the variable of interest and disease diagnostics

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
		diagnostic (e.g AD vs cognitively normal or with mild cognitive impairments) or disease outcome?			UC_ACC_01 to UC_ACC_08 MIP-BSD webapp univariate linear regression and/or multivariate inference methods (e.g Anova, MLM)	
A05	CLR	Clinical Validity: Test the predictive value and Performance of the test sensitivity (positive and negative predictive values)	Use case	Demonstration	UC_WEB_01 UC_WEB_02 UC_DTM_03 UC_DTM_04 UC_ACC_01 to UC_ACC_08 MIP-BSD webapp predictive models and machine learning tools (e.g. naïve Bayes, knn, rule based, tree classification)	Train, test and validate the model against the selected cohort data. The MIP-BSD provides the information about the performance of the test: <ul style="list-style-type: none"> • Accuracy • Sensitivity • Specificity
A06	CLR	Clinical Utility: test if the results confirm or change a diagnosis. Test and refine the model by adding other clinical scores such as education?	Use case	Demonstration	UC_ACC_01 to UC_ACC_08 Use the MIP-EE webapps to create a new model including the education variables. Use the MIP-BSD webapps to model comparisons Use the MIP-IA for further exploration	New model created. Compare the 2 models performance test and clinical utility measure (i.e. change in roc-curves, C-statistics ...). Post-hoc exploration of the miss-classified cases. The high-dimensional data can be summarized by dimension reduction methods (e.g. t-sne or parallel coordinates).

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A07	CLR	Publish results	Use case	Demonstration	UC_WEB_06 MIP-writing article webapp	Save the results and output of the model (graph, table). Model is available for use by other users.
Post-conditions		1) Scientific results and validation of the MIP methods (pre-processing, data quality, machine learning performance) 2) User feedback reports: feedback from the clinical users on the UI (data exploration/selection, model building/testing and results interpretation) 3) Recommendation reports: recommendation from the SP8 team and the users				

7.2.3 Differential Diagnostic: Fronto Temporal Dementia and Alzheimer's Disease

Validation Objectives	1) Compare patterns of brain atrophy in fronto-temporal dementia (FTD) and Alzheimer's Disease (AD). 2) Primary aim: compare patterns of brain atrophy in fronto-temporal dementia (FTD) and Alzheimer's Disease (AD). 3) Secondary aim: Create a classifier for discriminating AD and FTD cases. Test the classifier using the the data of the remaining hospital.					
Validation Actors	Neurologist (CLR)					
Pre-conditions	Data available and pre-processed in the MIP, federation in place					

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A01	CLR	Data Preparation: <ul style="list-style-type: none"> get the summary statics on all the variables of interest (number of 	Use Case Components	Demonstration	UC_WEB_01 UC_WEB_02 MIP-EE web App	Variables (all brain features, diagnostic) are selected. Select only pathologically diagnosed subjects.

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
		<p>patients/ mean and variance)</p> <ul style="list-style-type: none"> get information about the acquisition protocol and pre-processing methods filter to select the patients by setting inclusion and exclusion criteria 				<p>Population of interest (within one hospital and/or across hospitals) defined and described.</p> <p>Model of interest defined and built.</p>
A02	CLR	<p>Model building: create, compare and select the model that best discriminates fronto-temporal dementia (FTD) and Alzheimer's disease (AD).</p>	Use Case, Components	Demonstration	UC_WEB_02 MIP-EE web App	Test the predictive value and Performance of the model using the data from one hospital.
A03	CLR	<p>Model validation across hospitals: Apply the selected model to the data of the other hospitals. Test the predictive value and Performance of the test sensitivity (positive and negative predictive values)</p>	Use case Components	Demonstration	UC_DTM_02 UC_DTM_03 UC_DTM_04 UC_DTM_05 UC_ACC_01 to UC_ACC_08 MIP-interactive web-app	<p>Train, test and validate the model against the selected cohort data. The MIP-BSD provides the information about the performance of the test:</p> <ul style="list-style-type: none"> Accuracy Sensitivity Specificity
Post-conditions		<ol style="list-style-type: none"> Scientific results and validation of the MIP methods (pre-processing, data quality, machine learning performance) User feedback reports: feedback from the clinical users on the UI (data exploration/selection, model building/testing and results interpretation) Recommendation reports: recommendation from the SP8 team and the users 				

7.2.4 *Biological Signature of Alzheimer's Disease Using Pathological Measurements*

Validation Objectives	1) Build, test and validate an automated classifier using topographical markers extracted from structural MRI of clinically and pathologically diagnosed subjects 2) Applied the classifier to predict pathology in independent cohorts from other hospitals
Validation Actors	Neurologist (CLR)
Pre-conditions	Data available and pre-processed in the MIP, federation in place

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A01	CLR	Data Preparation: <ul style="list-style-type: none"> get the summary statics on all the variables of interest (number of patients/ mean and variance) get information about the acquisition protocol and pre-processing methods filter to select the patients by setting inclusion and exclusion criteria 	Use Case	Demonstration	UC_WEB_01 UC_WEB_02 MIP-EE web App	Variables (all biological features, diagnostic) are selected. Select only pathologically diagnosed subjects. Population of interest (within one hospital and/or across hospitals) defined and described. Model of interest defined and built.
A02	CLR	Model building and execution: create, compare and select the model that best discriminates the pathologically proven AD cases and other patients	Use Case	Demonstration	UC_WEB_02 UC_DTM_03 UC_DTM_04 UC_ACC_01 to UC_ACC_08 MIP-EE web App	Test the predictive value and Performance of the model using the data from one hospital.

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
A03	CLR	Model validation across hospitals: Apply the selected model to the data of the other hospitals.	Use case	Demonstration	UC_WEB_03 MIP-interactive web-app	Use the MIP-interactive web-app and select ANOVA or linear regression to compare variables from the clinic to the variable from research data (e.g. ADNI). Test the significance of the interaction between scanners and disease diagnostics
A04	CLR	Clinical Validity: Test if the variables are associated with the disease diagnostic (e.g AD vs cognitively normal or with mild cognitive impairments) or disease outcome?	Use case	Demonstration	UC_DTM_05 MIP-BSD webapp univariate linear regression and/or multivariate inference methods (e.g Anova, MLM)	Significance test of the association between the variable of interest and disease diagnostics
A05	CLR	Clinical Validity: Test the predictive value and Performance of the test sensitivity (positive and negative predictive values)	Use case	Demonstration	UC_ACC_01 UC_ACC_02 UC_ACC_03 UC_ACC_04 UC_ACC_05 UC_ACC_06 UC_ACC_08 MIP-BSD webapp predictive models and machine learning tools (e.g. naïve Bayes, knn, rule based, tree classification)	Train, test and validate the model against the selected cohort data. The MIP-BSD provides the information about the performance of the test: <ul style="list-style-type: none"> • Accuracy • Sensitivity • Specificity
A06	CLR	Clinical Utility: Test if the results confirm or change a diagnosis.	Use case	Demonstration	UC_CLU_01 UC_CLU_02 Use the MIP-EE webapps to create	New model created. Compare the 2 models performance test and clinical

Description of Validation Actions						
Action ID	Actor ID	Description	Item	Technique	Component	Validation Criteria
		Test and refine the model by adding other clinical scores such as education?			a new model including the education variables. Use the MIP-BSD webapps to model comparisons Use the MIP-IA for further exploration	utility measure (i.e. change in roc-curves, C-statistics, ...). Post-hoc exploration of the miss-classified cases. The high-dimensional data can be summarised by dimension reduction methods (e.g. t-sne or parallel coordinates).
A07	CLR	Publish results	Use case	Demonstration	UC_WEB_06 MIP-writing article webapp	Save the results and output of the model (graph, table). Model is available for use by other users.

Post-conditions	<ol style="list-style-type: none"> 1) Scientific results and validation of the MIP methods (pre-processing, data quality, machine learning performance) 2) User feedback reports: feedback from the clinical users on the UI (data exploration/selection, model building/testing and results interpretation) 3) Recommendation reports: recommendation from the SP8 team and the users
-----------------	---

Appendix I: Overview of MIP Use Case Model

Software Installation

The objective of this use case is to configure and install the Medical Informatics Platform software in a hospital's data centre.

The MIP microservices deployment architecture enables agile continuous integration and continuous component deployment developed or modified by different European-wide teams. This architecture enables efficient future upgrades of the platform with new technologies and new features needed to support evolved clinical needs. Automation of configuration and installation of the MIP software minimises IT efforts to keep the maximum focus on the scientific and clinical aspects of the projects.

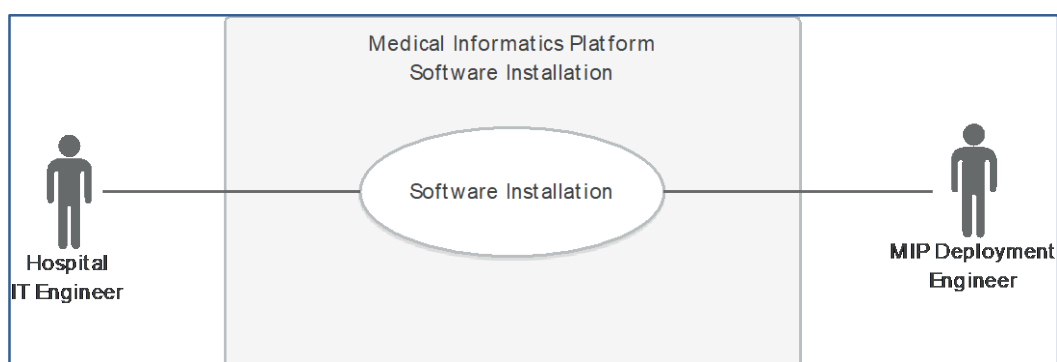


Figure 2: MIP Software Installation Use Case

Scientific Added Value

Hospital's data centre has a centralised platform for processing, storing and analysing de-identified and harmonised neuroimaging, neuropsychological, biological and demographic data of its patient population.

Efficient, configurable and automated end-to-end software installation, unifying operation system configuration, middleware installation and microservice building minimises the IT efforts to keep the focus on using the MIP platform for the scientific and clinical activities

Data Factory

The objective of the Data Factory use case group is to process patient data from different sources – hospitals and open research cohort datasets, EHR and PACS systems for:

- 1) Extraction of individual patient biomedical and health-related features
- 2) Transformation of source patient biomedical and health-related features to harmonised data structure and data vocabulary
- 3) Loading of transformed source datasets to permanent harmonised feature data store for federated multi-centre multi-dataset analytics

Patient source data from both hospitals and open research cohorts is typically structured and organised to capture the type and time of clinical observations, the type, modality, time and results of workups as well as the diagnoses. The Medical Informatics Platform is processing de-identified patient source data to extract biomedical and other health-related patient features, i.e. neuromorphometric, cognitive, biological, genetic, molecular and demographic, harmonises the extracted features across the different data sources, and permanently stores harmonised features for multi-centre, multi dataset clinical research studies.

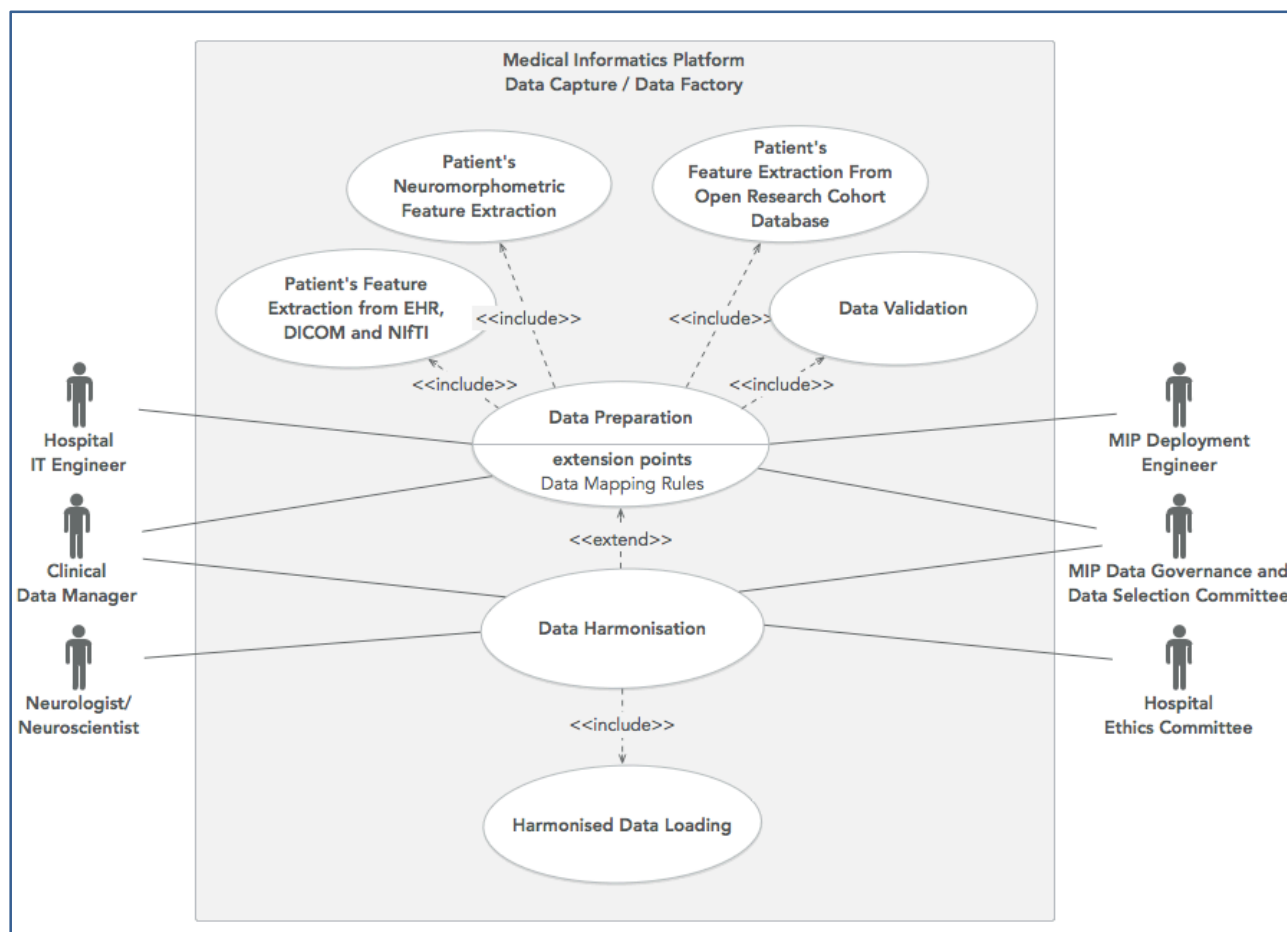


Figure 3: MIP Data Factory Use Cases

Clinical studies involving multiple open research cohort datasets and patient datasets from multiple hospitals are challenging because data sources have different structures and use different coding systems.^[24] The Medical Informatics Platform supports harmonisation of data from different sources and provides harmonised data to clinicians and researchers for further analysis. This process is becoming more and more significant since the need for multi-centre studies is rapidly growing and the volume of the available open research cohort data have a tendency to explode.

Scientific Added Value

Extraction and harmonisation of patient biomedical and other health-related features from the source patient data is a first step in the process of creation of a data model for comprehensive molecular-level data analysis of both individual patients and populations, including their brain features, DNA sequence, proteome, metabolome, microbiome, autoantibodies, etc. Unification of biomedical and other health-related data provides the best opportunity to discover new biological signatures of diseases, improve taxonomy of diseases, develop preventive strategies, and improve medical treatment. This approach shall support the development of individualised medicine and enable cross-comparison between the individual patients to make diagnosing of complex cases more efficient and precise.

Harmonisation of the full set of Medical Informatics Platform's patient biomedical and other health-related features enables large multi-centre, multi-data source studies, increasing the accuracy of analysis methods and the probability for new scientific discoveries.

Web Applications

A web sub-system provides a web portal and the following applications:

- **Collaboration Space** - landing page of the Medical Informatics Platform displaying a summary of statistics (users, available variables, written articles), and the latest three shared models and articles. It provides a link to the Article Builder web application
- **Data Exploration** - a statistical exploration of patient feature data (i.e. variables). It is possible to explore only statistically aggregated data, not an individual patient's information. This web application provides on-the-fly generation of the descriptive statistics and contains a caching mechanism to handle any future data import in an automated way. It uses information stored in a Metadata database to display additional information about the displayed statistical data, such as data acquisition methodology, units, variable type (nominal or continuous), etc. This web application provides the functionality to search, select and classify data elements as variables, co-variables and filters for configuration of the statistical or machine learning models
- **Model Builder** - configuration/design of statistical or predictive machine learning models. It also provides visualisation for searching the data element types, select and classify data elements as variables, co-variables (nominal and continuous) and filters. Once the model is designed, a design matrix is populated with the selected data. Model Builder provides a visual representation of the design matrix and the selected data for inspection before running a statistical, feature extraction or a machine learning algorithms. It also provides an option to save the designed models
- **Model Validation** - measuring machine-learning models' accuracy by calculating predictive error rate of the model trained on training data against a test dataset. The results guide the user to select the best-performing algorithm and fine-tune its parameters as well as to understand how well the model performs before it is used. The Model benchmark and Validation component from Algorithm Factory is used to measure machine-learning model accuracy. In MIP SGA1 it supports cross-validation method - data split using K-Fold cross-validation. This method of data sampling divides the complete dataset into K disjoint parts of roughly the same size. K different models are trained on K-1 parts each while being tested on the remaining one part of the data. That is done on all K parts exactly once to ensure that every data row is used equally often for training and exactly once for testing. Resulting K test errors are then averaged to get the final error estimate of the model, which was built on the complete dataset
- **Experiment Builder & Disease Models** - a selection of a statistical, feature extraction or machine learning method, the configuration of the method's parameters and the parameters for the trained model validation for supervised machine learning, as well as launching of the machine learning experiment. This application displays experiment validation results as bar charts and confusion matrices
- **Article Builder** - writing the articles using the results of the executed experiments
- **Third-party Applications and Viewers** - portal for accessing third-party web applications for data exploration and visualisation

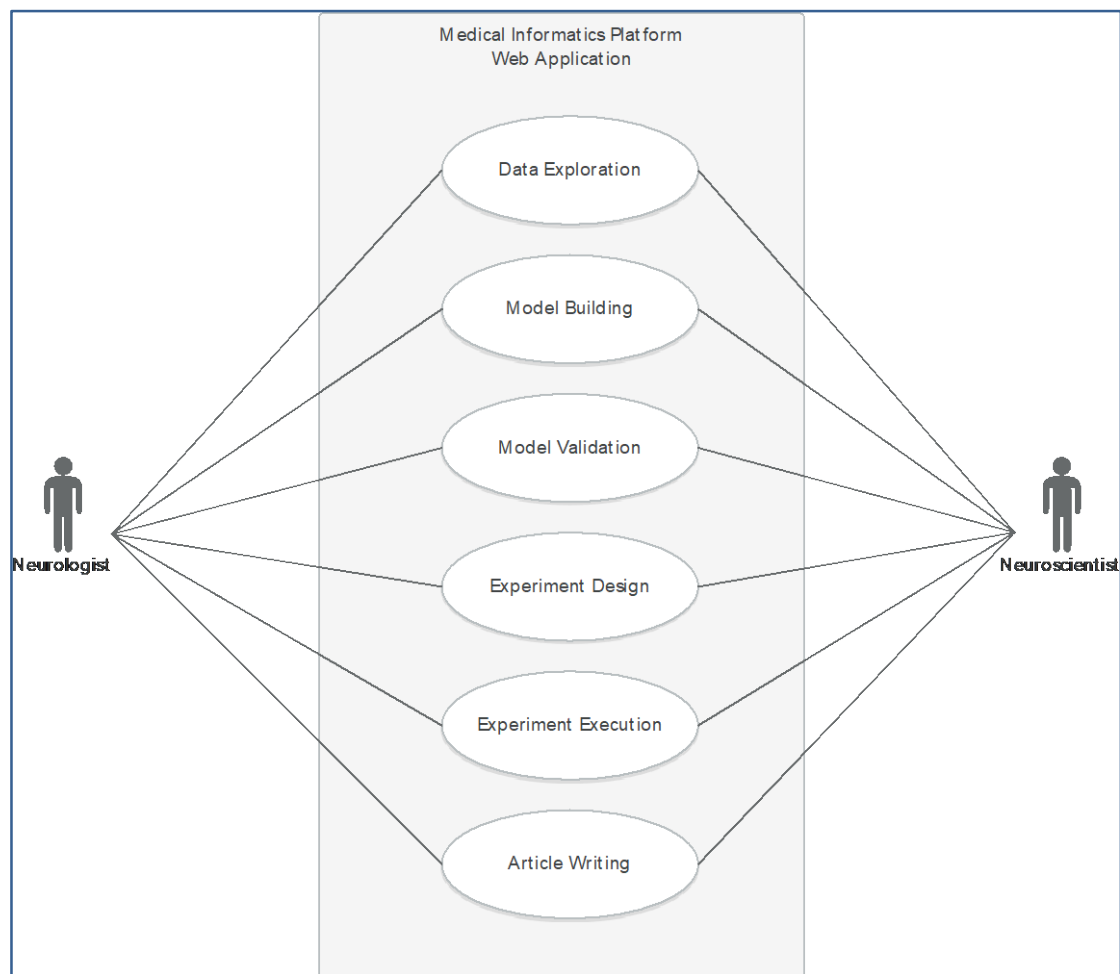


Figure 4: MIP Web Application Use Cases

Data Mining

The objective of data mining of a group of use cases is the discovery of properties of data in datasets. Out-of-the-box statistical and machine learning algorithms are used to realise MIP data mining use cases.

In case of using machine-learning algorithms for data mining, measurement of the learned model's accuracy and consequently the assessment of the accuracy of the discovered data properties is supported through using the algorithms from the Algorithm Factory's repository. Note that it is not possible to validate algorithms from the Distributed Query Processing Engine's repository in MIP SGA1.

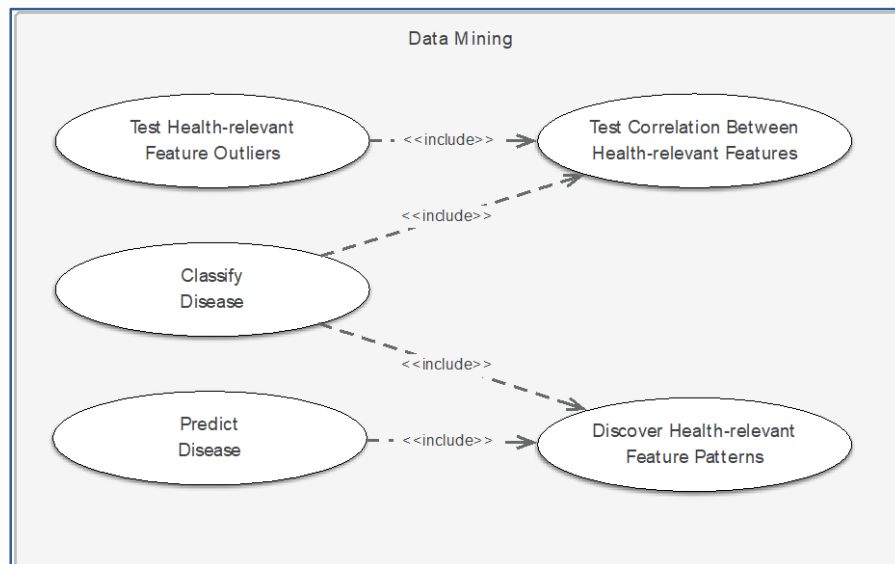


Figure 5: MIP Data Mining Use Cases

Scientific Added Value

This set of use cases specifies the core functionality of the MIP platform - data analytics. Any clinical / research operational scenario executes one or more of the data mining use cases. The four examples of scientific operational scenarios that execute all of the MIP data mining use cases are described in Chapter 8.

Example:

A correlation between brain volume and cognitive decline has been discovered. It was tested whether there are outliers: persons with brain volume decline but no cognitive decline. This gives the idea to include additional health-relevant features to discover whether they may correlate with the observed exceptions. For example, outliers have been discovered and with further data mining it was found that the age of the persons that have brain volume decline but no cognitive decline is in the same range - younger people who have brain volume decline do not have cognitive decline.

Data Analysis Accuracy Assessment

Analytical Validity

The MIP can be used to measure the analytical validity of tests, i.e. to measure the ability of the tests to accurately detect and measure patient health-related features of interest. MIP SGA1 can measure analytical validity of the following: brain MRI scans, scanning protocols, neuromorphometric feature extraction software applications, neuromorphometric feature extraction methods, neuropsychological instruments and methods, laboratory instruments and methods, etc.

The measured analytical validity using the MIP is the probability that the test results in a dataset chosen for the study will be in the same expected range with the results of the same test under the same conditions in different control datasets, i.e. other research cohorts with available data in the MIP. Analytical validity is a measurement of the MIP data quality.

When there are more data available in the MIP, meaning both the number of patients and the diversity of the test conditions and datasets, the measurement of analytical validity will be more accurate and reliable

The MIP can be used to measure analytical validity on its own, or to include measurement of analytical validity as a research dataset validation step prior to executing a scientifically relevant clinical or biomedical research study using that dataset.

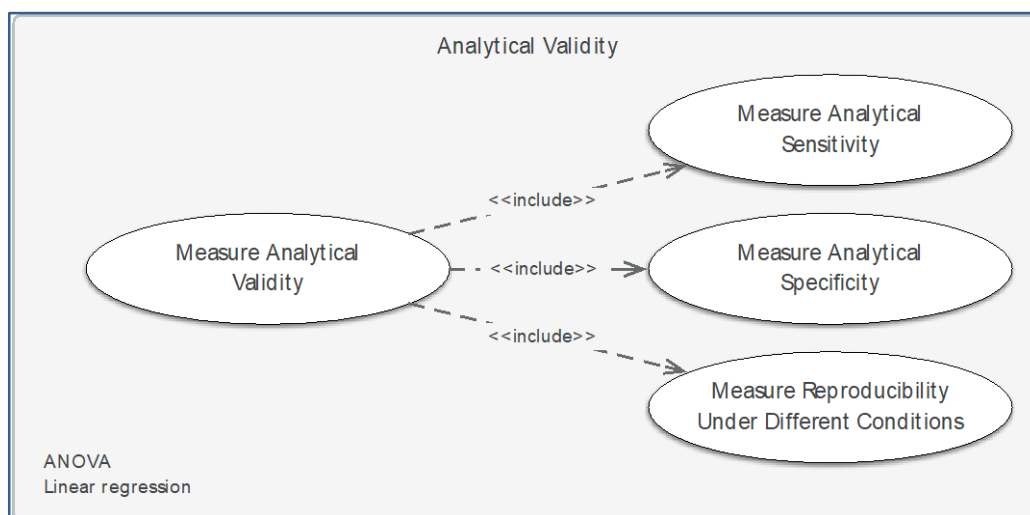


Figure 6: Analytical Validity Use Case

Analytical validity is the test's ability to accurately detect and measure the biomarker of interest (i.e. protein, DNA, RNA). Are the test results repeatable when performed under identical conditions? Are the test results reproducible when the test is performed under different conditions? Is the test sensitive enough to detect biomarker levels as they occur in a real-life setting?

For DNA-based tests, analytical validity requires establishing the probability that a test will be positive when a particular sequence (analyte) is present (analytical sensitivity) and the probability that the test will be negative when the sequence is absent (analytical specificity). In contrast to DNA-based tests, enzyme and metabolite assays measure continuous variables (enzyme activity or metabolite concentration). One key measure of their analytical validity is accuracy, or the probability that the measured value will be within a predefined range of the true activity or concentration. Another measure of analytical validity is reliability, or the probability of repeatedly getting the same result.

Clinical Validity

The MIP can be used to measure clinical validity of a biomarker or other health-relevant feature, i.e. to assess whether the biomarker or other health-relevant patient feature tested is associated with a disease or outcome or the response to a treatment.

Testing of whether a test is accurately detecting and measuring a biomarker or other health-relevant patient feature, i.e. the assessment of test's analytical validity, is a prerequisite for accurate and reliable measurement of the biomarker's or other health-relevant feature's clinical validity. To measure biomarkers' or other health-relevant features' clinical validity, the values for the tested biomarker or the other health-relevant feature, i.e. the data stored in MIP Feature Data Store, must be accurate and reliable. The MIP SGA1 can measure clinical validity of the following types of health-related features: neuromorphometric, cognitive, demographic, genetic, molecular and other biomedical metrics.

Assessment of clinical validity involves measurement of biomarker's or other health-relevant feature's clinical performance, including: (1) clinical sensitivity (ability to identify those who have or will get the disease), (2) clinical specificity (ability to identify those who do not have or will not get the disease), (3) positive predictive value (PPV) - the probability that a person with a positive test result for a predictor, i.e. a biomarker or other health-relevant feature, has or will get the disease, and negative predictive value (NPV) - the probability that a person with a negative test result for a predictor, i.e. a biomarker or other health-relevant feature, does not have or will not get the disease.

When there are more data available in MIP, meaning the number of patients and the diversity of their conditions and profiles, the measurement of clinical validity will be more accurate and reliable.

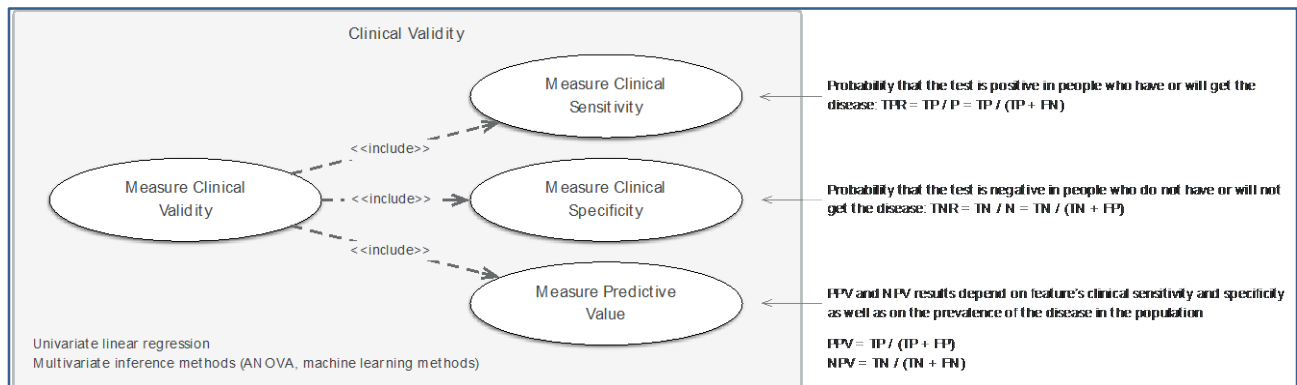


Figure 7: Clinical Validity Use Case

MIP can be used to measure clinical validity on its own, or to include measurement of clinical validity as a research dataset validation step prior to executing a scientifically relevant clinical or biomedical research study using that dataset.

Clinical Utility

Clinical utility is perhaps one of most important considerations when determining whether or not to order or cover a biomedical or other health-relevant feature test. While the meaning of the term has some variability depending on the context or source, there is a largely agreed-upon definition. Four factors are generally considered when evaluating the clinical utility of a test:

- **Patient outcomes** - do the results of the test ultimately lead to improvement of health outcomes (e.g. reduce mortality or morbidity) or other outcomes that are important to patients such as quality of life?
- **Diagnostic thinking** - does the test confirm or change a diagnosis? Does it determine the aetiology for a condition or does it clarify the prognosis?
- **Decision-making guidance** - will the test results determine the appropriate dietary, physiological, medical (including pharmaceutical), and/or surgical intervention?
- **Familial and societal impacts** - does the test identify family members at risk, high-risk race/ethnicities, and the impact on health systems and/or populations?

The development of tests to predict future disease often precedes the development of interventions to prevent, ameliorate, or cure that disease. Even during this therapeutic gap, benefits might accrue from testing. However, in the absence of definitive interventions for improving outcomes in those with positive test results, the clinical utility of the testing will be limited. To improve the benefits of testing, efforts must be made to investigate the safety and effectiveness of new interventions while the tests are developed.

Clinical utility is not always evident in testing for inherited disorders for which treatments have not yet been developed. The clinical utility of a genetic diagnosis for an incurable or untreatable disease, without knowing the outcome, just looking for a predisposition to disease, is not useful.

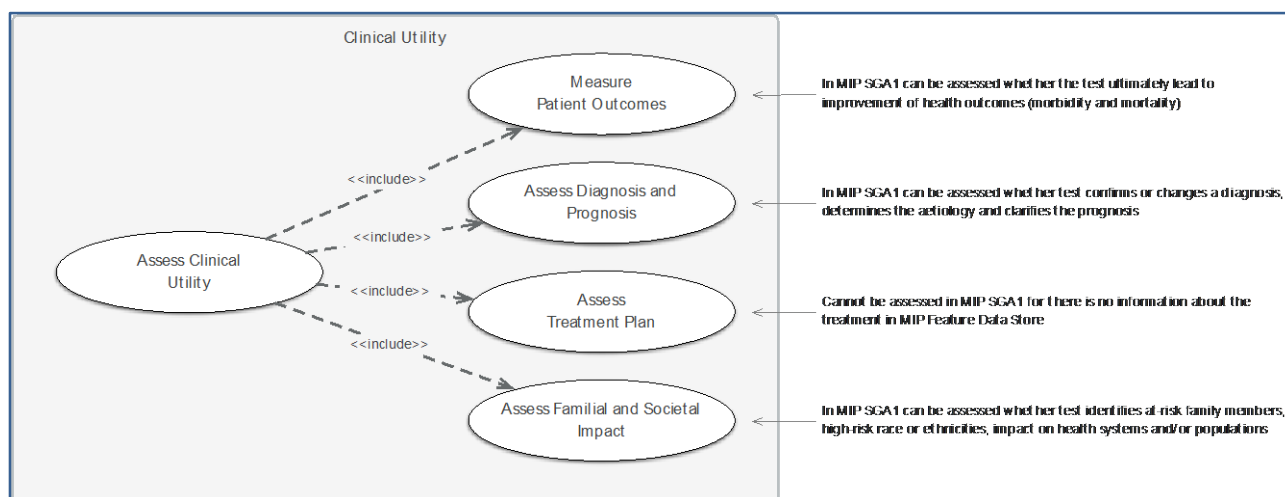


Figure 8: Clinical Utility Use Case

Appendix II: Medical Informatics Platform Component Model

The Medical Informatics Platform is a complex information system comprising numerous software components designed and integrated by different SP8 partners.

This Chapter provides an end-to-end functional overview of the Platform, describing the logical component architecture and the components' roles, showing how the functionality is designed inside the Platform, regarding the static structure of the Platform and the interaction between its components.

This Chapter also contains a brief overview of the key deployment architecture concepts, without providing a detailed specification of the deployment of components into the Platform's physical architecture. Some deployment terminology, such as "local hospital MIP" and "central MIP federation node" is used here only in the context of describing the function of relevant components.

Functional Architecture Overview

Data Capture Sub-system

The Data Capture sub-system provides a local interface to other hospital information systems. It is a single point of entry for all the data that contain personally identifiable information.

The purpose of the Data Capture sub-system is de-identification of patient data exported from hospital information systems (EHRs, PACS). De-identified data is uploaded to De-identified Data Version Control Storage, belonging to the Data Factory sub-system, for processing and feature extraction.

The flow of data between the Data Capture component (Data De-identifier) and, on one side, other local hospital information systems and, on the other side, the MIP Data Factory sub-system is as follows:

- 1) MIP captures personal health sensitive data from the following hospital information systems:
 - Electronic Health Record (EHR) Systems
 - Picture Archiving and Communication Systems (PACS)
- 2) Data De-identifier replaces the following personally identifiable information with pseudonyms:
 - Information exported from EHR systems in CSV format
 - Information from neuroimages stored in the headers of DICOM files
- 3) Data De-identifier saves the files with de-identified data to storage in the Data Factory sub-system

Anonymised patient cohort datasets (for example, ADNI, EDSD, PPMI) are stored directly in the De-identified Data Version Controlled Storage belonging to the Data Factory sub-system.

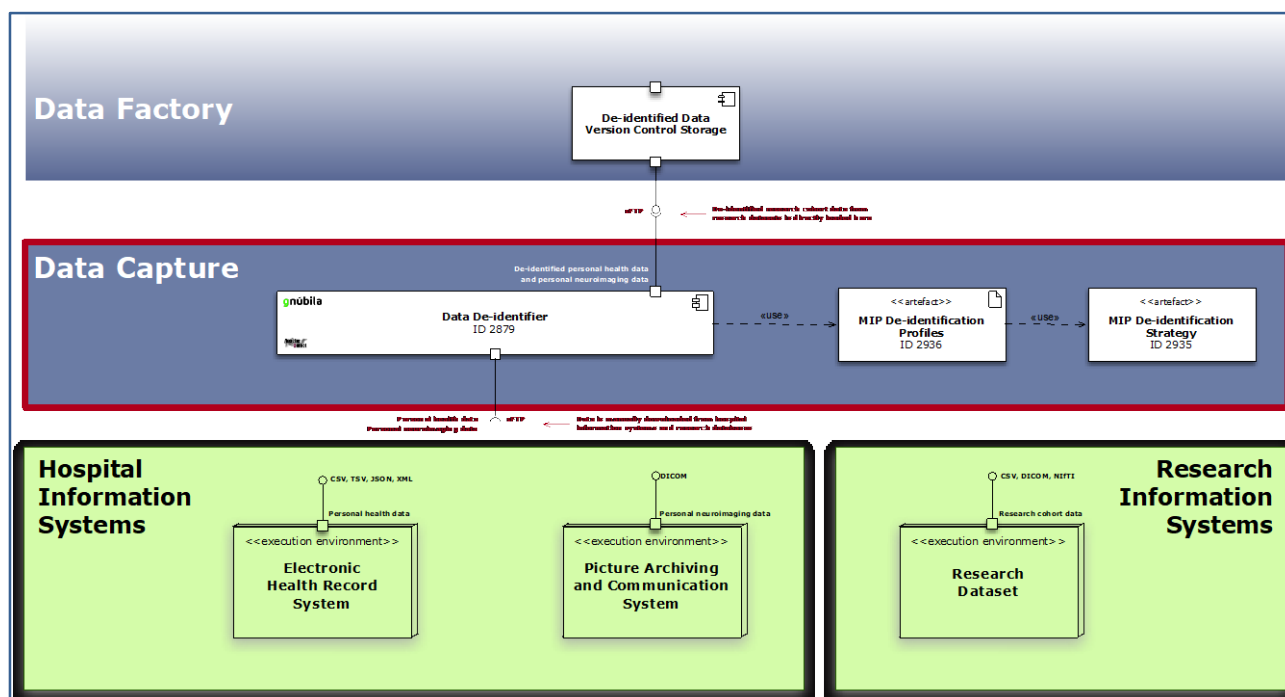


Figure 9: Data Capture Sub-system

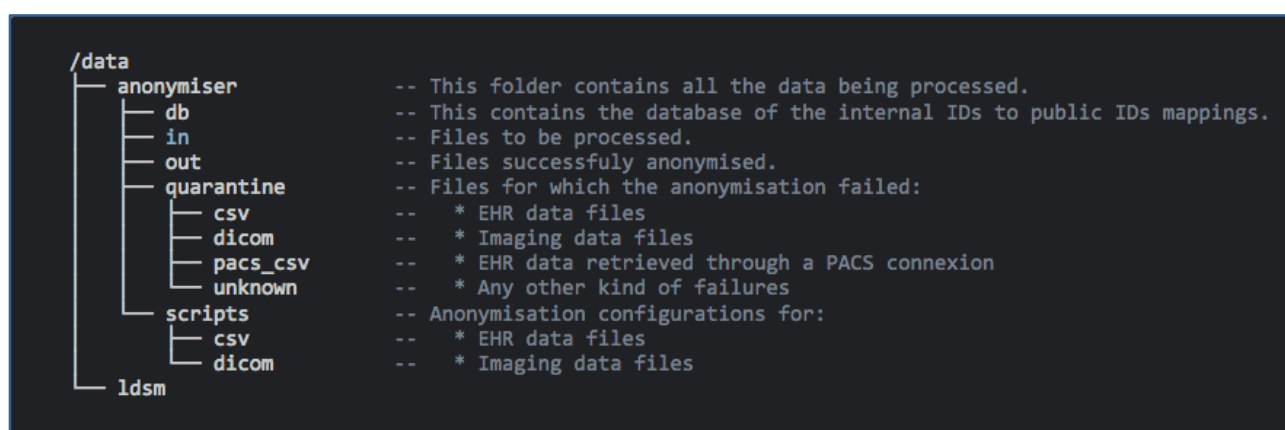


Figure 10: Data Folder Organisation for the De-identification Processing

The Electronic Health Record (EHR) is a collection of a patient health information stored by EHR systems in a digital format. EHR systems are designed for capturing and storing of patient data over time. Well-designed EHR systems are online transaction processing systems that collect and store patient data in a normalised database, therefore minimising data redundancy and improving data integrity.

Picture Archiving and Communication System (PACS) provides storage and access to digital images originating from multiple modalities (imaging machine types). The universal format for PACS image storage and transfer is DICOM (Digital Imaging and Communications in Medicine). Non-image data, such as image-related metadata and scanned PDF documents, can be encapsulated in DICOM files.

MIP captures patient personally identifiable demographic, diagnostic and biomedical data from EHR systems in CSV file format and neuroimaging MRI data from PACS systems in DICOM file format. Patient data are captured periodically for batch processing in the MIP.

Authorised hospital staff that exported the data, manually imports them into the MIP Data De-identifier component for de-identification.

In coordination with local hospital's data management team and ethics committee, the MIP data governance and data selection team (DGDS) is responsible for the specification of data de-personalisation rules in compliance with data protection regulations, such as EU/GDPR, CH/FADP

and US/HIPAA. The Data de-identifier component's rule engine is configured using configuration scripts derived from these rules.

The third-party GnuBila FedEHR Anonymizer data de-identification solution has been chosen for the Data De-identifier component. This component is a profile-based, rule-based asynchronous message-oriented mediation engine, developed using an Apache Camel framework. It can be extended to support new data formats and de-identification algorithms. It replaces all personally identifiable information from the captured data with pseudonyms using out-of-the-box data de-identification techniques, such as generalisation, micro-aggregation, encryption, swapping and sub-sampling.

Discussion About Data Re-identification

Data re-identification is not a feature of the Medical Informatics Platform. It is not possible to re-identify a patient using any of the designed functions of the MIP (data privacy by design). Administratively and organisationally, re-identification of patient data is the responsibility of their hospitals. Technically, for re-identifying patient data stored in the de-identified form in their hospitals' local MIP data storage, hospital IT staff needs to develop standalone lookup applications to map personally identifiable information with the pseudonyms at the point of de-identification. Those applications shall never be integrated with the MIP.

Data Factory Sub-system

The components of the logical Data Factory sub-system perform batch neuroimaging and EHR data pre-processing, extraction, transformation and loading into the normalised permanent data storage.

The ETL processes of the Data Factory sub-system are orchestrated as directed acyclic graphs (DAG's) of tasks in programmatically configurable pipelines using an open-source Apache Airflow workflow management platform. Additional components are built for data transformation and data provenance tracking, including the complex neuroimaging processing and brain feature extraction, brain scan metadata and EHR data extraction as well as data transformation and loading tasks.

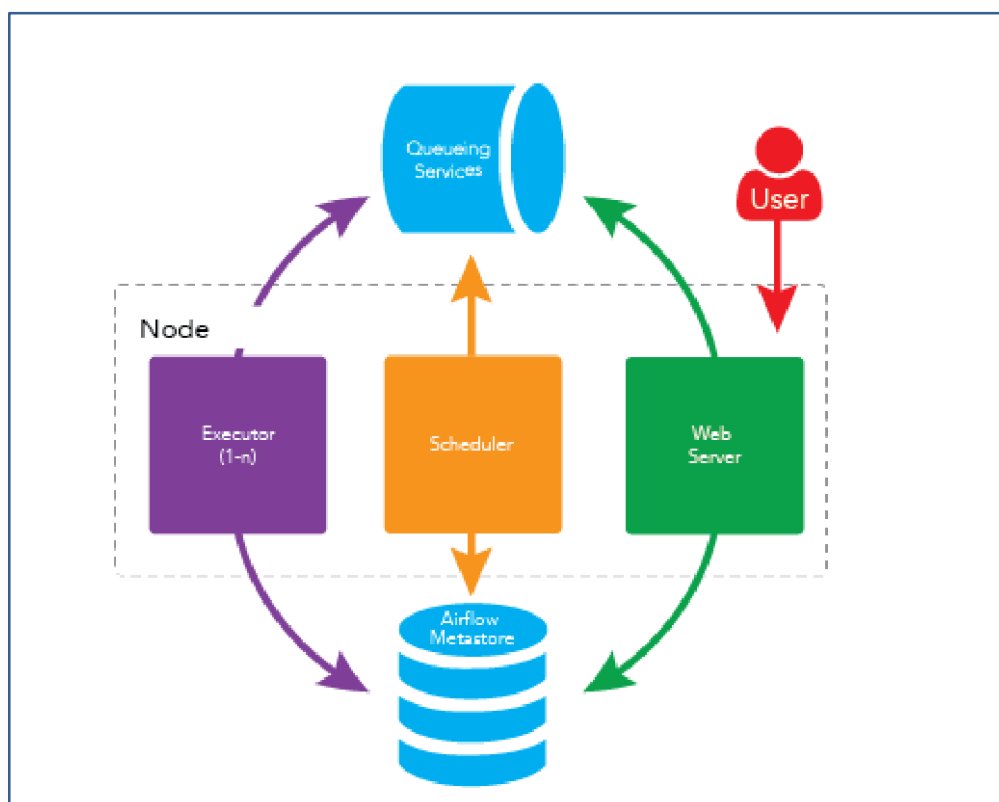


Figure 11: Apache Airflow Concept

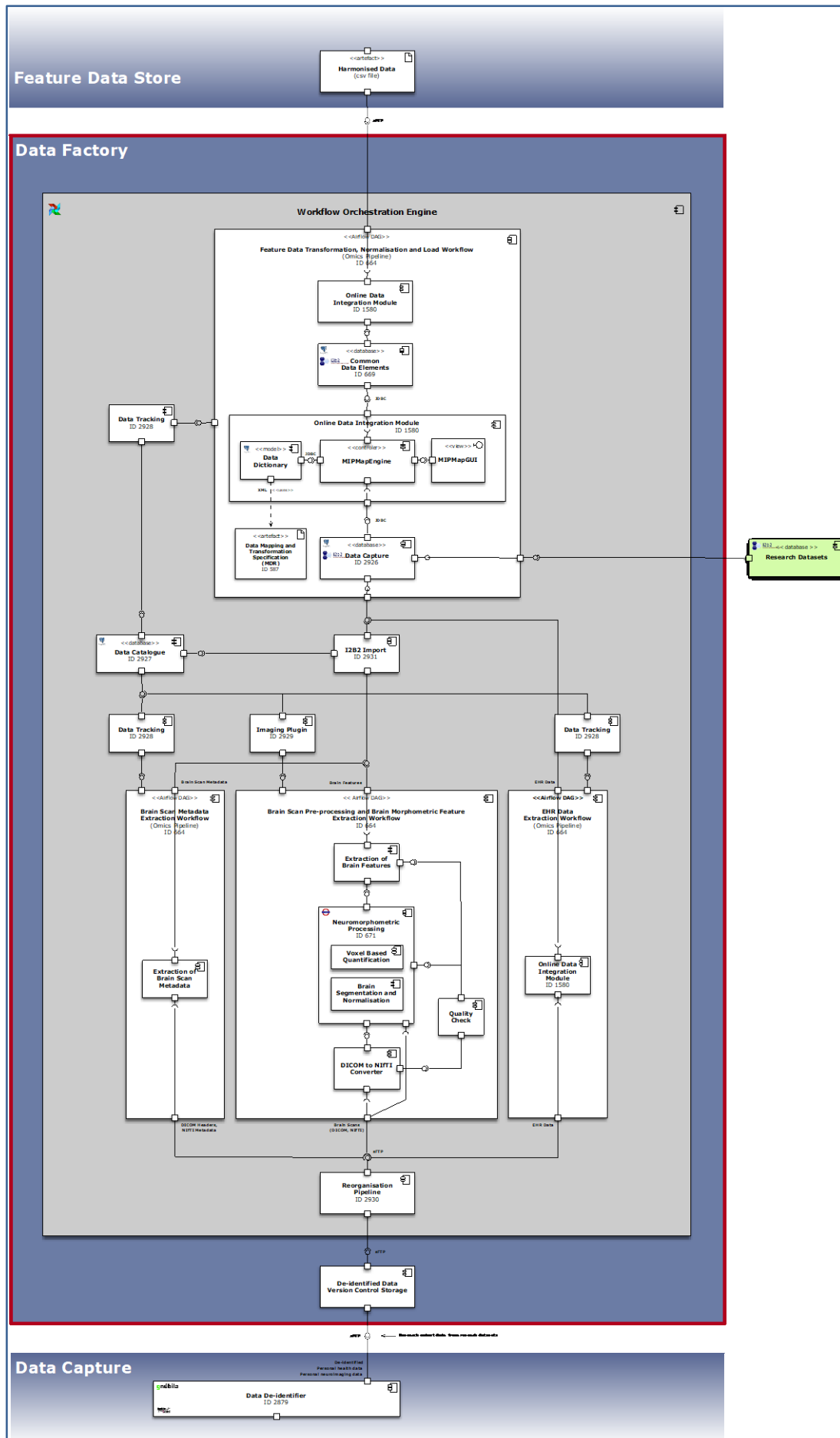
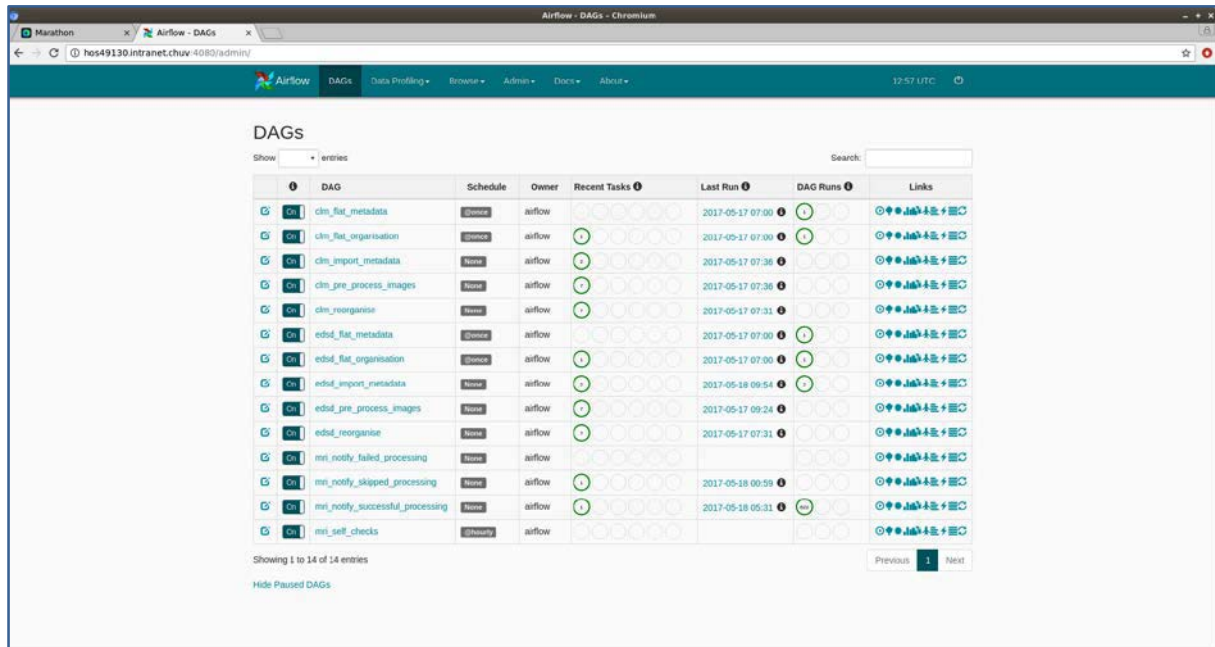


Figure 12: Data Factory Sub-system

Airflow is an open source solution for defining, scheduling, and monitoring of jobs. Pipelines are defined as a code using Python and the jobs are scheduled using cron expressions. The scheduler executes tasks on an array of workers according to the specified dependencies. The user interface makes it easy to visualise pipelines running in production, monitor progress, and troubleshoot issues when needed.



Icon	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	cmr_flat_metadata	Cron	airflow		2017-05-17 07:00		
	cmr_flat_organisation	Cron	airflow		2017-05-17 07:00		
	cmr_import_metadata	None	airflow		2017-05-17 07:36		
	cmr_pre_process_images	None	airflow		2017-05-17 07:36		
	cmr_reorganise	None	airflow		2017-05-17 07:31		
	edid_flat_metadata	Cron	airflow		2017-05-17 07:00		
	edid_flat_organisation	Cron	airflow		2017-05-17 07:00		
	edid_import_metadata	None	airflow		2017-05-18 09:54		
	edid_pre_process_images	None	airflow		2017-05-17 09:24		
	edid_reorganise	None	airflow		2017-05-17 07:31		
	mri_notify_failed_processing	None	airflow				
	mri_notify_skipped_processing	None	airflow		2017-05-18 09:59		
	mri_notify_successful_processing	None	airflow		2017-05-18 05:31		
	mri_self_checks	Cron	airflow				

Figure 13: Apache Airflow Dashboard

The Data Factory sub-system provides the following extraction, transformation and load functionality:

- 1) Pulling de-identified data out of the files stored in De-identified Data Version Control Storage

<pre> DICOM ├── 2016 │ └── 20161029 │ ├── scan_research_id │ │ ├── dicom_name_generated_01.dcm │ │ ├── dicom_name_generated_02.dcm │ │ └── dicom_name_generated_03.dcm └── EHR ├── 2016 │ └── 20161029 │ ├── table1.csv │ ├── table2.csv │ └── ... </pre>	<pre> -- yearly folder, date represents the date of export -- daily folder, date represents the date of export -- see description below -- set of DICOM files -- set of DICOM files -- set of DICOM files </pre>
<pre> EHR ├── 2016 │ └── 20161029 │ ├── table1.csv │ ├── table2.csv │ └── ... </pre>	<pre> -- yearly folder, date represents the date of export -- daily folder, date represents the date of export -- pre-defined name for 1st table containing EHR data, depends on hospital data -- pre-defined name for 2nd table containing EHR data, depends on hospital data -- more (or less) tables as needed, depends on hospital data </pre>

Figure 14: De-identified DICOM and EHR Data

<pre> NIFTI ├── 2016 │ └── 20161029 │ ├── scan_research_id │ │ ├── dicom_name_generated_01.nifti │ │ ├── dicom_name_generated_01.json │ │ ├── dicom_name_generated_02.nifti │ │ └── dicom_name_generated_02.json └── EHR ├── 2016 │ └── 20161029 │ ├── table1.csv │ ├── table2.csv │ └── ... </pre>	<pre> -- yearly folder, date represents the date of export -- daily folder, date represents the date of export -- see description below -- Nifti file -- metadata for the Nifti file -- Nifti file -- metadata for the Nifti file </pre>
<pre> EHR ├── 2016 │ └── 20161029 │ ├── table1.csv │ ├── table2.csv │ └── ... </pre>	<pre> -- yearly folder, date represents the date of export -- daily folder, date represents the date of export -- pre-defined name for 1st table containing EHR data, depends on hospital data -- pre-defined name for 2nd table containing EHR data, depends on hospital data -- more (or less) tables as needed, depends on hospital data </pre>

Figure 15 - De-identified NIFTI and EHR Data

- 2) Processing de-identified data to extract a patient's raw health-related features:
 - a) Brain morphometric features (grey matter volume, shape and dimensions)
 - b) Brain scan metadata
 - c) Data from EHR files (demographic, biomarkers, neuropsychological assessments, diagnoses)
- 3) Harmonising data types from different source datasets into a common data element (CDE) model
- 4) Transformation of the extracted feature data and its permanent storage into the CDE Database
- 5) Placing feature data into files accessible by Features Data Store sub-system components

In addition to the components for extracting personal health features, the Data Factory sub-system contains a set of quality assurance components:

- **Quality Check** for a computational check of the quality of processed and extracted data
- **Imaging Plugin** to track all data changes during brain scan data processing and extraction
- **Data Tracking** to track all data changes except during brain scan data processing and extraction
- **Data Catalogue** to store data provenance/data version information

Reorganisation Pipeline

The Reorganisation pipeline is a component conditional to reorganise datasets pulled from the De-identified data version control storage to prepare them to enter the workflows for processing and extracting brain scan metadata, brain scan pre-processing and brain morphometric feature extraction and EHR data extraction.

The configuration of this pipeline needs to be tailored to every new hospital and research data set. The structure of the brain scan files (DICOM or NIfTI), including the metadata in their headers, depends on the non-standardised procedures specific for each hospital. The structure and the content of EHR files also need to be inspected, and configuration of the pipeline tailored accordingly.

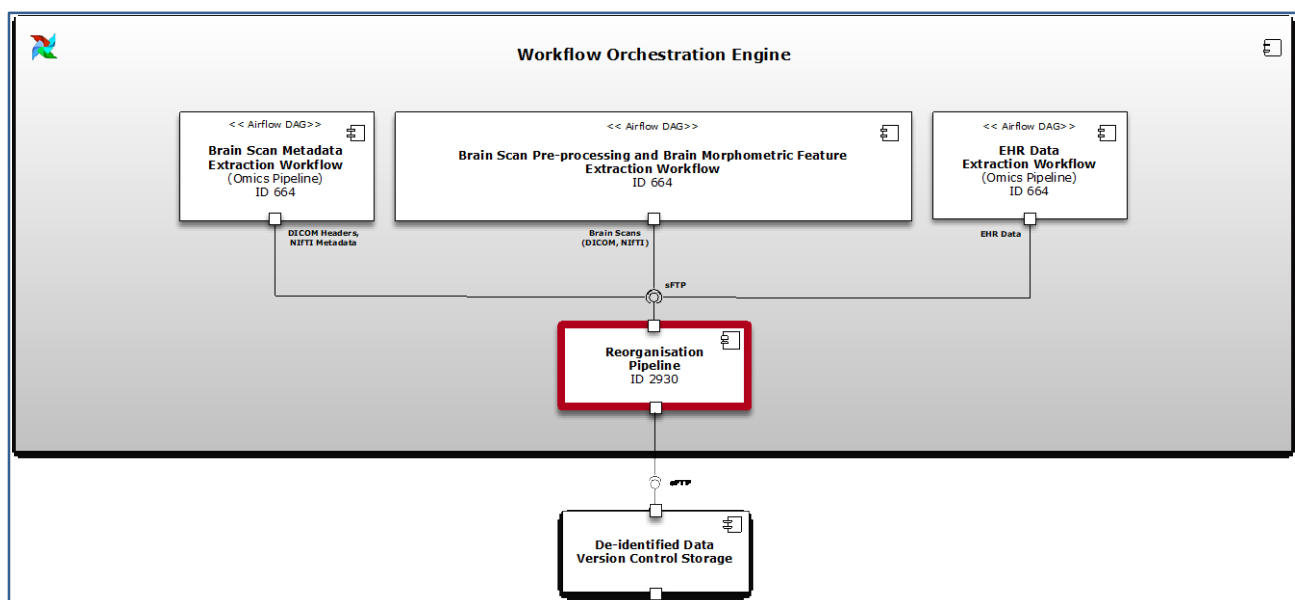


Figure 16: Reorganisation Pipeline

Brain Scan Pre-processing and Brain Morphometric Feature Extraction Pipeline

Software systems are essential in all stages of neuroimaging, allowing scientists to control highly sophisticated imaging instruments and to make sense of the vast amounts of generated complex data. For magnetic resonance imaging (MRI), software systems are used to design and implement signal-capturing protocols in imaging instruments, reconstruct the resulting signals into a three-dimensional representation of the brain, correct for and suppress noise, statistically analyse the data, and visualise the results. Collected neuroimaging data can then be stored, queried, retrieved and shared using PACS, XNAT, CBRAIN, LORIS or any other system. Neuro-anatomical data can be extracted from neuroimages, compared and analysed using other specialised software systems, such as SPM and FreeSurfer.

After capturing and de-identifying neuroimaging DICOM data from PACS systems, the MIP's Data Factory sub-system extracts neuroanatomical data from captured brain magnetic resonance images, permanently stores that data into the Feature Data Store sub-system where it is made available for data mining and analysis together with the rest of biomedical and other health-related information.

The flow of data between Brain Scan Pre-processing and Brain Feature Extraction pipeline components is as follows:

1) A visual quality check of the neuroimages performed by a neuroradiologist.

Pre-processing of magnetic resonance (MR) images strongly depends on the quality of input data. Multi-centre studies and data-sharing projects need to take into account varying image properties due to different scanners, sequences and protocols

Image format requirements:

- Full brain scans
- Provided either in DICOM or NIFTI format
- High-resolution (max. 1.5 mm) T1-weighted sagittal images.
- If the dataset contains other types of images (that is not meeting the above description, e.g. fMRI data, T2 images, etc.), a list of protocol names used and their compatibility status regarding the above criterion has to be provided
- Images must contain at least 40 slices

2) The DICOM to NIFTI Converter converts brain scan data captured in DICOM format to NIFTI data format

3) The Neuromorphometric Processing component (SPM12) uses NIFTI data for computational neuro-anatomical data extraction using voxel-based statistical parametric mapping of brain image data sequences:

- a) Each T1-weighted image is normalised to MNI (Montreal Neurological Institute) space using non-linear image registration SPM12 Shoot toolbox
- b) The individual images are segmented into three different brain tissue classes (grey matter, white matter and CSF)
- c) Each grey matter voxel is labelled based on Neuromorphometrics atlas (constructed by manual segmentation for a group of subjects) and the transformation matrix obtained in the previous step. Maximum probability tissue labels were derived from the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labelling". These data were released under the Creative Commons Attribution-Non-Commercial (CC BY-NC). The MRI scans originate from the OASIS project, and the labelled data were provided by Neuromorphometrics, Inc. under an academic subscription

4) The Voxel-Based Quantification (VBQ) component, through its sensitivity to tissue microstructure, provides absolute measures for neuroimaging biomarkers for myelination, water and iron levels comparable across imaging sites and in time

- 5) The I2B2 Import component stores extracted brain morphometric features in I2B2 Capture Database, alongside the brain scan metadata and patient EHR data

The Quality Check component evaluates essential image parameters, such as signal-to-noise ratio, inhomogeneity and image resolution. It evaluates images for problems during the processing steps. It allows comparing quality measures across different scans and sequences.

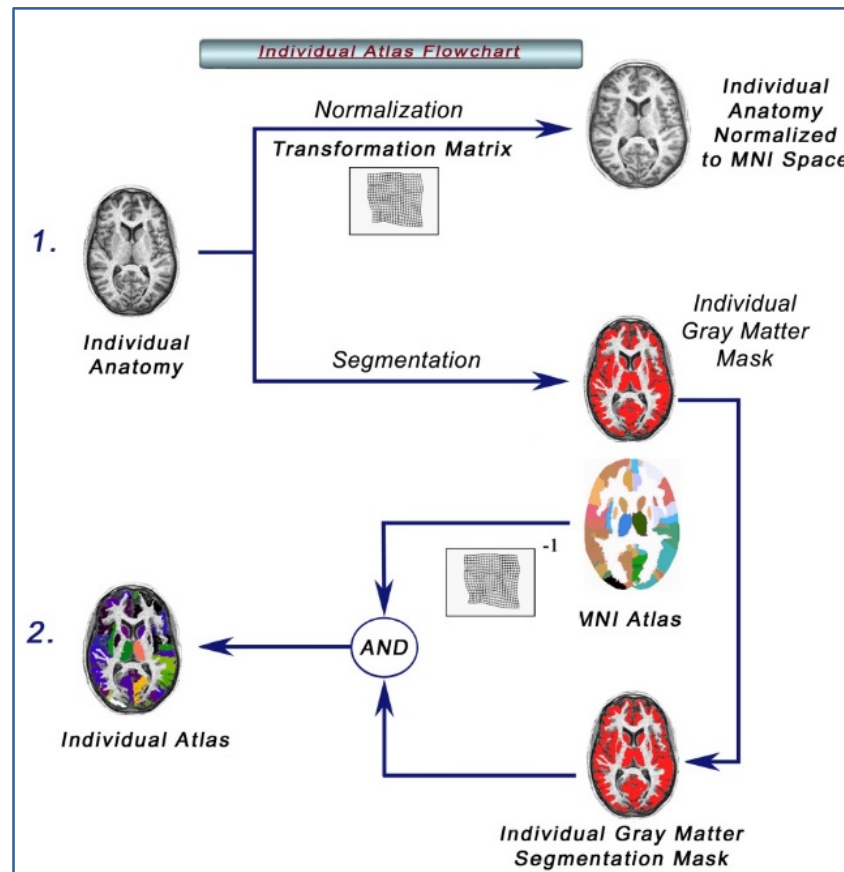


Figure 17: Neuromorphometric Processing

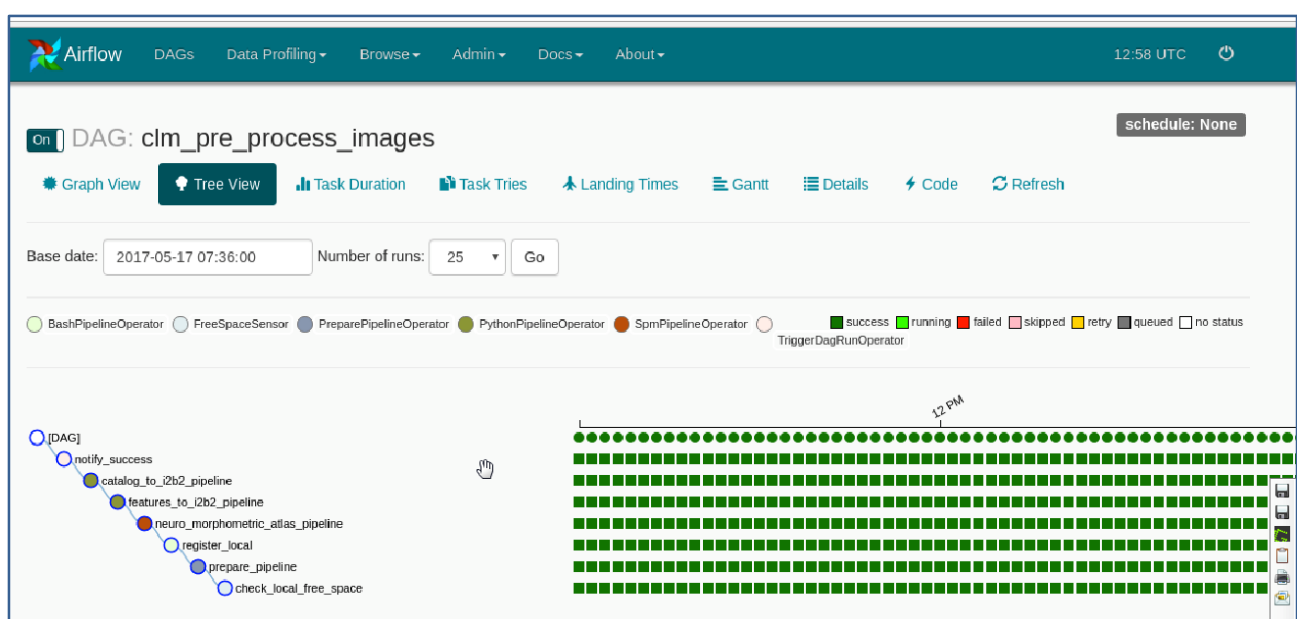


Figure 18: Apache Airflow Image Processing Pipeline Status

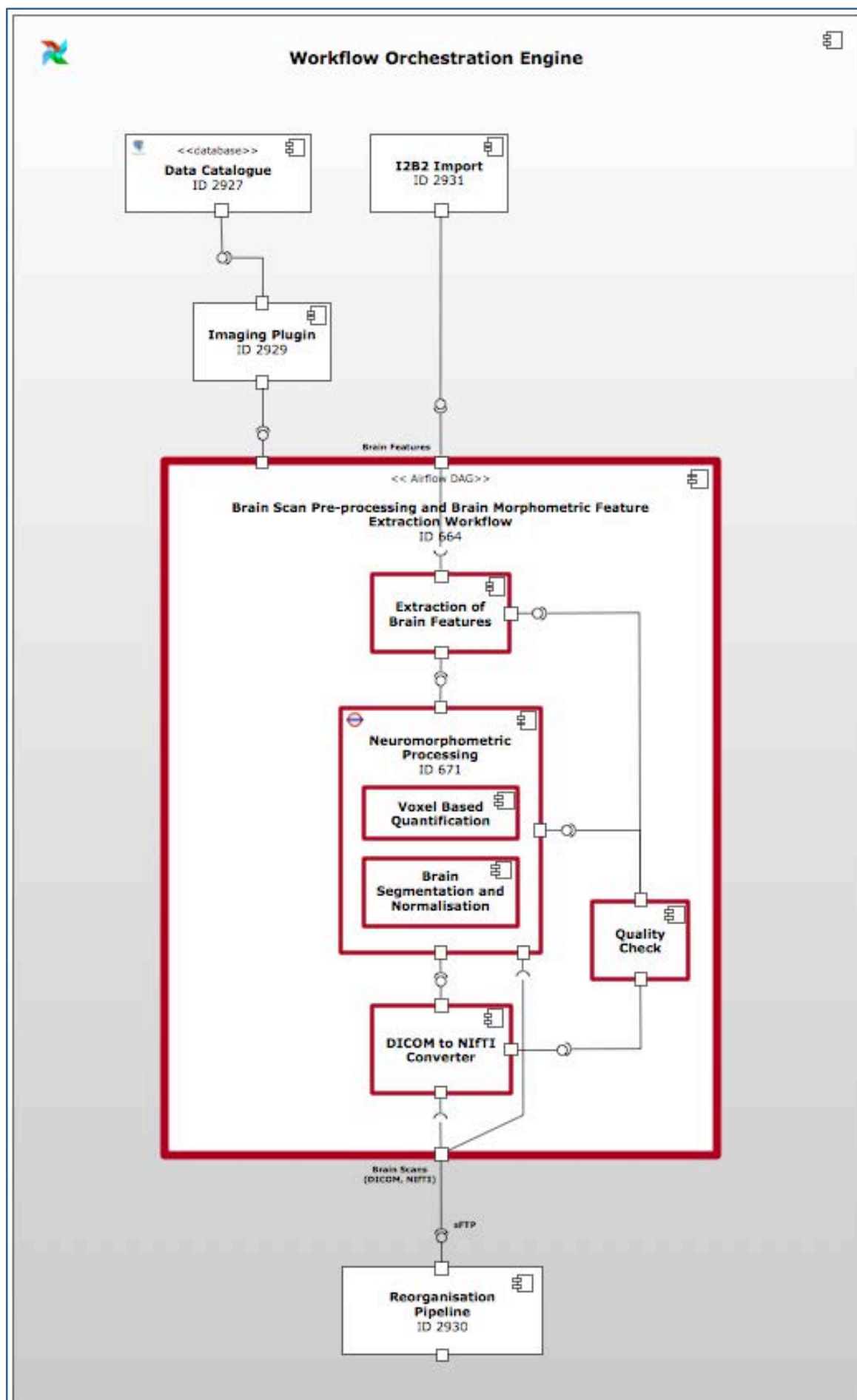


Figure 19: Brain Scan Pre-processing and Brain Feature Extraction Workflow

The Brain Scan Pre-processing and Brain Morphometric Feature Extraction pipeline contains components for processing T1-weighted brain image data sequences and extracting morphometric brain features - grey matter volume and shape - using voxel-based morphometry (VBM). VBM provides insight into macroscopic volume changes that may highlight differences between groups, be associated with pathology or be indicative of plasticity.

For neuromorphometric processing, the MIP uses SPM12 software running within the MATLAB software environment. For image pre-processing and morphometric feature extraction, SPM requires input data in a standard format used by neuromorphometric tools for computation and feature extraction: the NIfTI format.

The T1-weighted images are automatically segmented into 114 anatomical structures using the Neuromorphometrics atlas.

In addition to voxel-based neuromorphometric processing of T1-weighted images for classification of tissue types and measuring of macroscopic anatomical shape, the MIP uses a voxel-based quantification (VBQ) toolbox as a plugin for SPM12 that can analyse high-resolution quantitative imaging and can provide neuroimaging biomarkers for myelination, water and iron levels that are absolute measures comparable across imaging sites and in time.

Single NIfTI volumes of the brain are first partitioned into three classes: grey matter, white matter and background. This procedure also incorporates an approximate image alignment step and a correction for image intensity non-uniformities. This procedure uses the SPM12 Segment5 tool.

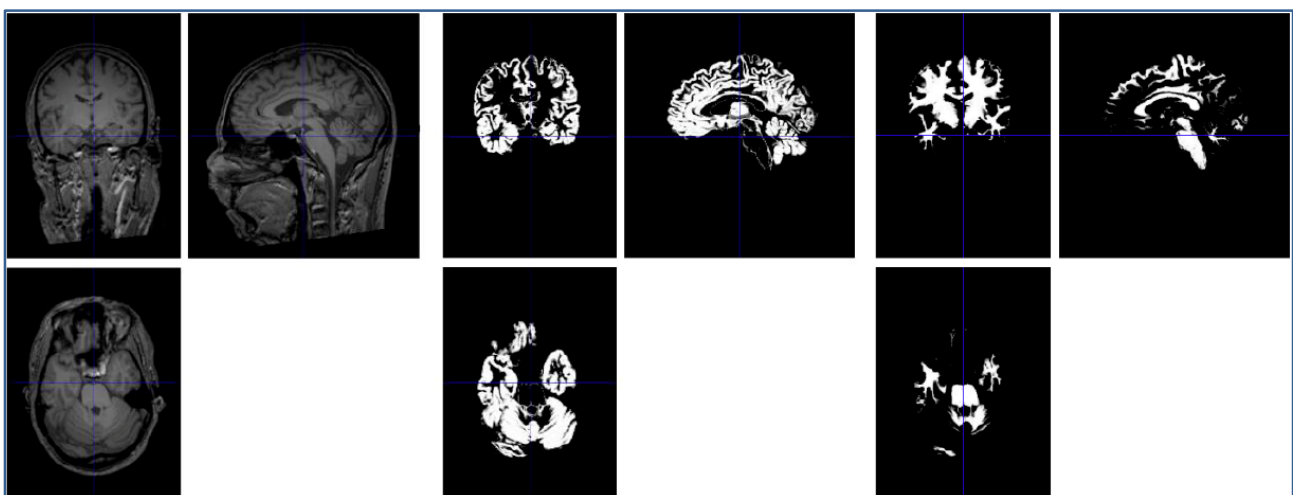


Figure 20: Original T1-weighted MRI scan (left), along with automatically extracted grey (middle) and white matter (right) tissue maps. The tissue maps encode the probability of each tissue type calculated using the given model and data

Tissue atlases, pre-computed from training data are then spatially registered with the extracted grey and white matter maps, using the Shoot5 tool from SPM12. The warps estimated from this registration step are then used to project other pre-computed image data into alignment with the original scans (and their grey and white matter maps).

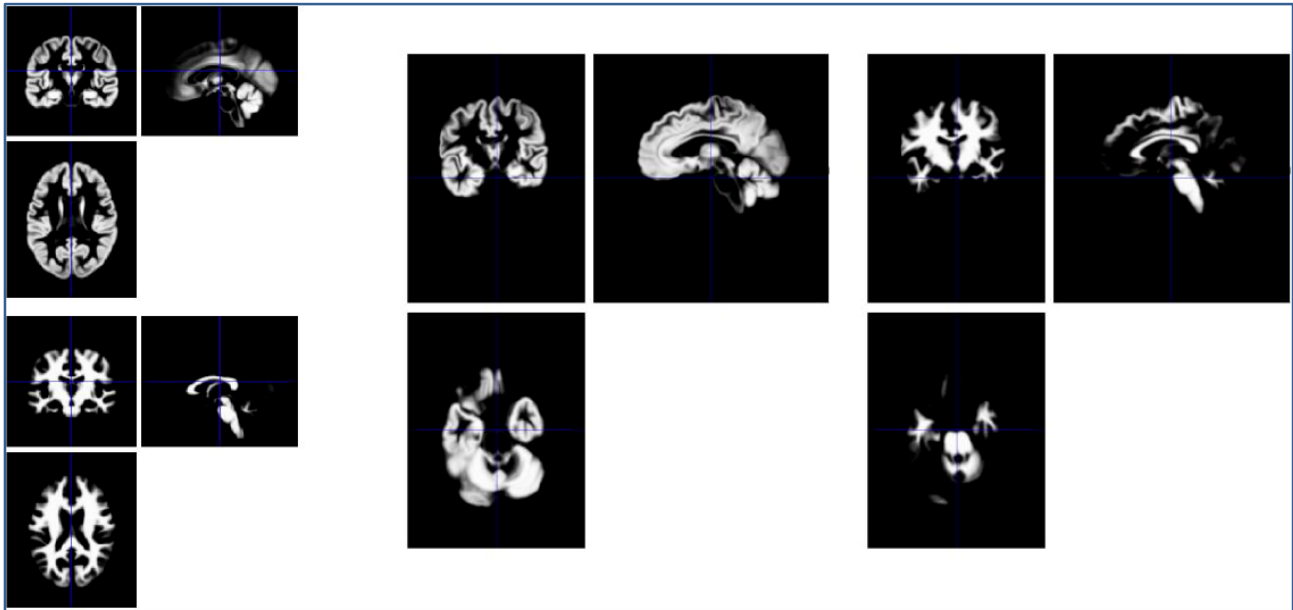


Figure 21: Grey and white matter from the original tissue atlases (left) along with registered versions (middle and right)

The rules of probability are then used to combine the various images to give a probabilistic label map for each brain structure. These probabilities are summed for each structure, to provide probabilistic volume estimates. These estimates are saved in the MIP platform as brain morphometric features.

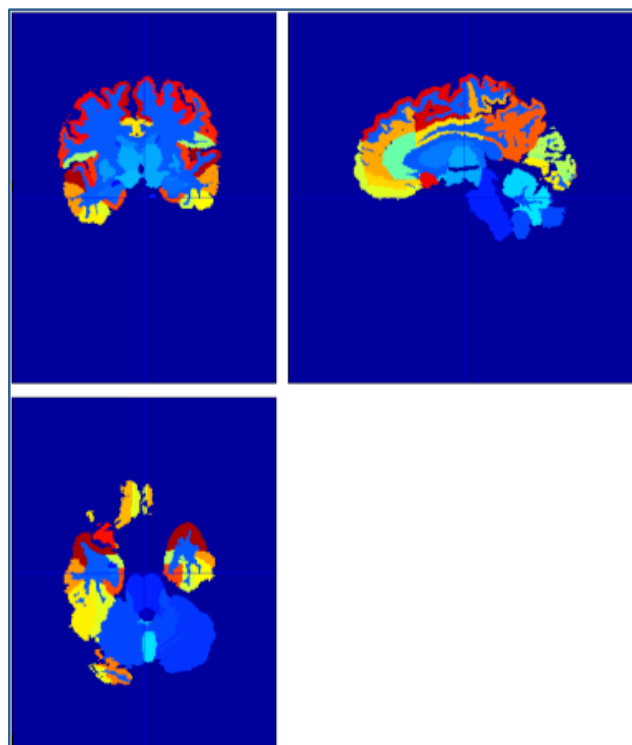


Figure 22: Automatically labelled image, showing most probable macro anatomy structure labels

While Voxel-based morphometry classifies tissue types and measures anatomical shape (Brain Segmentation and Normalisation component), the Voxel-Based Quantification component provides complementary information through its sensitivity to tissue microstructure. The Multi-parameter Mapping (MPM) imaging protocol is used to provide whole-brain maps of relaxometry measures ($R_1 = 1/T_1$ and $R_2^* = 1/T_2^*$), magnetisation transfer saturation (MT) and effective proton density (PD*) with the isotropic resolution of 1mm or higher.

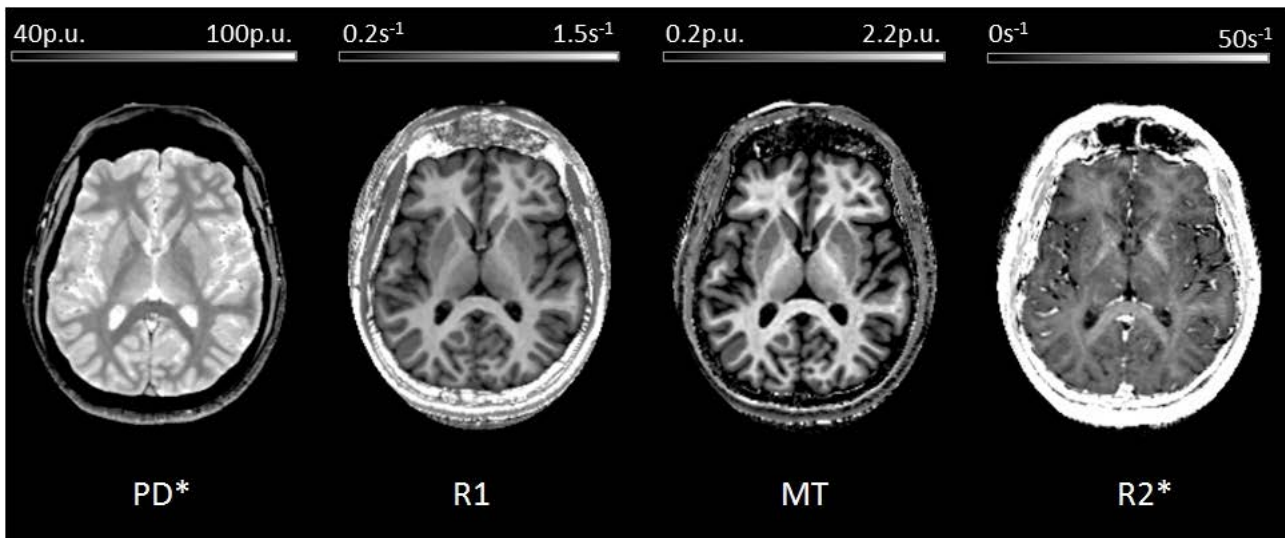


Figure 23: Multi Parameter Mapping high-resolution quantitative MRI acquisition protocol

MPM is a high-resolution quantitative imaging MRI protocol which, combined with VBQ data analysis, opens new windows for studying the microanatomy of the human brain *in vivo*. With T1-weighted images, the signal intensity is in arbitrary units and cannot be compared across sites or even scanning sessions. Quantitative imaging can provide absolute measures for neuroimaging biomarkers for myelination, water and iron levels comparable across imaging sites and in time.

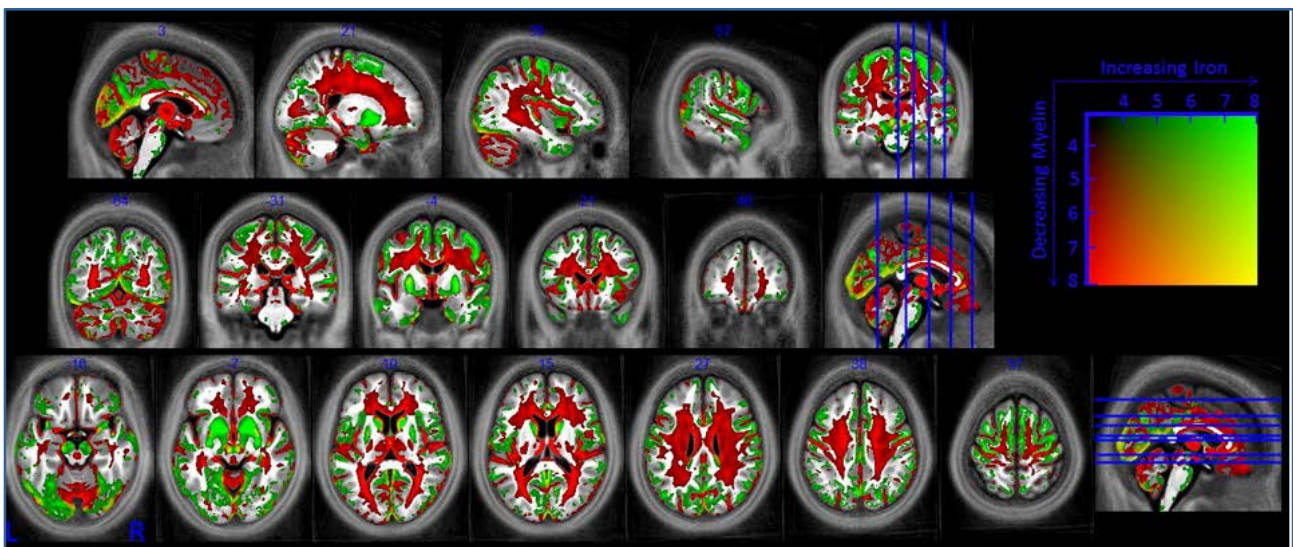


Figure 24: Voxel Based Quantification data analysis for studying microanatomy of the human brain *in vivo*

Brain Scan Metadata Extraction and EHR Data Extraction Pipelines

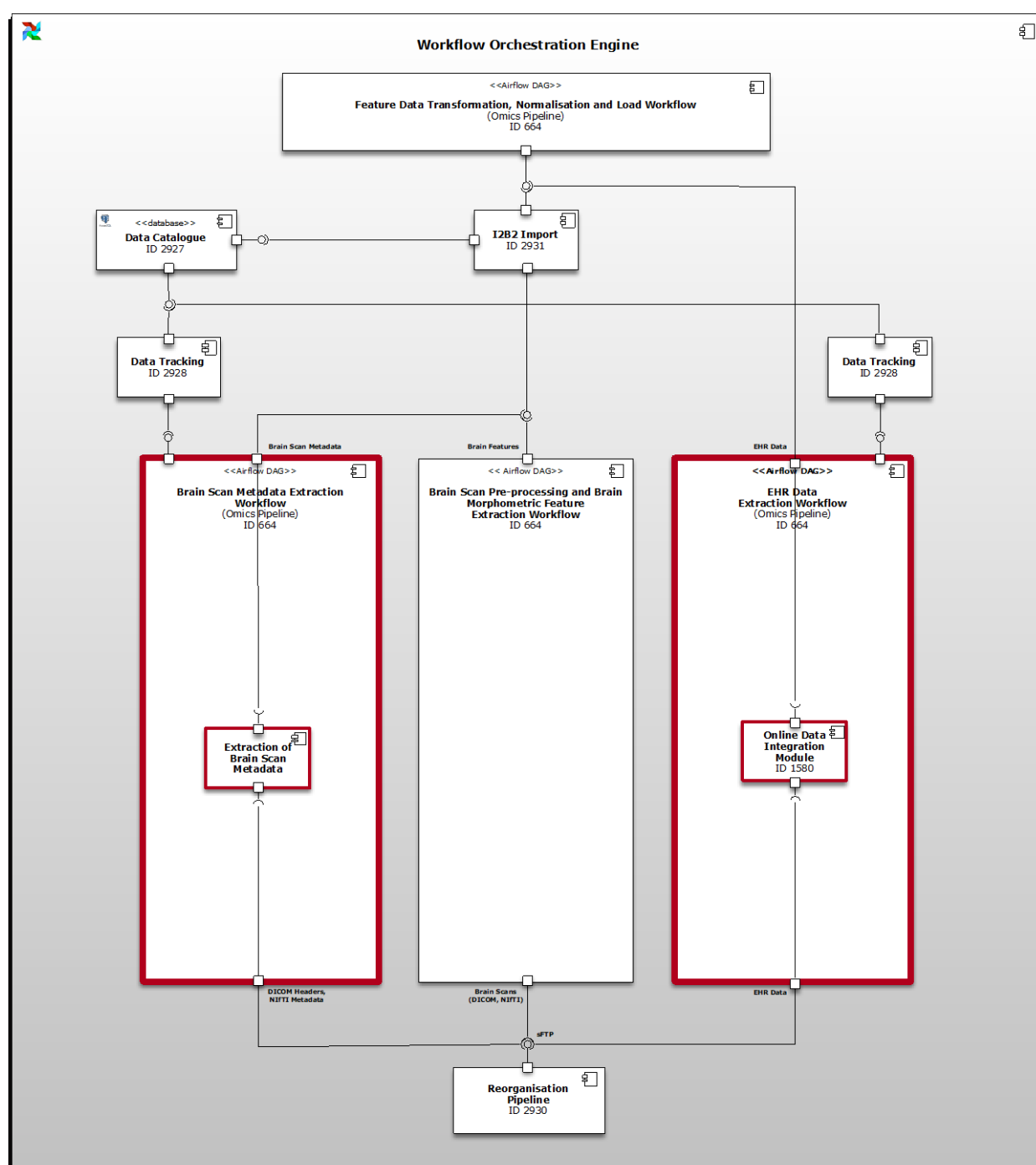


Figure 25: Brain Scan Metadata and EHR Data extraction pipelines

A patient's brain scan metadata and EHR data are extracted from the corresponding de-identified files and stored in I2B2 Capture Database alongside extracted brain morphometric features. Data provenance is stored in Data Catalogue.

Feature Data Transformation, Normalisation and Load Pipeline

This pipeline contains the following components:

- **Data Capture Database** - for storing patient health features extracted from brain scans and EHR files
- **Data Mapping and Transformation Specification** - data mapping rules - the results of harmonising data types from different source datasets into a common data element (CDE) model
- **Online Data Integration Module** - for transformation of the extracted patient feature data into the common data elements format, according to the Data Mapping and Transformation Specification rules. Also for exporting CDE Database to CSV file for storing the harmonised data into the local data store mirror (Features Table) in Features Data Store sub-system
- **Common Data Elements Database** - for permanently storing the transformed patient feature data into a normalised I2B2 schema

Data Capture Database

De-identified data, extracted from patient electronic health records and brain scans, is stored in the original data format in the Data Capture Database, implemented using I2B2 schema managed by PostgreSQL database management system.

The I2B2 schema allows for an optional direct update of Data Capture Database with data from a large number of I2B2-compliant anonymised patient cohort datasets. I2B2 is widely used for implementing clinical data warehouses as well as research data warehouses. Over the years, it became a de facto standard for bridging the gap between clinical informatics and bioinformatics, providing large datasets for clinical, biomedical and pharmaceutical research.

In cases when research datasets are stored in different formats, such as ADNI or BIDS files, they are initially saved in the Data Factory sub-system's version controlled storage before the data is extracted using the extraction pipelines and then finally stored in the Capture Database.

Data Mapping and Transformation Specification

The MIP Data Governance and Data Specification (DGDS) team receive information from hospitals about new data elements that shall be captured from patient EHR and brain scan datasets. In collaboration with hospital clinicians and data managers, the MIP DGDS team analyses new data types and harmonises them into a common data elements model. Data Mapping and Transformation Specification is updated with new harmonisation rules. This artefact is used for transformation of original data extracted from hospitals into the common data element format using the Online Data Integration Module.

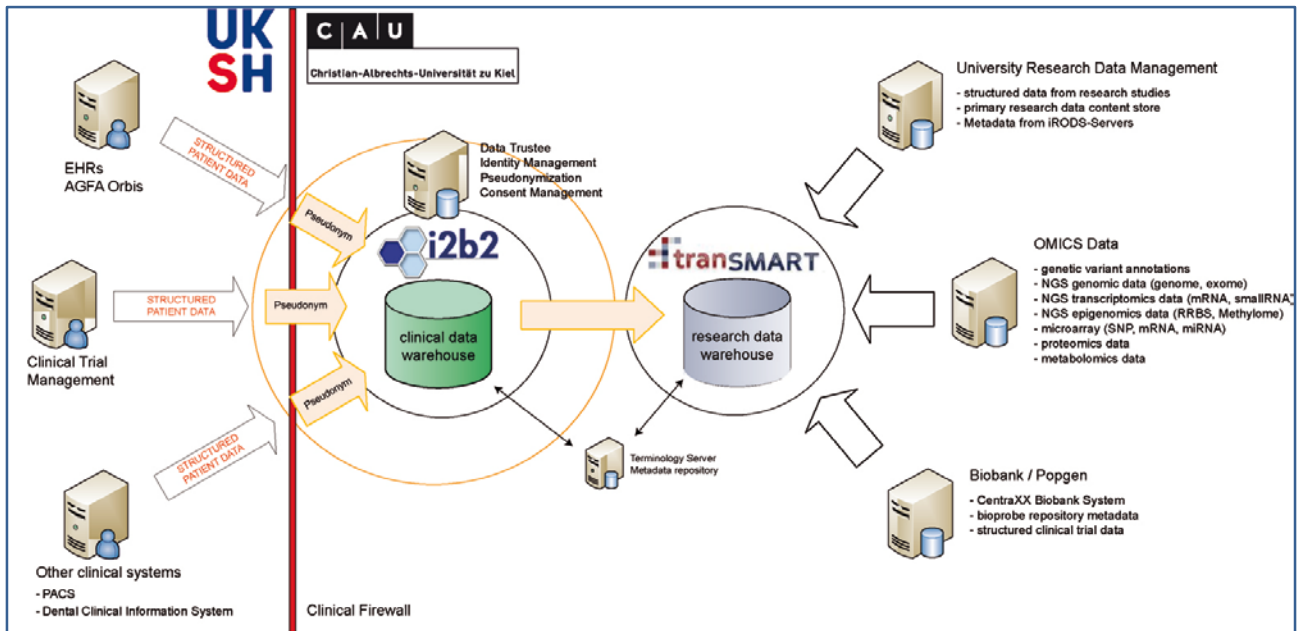


Figure 26: I2B2 tranSMART Foundation's research data warehouse for clinical, biomedical and pharmaceutical research

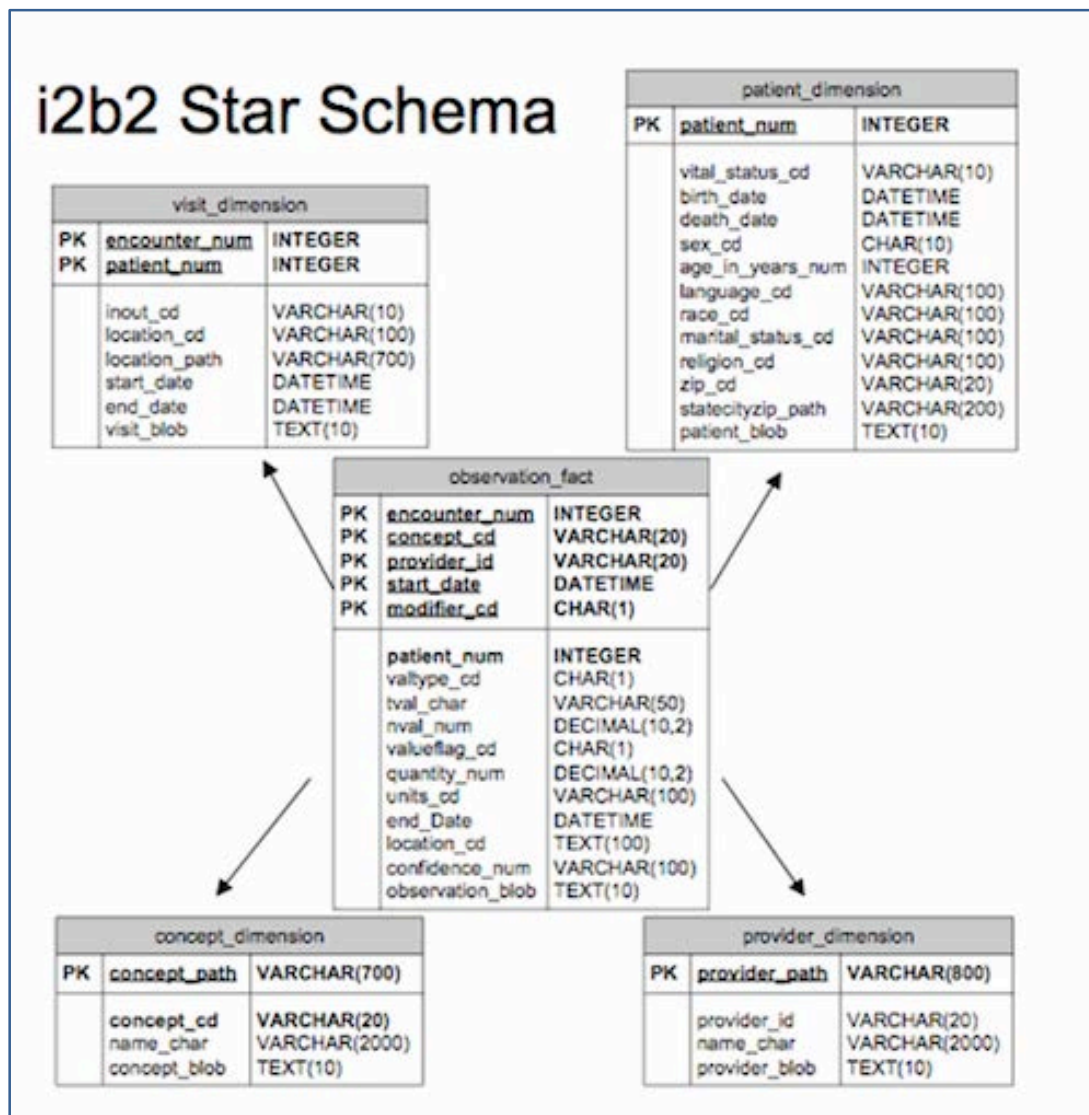


Figure 27: I2B2 Schema

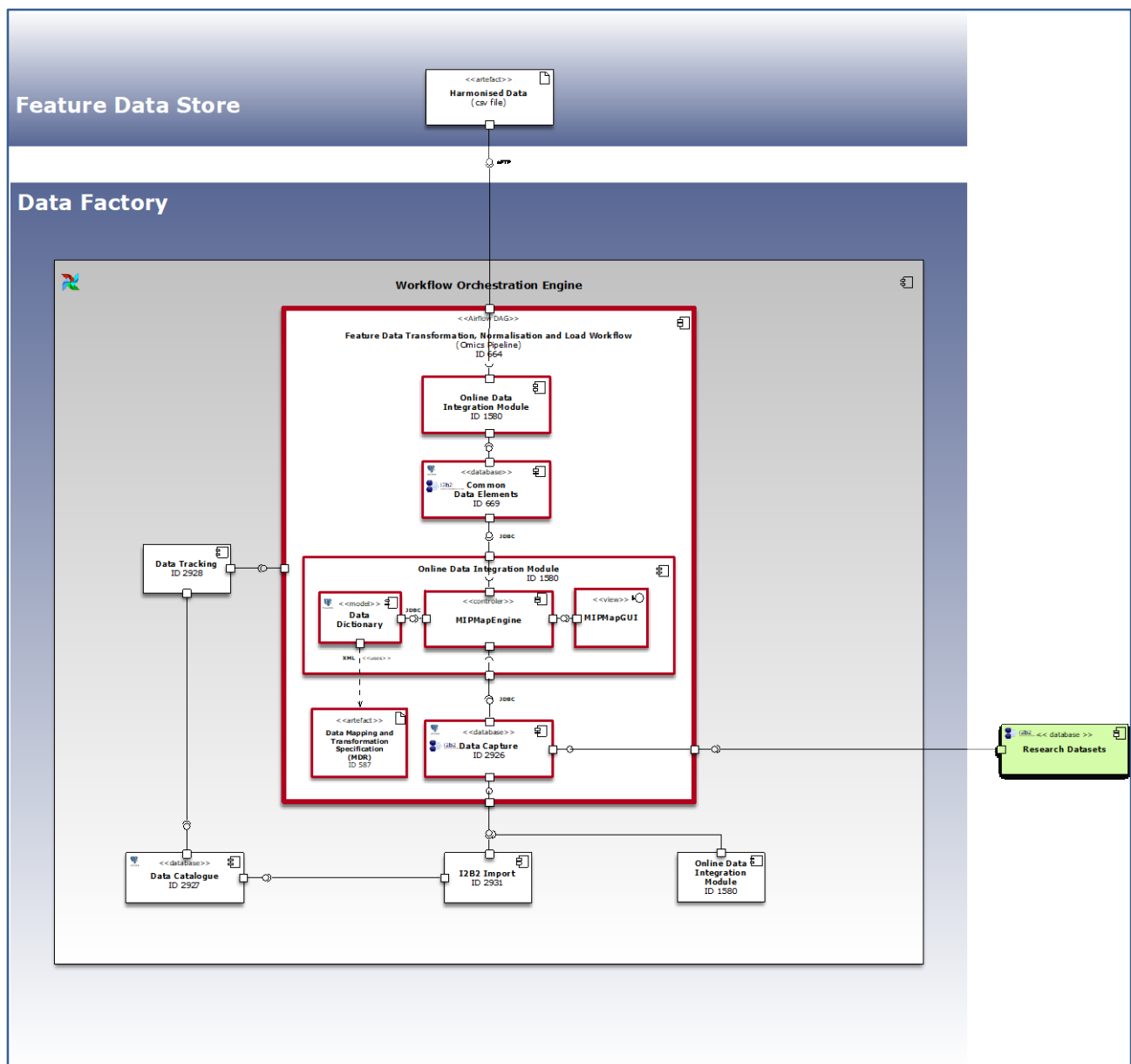


Figure 28: Feature Data Transformation, Normalisation and Load Pipeline

Online Data Integration Module for Data Transformation and Load to CDE Database

The Online Data Integration Module component is used for extracted data transformation, and loading into the normalised I2B2-compliant Common Data Element Database, managed by PostgreSQL database management system. This component is also used to export harmonised data from CDE Database to CSV files, out of which the Feature Table in the Feature Data Store subsystem is populated. The Online Data Integration Module is implemented using an open source ++Spicy data exchange tool. The adaptation of this application for the MIP is called MIPMap. This tool, which has been developed in Java using the NetBeans platform, applies Data Mapping and Transformation Specification rules for transformation of data stored in I2B2 Capture Database to the normalised I2B2 Common Data Elements Database.

MIPMap provides a graphical user interface where a hospital data manager or a MIP DGDS data manager can create mapping correspondences between source data elements and targets by drawing lines between them. This forms a mapping scenario that is stored in XML format. The mapping process is performed once for every hospital.

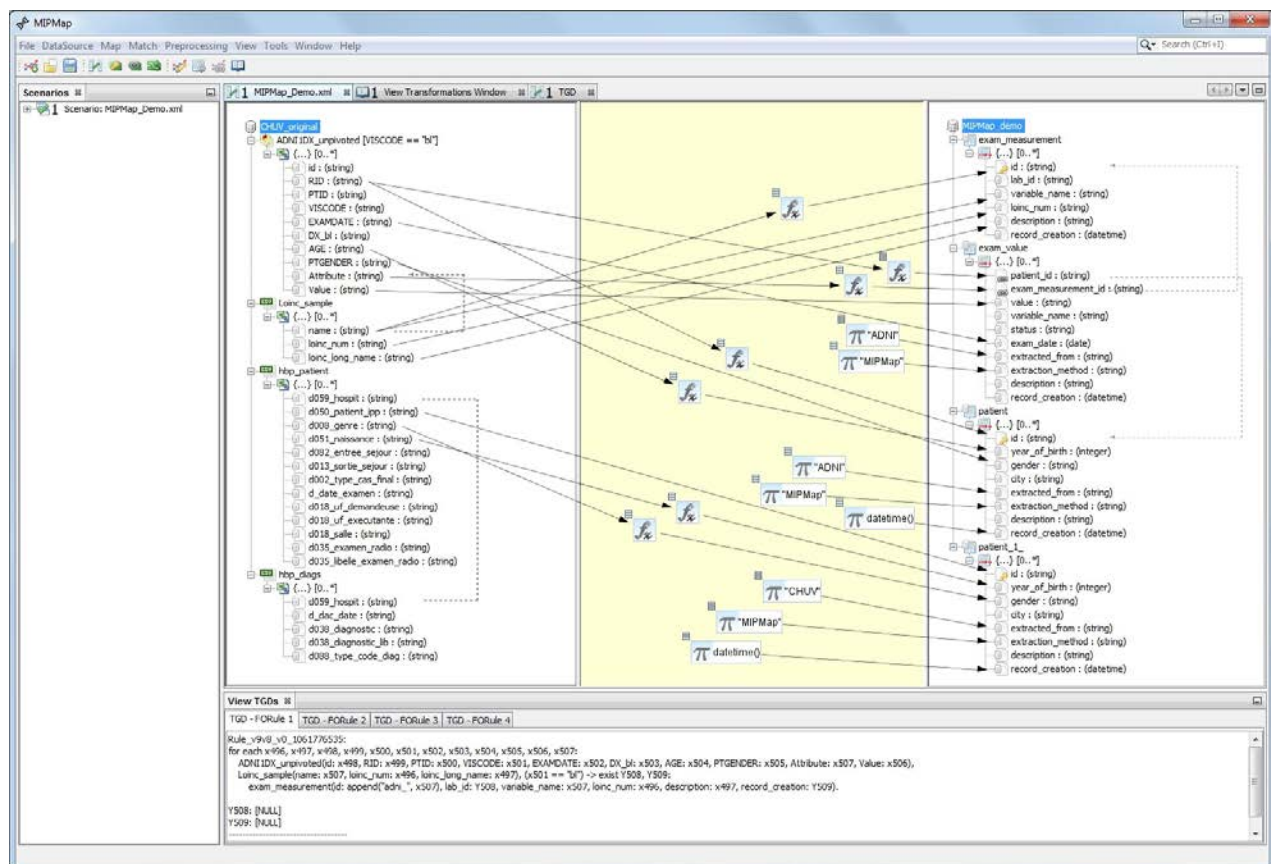


Figure 29: MIPMap user interface

Having created a mapping scenario, the MIPMap Engine generates an optimised SQL script that translates the data from the source (CSV file or a database schema) to the target database schema and then updates the target database.

Common Data Elements Database (delivered by CHUV team)

After the source data types have been mapped to the destination common data elements schema using the Data Mapping and Transformation Specification, the Online Data Integration Module loads the data from the Capture Database to the Common Data Elements Database.

An I2B2-compliant Common Data Elements (CDE) database schema is incorporated on top of the PostgreSQL database management system for permanently storing harmonised patient data from different hospitals and research datasets.

One of the key added-value characteristics of the MIP is the harmonisation of data elements from diverse source systems - EHR systems from different hospitals, imaging and PACS systems and research datasets. The harmonised data model is implemented as an I2B2-compliant database schema, which allows for a prospective easy integration with a large research datasets compliant with I2B2.

Online Data Integration Module for Transformation of CDE Database to Harmonised Data CSV File

Harmonised data from the CDE Database is transformed using the Online Data Integration Module component into a Harmonised Data CSV File in the Feature Data Store sub-system. The MIPMap Engine executes a pivoting script, for pivoting the variables and their values stored in the dimensional I2B2 (data mart) schema of CDE Database into a flat comma-separated value representation. The Harmonised Data CSV File is processed by the Query Engine and stored in the Feature Table to be available to the components of the Knowledge Extraction sub-system for data mining, statistical analysis and predictive machine learning.

Feature Data Store Sub-system

The Feature Data Store Sub-system contains components for mirroring harmonised patient data in the form appropriate for querying and using by machine learning algorithms. The components of this subsystem operate on and store the data belonging to one and only one hospital. The data is made available both for the local knowledge extraction MIP subsystem and to the remote, federated knowledge extraction MIP sub-system.

The components of the Feature Data Store sub-system are as follows:

- **Harmonised Data CSV file** - for mirroring harmonised CDE data exported from CDE database
- **Query Engine** - hospital DB back end, executing queries on extracted patient health sensitive data
- **Features Database** - hospital local data store mirror, data ready for querying and machine learning
- **PostgresRAW-UI** - user interface for Query Engine administration, including CSV files monitor

Harmonised Data CSV File

Using the Online Data Integration Module component, harmonised de-identified health-related patient data is exported from the CDE Database in the Data Factory sub-system into the CSV files accessible from the Feature Data Store sub-system components. The Query Engine component queries data stored in these files. The Query Engine also makes the data available for fetching by data mining and machine learning algorithms by storing it in the Hospital Dataset table of the Features Database.

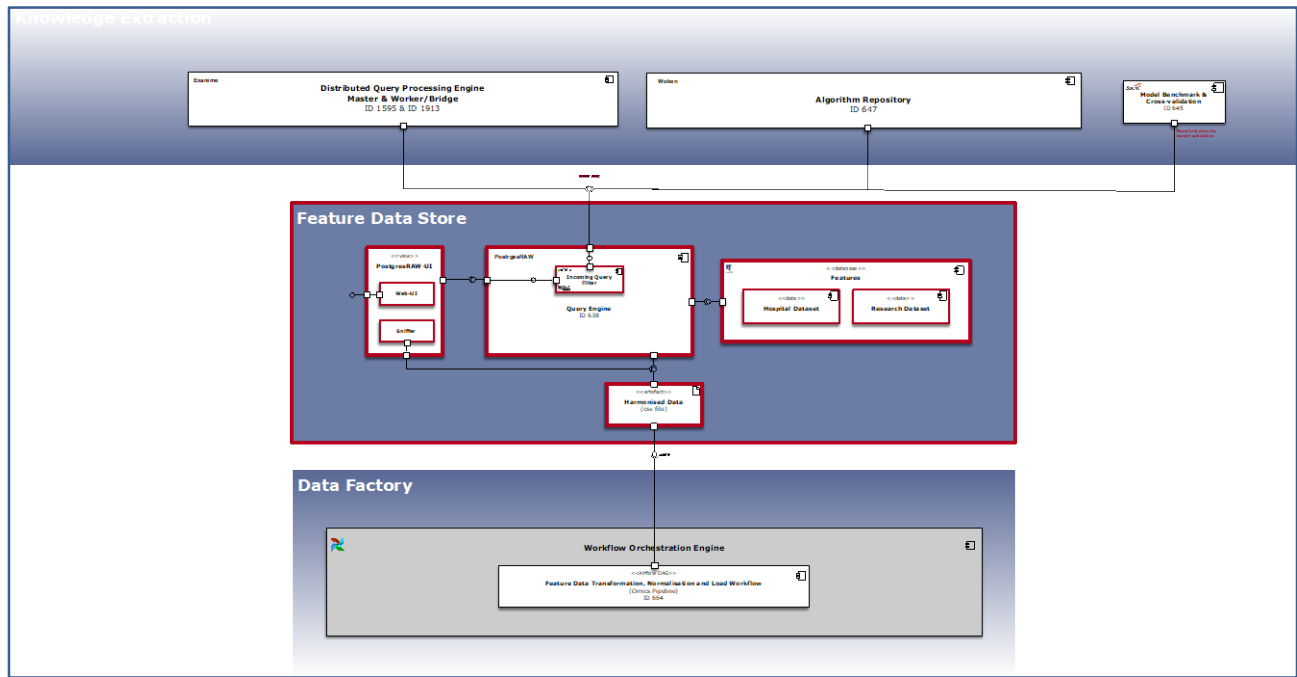


Figure 30: Feature Data Store Sub-system

Query Engine

The main purpose of the Query Engine component is to provide querying of the harmonised patient data stored in CSV files. The MIP Query Engine component is a database management system named PostgresRAW, based on PostgreSQL.

The input to the Query engine is data stored in CSV files. The output of the Query Engine is provided in JSON file format using REST services API or regular PostgreSQL connections.

Features Database

The Flat Hospital Dataset Table of the Features Database is updated with the data queried directly from the files. There it is made available for further querying by the Distributed Query Processing Engine or fetching by machine learning algorithms from the Algorithm Factory, both in the Knowledge Extraction sub-system.

The querying and fetching of data from the Feature Database is performed locally. For the privacy reasons, de-identified patient data is not allowed to be copied outside the hospital's MIP execution environment. The necessary computation is distributed throughout the hospital environments and only the results are fetched by the federation execution environment, either for visualisation or for further processing.

In addition to the Hospital Dataset flat table, the Features Database contains the Research Dataset flat table populated with the data captured from open research cohort datasets.

PostgresRAW-UI

PostgresRAW-UI automates detection and registration of raw files by providing a file monitor (Sniffer component). The folder containing the files with data that should update the Hospital Dataset table is provided as an argument when starting the database server.

Knowledge Extraction Sub-system

The components of the Knowledge Extraction sub-system are deployed both within the local hospital MIP execution environments and within the central MIP federation execution environment.

This MIP sub-system provides the functions for processing of the harmonised patient data, for local or distributed data mining and local or distributed execution of statistical inference and machine learning algorithms.

The two major complementary components of Knowledge Extraction sub-system are:

- **Algorithm Factory (Woken)** - orchestration of machine learning algorithm execution, including model benchmarking and cross-validation and storing of the trained models and their estimated predictive errors. Does not have out-of-the-box support for database query processing
- **Distributed Query Processing Engine (Exareme)** - query processing orchestration engine optimised for execution of distributed database queries extended with user-defined functions. Does not have out-of-the-box support for estimating trained machine learning model predictive errors

Algorithm Factory

Algorithm Orchestrator (Woken)

This component is a workflow orchestration platform, which runs statistical, data mining and machine learning algorithms encapsulated in Docker containers. Algorithms and their runtime are fetched from the Algorithm Repository, a Docker registry containing approved and compatible algorithms and their runtimes and libraries.

This component runs on top of the runtime environment containing Mesos and Chronos to control and execute the Docker containers over a cluster.

This component provides a web interface for on-demand execution of algorithms. It fetches the algorithms from the Algorithm Repository, monitors the execution of the algorithms also from the other execution environments in the cluster, collects the results formatted as a PFA document and returns a response to the web front end.

The Algorithm Orchestrator tracks data provenance information, runs model benchmarking and cross-validation of the models learned by the machine learning algorithms, using random K-Fold Sampling methods (Model Benchmark & Cross-validation), and stores PFA models in the Predictive Disease Model Repository.

Algorithm Repository

This component is a repository of Docker images that can be used by the Algorithm Orchestrator. It provides a workflow that allows contributors to provide new algorithms in a secured manner.

Algorithms, written in their native language (Python, MATLAB, R, Java, etc.), are encapsulated in a Docker container that provides them with the libraries and runtime environment necessary to execute this function. Currently, the MIP SGA1 platform supports Python-, Java- and R-based algorithms that are packaged in three Docker containers, respectively. The environment variables provided to the Docker container are used as algorithm parameters.

Algorithm Docker containers are autonomous:

- Connecting to the Features Database in the Features Data Store sub-system to retrieve feature data
- Processing data, taking into account Docker container environment variables
- Storing results into the Predictive Disease Model Repository

The Algorithm Registry database, implemented using PostgreSQL database management system, is used to keep track of results created by the execution of an algorithm.

New algorithms can be easily integrated with the others by packaging them in the relevant Docker container. The supported algorithm results format is PFA, described in YAML or JSON configuration file. PFA enables vendor-neutral exchange and execution of complex predictive machine learning models. For visualisations, MIP SGA1 supports different formats, including Highcharts, Vis.js, PNG and SVG.

Machine learning algorithms planned for integrated by the end of SGA1 phase are:

- k-nearest neighbours (Java)
- Naïve Bayes (Java)
- iSOUP-Tree-MTR (Java)
- Statistical Inference (Python)
- ANOVA (Python)
- Multivariate Linear Model (Python)
- t-SNE (Python)
- C3 (4)
- Heatmaply (R)
- ggparci (R)

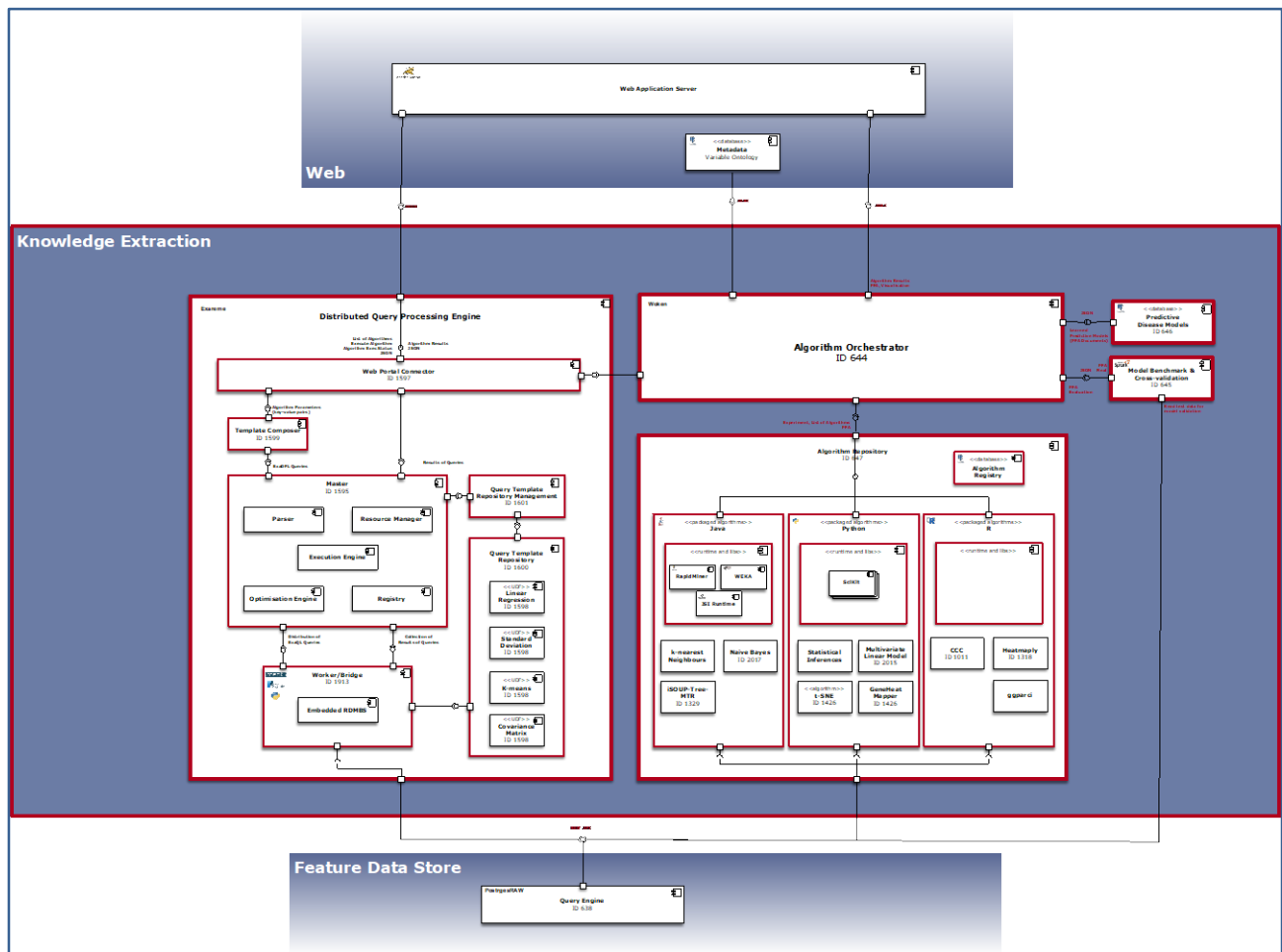


Figure 31: Knowledge Extraction Sub-system

Model Benchmark & Cross-validation

The Model benchmark and Cross-validation component is used to measure machine-learning models' accuracy. The results can guide the user to select the best-performing algorithm and fine-tune its parameters as well as to understand how well the model performs before it's used in production.

A model trained on training data needs to be validated. Its quality is measured by estimating its predictive error. Several techniques for assessing predictive errors exist, cross-validation being the most frequently used one. The predictive error is calculated by using the two disjoint datasets - training data set, to train the model, and test dataset to calculate the predictive error rate. The calculation of model predictive error rates is called validation.

Data used for both training and test datasets are stored in the Features Database, in the Features Data Store sub-system. The Model Benchmark & Cross-validation component performs data split using K-Fold cross-validation. This method of data sampling divides the complete dataset into K disjoint parts of roughly the same size. K different models are trained on K-1 parts each, while being tested on the remaining one part of the data. That is done on all K parts exactly once to ensure that every data row is used equally often for training and exactly once for testing. Resulting K test errors are then averaged to get the final error estimate of the model, which was built on the complete dataset.

The Algorithm Orchestrator stores the trained machine learning models and the results of cross-validations in the Predictive Disease Model Repository.

Figure 32 depicts the interaction between the Algorithm Factory components for a typical use case of running an experiment, ordered from the MIP Web sub-system.

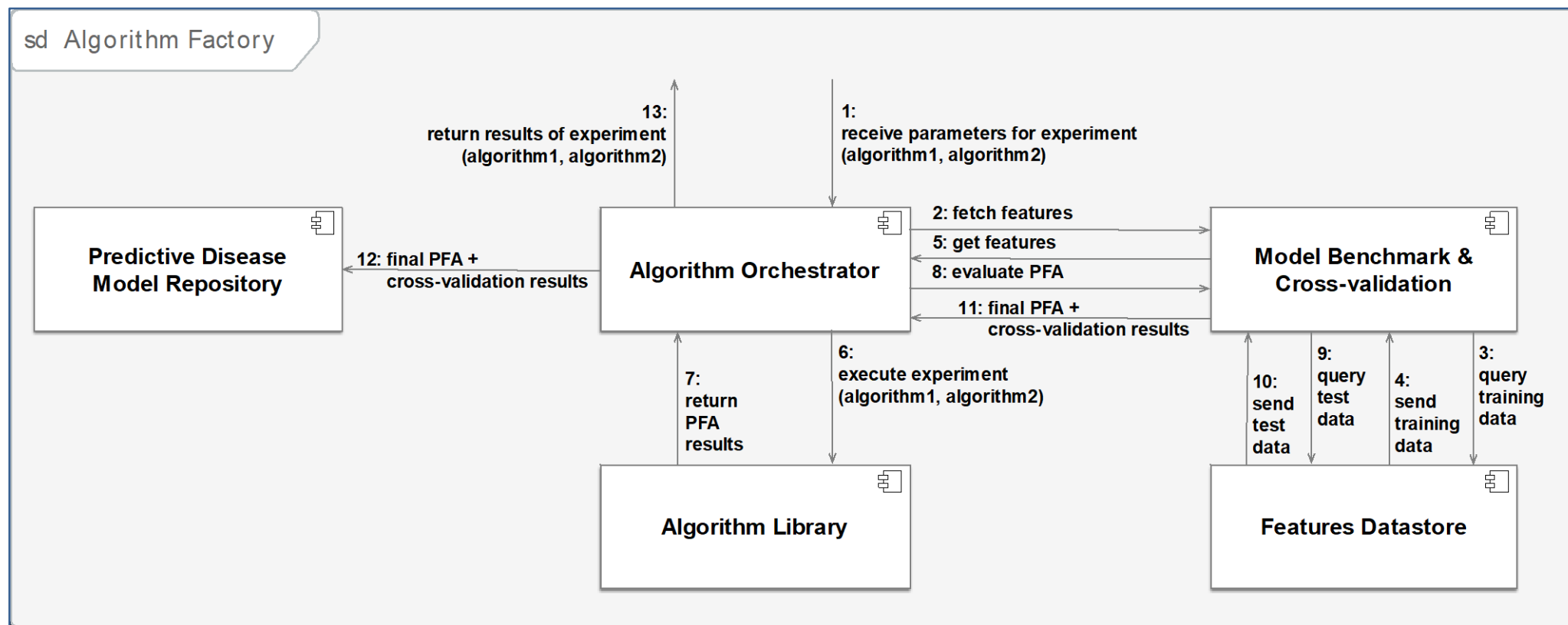


Figure 32: Algorithm Factory Communication Diagram

Predictive Disease Model Repository

This component serves as a permanent storage and search service for trained PFA models and their predictive error estimates.

Distributed Query Engine - Exareme

The Distributed Query Processing Engine plays a role in the Knowledge Extraction sub-system of the MIP platform. Master components deployed in the central federation node communicate with workers deployed in each of the hospitals, on one side, and with the Web sub-system components, on the other side. The Distributed Query Processing Engine does not allow direct communication between workers in different hospitals. Worker components, deployed in the hospitals, fetch the data from the local Feature Tables in the Features Data Store sub-system using the REST API and transfer the data to the master component for aggregation.

Systems Overview

The Distributed Query Engine, based on the open source project Exareme, is used as a traditional database system for: (1) data definition (creating, modifying, and removing database objects such as tables, indexes, and users), (2) data manipulation (data querying), and (3) external data import (from files or other databases). It is a distributed relational database management system extended with the support for complex field types - JSON, CSV and TSV.

The Distributed Query Engine uses a proprietary data manipulation language ExaDFL for specifying and orchestration of data processing. The Distributed Query Engine organises data processing in workflows designed as direct acyclic graphs (DAGs) - relational query operators are graph vertices, and the data flows between the operators are graph edges. ExaDFL is based on SQL extended with user-defined functions (UDFs) and data parallelism primitives. User-defined functions are used for specifying local data processing workflows and performing complex calculations on distributed data set partitions. ExaDFL primitives that support parallelism are declarative statements supporting parallel execution of partial queries on partitioned data sets.

The Distributed Query Engine translates ExaDFL queries to its internal declarative data manipulation language ExaQL, based on SQL-92 with extensions, for execution of query operators and user-defined functions on the distributed data set partitions.

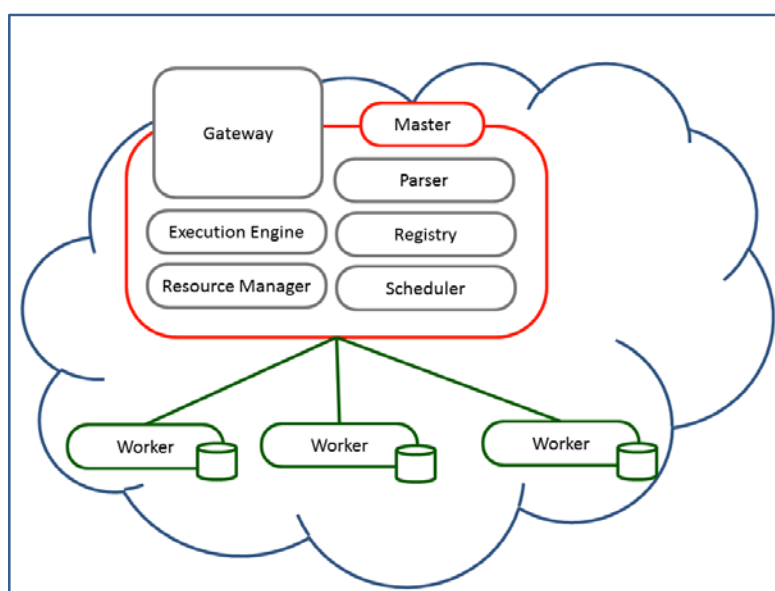


Figure 33: Distributed Query Engine Architecture Overview

The three main components of the Distributed Query Engine are:

- 1) **Worker** - an embedded SQLite relational database management system with Another Python SQLite Wrapper (APSW) - a Python API for SQLite running on the local hospital execution environments. It fetches local Feature Data Set sub-system's Features Table data set partitions needed for the execution of query operators and user defined functions and caches those data

set partitions to its local data storage for subsequent querying using ExaQL primitives. Worker is also a data processing system with functions for file import/export, keyword analysis, data mining tasks, fast indexing, pivoting, statistics and data processing workflow execution

- 2) **Master** - main entry point to the distributed query engine, running on the central federation execution environment, responsible for the coordination of the execution of other components. It aggregates query results transmitted by the Worker components distributed throughout the local hospital execution environments. Master consists of the following components:
 - **Registry** - stores all information about the data and allocated resources, i.e. allocated data set partitions and their execution environments
 - **Resource Manager** - allocates and de-allocates data processing resources on demand of the Execution Engine
 - **Execution Engine** - requests allocation of resources from the Resource Manager, resolves the dependencies between the query operators to create a schedule of their execution in direct acyclic graph-oriented workflows, monitors the execution of the workflows and handles failures
 - **Optimizer/Scheduler** - transforms ExaDFL queries into the distributed ExaQL statements and creates query execution plan by assigning operators to their respective workers
 - **Gateway (Web Portal Connector)** - provides web for the communication between the Master component and the Web sub-system components
- 3) **Query Template Repository** - version-controlled source code store for the query templates in the form of User Defined Functions (UDFs). It is used both by the Worker and the Master components

Supported Data Processing Workflow Types

The source code of each algorithm is split into a set of local queries executed in parallel by the Worker components on the local data sets and one or more global processing executed by the Master component on the central federation node. The source code of each local and global data processing is written in a form of a workflow of SQL queries extended with user-defined functions. The source code is stored in .sql files in the Query Template Repository component. Supported data processing workflow types are:

- 1) **Local-global workflow** - local data processing executed in local execution environments, the aggregated results merged on the master node followed by additional data processing steps, if needed
- 2) **Multistep local-global workflows** - data processing workflow of predefined number of local-global data processing
- 3) **Iterative local-global workflows** - execution of the local-global data processing until a convergence criterion is reached (under development)

Algorithm Execution Steps

- 1) The Gateway component receives a user request for running an algorithm with submitted parameter values
- 2) The Template Composer fetches the stored local and global query templates needed for executing the selected algorithm from the Query Template Repository and creates an Algorithm template using ExaDFL primitives. Each algorithm template has an associated JSON properties file that contains meta-information such as the algorithm's name, description, type, and parameters. Based on the type of the algorithm the type of the data processing workflow is determined
- 3) The Algorithm template that describes parameterized distributed workflows are forwarded to the Optimizer for generating the execution plan

- 4) The execution plan is forwarded to the Scheduler for determining partial algorithm execution plans, which are dispatched to Worker components running in local hospital execution environments
- 5) Each of the Workers executes the local data processing, then sends a confirmation of the successful execution to the Execution Engine in the central federation execution environment
- 6) Upon receiving success confirmations from all the Workers, the Scheduler determines global data processing plan and sends it to the Execution Engine
- 7) The Execution Engine then merges the aggregated results of all the workers, executes the global data processing plan and confirms its successful execution back to the Scheduler
- 8) The Scheduler checks if the complete local-global data processing plan has been completed
- 9) In case of the successful completion of the plan, it forwards the aggregated results to the user. In case the plan has not been completed, the Scheduler determines the next set of local data processing plans and the whole process of local-global plan execution is repeated until the successful completion of the algorithm

Overview of The Supported Features

The Distributed Query Processing Engine provides the following features:

- 1) List of the available algorithms
- 2) Requesting the execution of any of the available algorithms, and submission of relevant parameters
- 3) Execution status of the executing algorithms
- 4) Execution results of completed algorithms

The Distributed Query Processing Engine does not support automatic machine learning model validation. It does not provide out-of-the box predictive error estimation nor is there a component for recording the estimated accuracy of the trained machine learning models. The Algorithm Factory component can be used alongside the Distributed Query Processing Engine for trained model benchmarking and validation.

The MIP Distributed Query Processing Engine supports the following algorithms implemented as UDFs:

- K-Means
- Linear Regression

Web Sub-system

This section provides a brief overview of the functionality of the MIP Web sub-system. A detailed description of the front end functionality is provided in the MIP Web UI - User Guidelines, V2.0 Public Release:

(https://hbpmedical.github.io/documentation/HBP_SP8_UserGuide_latest.pdf).

The Web Sub-system provides a web portal and web applications for the end-users of the Platform. Users can explore only aggregated statistical data and perform data analysis using machine-learning methods provided by the Knowledge Extraction sub-system components. Web sub-system components have no direct access to the Feature Data Store sub-system where the individual patient de-identified health-related feature data are stored. For privacy reasons, the MIP allows exploration only of statistical data.



- **Collaboration Space** - the landing page of the Medical Informatics Platform, displaying a summary of statistics (users, available variables, written articles), and the latest three shared models and articles. It also provides a link to the Article Builder web application
- **Data Exploration** - a statistical exploration of patient feature data (i.e. variables). It is possible to explore only statistically aggregated data, not information from an individual patient. This web application provides on-the-fly generation of the descriptive statistics and contains a caching mechanism to handle any future data import in an automated way. It uses information stored in a Metadata database to display additional information about the displayed statistical data, such as data acquisition methodology, units, variable type (nominal or continuous), etc. This web application provides the functionality to search, select and classify data elements as variables, co-variables and filters for configuration of the statistical or machine learning models.
- **Model Builder** - configuration/design of statistical or predictive machine learning models. It also provides visualisation for searching data element types, select and classify data elements as variables, co-variables (nominal and continuous) and filters. Once the model is designed, a design matrix is populated with the selected data. The Model Builder provides a visual representation of the design matrix and the selected data for inspection before running a statistical, feature extraction or machine learning algorithms. It also provides an option to save the designed models
- **Experiment Builder & Disease Models** - a selection of a statistical, feature extraction or machine learning method, the configuration of the method parameters and the parameters for the trained model validation for supervised machine learning, as well as launching of the machine learning experiment. This application displays experiment validation results as bar charts and confusion matrices
- **Article Builder** - writing articles using the results of the executed experiments

- **Third-party Applications and Viewers** - a portal for accessing third-party web applications for data exploration and visualisation

The Web-Sub-system allows access to its back end services and the Knowledge Extraction sub-system's Algorithm Factory through Jupyter notebooks running in the Human Brain Project's Collaboratory environment.

The Web Sub-system's Authentication and Authorisation component is integrated with the HBP Collaboratory's OpenID authentication service. The User Management component maintains an access control list and logging of user activities on the Data Exploration, Model Builder and Experiment Builder web applications.

Google Analytics Dashboard is set up for monitoring the usage of the Platform web services: tracking users and their behaviour and keeping an audit log with all user activities to detect potential Platform abuse and take preventive measures.

Deployment Architecture Overview

This section contains a brief overview of the key MIP deployment architecture concepts, relevant to understand the context of the MIP Software Installation use case specification.

A detailed description of the deployment architecture and its components is out of the scope of this document. It will be provided in the Deployment Specification document, including the following:

- Deployment model (execution environments, deployment artefacts and runtime components)
- Use case specifications (Software Installation, Data Preparation and Data Harmonisation)
- Deployment project configuration guide

Microservice Architecture

Each of the SP8 teams was focusing on delivering software components in their specific area of expertise using different technology stacks - Java, Python, R, MATLAB, Scala. As opposed to a monolithic application architecture, microservice architecture allowed the teams to work independently in their specific functional areas: Web, distributed query processing, algorithm orchestration, data-mining, statistics and machine learning algorithms, integration and verification, local data store mirror, brain scan processing and ETL, data transformation and data harmonisation.

Another significant advantage of the microservice architecture is a possibility to adopt new technology and add new features incrementally. For example, encapsulated and loosely coupled permanent data storage can be replaced with a distributed big data-ready technology packaged and deployed in Docker containers, having no impact on the surrounding data processing, ETL and analytic software components.

Operating-system level virtualisation using Docker containers on top of Linux operating system has been chosen to build and deploy microservices and run corresponding processes. Software modules are packaged as Docker images and then integrated into a production version of the distributed MIP application using continuous integration software.

Docker Images As Microservices

MIP software developed by the HBP partners and 3rd party software components is packaged as microservices implemented as Docker images: independently deployable, small, loosely coupled services, each one running a unique process and communicating through a well-defined lightweight mechanism. Updating a component does not require redeployments of the entire application. MIP microservice deployment architecture supports a continuous integration and continuous deployment approach.

Docker image	Organisation	License	Build status	Image version	Image layers
docker hbpmp/flyway					
docker lren/xnat	CHUV LREN	license MIT	⊗ FAILED	version 1.6.5	993.6MB 34 layers
docker lren/labkey	CHUV LREN	license Apache-2.0		version latest	447.3MB 35 layers
docker hbpmp/woken	CHUV LREN	license Apache-2.0		version 2.1.4	144.9MB 25 layers
docker hbpmp/portal-backend	CHUV LREN	license AGPL-3.0	✓ PASSED	version 2.5.4	121MB 21 layers
docker hbpmp/portal-frontend	CHUV LREN	license AGPL-3.0		version 2.5.2	32.6MB 25 layers
docker hbpmp/mipmap	EPFL DIAS	license MIT		version latest	65.9MB 21 layers
docker hbpmp/webmipmap	EPFL DIAS	license MIT		version latest	87.2MB 30 layers
docker hbpmp/postgresraw	EPFL DIAS	license MIT		version v1.0	13MB 20 layers
docker hbpmp/postgresraw-ui	EPFL DIAS	license MIT		version v1.2	36.9MB 12 layers
docker hbpmp/exaremelocal	UOA madgik	license MIT		version latest	852.9MB 42 layers

Figure 35: List of MIP Docker Images

Automated Installation and Configuration of MIP Software

Platform for fast deployment of services on bare metal or preconfigured virtual machines supporting clustering, security and monitoring is based on Cisco's Mantl rapid deployment project. The MIP is deployed on Mesos stack with added support for automated deployment/upgrade of services managed by Mesos Marathon and hardened security of the Ubuntu operating system. The services are built using Ansible scripts, unifying operation system configuration, middleware and application software deployment.

The MIP Hospital Deployment use cases planned for demonstration are specified in the next Chapter (Medical Informatics Platform Software Installation Use Case Specification). Installation of the MIP in each new hospital is considered as a new git project, created as a clone of the generic Microservice Infrastructure project with configuration parameters updates tailored to a specific new hospital execution environment. Generic automatic MIP installation and configuration is stored and documented here:

<https://github.com/HBPMedical/mip-microservices-infrastructure>

Medical Informatics Platform Software Installation Use Case Specification

The MIP microservice deployment architecture enables agile continuous integration and continuous deployment of components developed or modified by different European-wide teams. This architecture enables efficient future upgrades of the Platform with new technologies and new features needed to support evolved clinical needs. Automation of configuration and installation of the MIP software minimises IT efforts to keep the maximum focus on the scientific and clinical aspects of the projects.

Table 5 - Use Case Specification: Medical Informatics Platform Software Installation

Actors		HIT: Hospital IT Engineer MIT: MIP Deployment Engineer
Use Case Objective		Installation of the Medical Informatics Platform hardware and software in a hospital data centre
Pre-conditions		1) Formally approved investment in infrastructure, software, time and material 2) Signed Medical Informatics Platform Deployment and Evaluation Agreement 3) Infrastructure, software, time and material procured by hospital
Main Flow of Events		
Event ID	Actor ID	Event Description
E01	HIT	Prepare data centre for installing and configuring new MIP servers, storage and network
E02	HIT	Install MIP all-in-one server or separate servers (typical hospital configuration is provided below): a. Data capture and de-identification server CPU: 2-core x64; RAM: 2 GB; Storage: 50 GB; Security level: Highly secure clinical network b. Pre-processing server CPU: 12-core x64; RAM: 32 GB; Storage: 16 TB; Security Level: Secure research network c. Knowledge extraction and web server CPU: 8-core x64; RAM: 32 GB; Storage: 2 TB; Security Level: Secure research network or DMZ
E03	HIT	Install operating system on MIP servers • recommendation: Ubuntu 16.04 LTS / RHEL 7.2+ / CentOS 7.2+
E04	HIT	Provide sudo access rights for each MIP server to MIP deployment engineer
E05	HIT	Configure IPv4/IPv6 settings for each MIP server
E06	HIT	Configure SSH VPN tunnelling for remote connection with the MIP deployment team environment a. Install and run OpenSSH server on each MIP server b. Configure TCP port 22 for ingress SSH traffic on each MIP server c. Open port 22 for ingress traffic through firewall(s) between each MIP server and the Internet
E07	HIT	Configure TCP port 443 for egress HTTPS traffic on MIP servers and open port in firewall(s) for: a. Software package repositories (Ubuntu, Mesosphere, PyPI) b. Source code repositories (GitHub, Bitbucket, Launchpad, CHUV git) c. Docker registries (Docker Hub, CHUV private Docker registry)

E08	MIT	<p>Install and configure MIP software automatically, using Ansible:</p> <ol style="list-style-type: none"> Clone the generic Microservice Infrastructure project to create a git project for storing the new MIP environment configuration Prepare a configuration for automatic installation: <ul style="list-style-type: none"> Install Python2 on the MIP servers - Ansible requires Python2 to run Install MATLAB 2016b - required by SPM software for neuromorphometric processing Server names and TCP/IP configuration Store the configuration in git, encrypt the passwords and confidential information Run a single Ansible script for the new MIP installation and configuration to: <ul style="list-style-type: none"> Install middleware - libraries, runtimes, DBMSs and open source software Deploy Docker images with software developed by MIP teams
E09	MIT	Confirm that all the processes are up and running from Marathon administrator's dashboard
E10	MIT	<p>Backup the installation and configuration scripts on external server:</p> <ul style="list-style-type: none"> MIP team uses a private storage space on Bitbucket.org Using the private repository, it is possible to safely and securely backup work, share it with other members of MIP for code review and receive upgrades of the platform
E11	MIT	Configure MIP backup for each MIP server in standard data centre backup environment
Special Requirements		Open relevant ports on firewalls, subject to the specific hospital IT security configuration
Post-conditions		<ol style="list-style-type: none"> MIP software is installed on all servers with all processes up and running MIP platform is ready for data processing, storing and analysis
Scientific Added-value		<ol style="list-style-type: none"> The hospital data centre has a centralised platform for processing, storing and analysing de-identified and harmonised neuroimaging, neuropsychological, biological and demographic data of its patient population Efficient, configurable and automated end-to-end software installation, unifying operation system configuration, middleware installation and microservice building minimises IT efforts to keep the focus on using the MIP platform for the scientific and clinical activities

Appendix III: Components: Old Name - New Name Mapping

Table 6 - MIP Data Components

New Name	Old Name	PLA ID	Task No	Team
Dataset Descriptions	Brain imaging-Genetic-Clinical (EHR)	455	T8.2.1	CHUV
Data Mapping and Transformation Specification	Common Variables & Metadata	587	T8.2.1	CHUV
Nifti test data files	Nifti test data files	1734	T8.1.1	EPFL
BIDS test data files	BIDS test data files	1733	T8.1.1	EPFL
ADNI Test Data	Features	2716	T8.5.2	CHUV

Table 7 - MIP Software Components

New Name	Old Name	PLA ID	Task No	Team
Data De-Identifier	Data Anonymizer	2879	T8.1.1	EPFL
Data Downloader	Data downloader	2865	T8.4.5	ICL
Data Uploader	Data uploader	2862	T8.4.5	ICL
Online Data Integration Module	Online Data Integration Module	1580	T8.1.4	AUEB
Neuromorphometric Processing	Omics Pipeline for feature engineering for Airflow	671	T8.5.2	CHUV
Omics Pipeline for feature engineering for Cbrain	Omics Pipeline for feature engineering for CBrain	670	T8.5.2	CHUV
Common Data Elements	Data Storage	669	T8.5.2	CHUV
Data Capture	<i>New Component</i>	2926	T8.5.2	CHUV
Data Catalogue	<i>New Component</i>	2927	T8.5.2	CHUV
WebMIPMap	Community Schema Curation	1581	T8.1.4	AUEB
BIDS Function Library	BIDS Data Library in Query Engine	1754	T8.1.1	EPFL
NIFTI Function Library	Nifti Library in Query Engine	1753	T8.1.1	EPFL

New Name	Old Name	PLA ID	Task No	Team
Plug-in for BIDS Data	Query plug-in for BIDS Data	1752	T8.1.1	EPFL
Imaging Data Plug-in	Query plug-in for Medical Imaging Data	1751	T8.1.1	EPFL
Extended Array Query Support	Extended Array Query Support	1750	T8.1.1	EPFL
Query Engine	Query Engine	638	T8.1.1	EPFL
Distributed Query Engine Over HPC	Distributed Query Engine Over HPC	1755	T8.1.2	EPFL
Ontology Based Data Access	Ontology Based Data Access	1579	T8.1.4	AUEB
Access Right Module	Access Right Module	1578	T8.1.4	AUEB
Web portal connector	Web portal connector component	1597	T8.1.5	UoA
Distributed Query Processing Engine Worker/Bridge	Worker / Bridge Component	1913	T8.1.5	UoA
Distributed Query Processing Engine Master	Master component	1595	T8.1.5	UoA
Template composer	Template composer component	1599	T8.1.6	UoA
UDFs component	UDFs component	1598	T8.1.6	UoA
Management	Management component of query template repository	1601	T8.1.7	UoA
Query template repository	Query template repository	1600	T8.1.7	UoA
Algorithm Repository	Algorithm repository	647	T8.5.2	CHUV
Predictive Disease Models	PFA model store	646	T8.5.2	CHUV
Model Benchmark and cross-validation	Cross-validation module	645	T8.5.2	CHUV
Algorithm Orchestrator	Woken	644	T8.5.2	CHUV
iSOUP Distributed Rule-based Methods	Disease signature: Distributed rule-based methods	1329	T8.3.5	JSI
Naive Bayes	Bayesian methods and deep learning tools for identification of homogeneous disease using the Biological signatures	2017	T8.4.2	CHUV
Multivariate linear models	Mathematical methods for predicting multi-level features of diseases	2015	T8.4.2	CHUV

New Name	Old Name	PLA ID	Task No	Team
VBO	Quantification of tissue properties from qMRI	1287	T8.4.3	CHUV
Longitudinal disease progression model	Tools to build disease progression models from scalar measurement	2416	T8.3.12	ICM
3C	3-C (Categorize, Cluster & Classify)	1011	T8.3.1	TAU
Web Application > Knowledge Base > Research Dataset List	Database containing information of MIP solutions adopted at hospital level	2286	T8.2.2	UNIGE
Web Application > heatmaply	Methods for high-dimensional data with possible missing values	1318	T8.3.2	TAU
Web Application > Brain insight > GeneHeatMapper	GeneHeatMapper	1426	T8.3.10	LUMC
Web Application > Web Exploration & Analytics	Web-based Modeling and Visualisation	2395	T8.5.1	CHUV
Web Application > Portal DB (articles, experiments, models)	Research Object wrapper	633	T8.2.3	CHUV
Web Application > Knowledge Base	Knowledge Base Content Development, User Guide	1541	T8.2.3	CHUV
Remote Starting of Services	Remote Starting of Services	1759	T8.1.3	EPFL
Encrypted Overlay Network	Encrypted Overlay Network	1760	T8.1.3	EPFL
MatLab	Workflow tools	665	T8.5.2	CHUV
MIP microservice infrastructure	MIP microservice infrastructure	102	T8.5.2	CHUV
Algorithm Factory, Data Factory and Web Analytics Integration, Collaboratory integration	Algorithm Factory, Data Factory and Web Analytics Integration, Collaboratory integration	1508	T8.5.2	CHUV
Airflow DAGs	Workflow engine	664	T8.5.2	CHUV
Imaging Plugins	<i>New Component</i>	2929	T8.5.2	CHUV
Reorganization Pipeline	<i>New Component</i>	2930	T8.5.2	CHUV
I2B2 Import	<i>New Component</i>	2931	T8.5.2	CHUV
Ansible Airflow	<i>New Component</i>	2932	T8.5.2	CHUV
Data Tracking	<i>New Component</i>	2928	T8.5.2	CHUV
Platform Usage Monitoring	Platform Usage Monitoring	685	T8.5.1	CHUV



New Name	Old Name	PLA ID	Task No	Team
User Management	User Management service	686	T8.5.1	CHUV
Security, Load balancing, Clustering and Recovery Services	Security, Load balancing, Clustering and Recovery Services	684	T8.5.2	CHUV

Appendix IV: Answers to Experts Review Report

This appendix contains a structured list of the answers to the questions and comments documented in the Experts Review Report received at the end of January 2018.

No.	Expert's Comment	Page, Paragraph	SP8 Answer
1	The MIP is presented on page 7 as "a complete solution for descriptive and predictive disease diagnosis because it provide a complete data analytics information, including the assessment of the accuracy of predictive errors". The sounds still far-fetched and perhaps not realistic at this time. "is designed to become" or "aims to result in" or likewise would be probably better describing the actual and near future situation.	Page 3, paragraph 1	The last paragraph of Sub-Chapter 4.1 on page 7 of this document is rephrased as follows: "The Medical Informatics Platform is, therefore, designed with the objective to become a complete solution for descriptive and predictive disease diagnosis because it provides a complete data analytics information, including the assessment of the accuracy of predictive errors ^{[27][28]} ."
2	... the validation approach appears still rather abstract and hard to grasp in places. At this stage of the project one would expect examples that showcase the process as well as measures to assess the quality and usefulness of the MIP. For instance, regarding AD signatures, how much data are there to drive the validation? What is actually evaluated? Given the currently available data: what is the classification accuracy? How is the test-retest reliability? Which features are available and can actually be used; how sparsely sampled are data across hospitals? etc.	Page 3, paragraph 3	<p>As discussed in the first paragraph of Chapter 1 on page 5 of this document, "the main objective of the document is to provide a system validation plan for collecting the objective evidence that the Medical Informatics Platform fulfils its strategic and operational objectives and the needs of the clinicians and researchers." Furthermore, paragraphs 4 and 5 of Chapter 6, on page 11 of this document, "The objective of the MIP system validation is to prove satisfaction of the desired operational capabilities by showing through execution of operational scenarios that user needs are met. Platform's operational capabilities and user needs are formally defined using use case modelling approach. MIP operational scenarios selected for MIP system validation include the execution of one or more MIP use cases..."</p> <p>In Chapter 7.2 of this document, in each of the high-level specifications of clinical system validation scenarios, the measures for assessing the usefulness of the MIP are provided as a validation criterion for each of the system validation action ("Validation Criteria" column).</p> <p>In summary, the measures to assess the quality and usefulness of the MIP, in the scope of the planned system validation are qualitative, rather than quantitative. No quantitative measurements of the usefulness of the MIP are projected. The scope of the system validation is qualitative, as stated in ISO/IEC/IEEE 15288 - "[6.4.11.1] The purpose of the Validation process is to provide objective</p>

No.	Expert's Comment	Page, Paragraph	SP8 Answer
			<p>evidence that the system, when in use, fulfils its business or mission objectives and stakeholder requirements, achieving its intended use in its intended operational environment."</p> <p>However, the hospitals selected for execution of the system validation are chosen for the expected value of their datasets.</p> <p>The detailed information about the cross-hospital dataset profiles, the measures, and the results have been planned and are documented in Deliverable D8.6.3. Some details are provided in the discussion of the key results in Chapter 2 of D8.6.3. The complete details, including the analysis of the system validation results, are provided in Chapter 3 of D8.6.3.</p>
3	It is not clearly defined on what data and on how much data clinical validation will be performed. What are the validation criteria?	Page 3, paragraph 4, bullet point 1	<p>Please see the discussion on the previous point.</p> <p>The qualitative, rather than quantitative validation criteria is formally provided in the high-level specification of the system validation scenario actions</p> <p>Also, find in Appendix V: Hospital Selection Criteria and Expected Dataset Size, a high level overview of the system validation plan and hospital dataset size, known at the time of creating this plan during December 2017.</p> <p>At the end, the detailed information and analysis of the data and system validation results for each of the three hospitals is provided in Deliverable D8.6.3.</p>
4	Clinical validation will be performed using only a fraction of the potentially available algorithms (linear regression and ANOVA). Validation of more sophisticated methods remains open	Page 3, paragraph 4, bullet point 2	<p>The following statistical methods and machine learning algorithms have been used - ANOVA, linear regression, Naive Bayes, k-means and knn.</p> <p>More precise details are documented in Deliverable D8.6.3</p>
5	It is also unclear if the MIP-local testing will be done for each participating clinic separately or if these tests will be performed only for one hospital	Page 3, paragraph 4, bullet point 3	<p>As mentioned in Table 8 - Selection Criteria for Hospitals and Clinical Use Scenario Demonstrations in the additional Appendix V: Hospital Selection Criteria and Expected Dataset Size, the MIP system validation shall be performed separately on the datasets in each of the three participating hospitals as well as cross-hospital on multiple datasets</p>
6	The described clinical validation criteria are mostly not measurable, e.g. it remains unclear what threshold will be used for the provided quantitative	Page 3, paragraph 4, bullet point 4	<p>No thresholds have been set up. The validation criteria are qualitative, not quantitative. Please see the answers to Points 2 and 3 above for a more elaborate discussion</p>

No.	Expert's Comment	Page, Paragraph	SP8 Answer
	measures to consider them as successful		
7	A clear specification of the configuration of the MIP Local instances for each testing site is still needed as well as a description and quantification of the aggregated data on which tests will be performed. Thus, it remains unclear what exactly will be evaluated and what and how much data will be used for testing and validation.	Page 3, paragraph 5	Please see Table 8 in the new Appendix V: Hospital Selection Criteria and Expected Dataset Size for more details. Even more detailed information about the cross-hospital dataset profiles, the measures, and the results have been planned and are documented in Deliverable D8.6.3.
8	<p>Based on a recently published paper (Frissoni et al) three important MIP applications have been identified (section 4.2):</p> <p>Computing, testing and validating biomarkers</p> <p>Improving the classification of dementia subtypes using (1) differential patterns of cortical atrophy associated with cognitive decline or (2) using neuro-pathological examination (This data is not present in MIP?).</p> <p>Although the selected use-cases seem to fit here, scientific novelty behind the choice of the use-cases is unclear. Are they suitable and publishable to showcase and advertise MIP functionality? What publications are planned in this context (as requested as part of the showcase)?</p>	Page 3, paragraph 6	<p>The novelty of the potential scientific value is the possibility to validate based on real clinical datasets whether candidate biological data are indeed biomarkers of particular dementia-type disorders.</p> <p>Furthermore, the scientific utility (value) of the Platform, when used for dementia data analysis is discussed in Deliverable D8.6.3, Chapter 2.5.</p> <p>Last but not least, as stated in Chapter 2.5 of Deliverable D8.6.3: "The primary scientific objective is to understand the causes, mechanisms and progression of these disabling diseases. MIP aims to bridge the gap between fundamental research and real-world clinical data by integrating and statistically comparing clinical datasets with reference research cohort databases (i.e., "gold research standards"). The platform serves clinical and fundamental research worldwide community as a globally accessible distributed information system for dementia, supporting the studies by providing advanced federated analytics of diverse clinical and research datasets. As such, the MIP should be seen as the implementation of target action areas 6 and 7 of the WHO's Global Action Plan on the Public Health Response to Dementia 2017 - 2025."</p>
9	...from a research point of view, one can be rather skeptical of success. This classification (stated in the deliverable) is meant to be mostly imaging-guided and will then additionally be informed by neuropsychological and clinical measures. To our knowledge, the highest accuracy obtained to date by such approaches (and the authors of the mentioned paper, Frissoni et al., are no exception) is in the order of 70%. This is by	Page 4, paragraph 1 (cont. from the previous page)	<p>The advantage of the MIP is the possibility to make available additional patient biomedical, environmental, risk factors and other types of information, currently not available through any analytic solution, that may explain misdiagnosed cases and provide better and more accurate insight into the biological signatures of dementia diseases.</p> <p>This view is further supported in WHO's Global Action Plan on the Public Health Response to Dementia 2017 - 2025 in</p>

No.	Expert's Comment	Page, Paragraph	SP8 Answer
	far not sufficient to guide dementia diagnostics or treatment. It is a rather high risk to pursue this work in order to provide more reliable and perhaps better validated methods, in the hope that researchers (or clinicians) will find in the future additional criteria that can provide more robustness in personalised computer-aided diagnostics (or treatment) for dementia.		which they explicitly specified development of information systems for dementia as action point 6. Please, see the discussion about the clinical utility of the MIP as a key SP8 result in Chapter 2.5 of Deliverable D8.6.3,.
10	It is not clear which task is for <i>verification</i> i.e. acceptance and suitability with external users (<i>i.e. fulfils its user needs and expectations in its intended environment</i> - page 16), versus which task is for <i>validation</i> i.e. that <i>the system fulfils its mission objectives</i> (page 16) and complies with regulations, requirements and specifications (and <i>users can mechanically perform the task</i>). These appear to be confused under point 7 on page 20 for instance.	Page 4, paragraph 2, bullet point 1	The scope of the MIP system validation process is fully compliant with the standard definition of the process, as stated in ISO/IEC/IEEE 15288 - "[6.4.11.1] The purpose of the Validation process is to provide objective evidence that the system, when in use, fulfils its business or mission objectives and stakeholder requirements, achieving its intended use in its intended operational environment." Also, please do refer to the answers to previous points that addressed the same concern. It's perhaps worth noting here, that "system validation" is a standardised term (ISO/IEC/IEEE 15288) derived from the older term "users acceptance test"
11	It remains unclear if validation will be applied only to a single demo system or to all hospitals with a deployed MIP.	Page 4, paragraph 2, bullet point 2	Please see the answer to Point 5 and Appendix V: Hospital Selection Criteria and Expected Dataset Size. The scope of the system validation is the assessment of the value by users in their environment. Therefore, system validation is applicable to all three hospitals, both individually and across-hospital
12	It remains unclear what the amount of data and kind of features to be extracted per hospital required for a valid test of MIP local deployment and operational qualification is.	Page 4, paragraph 2, bullet point 3	Required features are implicitly specified in the high-level clinical system validation specifications. The amount of expected data is provided in Appendix V: Hospital Selection Criteria and Expected Dataset Size, Table 8. All the details, results and analysis of the system validation clinical scenarios are documented in Deliverable D8.6.3
13	It is not clear how comprehensively each operational scenario is being user tested. One is being asked to trust that the elements in pages 21ff are all that could be done. Has this breakdown of scenarios by use-case been peer-reviewed by others in HBP?	Page 4, paragraph 2, bullet point 4	The specified system validation scenarios, planned for execution by clinicians and researchers in three participating hospitals, have been demonstrated using open research cohort data (ADNI and ESDS) on numerous occasions both in the scope of the MIP Ramp-up phase and during the SGA1. The functionality has been tested locally in

No.	Expert's Comment	Page, Paragraph	SP8 Answer
			the MIP lab. No formal peer-review has been performed on the cross-SP HBP level. Instead, the demonstrations have been presented
14	Similarly looking at the preceding point the other way around - some use-cases get more tested than others over the 5 operational scenarios. Some only get tested once. This does not seem to be a fair assessment overall of the MIP. One would like to see more than one test for each use-case.	Page 4, paragraph 2, bullet point 5	In fact, each system validation scenario executes all specified "atomic" use cases, including the Web Application group, the Data Mining group, the Analytical Validity group, the Clinical Validity group and the Clinical Utility group (see Appendix I: Overview of MIP Use Case Model).
15	Insufficient detail is provided on what data analytics and model building methods will be included in the installations and tests,... This should be provided at a higher granularity and in the context of the clinical use-cases.	Page 4, paragraph 2, bullet point 6	That is correct observation. The reason more details regarding the model building and data analytics have not been provided is the purpose of the document. It is just a strategic-level plan for the system validation. All the details, including the discussion of the results, are provided in Deliverable D8.6.3
16	Validation of the quality of extracted data is currently planned to be performed through inspection and outlier detection by a MIP-deployment engineer. This might be sufficient for a quick check, but allows no-one to ensure high data quality standards. Data quality assessment should involve an expert using clearly defined and objective quality measures tailored to each respective data type.	Page 4, paragraph 2, bullet point 7	This is a correct observation. The specification provided in this document has been further refined and corrected before the execution of the deployment scenarios in each of the three hospitals. Furthermore, the detailed description of the actual process including the quality control part, shall be provided in Chapter 2.4 of Deliverable D8.6.3. The correction of the Validations Action table in Chapter 7.1 on pages 19-21 is done in accordance with this comment.
17	Data accuracy measurement (p14) – it is good to do validation MIP vs cohort data to test „wild“ datasets against datasets collected under controlled conditions, but also all computational methods should be tested and compared inside MIP vs outside MIP to ensure their correct computation	Pages 4 & 5, paragraph 2, bullet point 8	The "standard" method of data exploration and validation is indeed comparison of the distribution of characteristic variable values against the "gold standard" open research cohort datasets, such as ADNI. Computational methods are based on standard Python, R and Java libraries exactly to ensure that the implementations are comprehensively tested. The exceptions are novel algorithms developed in the scope of the SP8 project by different partners. These algorithms have been integrated and thoroughly tested by in collaboration with authors.
18	Deployment of the MIP-Federate components necessary to address point 4) of the reviewers request above (<i>To showcase an understandable interface for MIP-Local and MIP-Federate</i>) are not apparently covered.	Page 5, paragraph 2, bullet point 9	The scope of the planned MIP system validation is end-users assessment of the value of the platform for their need. In that context, the users shall assess the federated cross-centre, multiple-dataset analytics by performing planned and specified "federated" clinical studies (UC3 and UC4). The deployment scenario

No.	Expert's Comment	Page, Paragraph	SP8 Answer
			in the scope of the planned systems validation covers, indeed, only deployment of the system in a "private" hospital execution environment. The deployment of MIP "community" execution environment is executed and operated by internal SP8 team(s). Hence, it is out of the scope of the end-users acceptance test. System verification (ISO/IEC/IEEE 15288 [6.4.9.1]) has, however, been performed in this case.
19	It is not clear to what degree the verification and validation staffs involved are independent of the SP8 MIP project team to ensure complete objectivity.	Page 5, paragraph 2, bullet point 10	System validators are end-users -from the three participating hospitals. They are independent both from the SP8 MIP and all other HBP projects
20	Although the individual validation IDs are allocated to a validation role or roles (Table 2 page 17), who is the person who will recount the outcome of that validation task within the Use-Case, there is no definition in advance of who makes the final decision for any task or for any one use-case across multiple different users (page 22ff). Furthermore, it is not clear who decides whether the 5 operational scenarios have 'passed' or not.	Page 5, paragraph 2, bullet point 11	The standard professional systems engineering approach shall be applied - the senior research and software engineering staff of SP8 MIP shall analyse users feedback, confirm understanding with the users, and provide assessment of the system validation. Accordingly, they will provide assessment and proposal for the system-level technology readiness level of the Platform. This is a standard process in the high-tech industry which develops technology, such as telecommunications vendors, aeronautics, space programs, etc.
21	Section 6 states that the successful execution of the 5 operational scenarios will show user needs are met. However there are no criteria in advance regarding the degree of reproducibility across the 3 hospitals when results are expected to be the same i.e. what results would be 'acceptably near each other'? Nor is there specified in advance the degree of agreement necessary between individual hospital-derived results and either literature or research data where locale-specific results are to be expected i.e. what is the scale of consilience required?	Page 5, paragraph 2, bullet point 12	<p>Some of the studies are done for the first time. There has been no solution that connects and uses multiple patient datasets originating from the participating hospitals.</p> <p>Validation criterion is, hence, not the measure of the scientific value of the results or measure of the similarity of the results obtained separately on the single-hospital datasets. Conceptually, and according to the standard definition of the system validation process (see discussions and references in multiple previous points), the validation criterion is users assessment of the value of the Platform for their needs.</p> <p>The level of comparability of the results and assessment against the literature or other available research data cannot be used as only measure of the value of the Platform for the users. For example, the Platform could have a value for a clinician not only for research purposes, but also by discovering inconsistencies in data (outliers) and referring back to individual patients for further improvement of the precision of the</p>

No.	Expert's Comment	Page, Paragraph	SP8 Answer
			diagnosis and consequently the treatment.
22	There appears to be no explicit testing of UC_CLU_03 and UC_CLU_04 in pages 25ff.	Page 5 paragraph 2, bullet point 13	These two use cases are included in UC-CLU_01. They are implicitly tested. Furthermore, the purpose of the standard system validation process (please, see discussions and references above) is not to test individual "atomic" use cases, or functional requirements, but the value i.e. usefulness of the system as a whole. And that test is done by users themselves.
23	There appears to be no explicit testing of UC_DFY-07 in the MIP deployment validation.	Page 5, paragraph 2, bullet point 14	The loading of the data analytics ready datasets has been performed during deployment system validation scenario in each of the three hospitals. The results and the discussion shall be presented in Deliverable D8.6.3
24	It is not clear how representative of the likely user community of MIP users the 3 hospitals are.	Page 5 paragraph 2, bullet point 15	The three participating hospitals are chosen exactly because they are judged as representative. The criteria are provided on page 16, within Chapter 6: Diversity, size, clinical excellence available resources and influence.
25	The issue that remains here is the "coherent clinical protocol". The authors describe on page 57 that "the images must be high-resolution (max 1.5 mm) T1weighted sagittal images (Page 66). Detailed anatomy scans are made in some hospitals but these would still mostly be made for research purposes at this time. Indeed, as far as one is aware, 1.5mm T1w images are not standard routine for scanning in case of possible dementia since other scans are in place for such diagnostic purposes. Also, why only sagittal scans are included is not clear. Although direction of scanning has its influence on the result, this narrows down the potential for clinical future implementation of already existing data? Even if research data across several T1w acquisition protocols is being shared (thus including sagittal, axial and coronal directions). It seems that the image selection is narrowed in such a way that only specific research protocols can be included? This should really be made clear in the demo-plan and subsequent results reporting so as	Page 5 & 6, Chapter 2.3, bullet point 2	The observation is fully correct. The data factory, neuromorphometric feature extraction pipeline and implementation of the SPM12 tool actually does not impose such a strict set of requirements. These should be taken as recommendations. The only strict requirement is the T1W protocol. System validation demonstrated that the brain scans with different resolution are processed successfully.

No.	Expert's Comment	Page, Paragraph	SP8 Answer
	to avoid mixing up clinical and research data.		
26	A point to be careful with is the claimed scientific added value shown in the example on page 46, where it is stated that brain volume and cognitive functions declined differently dependent on age. Based on clinical data in particular this is not a very straightforward conclusion to be drawn since data includes patients with (a risk for) brain disease and due to such sometimes unforeseen selection biases findings cannot one on one be extended towards the healthy population. This point should be taken into account.	Page 6, Chapter 2.3, bullet point 3	<p>We agree with the observation about the risk to have unforeseen selection biases as the selected population is already at risk for brain disease. The example on page 37 (in the Sub-Chapter Data Mining, in Appendix I) is just an example with intention to provide an simplistic explanation of a potential value of the use case group.</p> <p>In Chapter 2.5 of Deliverable D8.6.3, we discuss in a much more systematic way where and why a potential clinical utility and scientific value of the platform can be found.</p>
27	A second smaller point - only PACS systems are now mentioned (Page 56). Is it indeed realistic that all the phases of the MIP-local and MIP-federated are being done on PACS? Should alternative platforms such as XNAT (widely used in the imaging research community be mentioned as possible alternatives?	Page 6, Chapter 2.3, bullet point 4	<p>It is a mistake to mention just PACS systems. XNAT, CBRAIN, LORIS, etc., can be used too.</p> <p>The text on pages 46-47, in the Sub-Chapter Data Factory Sub-system in Appendix II, has been updated with corrections as suggested in this comment.</p>
28	It is not specified what methods will be used to assess and quantify usability of the system.	page 6, Chapter 2.3, bullet point 5	A usability test is one of the standard types of system verification and is not in the scope of systems validation. Please, refer to the discussions to the previous points, especially the ones with references to standard process definitions
29	The described methodology for testing how well the generated models generalize is only partially reflecting best practice.	Page 6, Chapter 2.3, bullet point 6	It has been highlighted that the models are benchmarked and their accuracy measured using k-fold cross-validation methodology. It is possible to select validation both within and outside the same cohort. Also, the standard statistical accuracy measurements are performed, including PPV, NPV, standard deviation, z-score etc.
30	In general: KPIs providing a clear framework for measuring success of the MIP validation are not provided. Furthermore, it remains unclear what algorithms will be really included in the Algorithms factory of MIP-Local deployments and how the correct function of these algorithms will be evaluated. The requested MIP-Federate Showcase for MIP-Federate doesn't appear to have addressed.	Page 6, Chapter 2.3, bullet point 7	<p>Validation criteria of the system validation process do not usually include performance measurements (KPI), unless they are specified by the users who are executing the system validation. The validation is usually qualitative, and the MIP one is planned that way.</p> <p>MIP federation system validation is planned - see specifically clinical system validation scenarios 3 and 4. Also listed in the table of Appendix V: Hospital Selection Criteria and Expected Dataset Size.</p>

No.	Expert's Comment	Page, Paragraph	SP8 Answer
31	Data Privacy (p 9): it is unclear what exactly is meant with „De-identified patient data never leaves local hospital's MIP execution environments.“ It is claimed that „The federated analytic results are visualized in the central federation node's web-based user interface, by means of aggregation, meta-analysis and cross-hospital validation.“ However, for verification of models or predictions, it is essential to have access to all data available for validation and it is unclear how trust in generated models can be built if a user does not have the chance to inspect the raw data.	Page 6, Chapter 2.3, bullet point 8	Users can explore statistics of patient data (distribution of values etc) in any hospital or cross-hospitals, but cannot query individual patient data. The models are built centrally after the examination of the variable values, but are executed in “private” hospital MIP execution environments where local data are also stored. The results (aggregate data) are shared, and further aggregated, for cross-dataset analytics. The users cannot inspect individual patient data.
32	The MIP-Federated analytics will be applied locally which is a first step and on a multicentre bases when ethical approval is arranged. On page 7 it reads “once federated, the data stored in the local hospital MIP deployments become accessible for multicentre, multi-dataset studies.” Several benefits are being described but no clear plan is presented for show-casing MIP-federated. It seems that at this point the comparison datasets will be imported to the local hospitals for comparison to their local data (page 10). Once ethical approval is given, the transformed source datasets will be loaded to a permanent harmonised feature data store for federated analytics (page 43). This scenario seems to be a realistic one albeit a rather limited one. More proposals are needed.	Pages 6 & 7, Chapter 2.3, bullet point 10	The system validation plan includes so-called “MIP -federated” clinical studies. Please refer to Appendix V: Hospital Selection Criteria and Expected Dataset Size, Table 8. The results shall be presented in Deliverable D8.6.3 and analysed.
33	A web subsystem is shown in the demo-plan (page 77) with applications for end-users of the platform. For privacy reasons only exploration of statistical data is possible. Several aspects are then outlined. This may work but holds the potential problem at this stage of being caught in the middle: at this stage of researchers on the development side the information may be too generic. For researchers who want to apply the output of the data this may be too experimental	Page 7, Chapter 2.3, bullet point 10	We fully agree with this observation and hope that the functional improvement of the platform shall be planned and approved for the SGA2 phase of the SP8 MIP project

No.	Expert's Comment	Page, Paragraph	SP8 Answer
	still. However, if the interface is indeed going to work it can present a starting point that can be built on further.		
34	<i>(Editorial Suggestions)</i> The bulleted lists on page 11 in section 6 should be explicitly cross-referenced to 7.1 and 7.2.1 through 7.2.4. That way the use-cases can be better linked to the 5 operational scenarios. The last bullet point should read "Biological signature of Alzheimers disease using pathological measures".	Page 7, Chapter 2.4, point a)	The document has been updated with the suggestions. The updated bulleted list is in Chapter 6 on page 11 of this document.
35	<i>(Editorial Suggestions)</i> Table 1 page 12ff needs to be split into 2 parts so that UC-ITL-01 through to UC_DFY-07 is clearly about the MIP deployment scenario (i.e. <i>the software installation and data factory use-cases for preparing patient data for analytics</i>). The other part of the table (UC_WEB_01 through UC_CLU_04) being clearly linked to the clinical scenarios (i.e. <i>all the web applications and data analytics use-cases</i>).	Page 7, Chapter 2.4, point b)	The document has been updated with the suggestions. Table 1 has been divided into two tables: Table 1 - Medical Informatics Platform Deployment Use Cases and Table 2 - Medical Informatics Platform Web Applications and Data Analysis Use Cases. These tables are on pages 12-15 of this document.
36	<i>(Editorial Suggestions)</i> The bullet points on page 20 under 7 should use the same abbreviations as in Table 3 on page 19 and be in the same order as the columns in the Table.	Page 7, Chapter 2.4, point c)	The document has been updated with the suggestions. The updated bulleted list is on the page 19 of the updated document. The reference made in the editorial suggestion is not in Table 3 but in Table 4.
37	<i>(Editorial Suggestions)</i> Table 3 should have the role ID (pages 22ff) as an extra column so as to relate the results to the person who did them.	Page 7, Chapter 2.4, point d)	The document has been updated with the suggestions, in Table 4 - Requirements and Validation Traceability Matrix. An additional column named "Actor Role ID" is added as a third one from the left.

Appendix V: Hospital Selection Criteria and Expected Dataset Size

Table 8 - Selection Criteria for Hospitals and Clinical Use Scenario Demonstrations

Use Case Name	Demonstration Sites				
	CHUV Switzerland local study	CHRU Lille France local study	Brescia Italy local study	Research cohorts (ADNI, EDS)	Multi- centre study
measuring the clinical utility of the hippocampal volume for diagnosing alzheimer's disease	YES	YES	YES	YES	YES
measuring clinical utility of csf markers for alzheimer's disease: total tau, phosphorylated tau and aB42	NO	NO	NO	YES	YES
differential diagnostic between fronto-temporal dementia and Alzheimer's disease	YES (apply model)	YES (train model)	YES (apply model)	YES (apply model)	YES
biological signature of Alzheimer's disease using pathological measurements	YES (train model)	YES (train model)	YES (apply model)	YES (apply model)	YES
Dataset	Total patients: 700 Type of data: Brain features, Socio-demographic, Bio-specimen, Genetic, Clinical scores	Total patients: 1000 Type of data: Brain features, Socio-demographic, Bio-specimen, Genetic, Clinical scores	Total patients: 2000 Type of data: Brain features, Socio-demographic, Bio-specimen, Genetic, Clinical scores	Total patients: 1500 (including healthy controls) Type of data: Brain features, Socio-demographic, Bio-specimen, Genetic, Clinical scores	Trained machine learning models on all three hospital datasets
Input to D8.6.3 Deliverable	Study reports: Scientific results and validation of the MIP methods (pre-processing, data quality, machine learning performance) User feedback reports: Feedback from the clinical users on the UI (data exploration/selection, model building/testing and results interpretation) Recommendation reports: Recommendation from the SP8 team and the users				

Appendix VI: Acronyms and Abbreviations

ADNI	Alzheimer's Disease Neuroimaging Initiative
AUEB	Athens University of Economics and Business, Greece
BIDS	Brain Imaging Data Structure
CD	Continuous Deployment
CHUV	Centre hospitalier universitaire vaudois (Lausanne University Hospital)
CI	Continuous Integration
CSF	Cerebrospinal Fluid
CSV	Comma-Separated Value
DAG	Directed Acyclic Graph
DB	Database
DGDS	Data Governance and Data Specification
DICOM	Digital Imaging and Communications in Medicine
EHR	Electronic Health Record
EMR	Electronic Medical Record
EPFL	École polytechnique fédérale de Lausanne (Federal Institute of Technology in Lausanne)
ETL	Extraction, Transformation and Load
HW	Hardware
I&V	Integration and Verification
ICM	Institut de Cerveau et de la Moelle épinière, France (Brain and Spine Institute)
JSI	Jožef Stefan Institute, Slovenia
JSON	JavaScript Object Notation
LUMC	Leiden University Medical Center, Netherlands
MDR	Meta Data Register
MIP	Medical Informatics Platform
MPM	Multi Parameter Mapping
MRI	Magnetic Resonance Imaging
MRI	Magnetic Resonance Imaging
MW	Middleware
NIFTI	Neuroinformatics Technology Initiative
OLAP	Online Analytics Processing
OLTP	Online Transaction Processing
OS	Operating System



PACS	Picture Archiving and Communication System
PFA	Portable Format for Analytics
PII	Personal Identifiable Information
PPMI	Parkinson's Progression Markers Initiative
RHEL	Red Hat Enterprise Linux
SPI	Sensitive Personal Information
SW	Software
TAU	Tel Aviv University, Israel
TGD	Tuple Generated Dependencies
UNIGE	University of Geneva
UoA	National and Kapodistrian University of Athens, Greece
VBM	Voxel-Based Morphometry
XML	Extensible Mark-up Language
UDF	User Defined Function
RDBMS	Relational Database Management System
HBP	Human Brain Project

Appendix VII: References

- [1] Gnubila FedEHR Anonymizer
<https://www.gnubila.fr>
- [2] Apache Airflow
<https://airflow.apache.org/>
- [3] Neuroimaging Informatics Technology Initiative
<https://nifti.nimh.nih.gov/>
- [4] Feature Extraction Framework, John Ashburner
<http://www.fil.ion.ucl.ac.uk/~john/LabelProp/Label%20Propagation%20Framework.pdf>
- [5] Quantitative MRI and Voxel Based Quantification (VBQ), Martina Callaghan
http://www.fil.ion.ucl.ac.uk/Research/physics_info/QuantMRI_VBM.html
- [6] MIPMap Tutorial
<https://github.com/HBPMedical/MIPMap/blob/master/MIPMap%20Tutorial.docx>
- [7] ++Spicy Mapping System
<https://github.com/dbunibas/spicy/blob/master/spicyManual/manual.pdf>
- [8] Medical Informatics Platform Knowledge Base
<https://hbpmmedical.github.io/>
- [9] PostgresRAW-UI
<https://github.com/HBPMedical/PostgresRAW-UI>
- [10] Specification of the Hospital Bundle Functionality
SP8_D8.6.4_Resubmission_Final_annex-
I_HBP_hospital_bundle_specification_v0.7_final.pdf
- [11] A Relational Approach To Complex Dataflows
EDBT/ICDT 2016 Joint Conference Workshop Proceedings, Yannis Ioannidis et al.
- [12] madIS - The Core Engine of Exareme
<https://github.com/madgik/madis>
- [13] Recommended Operating Systems
<https://hbpmmedical.github.io/deployment/os/#>
- [14] External Services for Software Installation, Configuration, Upgrade and Maintenance
<https://hbpmmedical.github.io/deployment/services/#>
- [15] Microservice Infrastructure - automatic installation, configuration and management of MIP SW
<https://github.com/HBPMedical/mip-microservices-infrastructure>
- [16] Installation of MIP Local
<https://github.com/HBPMedical/mip-microservices-infrastructure/blob/master/docs/installation/mip-local.md>
- [17] External Software Packaged Into MIP
<https://hbpmmedical.github.io/deployment/software/#external-software>
- [18] Software Developed by MIP Teams
<https://hbpmmedical.github.io/deployment/software/#sp8-made-software>



- [19] MIP Software Packaged in Docker Images
<https://hbpmmedical.github.io/software-catalog/#docker-images>
- [20] D8.6.2 SP8 Medical Informatics Platform - Results for SGA1 Period 1
SP8 D8.6.2 FINAL.pdf
- [21] Data Mapping and Transformation Specification
<https://drive.google.com/drive/folders/0B5K3IDNQ5PbrbFdnamE0UWoycE0>
- [22] MIP Web UI - User Guidelines, V2.0 Public Release
https://hbpmmedical.github.io/documentation/HBP_SP8_UserGuide_latest.pdf
- [23] Data Sharing and Demographic Research (DSDR) Project
<https://www.icpsr.umich.edu/icpsrweb/content/DSDR/harmonization.html>
- [24] European Medical Information Framework (EMIF)
<http://www.emif.eu/>
- [25] SP8_Medical Information Platform - Architecture and Deployment Plan
SP8 D8.6.1
- [26] Systems and software engineering -- System life cycle processes
ISO/IEC/IEEE 15288
- [27] Multiple Linear Regression: Bayesian Inference for Distributed and Big Data in the Medical Informatics Platform of the Human Brain Project
<https://www.biorxiv.org/content/early/2018/01/05/242883>
- [28] Medical informatics tutorial
<https://www.youtube.com/watch?v=2jMbElcwKnA>
- [29] Experts Review Report
European Commission, Brussels, end of January 2018