SP8 MIP - Results for SGA2 Year 2 (D8.1.2 – SGA2)

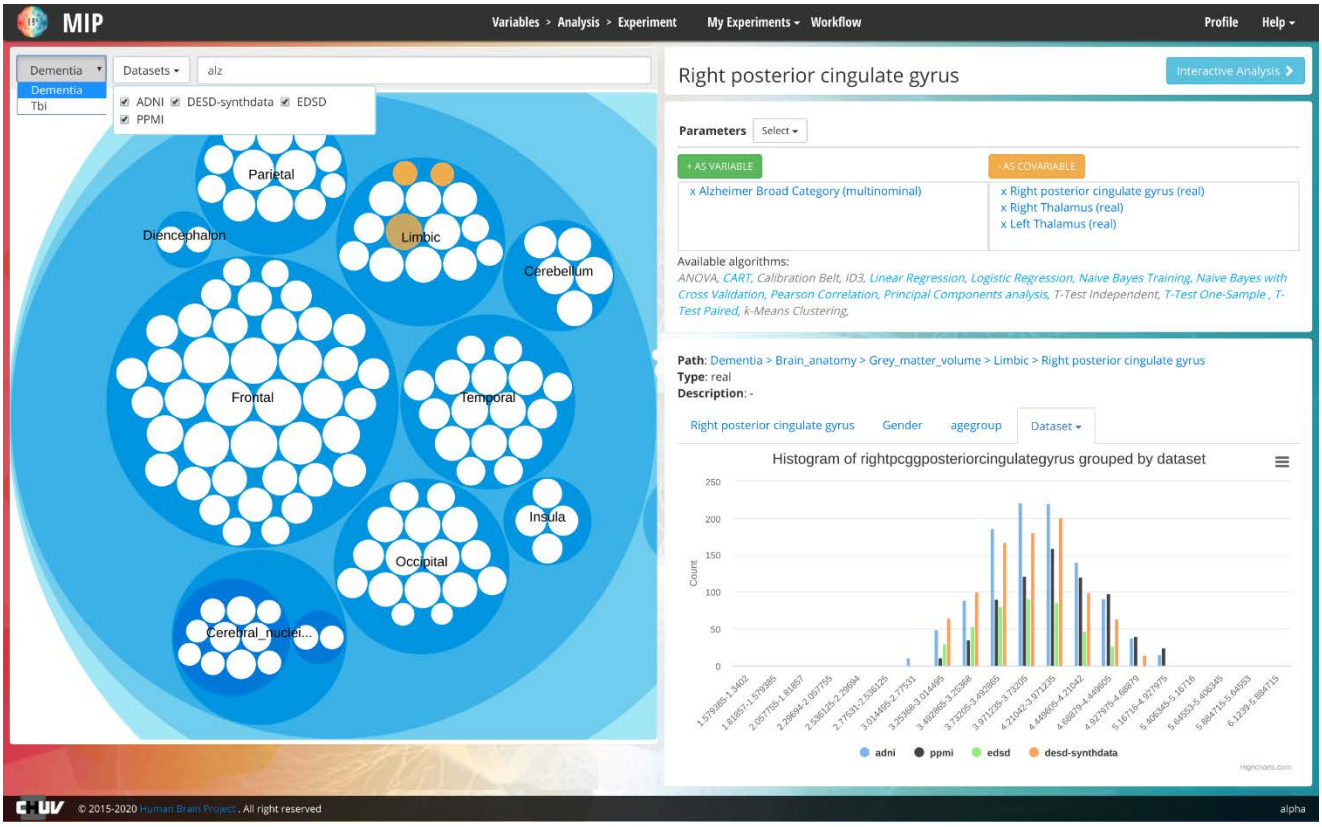

Figure 1: Front page of the new MIP new user interface

The MIP new user interface is now adapted to various pathologies (dementia, epilepsy, TBI, mental health).

Project Number:	785907	Project Title:	Human Brain Project SGA2

Document Title:	SP8 MIP - Results for SGA2 Year 2)
Document Filename:	D8.1.2 (D48.2 D44) SGA2 M24 ACCEPTED 201005.docx
Deliverable Number:	SGA2 D8.1.2 (D48.2, D44)
Deliverable Type:	Report
Work Packages:	WP8.1, WP8.2, WP8.3, WP8.4, WP8.5, WP8.8, WP8.9, WP8.10 WP8.6, WP8.7 are reported in Compound Deliverables of CDP6 and CDP8 resp.
Key Result(s):	KR8.1, KR8.2, KR8.3, KR8.4, KR8.5
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA2 M24 / 31 Mar 2020
Actual Delivery Date:	SGA2 M26 / 29 May 2020; resubmitted 25 Sep 2020; approved 25 Sep 2020; resubmitted 2 Oct 2020; accepted 5 Oct 2020
Author(s):	Jacek MANTHEY, CHUV (P27), Philippe RYVLIN, CHUV (P27)
Compiled by:	Jacek MANTHEY, CHUV (P27)
Contributor(s):	Erika BORCEL, CHUV (P27), Philippe RYVLIN, CHUV (P27), Sandra SCHWEIGHAUSER, CHUV (P27), Tomas TEIJEIRO, EPFL (P1), David STEINBERG, TAU (P57), Evita MAILLI, UoA (P43), Olivier DAVID, UGA (P125), Sumit MADAN, FG (P22), Pegah SARKHEIL, UKAACHEN (P73); all sections
SciTechCoord Review:	Mehdi SNENE, EPFL (P1)
Editorial Review:	Annemieke MICHELS, EPFL (P1)
Description in GA:	For consistent presentation of HBP results, SGA2 M24 Deliverables describing the accomplishments of an entire SP, WP or CDP have been prepared according to a standard template, which focuses on Key Results and the outputs that contribute to them. Project management elements such as Milestones and Risks will be covered, as per normal practice, in the SGA2 Project Periodic Report.
Abstract:	This Deliverable is the Y2 annual compound of SP8 'Outputs' organised by SP8's Key Results. The main Outputs are the continuation of the consolidation and completion of the SGA2 deployment in hospitals of the Medical Informatics Platform, according to SP8 Work Plan, as well as the completion of the dementia and epilepsy use cases. Outputs from new Work Packages selected through Calls of Expression of Interest are available.
Keywords:	SP8, SGA2, MIP, Medical Informatics Platform, ethics, MIP operations, MIP helpdesk, MIP deployment, MIP maintenance, MIP software, MIP software development, Data Catalogue, MIP algorithms, privacy-aware, IMAGEN, HIP, ontologies, ROHAN, dementia, epilepsy, use cases
Target Users/Readers:	Clinicians, computational neuroscience community, Consortium members, experts in neuroscience, funders, neuroimaging community, neuroinformaticians, neuroscientific community, neuroscientists, Platform users, policymakers, researchers, scientific community, students, other potential users of HBP results

Table of Contents

Table of Tables

Table of Figures

History of Changes made to this Deliverable (post Submission)

Date	Change Requested / Change Made / Other Action
29 May 2020	Deliverable submitted to EC
25 Sep 2020	Revised draft sent by SP8 to PCO. Main changes made, with indication where each change was made: • Correction of links (see Section 5.1.4 Output 3)
25 Sep 2020	Revised version resubmitted to EC by PCO via SyGMa
25 Sep 2020	Deliverable approved by EC
2 Oct 2020	Minor editorial change by PCO
2 Oct 2020	Revised version resubmitted to EC by PCO via SyGMa

1. Overview

The Medical Informatics Platform (MIP) is an innovative digital solution that aims at leveraging data sharing for medical research, while ensuring patients' data privacy. The MIP connects databases from various hospitals in order to analyse them simultaneously (so-called "federated analyses"), without displacing any data from their hospital of origin.

During the last period of SGA2, SP8 has achieved the following:

1) Release of improved versions of the MIP which offer more robust, functional, ergonomic and easy to install platforms. These improved versions include novel federated analytical tools, including machine learning algorithms.
2) Ensured that the MIP complies with all European ethics and data privacy/security regulatory requirements. To this end, the MIP only operates on anonymised data and only provides aggregate findings. Access to, and re-identification of, individual data through the MIP is deemed impossible.
3) Deployed the MIP over a large network of hospitals to leverage data sharing at the European scale. The MIP is currently deployed in 30 European centres with nine more having signed an installation agreement and 20 additional hospitals that have declared being interested.
4) Developed partnerships with consortia of physicians and researchers providing data of interest. Data from more than 20,000 patients have now been integrated into the MIP covering the field of Alzheimer's Disease, Mental Health, Epilepsy and traumatic brain injury.
5) Delineate a roadmap for the future development and sustainability of the MIP. To this end, the MIP has been embedded into EBRAINS, the global infrastructure of the Human Brain Project.

Based on all above achievements, the MIP and its federated machine learning algorithms are now being used on the large shared databases to develop predictive models of various brain diseases.

During SGA3, the MIP will continue to be improved and deployed over more major European hospitals, with a target of 60 equipped medical centres by the end of SGA3. Furthermore, the MIP shall evolve from a purely research tool to a diagnostic tool with the aim to promote the early identification of rare diseases.

2. Introduction

The MIP offers a unique privacy-aware digital solution to perform federated analyses, including machine learning, over large datasets distributed across hospitals. In doing so, the MIP enables to reach out for medical data collected in clinical routine that would otherwise not be available for research. Those data never leave their hospital of origin and can neither be copied, uploaded or even investigated at an individual level, thus preserving patients' data privacy.

This Deliverable presents the SGA2 year 2 annual compound of SP8 'Outputs' organised around SP8's Key Results (KRs).

In the second year of SGA2, the main achievements are the MIP consolidation as a network and as a software solution. The MIP network has reached its ambitious SGA2 objective of 30 MIP instances installed in hospitals and research centres. More are in the pipeline having already signed an installation agreement. The MIP software solution, gone through releases 4.0, 5.0 to 6.0, benefits from fundamental architectural enhancements, impressive both in terms of ease of deployment and system stability. New algorithms and new functional features have been introduced, to respond to various end users' needs. One of these needs was to adapt the MIP front end and its data factory (in particular the data catalogue that manages meta-data) to several brain disorders, since it is now covering the fields of Dementia, Mental Health, Traumatic Brain Injury (TBI) and Epilepsy.

The first Key Result (KR8.1) concerns MIP compliance with ethics, regulatory and privacy-preserving requirements. The latest MIP releases include a revised architecture of MIP federated nodes, an anonymisation tool and privacy-aware algorithms ensuring that MIP federated analyses are restricted to anonymised data and that they can only retrieve aggregated findings calculated on at least 12 subjects.

The second KR (KR8.2) describes the successful deployment of the MIP in 30 centres, in line with its original target. For such endeavour, a number of administrative and technical challenges had to be overcome, offering opportunities to further deploy the MIP during SGA3 in another set of about 30 EU hospitals that have already signed the MIP installation agreement or have declared their interest in doing so. The partnership with the European Reference Networks (ERNs), and in particular EpiCare, proved instrumental in achieving this KR.

The third KR (KR8.3) aimed at integrating as many data of interest as possible into the MIP to further showcase the value of the Platform and its federated analyses. It is noteworthy that we cannot monitor data installed in MIP Local and we can only report on data integrated into MIP federated nodes or reported to us by MIP local data users. In total, datasets from more than 20,000 patients have been integrated into the MIP. These allow to test use cases in the field of dementia, mental health and TBI, though the latter have been delayed by the major impact of the COVID-19 pandemic on many MIP-equipped hospitals.

The fourth KR (KR8.4) aimed at providing the MIP with advanced federated analytical tools enabling multidimensional analyses of longitudinal data. This objective has not yet been fully reached, though federated Kaplan Meyer analyses have been integrated into the MIP.

The fifth KR (KR8.5) concerns automatic extraction of data from electronic health records to facilitate their integration into the MIP. As agreed from the very beginning of SGA2, this objective was revised to focus on the automatic extraction of data and metadata from the most widely used eCRF, REDCap.

In parallel to the KR-related Work Plan, SP8 has conducted several use cases and coordinated the activity of its Open Calls. The latter resulted in additional valuable Outputs such as the integration of novel pipelines in the data factory to manage Human intracerebral EEG data, and access through the MIP front end to comprehensive ontology resources.

3. KR8.1 MIP infrastructure and operational activities comply with EU ethics and data privacy/security regulatory requirements

3.1 Outputs

3.1.1 Overview of Outputs

3.1.1.1 List of Outputs contributing to this KR

- Output 1: Data Governance Steering Committee
- Output 2: Data Privacy Impact Assessment
- Output 3: Privacy aware MIP software and algorithms
- Output 4: MIP federation infrastructure compliant with requirements

3.1.1.2 How Outputs relate to each other and the Key Result

Output 1: *Data Governance Steering Committee* and Output 2: *Data Privacy Impact Assessment* are complementary. While the latter describes and analyses the processing of data within the MIP to ensure it complies with the General Data Protection Regulation (GDPR), the former reaches further out beyond the MIP, providing ethical and legal guidelines to providers of clinical data, identifying and solving cross-cutting data governance issues that affect the MIP. It also provides guidelines for information produced using the MIP, by including the authorship and acknowledgement policy for publications. Both are inputs for delivering Output 5: Fulfilling regulatory requirements. The Output 3: *MIP Software* serves to build the Output 4: *MIP Federated Infrastructure*.

All the Outputs contribute to KR8.1.

3.1.2 Output 1: Data Governance Steering Committee (C2970, C2971, C3053)

Data Governance Steering Committee mission (DGSC) is to ensure that all MIP-related activities (curating, sharing and analysing data) are governed by sound principles validated by all data owners and controllers, and comply with applicable legislation. It aims to protect the privacy of research subjects and the confidentiality of their personal information, and to ensure that participating hospitals comply with the current ethics rules, by producing documents for ethics approval gathering and storing all compliance documents. The DGSC also ensures that the appropriate roles and access rights are properly defined in the data processing in various MIP instances (Local, Federated, Public, Central) and proper anonymisation procedures are put in place, that agreements with data providers correspond to the data protection requirements for data use, and that data sharing within the MIP respects the data providers' rights. Publication policy is included. DGSC activities are described in the charter that was developed and validated by DGSC members.

The DGSC charter is available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Data%20Governance%20Steering%20Committee%20charter/ .

DGSC itself was formed and put into force during SGA2, with diseases-specific sub-committees, all of which had regular visio-conference meetings.

Table 1: Output 1 Links

Component	Link to	URL
C2970	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Data%20Governance%20Steering%20Committee%20charter/
	Technical Documentation	Not applicable-
	User Documentation	Not applicable
C2971	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Data%20Governance%20Steering%20Committee%20charter/
	Technical Documentation	Not applicable
	User Documentation	Not applicable
C3053	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Data%20Governance%20Steering%20Committee%20charter/
	Technical Documentation	Not applicable
	User Documentation	Not applicable

3.1.3 Output 2: Data Privacy Impact Assessment

Data Privacy Impact Assessment on the MIP has been conducted, analysing data processing in the MIP from the point of view of the fundamental principles of the GDPR, also including elements like illegitimate access to data, unwanted modification of data, data disappearance, controls to protect the personal rights of data subjects, etc. Risks analysis is included. MIP is considered GDPR-compliant.

The MIP DPIA (Data Privacy Impact Assessment) document is confidential.

3.1.4 Output 3: Privacy aware MIP software and algorithms (C3000)

The MIP software is privacy-aware by enforcement of privacy and security standards that were set in the Anonymisation process in local/federated MIP and the lower limit of number of datapoints that can be used by an algorithm. Testing is conducted so that they are compliant with the privacy requirements described in SGA2 Deliverables '*Software requirements specifications'* [D8.5.1 (D52.1 D4)] and '*ETHICS: Ethics requirements'* [D12.4.8 (D1.7 D118)].

During the 2nd year of SGA2 additional algorithms for federated analysis were developed (C3000-Federated Data Processing Engine). Their respective testing and validation methodology ensure privacy awareness while MIP's analytical capabilities increase and become more complex. MIP release 5.0 included additional algorithms ANOVA, Cross Validation, Descriptive Statistics, Histograms, Logistic Regression, and Paired and Simple T-test. MIP release 6.0, additionally contains Calibration Belt Algorithm, an assessment of quality of care algorithm from Bergamo hospital, CART, an algorithm for decision-tree learning, and Kaplan Meier, an algorithm for survival analysis of longitudinal data. An important MIP feature is that it only provides access to aggregated findings calculated on at least 12 subjects.

The documentation for C3000 is at https://github.com/madgik/exareme/tree/master/Documentation (user doc.) and https://github.com/madgik/exareme/releases/tag/22.0.0 (technical doc.).

Table 2: Output 3 Links

Component	Link to	URL
C3000	Software Repository	https://github.com/madgik/exareme/releases/tag/22.0.0
	Technical Documentation	https://github.com/HBPMedical/mip-federation/tree/master/Documentation
	User Documentation	-

3.1.5 Output 4: MIP-Federation infrastructure

The MIP-Federation infrastructure, thanks to the anonymisation module (C3088), ensures that only anonymised data are stored at Federation nodes, warranting that only anonymised data are used for MIP-Federated analyses and respecting data safety and privacy. The MIP-Federation is accessible via the MIP portal 'https://mip.ebrains.eu/' and via its direct URL 'https://mip.humanbrainproject.eu' to users accredited by the DGSC for a particular pathology.

Table 3: Output 4 Links

Component	Link to	URL
C3088	Software Repository	https://github.com/aueb-wim/anonymization-4-federation
	Technical Documentation	https://github.com/aueb-wim/anonymization-4-federation
	User Documentation	https://github.com/aueb-wim/anonymization-4-federation/blob/master/README.md

3.2 Validation and Impact

3.2.1 Actual and Potential Use of Output(s)

Output 1: *DGSC* is the data governance body allowing hospitals to adapt their data curation to the new requirements and challenges in an efficient way. It plays an essential role to ensure that all MIP data providers and users follow appropriate guidelines in data handling and curation, and to promote and to organise the participation of new data providers and users, including those involved in brain diseases not yet covered by the MIP.

Output 1: *DGSC* and Output 2: *DPIA* are used in the present MIP-network and pave the way to further increase the MIP data providers and users, as well as the volume of clinical data available in the MIP. These Outputs, once validated by the communities of users, correspond to their requirements, and allow to deliver upfront answers to common questions, thus speeding the data process.

Output 3: *Privacy aware MIP software* and Output 4: *MIP federation infrastructure* make possible the present use of the MIP by researchers and other users and create the base for its future use and development. As more diverse algorithmic capabilities are available through the MIP, important use cases may be verified by clinical data, such as advanced phenotyping of the ageing brain cognitive diseases at early stage, and discovery of novel disease models and their application to Parkinson's Disease.

The validation measures are: *Final DPIA results* [MS8.1.4], Software validation was done internally from the development teams, using extensive tests designed with the help of statisticians.

3.2.2 Publications

None.

4. KR8.2 MIP is operated over a large network of European Hospitals (≥ 30)

4.1 Outputs

4.1.1 Overview of Outputs

4.1.1.1 List of Outputs contributing to this KR

- Output 1: Ensuring continuing operations and maintenance of the MIP
- Output 2: Operating MIP within the HBP EBRAINS infrastructure
- Output 3: Consecutive releases of MIP software (C2967, C2968, C2969)
- Output 4: Clinical Data Catalogue (C2990)
- Output 5: MIP installed in 30 hospitals
- Output 6: Expanding MIP network build-up

4.1.1.2 How Outputs relate to each other and the Key Result

Output 1 *"Operations"* together with Output 2 *"EBRAINS infrastructure"* provide the basis for the MIP-network, into which are deployed Output 3 *"Consecutive releases"*.

Output 3 in combination with the introduction of pathologies in Output 4 *"Data Catalogue"* contribute and will further contribute to the adoption of MIP by growing number of hospitals, yielding already by now Output 5 *"MIP installed in 30 hospitals"*. They are instrumental for adoption of the MIP by networks of hospitals and increasing the number of installed MIP instances.

All these Outputs thus logically fit in together, from providing the software and its upgraded versions, to deploying it to the various MIP instances, ensuring its availability to the end users and providing support, with continuous improvement of the software robustness, functionalities and deployment and upgrades. Together with Output 6 *"Expanding MIP network build-up"*, all these elements concur to growing installation base and thus to KR8.2.

4.1.2 Output 1: Ensuring continuing operations and maintenance of the MIP

Running MIP instances are being operated and monitored to ensure their smooth running. Software and hardware errors are diagnosed and corrected when applicable (C2969).

The helpdesk is available to MIP users at support@humanbrainproject.eu and operates within the framework of the HBP High Level Support Team (HLST) (C2962, C2963, C2964). It provides first-level support as well as direct support to MIP users.

Based on the operations of the MIP, if needs for improving procedures and for fixing software bugs are identified, the operational procedures are improved and bug fixes issued, and then implemented in MIP operations.

For major releases of MIP software, i.e. 4.0, 5.0 and 6.0, a plan for MIP upgrade deployment was scheduled and executed in hospitals, MIP central infrastructure, development and test machines. Due to the COVID-19 situation, the last upgrade (6.0) has not yet been possible in most hospitals, but will be performed in the near future.

Table 4: Output 1 Links

Component	Link to	URL
C2969	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20QC%20tool%20for%20SW%20and%20infrastructure/
	Technical Documentation	-
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20QC%20tool%20for%20SW%20and%20infrastructure/
C2962	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
	Technical Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
C2963	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
	Technical Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Helpdesk%20and%20Support/
C2964	Software Repository	https://enterprise.frontline.ca/fully-managed-it-services/remote-on-site-service-desk/
	Technical Documentation	-
	User Documentation	https://enterprise.frontline.ca/fully-managed-it-services/remote-on-site-service-desk/

4.1.3 Output 2: Operating MIP within the HBP EBRAINS infrastructure

The MIP central instance has been moved during SGA2 from CHUV into the EBRAINS Infrastructure (http://148.187.97.159 in the CSCS supercomputing centre). This is an important step towards ensuring MIP federation sustainability within EBRAINS. We envision that in the near future, hospitals might wish to have their MIP federated node (with anonymised data only) installed on a "private" secured virtual machine in EBRAINS (one per hospital) rather than in their own IT system. This will require changes in the DPIA and data safety insurance from EBRAINS infrastructure, but will greatly facilitate future MIP federation deployment.

4.1.4 Output 3: Consecutive releases of MIP software (C2967, C2968, C2969)

Consecutive releases of the MIP software (latest is 6.0) bring with them increased functionality and robustness, as new versions of various modules were developed and integrated. They are described in the *Medical Informatics Platform Releases - SOFTWARE & REPORT* [SGA2 Deliverables D8.5.2 (D52.2 D8) and D8.5.3 (D52.3 D9), two Deliverables with the same title].

Of particular importance are enhancements of Federated Data Processing engine "Exareme" (C3000), EXAREME is a query processing engine optimised for execution of federated database queries extended with user-defined functions processing data and exchanges of aggregated results when crossing data ownership boundaries. Some aspects are described above in Output 3 under KR8.1 (C3000). Other enhancements are related to improvement in local execution of algorithms. In previous releases, local processing was handled by the "Woken" component, an orchestrator platform for Docker containers relying on Mesos and Chronos to control and execute the containers

over a cluster. This cluster-based analysis approach could not be utilised for federated analysis as the two approaches have fundamental semantic differences with respect to data boundaries. For these reasons, the Federated / distributed data processing engine ("EXAREME") component, already used for federated analysis, was adapted to allow restricting its processing pattern to single node execution to perform local analysis, thus replacing 'Woken" and improving performance. Another facet of the Federated Data Processing engine "Exareme" (C3000) is described in Output 3 under KR8.1.

Table 5: Output 3 Links

Component	Link to	URL
C2967	Report Repository	D8.5.2 (D52.2 D8) and D8.5.3 (D52.3 D9): Medical Informatics Platform Releases – Software & Report
	Technical Documentation	-
	User Documentation	D8.5.2 (D52.2 D8) and D8.5.3 (D52.3 D9): Medical Informatics Platform Releases – Software & Report
C2968	Report Repository	Deliverable D8.2.1 (D49.1 D31) - MIP installed in 30 hospitals
	Technical Documentation	-
	User Documentation	Deliverable D8.2.1 (D49.1 D31) - MIP installed in 30 hospitals
C2969	Report Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20QC%20tool%20for%20SW%20and%20infrastructure/
	Technical Documentation	-
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20QC%20tool%20for%20SW%20and%20infrastructure/
C3000	Software Repository	https://github.com/madgik/exareme/releases/tag/22.0.0
	Technical Documentation	https://github.com/HBPMedical/mip-federation/tree/master/Documentation
	User Documentation	-

4.1.5 Output 4: Clinical Data Catalogue (C3290)

Data Catalogue is a web application for metadata visualisation and management. For the MIP network with a large number of hospitals and various pathology data models, its serves as a single point of truth presenting the structure and the semantics of the MIP latest set of the Common Data Elements, as well as hospital's local data models. It allows to cope with possible metadata inconsistencies and variations.

Data Catalogue's latest release v2.0 encapsulates and depicts the notion of multiple Medical Conditions (various pathologies). Every Medical Condition has its own data model definition which consists of the Common Data Elements (CDEs) and organised in a multilevel taxonomy. Each hospital has a dataset related to -at least- one Medical Condition but in most cases does not conform directly to the CDEs as it has its own local data model. Data Catalogue v2.0 illustrates the following:

- Medical Conditions and related hospitals
- Medical Conditions' CDEs data models
- Hospitals' local data models
- Correspondences from local to CDEs data models

The metadata management is done by each Medical Condition's authorised data manager personnel. The authorisation system is integrated with HBP's tool Keycloak. The documentation for C3290 is under https://github.com/HBPMedical/DataCatalogue/blob/master/README.md (user doc.) and https://github.com/HBPMedical/DataCatalogue (technical doc.).

Table 6: Output 4 Links

Component	Link to	URL
C3290	Software Repository	https://github.com/HBPMedical/DataCatalogue
	Technical Documentation	https://github.com/HBPMedical/DataCatalogue
	User Documentation	https://github.com/HBPMedical/DataCatalogue/blob/master/README.md

4.1.6 Output 5: MIP installed in 30 hospitals

Following signature of installation agreements, all necessary actions were conducted to deploy the MIP according to plan (instructions, configuration of the infrastructure and the software, training and documentation). The MIP is now installed in 30 centres, including 28 hospitals. Within this number networks of hospitals specialising in different brain disorders are defined (dementia, epilepsy, Traumatic Brain Injury (TBI), mental health,…).

Procedures for MIP Deployment to hospitals are continuously improved and simplified. Huge improvement was achieved to ease and speed the installation. The execution of a downloaded installation script makes a running MIP available in less than 15 min. Also, Vagrant configuration files are ready for any user to go from no Virtual Machine (VM) to a full environment with a running MIP within no more than five minutes. Output 3 *"Consecutive releases"* has also contributed to streamlining MIP deployment. The deployment procedures are now executable in EBRAINS infrastructure. The documentation is available under https://github.com/crochat/mip-deployment.

4.1.7 Output 6: Expanding MIP network build-up

To continue building up the MIP network beyond the SGA2 target of 30, further MIP installation agreements have been signed (38 in total up to end of SGA2). An additional list of more than 20 major EU hospitals have declared their interest to install the MIP.

To assist the build-up of the network and make effective the deployment process, specific documentation for deployment and training videos for MIP users are available at https://mip.ebrains.eu/documentation. The MIP Flyer for the general public is available at the following link: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/MIP%20Flyer/.

4.2 Validation and Impact

4.2.1 Actual and Potential Use of Output(s)

The effective deployment of MIP represents the first mandatory step towards an effective use of the MIP by HBP and non-HBP users. It proved to be successful in SGA2 with the target of 30 centres being reached. However, usage of the MIP also depends on the data availability for analysis which typically needs to be provided and curated by MIP users, as detailed in the KR8.3 Section. Furthermore, in line with SP8 strategy, a majority of MIP-equipped hospitals have chosen to first install and test a MIP local before entering into a MIP federation. We cannot monitor the activity of MIP local. Nevertheless, all MIP users refer to the Data Catalogue for the MIP dataset's metadata definitions.

The growing number of hospitals and centres that joined the MIP network demonstrates its attractiveness. This network will continue to grow as many other hospitals have already signed the MIP Installation Agreement or declared their interest to do so. This growth is promoted by the increased effectiveness of MIP functionalities and easiness of installation. The growing collaboration within this network provides a perspective of third party funding for multicentre research based on

federated MIP framework. The future development of MIP federation within EBRAINS is likely to leverage its development.

The following Deliverables and Milestones serve as validation measures: *Release of quality control tool for software and infrastructure of SP8* [MS8.1.2], *Implementation of Help-Desk* [MS8.2.1], *MIP deployed in 10 new hospitals as compared to SGA1* [MS8.2.3], *Contracts signed with 20 new hospitals for installing the MIP* [MS8.3.4], *Software requirements specifications* [SGA2 Deliverable D8.5.1 (D52.1 D4)], *Medical Informatics Platform Releases - SOFTWARE & REPORT* [SGA2 Deliverables D8.5.2 (D52.2 D8) and D8.5.3 (D52.3 D9), two deliverables with the same title], *MIP test runs in 3 partner sites* [MS8.10.2] and especially *MIP installed in 30 Hospitals* [SGA2 Deliverable D8.2.1 (D49.1 D31)].

4.2.2 Publications

No publications.

5. KR8.3 Established large-scale network of MIP-data providers, including clinical departments and research consortia, collating data from more than 30,000 patients with various brain diseases

5.1 Outputs

5.1.1 Overview of Outputs

5.1.1.1 List of Outputs contributing to this KR

- Output 1: Signed partnership with EpiCare network
- Output 2: Clinical Data Catalogue
- Output 3: MIP datamodels for several brain disorders
- Output 4: Clinical datasets from several brain disorders
- Output 5: Use Case Dementia
- Output 6: Use Case Epilepsy
- Output 7: Shareable data format for intracranial EEG
- Output 8: BIDS-iEEG manager
- Output 9: BIDS-iEEG pipeline
- Output 10: F-TRACT CCEP database
- Output 11: Android mobile app for acquisition of multimodal data in outpatient settings (updated version)
- Output 12: Systematic assessment of novel data type integration into the MIP (C2975, C2976)
- Output 13: Use Case Post Traumatic Stress Disorder (PTSD)

5.1.1.2 How Outputs relate to each other and the Key Result

Together with outputs related to KR8.2, which maintain and grow the MIP Network, the above Outputs extend the MIP Network work into the domain of the clinical data and their usage.

Output 1 opens the possibility to share data within a large consortium of top EU medical centres specialised in epilepsy, one of the primary brain diseases considered in HBP. Output 3 and Output 4 demonstrate how MIP moved from a single pathology in SGA1 (dementia) to several brain disorders. Finally, Output 5, 6 and 13 provide examples of analysis done for two use cases.

Outputs 7, 8, 9 and 10 are used to provide more data for the MIP for the epilepsy pathology.

All above Outputs contribute clearly to the KR8.3 of establishing large network of data providers.

Output 11 helps to accelerate the deployment of new use cases for data acquisition in outpatient environments, targeting different brain diseases. Both components in Output 12 complement each other in analysing the future evolution of the MIP to allow the integration of novel types of data. C2975 is more focused on the requirements perspective, while C2976 has a pure technical viewpoint.

5.1.2 Output 1: Signed partnership with Epicare network

A partnership was signed between HBP and the European Reference Network EpiCare to deploy the MIP in this network to promote data sharing and leverage clinical research within the network. Twelve of these EpiCare centres have already signed the MIP installation agreement, and nine of them have already installed the MIP. Furthermore, 26 of the 28 EpiCare centres have indicated their willingness to participate to the related Human Intracerebral EEG Platform (HIP) in SGA3. Given the link envisioned between HIP and MIP, we anticipate that the same centres shall also install the MIP in the near future.

5.1.3 Output 2: Clinical Data Catalogue

This Output is described in Output 4 under KR8.2 as it contributes to growing the MIP network.

It also contributes to KR8.3 by facilitating data provision to the MIP, especially from the networks of hospitals specialising in major brain disorders at stake (i.e. epilepsy, dementia, traumatic brain injury, mental health), and by allowing all types of MIP users to refer to the Data Catalogue for the MIP dataset's metadata definitions.

5.1.4 Output 3: MIP datamodels for several brain disorders

The MIP has implemented datamodels for several brain disorders (dementia, TBI, epilepsy, mental health). These were elaborated by experts of these conditions responsible for the corresponding SP8 Use Cases. The data models are available in the Data Catalogue and their descriptions and links to their Knowledge Graph cards are accessible at:

Dementia:
https://kg.ebrains.eu/search/instances/Dataset/42b69ce8-9a73-4522-9be9-07f1ea8fb60d
Depression:
https://kg.ebrains.eu/search/instances/Dataset/5390fd8f-b095-4963-addc-f984df119265
Mental Health:
https://kg.ebrains.eu/search/instances/Dataset/e5aad7e8-b42d-4cd2-8cac-90998a4dd6ff
TBI: https://kg.ebrains.eu/search/instances/Dataset/75b01958-b59a-42a1-9e1d-326532beabee

5.1.5 Output 4: Clinical datasets from several brain disorders (C2978, C2979)

Various datasets from populations affected by dementia, TBI, mental health and epilepsy have been processed, curated and integrated into various MIP instances (both local and federated nodes). Dementia datasets placed in MIP federated nodes include those from Lausanne CLM, Lille and Brescia as well as the ADNI, EDSD and PPMI public databases. The TBI datasets are those from the EU-funded centre-TBI and CREACTIVE cohorts from Karolinska and Bergamo hospital, respectively. Mental health data are those from the Imagen cohort secured by the network of the hospitals UKAachen, MPIP Munich and Charite and distributed over the MIP federated nodes of Aachen and CHUV. Two epilepsy databases are installed on MIP instances from Grenoble and Lausanne. We are aware of other datasets placed on MIP local (e.g. Plovdiv). These represent a total of more than 20,000 patients.

The dataset descriptions are available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Datasets/.

Table 7: Output 4 Links

Component	Link to	URL
C2978	Data Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-collab-for-sga2-m24-report-supportin/WPs%20chapter/WP8.3/Datasets/
	Technical Documentation	-
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-collab-for-sga2-m24-report-supportin/WPs%20chapter/WP8.3/Datasets/
C2979	Data Repository	https://wiki.ebrains.eu/bin/view/Collabs/sp8-collab-for-sga2-m24-report-supportin/WPs%20chapter/WP8.3/Datasets/
	Technical Documentation	-
	User Documentation	https://wiki.ebrains.eu/bin/view/Collabs/sp8-collab-for-sga2-m24-report-supportin/WPs%20chapter/WP8.3/Datasets/

5.1.6 Output 5: Use Case Dementia

The Use Case Dementia, *"Advanced phenotyping of the ageing-brain cognitive diseases at early stage"*, showcases research performed using the MIP-in the field of dementia using data from several clinical centres as well as relevant public databases. The report is available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Use%20Cases/UC%20Dementia/.

5.1.7 Output 6: Use Case Epilepsy

The Use Case Epilepsy, *"Federated Analysis of large-scale intracerebral EEG data from patients with epilepsy"* developed tools to process and extract features from Human intracerebral EEG recordings to be further analysed with the MIP, including a BIDS pipeline that was incorporated into the MIP Data Factory. This UC also provides data from the F-TRACT project to be integrated into the BigBrain Atlas of EBRAINS. The report is available at : https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Use%20Cases/UC%20Epilepsy/.

5.1.8 Output 7: Shareable data format for intracranial EEG (iEEG)

A sharable data format for Human intracranial EEG (iEEG), which facilitates data sharing between hospitals was published in its final version at M15, see P2033 below under Publications. This format is based on the BIDS (Brain Imaging Data Sharing, http://bids.neuroimaging.io) initiative already

developed for MRI and MEG. An international group of experts of intracranial data, to which UGA and UCBL belong, defined BIDS for iEEG with full compatibility for WP8.8 objectives.

5.1.9 Output 8: BIDS iEEG manager (C3068, rel.1)

BIDS-iEEG manager has been developed by AMU to efficiently manage BIDS-iEEG databases (https://github.com/Dynamap/BIDS_Manager). Data are collected in raw format (e.g. DICOM for Anatomical and Micromed for iEEG) and converted in format specified by BIDS and BIDS-iEEG.

BIDS-iEEG manager has been implemented at CHUGA by UGA for the HBP Medical Informatics Platform based on the BIDS Pipeline Tool (Output 8). It is also used by AMU in other multi-centre studies such as RHU EPINOV and PHRC SPREAD.

In addition, a BIDS-iEEG extractor of the F-TRACT database has been developed by AMU and UGA to create the BIDS-iEEG CCEP database (see Output 10 below).

Table 8: Output 8 Links

Component	Link to	URL
C3068	Software Repository	https://github.com/Dynamap/BIDS_Manager
	Technical Documentation	in process of writing
	User Documentation	user is informed on-line during software execution

5.1.10 Output 9: BIDS iEEG pipeline (C3068, rel.2)

BIDS Pipeline is an extension of BIDS Manager and a new module of the MIP data factory which allows to launch process on BIDS dataset. Different software can be integrated to this BIDS Pipeline, either locally created or shared by the neuroscience community. It has been developed by AMU and used/tested at UGA. Bids pipeline allows to do Big Data analyses on Bids dataset.

UGA has used Bids Pipeline to implement a local MIP for SEEG at CHUGA.

Table 9: Output 9 Links

Component	Link to	URL
C3068	Software Repository	https://github.com/Dynamap/BIDS_Manager
	Technical Documentation	in process of writing
	User Documentation	user is informed on-line during software execution

5.1.11 Output 10: F-TRACT CCEP database (C3067)

A BIDS-iEEG extractor of the F-TRACT database has been developed by AMU and UGA to produce the fully anonymised CCEP database (data of 217 patients included). It comprises a fully anonymised extract of a part of the F-TRACT database on intracortical stimulations in epileptic patients implanted with SEEG (stereotactic electroencephalography) electrodes (project funded by ERC 2014-2019, https://f-tract.eu/) in BIDS-iEEG format. This database can thus contribute to the development and validation of human brain atlases and modelling by bringing relevant information from CCEPs. The DOI of the database is 10.25493/SV5Z-FSB. It is available in in the Knowledge Graph at https://kg.ebrains.eu/search/instances/Dataset/ebe50517-41d5-4029-9355-04f1e49e23c8 . The database is documented in SGA2 Deliverable D8.8.2 (D55.2 D125).

This dataset has been produced by UGA, CHUGA, AMU, CERCE and UCBL and has been used by UKLFR, AMU, UGA, UCBL and CNR to develop and the software integrated in the SEEG MIP.

Table 10: Output 10 Links

Component	Link to	URL
C3067	Data Repository	https://kg.ebrains.eu/search/instances/Dataset/ebe50517-41d5-4029-9355-04f1e49e23c8
	Technical Documentation	-
	User Documentation	-

5.1.12 Output 11: Android mobile app for multimodal data acquisition

The Android app for the acquisition of multimodal data, oriented to the follow-up of patients with epilepsy, has been updated to version 1.0, solving many stability issues in the connectivity with different commercial wearable devices. With respect to the state of the art, there are currently no open-source solutions like this, that allow an easy customisation for new use cases with minor development efforts.

The documentation for the app is at https://c4science.ch/diffusion/9555/ (software repository) and https://c4science.ch/diffusion/9555/ (technical doc.).

5.1.13 Output 12: Systematic assessment of novel data type integration (C2975, C2976)

The C2975 document (https://infoscience.epfl.ch/record/276487?&ln=en) summarises from a technical perspective the main challenges faced by the current MIP for the integration of three new types of data: complex neuroimaging, omics, and data from wearable devices. Details are discussed up to the implementation level, considering risks, costs, and possible ethical and data privacy issues.

On the other hand, the C2976 document (https://infoscience.epfl.ch/record/276489?&ln=en) collects the specific technical recommendations for the future development of the MIP derived from the potential integration of new types of data into the Platform. It begins with a high-level description of the MIP architecture and of its different software components, and then the evolution of each of them is analysed from a technical perspective.

Table 11: Output 12 Links

Component	Link to	URL
C2975	Report Repository	https://infoscience.epfl.ch/record/276487?&ln=en
	Technical Documentation	https://infoscience.epfl.ch/record/276487?&ln=en
	User Documentation	https://infoscience.epfl.ch/record/276487?&ln=en
C2976	Report Repository	https://infoscience.epfl.ch/record/276489?&ln=en
	Technical Documentation	https://infoscience.epfl.ch/record/276489?&ln=en
	User Documentation	https://infoscience.epfl.ch/record/276489?&ln=en

5.1.14 Output 13: Use Case Post Traumatic Stress Disorder (PTSD)

The objective of this Use Case "Potential Neurocognitive Biomarkers for PTSD Severity in Recent Trauma Survivors" is to computationally derive potential biomarkers that could efficiently differentiate PTSD subtypes, based on an observational cohort study of recent trauma survivors. A three-staged semi-unsupervised method ("3C") was used to categorise trauma survivors based on current PTSD diagnostics, derive clusters of severe and mild PTSD based on features related to symptom load, and to classify participants' cluster membership using objective features. A total of

256 features were extracted from psychometrics, cognitive, structural and functional neuroimaging data. Multi-domain biomarkers revealed by the 3C analytics offer objective classifiers of post-traumatic morbidity shortly following trauma, and also map onto previously documented neurobehavioural PTSD features, supporting the future use of standardised and objective measurements to more precisely identify psychopathology subgroups shortly after trauma. The UC report is available at : https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Use%20Cases/Use%20Case%20PTSD/, and a pre-print of this work on BioRxiv: https://www.biorxiv.org/content/10.1101/721068v2, see also P1954. The full article is currently under second revision before being published.

5.2 Validation and Impact

5.2.1 Actual and Potential Use of Output(s)

Output 1 *"EpiCare partnership"*: EpiCare is one of the two EU-funded European Reference Networks (ERN) dedicated to brain diseases. ERNs represent the most important EU-funded initiative in the field of health, covering 24 categories of rare and complex diseases. Each of the 24 consortia gathers about 30 public health care providers across Europe, representing the most expert centres in the corresponding medical conditions. EpiCare includes 28 such centres focusing on rare and complex epilepsies, including those considered for intracerebral EEG pre-surgical investigation and epilepsy surgery, a population highly relevant for several core developments in HBP (clinical applications of The Virtual Brain, HIP platform). The signed partnership between HBP and EpiCare offers unique opportunity to deploy the MIP and share data within a strong network of EU hospitals. It will also enable to develop the Human intracerebral EEG platform planned in SGA3.

Output 2 *"Clinical Data Catalogue"* enables the MIP data providers to manage metadata of their datasets according to pre-defined and validated data models.

Outputs 3 *"MIP datamodels"* and 4 *"Clinical datasets"*, provide the material to show case the usefulness of the MIP across various research networks and conditions, which are illustrated by Outputs 5 "*Use Case Dementia*" and 6 "*Use Case Epilepsy*".

The Outputs 7 *"Shareable BIDS iEEG data format"*, 8 "*BIDS-iEEG manager*", 9 *"BIDS-iEEG pipeline"* and 10 "*CCEP database"* are participating to the development of the MIP epilepsy use case and will also be used for the Human Intracerebral Platform to be developed in SGA3. Output 10 *"CCEP database"* is available on EBRAINS Knowledge Graph at https://kg.ebrains.eu/search/instances/Dataset/ebe50517-41d5-4029-9355-04f1e49e23c8.

Output 8 *"BIDS-iEEG manager"* has been selected by the OHBM (Organization for Human Brain Mapping) 2020 Program Committee for a software demo in Montreal in June 2020.

Output 11 *"Android mobile app"* was initially developed within a pilot study for the integration of wearable data into the MIP during SGA2. This pilot was cancelled during the second year of the Project due to resources reallocation, but the app has been successfully adopted by external projects, including: 1) The DeepHealth Project (H2020 GA No. 825111, https://deephealth-project.eu/) for monitoring patients with multiple sclerosis; and 2) the MyPreHealth project (Hasler Foundation project No. 16073) for monitoring patients with migraine. The app is expected to be used to capture monitoring data from around 60 patients in the next 6 months. These data may be of future interest for the MIP.

Output 12 *"Systematic assessment of novel data types"* has two main purposes: 1) to help in the decision-making process for the next major steps in the development of the MIP Platform regarding the integration of new types of data, and 2) to define the main architectural changes that will be necessary to implement from the software development perspective. It therefore naturally connects to KR8.3.

The following Deliverables and Milestones served as validation measures: *Signed partnership with one European reference network* [D8.3.1 (D50.1 D7)], *Intermediate report of the "ageing-brain*

cognitive disease" use-case [MS8.3.5, *Data from case-control studies linked to MIP local* [MS8.10.3], *CCEP database* [D8.8.2 (D55.2 D125)].

Together with the effective delivery and deployment of a robust infrastructure compliant with EU ethics and data privacy/security regulatory requirements (KR8.1, KR8.2), KR8.3 offers the third pillar to MIP's successful development and usage, i.e. its implementation over active clinical research networks.

Output 13 "*Use Case Post Traumatic Stress Disorder*", in addition to advancement in the field of disease signatures for PTSD, is a very timely work, as with the disruption around the world due to COVID-19 pandemic we foresee increasing incidence of PTSD, for which this Use Case has a great potential importance.

5.2.2 Publications

Output 7

- Holdgraf C, Appelhoff S, Bickel S, Bouchard K, D'Ambrosio S, David O, Devinsky O, Dichter B, Flinker A, Foster BL, Gorgolewski KJ, Groen I, Groppe D, Gunduz A, Hamilton L, Honey C, Jas M, Knight R, Lachaux JP, Lau JC, Lundstrom BN, Miller KJ, Ojemann JG, Oostenveld R, Petridou N, Piantoni G, Pigorini A, Pouratian N, Ramsey NF, Stolk A, Swann N, Tadel F, Voytek B, Wandell BA, Winawer J, Zehl L, Hermes D. BIDS-iEEG: an extension to the brain imaging data structure (BIDS) specification for human intracranial electrophysiology. Sci Data. 2019 Jun 25;6(1):102. (P2033)

  Significance: The importance of the shareable BIDS iEEG data format is underlined by the favourable metrics on the article ranking, as per the publisher website (https://www.nature.com/articles/s41597-019-0105-7/metrics): 2,592 article accesses, 3 citations in Web of Science and CrossRef. Altmetric score is 83. This article is in the 97th percentile (ranked 7,190th) of the 257,813 tracked articles of a similar age in all journals.

6. KR8.4 MIP analytical tools enable federated analyses of multidimensional longitudinal data for advanced biological disease signature

6.1 Outputs

6.1.1 Overview of Outputs

6.1.1.1 List of Outputs contributing to this KR

- Output 1: Guidelines on model validation.
- Output 2: Guidelines on analysis of longitudinal data.
- Output 3: Algorithms for federated longitudinal analysis (Kaplan Meier) (C3000)

6.1.1.2 How Outputs relate to each other and the Key Result

Outputs 1 and 2 are independent, but interrelated: the methods for model validation are relevant to the analysis of longitudinal data and contribute to the development of tools for federated analyses of multidimensional longitudinal data.

Output 3 provides a survival analysis algorithm (Kaplan–Meier), broadly used in longitudinal analysis.

Output 4 will enable the processing of very long sequences of longitudinal ECG signals while reducing by an order of magnitude the amount of data to capture, process, transmit and store.

All of the above Outputs contribute to KR8.4.

6.1.2 Output 1: Guidelines on model validation

The guidelines on model validation address problems that arise when federating multidimensional and longitudinal data. The fact that data are nested within subjects requires more careful application of conventional validation methods. The document is available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Guidelines%20for%20model%20validation%20-%20C3054%20Statistical%20assessment%20of%20predictive%20capability/.

6.1.3 Output 2: Guidelines on analysis of longitudinal data

The guidelines on statistical tools for the analysis of longitudinal data summarise the leading methods and models for such research. The document is available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Guidelines%20on%20analysis%20of%20longitudinal%20data/ and was used to validate the milestone MS8.4.2.

6.1.4 Output 3: Algorithms for federated longitudinal analysis (Kaplan Meier) (C3000)

This Output is another facet of a part of the Federated / distributed data processing engine (C3000) described in Output 3 under KR8.1 and in Output 3 under KR8.2.

With respect to KR8.4, this Output contributes with Kaplan Meier algorithm, which is tailored for longitudinal survival analysis and has been integrated into the newest release (MIP 6.0) of the MIP.

Table 12: Output 3 Links

Component	Link to	URL
C3000	Software Repository	https://github.com/madgik/exareme/releases/tag/22.0.0
	Technical Documentation	https://github.com/HBPMedical/mip-federation/tree/master/Documentation
	User Documentation	-

6.2 Validation and Impact

6.2.1 Actual and Potential Use of Output(s)

Both Output 1 *": Guidelines on model validation"* and Output 2 *"Guidelines on analysis of longitudinal data"* contribute to the HBP infrastructure by providing up-to-date information on common concerns in analysing federated data. They have not had direct use but have potential value to many groups conducting neuroscience research. Model validation is an issue in almost every empirical research endeavour, so these guidelines have potential for very widespread use. Longitudinal data is common in neurological diseases, with subjects tracked over time to observe trends in motor and cognitive functioning; such research can benefit from the guidelines we have prepared.

Output 3 *"Algorithms for longitudinal analysis"* is now used to explore potential for longitudinal analysis with current available datasets, and to establish the roadmap for adoption.

The Outputs were internally validated by their teams. Output 2 is also documented by *Validation of multi-domain model across hospitals* [MS8.4.2]

6.2.2 Publications

None.

7. KR8.5 MIP performs advanced automatised extraction of clinical relevant data from hospitals electronic health records (EHR)

7.1 Outputs

7.1.1 Overview of Outputs

7.1.1.1 List of Outputs contributing to this KR

- Output 1: Automated clinical data extraction pipeline (C2998)
- Output 2: Brain Disease Ontologies (3071)
- Output 3: OLS-Neuro (C3072)
- Output 4: SCAIView-Neuro (3073)

7.1.1.2 How Outputs relate to each other and the Key Result

Output 1 contributes directly to KR8.5 by providing the configuration to our Data Factory pipeline to adapt it to one of the most frequently used electronic Clinical Report Form (eCRF), REDCap. REDCap is installed in more than 4,000 institutions and has been used to collect and structure research data in about 900,000 projects worldwide (https://www.project-redcap.org/). It thus contributes to KR8.5 by automating data extraction.

Output 2 provides curated ontologies for brain disease that are relevant to the HBP. The ontologies comprise the majority of relevant entities and their relationships that are necessary to describe data and knowledge about brain diseases. The ontologies are essential for creating datamodels and harmonising data (KR8.3). The developed ontologies are available within an ontology store, which is part of Output 3. It allows end users to browse and search these resources.

Output 4 provides a dedicated literature mining environment for brain research, which may be used in the future for automated extraction of data from textual clinical records.

7.1.2 Output 1: Automated clinical data extraction pipeline (C2998)

Hospitals use a variety of information systems and databases to store and manage their data. The heterogeneity in data sources is the biggest impediment to a global approach for data integration and import.

We designed, created and incorporated in our pipeline a REDCap data exporter that facilitates the task of exporting data from REDCap, for further processing and integration into the MIP. In addition, we took advantage of the flat nature of all data derived from REDCap and facilitated the schema

mapping configuration, which is essential to the execution of the whole data integration process. The behaviour of the pipeline and its containing tools has been proven to be consistent and expected as long as the mapping task configurations are designed correctly.

The documentation is available at https://github.com/aueb-wim/ansible-datafactory (user and technical doc.)

Table 13: Output 1 Links

Component	Link to	URL
C2998	Software Repository	https://github.com/aueb-wim/ansible-datafactory/releases/tag/v.1.0
	Technical Documentation	https://github.com/aueb-wim/ansible-datafactory
	User Documentation	https://github.com/aueb-wim/ansible-datafactory

7.1.3 Output 2: Brain Disease Ontologies (3071)

Four brain-specific ontologies have been developed to help structuring and curating clinical data to be extracted from EHR. They are:

- Epilepsy ontology, which represents a semantic assembly of structured knowledge on various aspects of epilepsy.
- Schizophrenia ontology, which represents a semantic assembly of structured knowledge on various aspects of schizophrenia.
- AMDP-System terminology: The AMDP-System allows the documentation of psychopathological reports in a standardised manner in the psychiatric field. In this work, the terminology has been extracted from the AMDP-System and encoded in a structured OWL (Ontology Web Language) file. Mining of EHR documents, concretely the psychopathological reports, has been defined as an application case for the usage of the terminology.
- Clinical Trial NDD-specific Ontology (CTO-NDD): The neurodegeneration disease-specific version of clinical trial ontology has been revised.

The ontologies are accessible from the MIP and from the OLS-NEURO service, see below.They are documented in SGA2 Deliverable D8.9.2 (D56.2 D126).

Table 14: Output 2 Links

Component	Link to	URL
C3071	Report Repository	D8.9.2 (D56.2 D126) Curated and tested brain disease ontologies integrated within application services -
	Technical Documentation	-
	User Documentation	-

7.1.4 Output 3: OLS-Neuro (C3072)

The application service OLS-NEURO serves as an ontology and terminology hub for neurodegenerative disease research. The service is accessible under http://rohan.scai.fraunhofer.de/ols/

The documentation is available under https://www.ebi.ac.uk/ols/docs/index (user doc.) and https://github.com/EBISPOT/OLS/blob/master/README.md (technical doc.).

Table 15: Output 3 Links

Component	Link to	URL
C3072	Data/Model/Software/Report Repository	http://rohan.scai.fraunhofer.de/ols/
	Technical Documentation	https://github.com/EBISPOT/OLS/blob/master/README.md

	User Documentation	https://www.ebi.ac.uk/ols/docs/index

7.1.5 Output 4: SCAIView-Neuro (C3073)

The application service SCAIView-NEURO allows to retrieve semantically-enhanced information from scientific literature. The service is accessible to authorised users at https://ui.scaiview.com/

Table 16: Output 4 Links

Component	Link to	URL
C3073	Software Repository	https://neuro.scaiview.com/
	Technical Documentation	Technical documentation is available in code, only available for developers.
	User Documentation	In the process of writing. It will be available on the platform soon.

7.1.6 Validation and Impact

7.1.7 Actual and Potential Use of Output(s)

Output 1 *"Automated clinical data extraction pipeline"* has been thoroughly tested with data exported from REDCap data sources. We have now the knowledge and the framework to process and import data from such sources quickly and efficiently and feed them into the Data Factory pipeline. Configuring our Data Factory pipeline so as to process data coming from popular storage solutions is a step towards standardising hospital data ingestion into the MIP. Especially in view of the partnership with the EpiCare network, which keeps its data in REDCap, will Output 1 be of major importance and clearly shows the relation with KR8.3.

We have utilised Outputs 2, 3, 4 (all ontology-related) for data annotation to prepare training datasets for machine learning algorithms in EHR text mining. The textual data has been annotated with the concepts from ontology. Furthermore, the ontology has been also used to perform a data-mining task. In the future, these ontology-related outputs can be used for various purposes such as data integration, data annotation, data curation, data harmonisation, data mining and more.

Outputs 2, 3, and 4 have together been validated by *"Curated and tested brain disease ontologies integrated within application services"* [SGA2 Deliverable D8.9.2 (D56.2 D126)]

7.1.8 Publications

None.

8. Main Outputs Not Directly Linked to KR

8.1 Outputs

8.1.1 Overview of Outputs

8.1.1.1 List of Outputs contributing to this chapter

- Output 1: Analysis of predictive markers in PTSD.
- Output 2: Analysis of genetic markers in Parkinson's Disease (PD).

- Output 3: Statistical methods for computer experiments.
- Output 4: Methods for clustering trees for multi-label classification.
- Output 5: Ensembles for multi-target regression.
- Output 6: Defining disease signatures.
- Output 7: Application of emulator methodology to neuroscience simulation.
- Output 8: Guidelines for clustering.
- Output 9: European health research and innovation cloud.
- Output 10: Identifying psychiatric conditions from fMRI images.
- Output 11: Markers for early stage in Alzheimer's Disease.
- Output 12: Neuro-clinical signatures following acute stroke.
- Output 13: Neuro-clinical signatures of language impairments.
- Output 14: Brain remodeling in temporal lobe epilepsy.

8.1.1.2 How Outputs relate to each other

The Outputs described here fall naturally into 4 groups.

1) Outputs 1, 2, 6, 10, 11, 12, 13 and 14 – related to all others, as they can use their conclusions: Predictive markers have been analysed in PTSD, PD, AD, psychiatric disorders, epilepsy, language impairments and acute stroke. All the analyses applied advanced techniques for multivariate data, including clinical, genetic and imaging features. All studies are concerned with the development of disease signatures, and thus relate to the article that summarises the literature on the meaning and use of this term.
2) Outputs 4, 5, 8 – related to the first group, as they can be used as analysis methods: Methods have been developed using clustering techniques to address problems in multi-label classification and in multi-target regression. Here similar ideas guide the methodology for both problems. Pseudo-code was provided for the clustering-based 3C strategy.
3) Outputs 3, 7– related to the first group, as they can be used as analysis methods: The use of statistical methods for computer experiments has been explored. These methods are summarised in a review paper and then applied to neuroscience simulation.
4) Output 9 – related to the first group, as they have implications for data analysis: A roadmap was proposed for GDPR-compliant sharing and analysis of medical data. The other Outputs develop methodology for data analysis and/or analysis of medical data, so that the roadmap is an umbrella that covers them all.

8.1.2 Output 1: Analysis of predictive markers in PTSD

Our analysis identifies prognostic markers for Post-Traumatic Stress Disorder (PTSD), combining features from clinical, imaging and functional domains. The 3C strategy was applied to identify clusters of patients and relate them to the markers. The research demonstrates the ability of multi-domain analytic assessment using standardised and objectively measured neuro-behavioural features to differentiate PTSD subgroups at the early aftermath of traumatic events.

8.1.3 Output 2: Analysis of genetic markers in PD

We demonstrate application of a data-driven analytical approach to relate mutations in the LRRK2 and GBA genes to disease phenotype in a cohort of Parkinson's Disease patients. The analysis exploits the false discovery rate to screen a large collection of features while controlling error rates. The

results portray a possible unique disease phenotype based on genotype among patients with these mutations. The findings could help direct a more personalised therapeutic approach.

8.1.4 Output 3: Statistical methods for computer experiments

This Output summarises advanced statistical techniques developed for collecting and analysing data from computer simulation platforms and shows how they can be applied in the context of neuroscience.

8.1.5 Output 4: Methods for clustering trees for multi-label classification

We developed an algorithm developed for learning option predictive clustering trees (OPCTs) for multi-label classification, based on the predictive clustering framework.

8.1.6 Output 5: Ensembles for multi-target regression

Multi-target regression concerns problems with multiple outputs that require prediction. We developed methods for the prediction problem that exploit ensembles of predictive clustering trees (PCTs), using ideas from bagging and random forests. The methods are tested on a variety of different real examples. The results are available at : https://zenodo.org/record/3715018.

8.1.7 Output 6: Defining disease signatures

We provide a comprehensive review of the concept of "disease signature". The scientific literature shows that this term has not been properly defined, leading to inconsistency and confusion. A novel hierarchical multidimensional concept is suggested for this term that would combine both current approaches for identifying diseases (one focusing on undesired effects of the disease and the other on its causes). This model could lead to developing statistical confidence in a disease signature that would allow physicians/patients to estimate the precision of the diagnosis, which, in turn, may have important implications for patients' prognosis and treatment. The results are described at: http://dx.doi.org/10.1007/s12031-019-01269-0 .

8.1.8 Output 7: Application of emulator methodology to neuroscience simulation

Emulators are statistical surrogates that replace complex simulation models with empirical predictors. When simulator runs require extensive resources, which is often the case in neuroscience, the use of an emulator has the potential to accelerate exploration of input parameter spaces. We apply the emulator methodology in the context of multi-objective optimisation for tuning a simulation model to be consistent with observed laboratory data. The results are available at: https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Application%20of%20emulator%20methodology%20to%20neuroscience%20simulation/.

8.1.9 Output 8: Guidelines for clustering

Clustering is a useful method for identifying and characterising commonalities and subgroups in multidimensional data. We show how to approach common challenges in clustering using predictive clustering trees. The results are available at : https://wiki.ebrains.eu/bin/view/Collabs/sp8-public-collab-for-sga2-m24-compound-/Guidelines%20for%20clustering/

8.1.10 Output 9: European health research and innovation cloud

Creating a European Health Research and Innovation Cloud (HRIC) within the framework of the European Union (EU) initiative on the Digital Transformation of Health and Care should enable GDPR compliant data sharing and analysis for health research across the EU. We describe the vision and expected benefits of digital data sharing in health research activities and presents a roadmap for doing so. The roadmap proposal is built around five specific recommendations and action points.

8.1.11 Output 10: Identifying psychiatric conditions from fMRI images

We applied unsupervised machine learning to identify brain signatures from three principal components based on activations from three kinds of diagnostically relevant stimuli. The components are the basis for cross-validation markers for clinical populations and eventually delineate diagnostic and classification groups. The markers separate the two investigated clinical entities – schizophrenia and recurrent depression. The analysis identified the three brain patterns that summarised all the individual variabilities of the individual brain patterns. The outcome confirms the possibility to use brain signatures to achieve bottom-up classification of mental disorders.

8.1.12 Output 11: Markers for early stage in Alzheimer's Disease

The multifaceted nature of Alzheimer's Disease (AD) and Mild cognitive impairment (MCI) can lead to wide inter-individual differences in disease manifestation in terms of brain pathology and cognition. The lack of understanding of phenotypic diversity in AD arises from a difficulty in understanding the integration of different levels of network organisation (i.e. genes, neurons, synapses, anatomical regions, functions) and in inclusion of other information such as neuropsychiatric characteristics, personal history, general health or subjective cognitive complaints in a coherent model. Non-cognitive factors, such as personality traits and behavioural and psychiatric symptoms, can be very informative markers of early disease stage. It is known that personality can affect cognition and behavioural symptoms. In this Output we reviewed the different types of interactions existing between personality, depression/anxiety and cognition and cognitive disorders at behavioural and brain/genetic levels in order to contribute to understanding of phenotypic diversity in AD.

8.1.13 Output 12: Neuro-clinical signatures following acute stroke

Ischemic stroke affects language production and/or comprehension and leads to devastating long-term consequences for patients and their families. We investigated the use of Voxel-Based Morphometry (VBM), multivariate modelling and native Computed Tomography (nCT) scans routinely acquired in the acute stage of stroke for identifying biological signatures that explicate the relationships between brain anatomy and types of impairments. Individual patient's nCT scans were compared to a group of controls. Consistently, the regions that presented significant difference GM and WM values overlapped with known areas that support language processing. The method applied to nCT scans provided robust and accurate information about brain lesions' location and size, as well as quantitative values. We found that nCT and VBQ analyses are effective for identifying neural signatures of concomitant language impairments at the individual level, and neuroanatomical maps of aphasia at the population level. Analyses with larger cohorts could lead to a more integrated multimodal model of behaviour and brain anatomy in the early stage of ischemic stroke.

8.1.14 Output 13: Neuro-clinical signatures of language impairments

There are complex and combinatorial relationships between language functions and anatomical brain regions. We reviewed recent attempts to understand the underlying principles of this complex mapping, which is important for the identification of the brain signature of language. We also described Neuro-Clinical signatures that explain language impairments and predict language recovery after stroke. The different concepts of mapping (from diffeomorphic one-to-one mapping to many-to-many mapping) have been introduced and they provide a basis for deriving a theoretical framework that describes the current principles of brain architectures including redundancy, degeneracy, pluri-potentiality and bow-tie network.

8.1.15 Output 14: Brain remodelling in temporal lobe epilepsy

Our aim was to test for the hypothesised bidirectional pattern of epilepsy-associated brain remodelling in the context of the presence and absence of mesial temporal lobe sclerosis. Using MRIs from a large cohort of mesial temporal lobe epilepsy patients, we compared patients with or without hippocampus sclerosis and healthy controls, testing for common and differential brain patterns. The main effect of disease was associated with continuous hippocampus volume loss ipsilateral to the seizure onset zone in both temporal lobe epilepsy cohorts. The post hoc tests demonstrated bilateral hippocampus volume increase in the early epilepsy stages in patients without hippocampus sclerosis. Early age of onset and longer disease duration correlated with volume decrease in the ipsilateral hippocampus. The findings of seizure-induced hippocampal remodelling are associated with specific patterns of mesial temporal lobe atrophy. Directionality of hippocampus volume changes strongly depends on the chronicity of disease. Specific anatomy differences represent a snapshot within a progressive continuum of seizure-induced structural remodelling.

8.2 Validation and Impact

8.2.1 Actual and Potential Use of Output(s)

The analyses of PD and PTSD (Outputs 1, 2 and 6) point to novel biomarkers and have potential to guide, and even to personalise, treatment. The results from analysis of stroke, AD and epilepsy ((Outputs 10, 11, 12, 13 and 14) highlight the usefulness of the Panomic model, which can be applied to additional diseases.

Multi-label classification and multi-target regression are problems that arise in a great deal of neuroscience research. Thus the methods developed here (Outputs 4, 5 and 8) have great potential for application.

Statistical methods developed for use with computer simulation platforms have not been widely used in HBP research. They have the potential to accelerate research using these platforms by effective schemes for data collection and emulation of slow and expensive simulators. The review paper (Output 3) and subsequent application (Output 7) show the benefits of employing these methods.

The Deliverable *"Reports on disease signatures via sub-type identification, on development and application of methods, and tools"* [SGA2 Deliverable D8.4.1 (D51.1 D24)] contributes to validation of the above Outputs.

Output 9, involving numerous institutions already, has a potential to contribute to defining the EC policy on digital data sharing in health research activities.

8.2.2 Publications

Output 1: Analysis of predictive markers in PTSD -–(P1954): Ziv Ben-Zion; Yoav Zeevi; Nimrod Jackob Keynan; Roee Admon; Haggai Sharon; Pinchas Halpern; Israel Liberzon; Arieh Y. Shalev; Yoav Benjamini; Talma Hendler, "Multi-Domain Potential Biomarkers for Post-Traumatic Stress Disorder (PTSD) Severity in Recent Trauma Survivors", in *Biorxiv*, 2019

Output 2: Analysis of genetic markers in PD - (P1953): Tal Kozlovski; Alexis Mitelpunkt; Avner Thaler; Tanya Gurevich; Avi Orr-Urtreger; Mali Gana-Weisz; Netta Shachar; Tal Galili; Mira Marcus-Kalish; Susan Bressman; Karen Marder; Nir Giladi; Yoav Benjamini; Anat Mirelman, "Hierarchical Data-Driven Analysis of Clinical Symptoms Among Patients With Parkinson's Disease", in *Frontiers in Neurology*, Vol. 10, p.531, 2019.

Output 3: Statistical methods for computer experiments – (P1952): Gilad Shapira; Ella Shaposhnik; David M. Steinberg, "Statistical Methods for Computer Experiments with Applications in Neuroscience", *preprint*.

Output 9: European health research and innovation cloud – (P2400): F. M. Aarestrup; A. Albeyatti; W. J. Armitage; C. Auffray; L. Augello; R. Balling; N. Benhabiles; G. Bertolini; J. G. Bjaalie; M. Black; N. Blomberg; P. Bogaert; M. Bubak; B. Claerhout; L. Clarke; B. De Meulder; G. D'Errico; A. Di Meglio; N. Forgo; C. Gans-Combe; A. E. Gray; I. Gut; A. Gyllenberg; G. Hemmrich-Stanisak; L. Hjorth; Y. Ioannidis; S. Jarmalaite; A. Kel; F. Kherif; J. O. Korbel; C. Larue; M. Laszlo; A. Maas; L. Magalhaes; I. Manneh-Vangramberen; E. Morley-Fletcher; C. Ohmann; P. Oksvold; N. P. Oxtoby; I. Perseil; V. Pezoulas; O. Riess; H. Riper; J. Roca; P. Rosenstiel; P. Sabatier; F. Sanz; M. Tayeb; G. Thomassen; J. Van Bussel; M. Van den Bulcke; H. Van Oyen, "Towards a European health research and innovation cloud (HRIC)", in *Genome Medicine*, Vol. 12, No. 1, 2020.

Output 10: Identifying psychiatric conditions from fMRI images - (P2387): Drozdstoy Stoyanov; Sevdalina Kandilarova; Rositsa Paunova; Javier Barranco Garcia; Adeliya Latypova; Ferath Kherif, "Cross-Validation of Functional MRI and Paranoid-Depressive Scale: Results From Multivariate Analysis", in *Frontiers in Psychiatry*, Vol. 10, p.869, 2019.

Output 11: Markers for early stage in Alzheimer's Disease – (P2469 – *in validation process*): Valérie Zufferey; Armin von Gunten; Ferath Kherif, "Interactions between personality, depression/anxiety and cognition to understand early stage of Alzheimer's Disease", in *Current Topics in Medicinal Chemistry*, Vol. 20, p.782–791, 2020.

Output 12: Neuro-clinical signatures following acute stroke – (P2386 – *in validation process*): Sandrine Muller; Kaisar Dauyey; Anne Ruef; Sara Lorio; Ashraf Eskandari; Laurence Schneider; Valérie Beaud; Elisabeth Roggenhofer; Bogdan Draganski; Patrik Michel; Ferath Kherif, "Neuro-Clinical signatures of language impairments after acute stroke: a VBQ analysis of quantitative native CT scans", in Current Topics in Medicinal Chemistry, Vol. 20, p.792–799, 2020.

Output 13: Neuro-clinical signatures of language impairments – (P2468): Ferath Kherif; Sandrine Muller, "Neuro-Clinical signatures of language impairments: A theoretical framework for function-to-structure mapping in clinics", in *Current Topics in Medicinal Chemistry*, Vol. 20, p.800–811, 2020.

Output 14: Brain remodeling in temporal lobe epilepsy – (P2467): Elisabeth Roggenhofer; Emiliano Santarnecchi; Sandrine Muller; Ferath Kherif; Roland Wiest; Margitta Seeck; Bogdan Draganski, "Trajectories of brain remodeling in temporal lobe epilepsy", in *Journal of Neurology*, Vol. 266, No. 12, p.3150–3159, 2019.

9. Conclusion and Outlook

1) Achievement vs work plan

MIP consolidation: Two new versions of the MIP (5.0, 6.0) were released during year 2. These represent very significant evolutions from the previous versions inherited from SGA1, which suffered serious issues. Most importantly, the privacy-preserving model of the MIP was fully integrated, facilitating deployment and usage of the MIP. As a result, a number of novel end users' needs were delineated, in particular regarding the management of datasets covering different brain disorders within the same federation. MIP algorithms underwent comprehensive validation, while new analytical tools were integrated, including the Calibration Belt required for the TBI use case and Kaplan-Meier estimator for longitudinal data analysis. Finally, the functionality of the MIP data factory was expanded with the data catalogue, the BIDS-SEEG pipeline and the REDCap data extractor. Overall, thanks to the strong synergy operating between CHUV and Athens' groups, most of the MIP consolidation work plan was executed during SGA2.

MIP deployment: Our ambitious objective to install the MIP in 30 hospitals by the end of SGA2 was met, with another 30 being planned. This was made possible by a significant improvement in the overall MIP installation process, expanding the diversity of operating environments that can run the MIP and dramatically reducing the efforts required from hospital IT departments to install the Platform. Datasets from more than 20,000 patients were integrated into the MIP, enabling to run use-cases in the fields of Dementia, TBI and mental health.

Integration of CEoIs: A large part of the CEoIs objectives were achieved, including the BIDS-SEEG (WP8.8) and ontology (WP8.9) Deliverables, as well as the development of a mental health use case (WP8.10). The BIDS-SEEG offers novel tools to share iEEG data according to the international BIDS format. Brain diseases ontologies were curated together with enhancing interoperability and applicability of services, including the Referential Ontology Hub for Applications within Neurosciences (ROHAN) which is now directly accessible through the MIP front page. The level of CEoIs integration was partly hampered by their late starting date (November 2018) and the major downgrade of HBP-SGA3 medical objectives decided in February 2019, which clearly impacted the motivation of some CEoIs partners. This largely accounted for the lower number of datasets brought to the MIP by these partners.

Other activities: On side of the above MIP-centered activities, methodological research was continued to guide the future integration of additional analytical tools into the MIP, including clustering and prediction algorithms, model validation, longitudinal data models, and statistical modelling of simulation output. These resulted in several guidelines and use cases performed in the field of PTSD with the 3C approach.

2) Impact and significance

As delineated in the original SGA2 DoA, we strongly believe that the success of the MIP project requires to build in parallel a reliable platform and a solid network of end users. This strategy drove our SGA2 action plan and proved effective in engaging a large number of hospitals. We see this outcome as a major milestone for the MIP to become one of the primary digital solutions to federate data across hospitals in Europe. Recent collaborations with the COVID research community demonstrate both the relevance of such a solution, but also the paucity of current alternatives. Indeed, despite many organisations claiming an ability to perform federated analyses, we are not aware of other comparably mature platforms already installed in ≥ 30 hospitals. This was confirmed by a comprehensive report ordered to Gartner.

3) Lessons learnt

Developing a Platform that enables clinicians and researchers to perform federated analyses of clinical/medical data across hospitals is a very challenging task, definitely more complex and difficult than originally thought of. In parallel to this greater complexity, we also observed an increase in the interest and relevance of the medical and scientific communities for such a Platform, reaching a momentum in the context of COVID. This is extremely encouraging for the future of the MIP.

We also experienced that the level of interactions required between all relevant stakeholders (IT staff, computer and data scientists, neuroscientists, data owners and providers and potential end users) to develop such a Platform needs to be much higher than that usually required for successful neuroscientific collaborations. That level of interaction between SP8 historical partners was missing in SGA1 and was progressively reached during SGA2.

4) Changes made to the work plan

The main deviations reported in detail in this document and elsewhere concern the total number of datasets currently available in the MIP (over 20,000 versus 30,000 expected), the number of analytical tools for longitudinal data analysis integrated in the MIP (primarily the Kaplan Meier estimator), the shift from EHR data extraction to REDCap eCRF data and metadata extraction, and the pending analyses of the TBI and mental health use-cases due to COVID-19 related delays. We do not see any of these deviations as having a significant impact on the global MIP and SP8 objectives.

5) MIP in SGA3

Adoption of the MIP current model by hospitals and clinical researchers, while much improved, remains cumbersome. This has led us to propose the two following developments in SGA3: 1) a light version of MIP local to be easily downloaded and installed together with its virtual machine on an end user desktop, and 2) MIP federated nodes installed and pre-configured on an EBRAINS "private" and secured virtual machine (one per hospital) as an alternative to in-hospital MIP. We believe that these developments are likely to boost adoption and usage of the MIP.

A sustainable business model for the MIP requires the development of diagnostic applications. We strongly believe that the diagnostic value of a large federated infrastructure across EU hospitals primarily lies in the early identification of rare diseases using federated AI-based approach (e.g. digital twins). The EU and DG Sante have developed an ambitious long-term strategy to promote harmonisation of care for rare and complex diseases across Europe, by setting dedicated European Reference Network (ERN). HBP, SP8 and the MIP have signed a partnership with one brain-oriented ERN and we are currently discussing collaborations with several other ERNs within the framework of an IMI call for "*Shortening the path to Rare Disease diagnosis by using new born genetic screening and digital technologies*". To increase the chance that the MIP will be selected through that call, we have planned to undertake a major MIP upgrade during SGA3, in order to shift its current privacy-aware model to one enabling individual diagnosis.

6) Legacy to the scientific community

By the end of SGA3, HBP will have provided the scientific community a robust and versatile MIP that should accelerate and leverage clinical research, and might later offer valuable diagnostic tools. The 10-year MIP roadmap will have been bumpy, marked by both successes and inevitable failures and errors from which we will have learned many lessons. Though suffering from unrealistic expectations, the original project was truly visionary, leading a path towards a unique research infrastructure which potential continues to grow as more and more stakeholders appreciate the value of federating clinical research data. HBP's double requirement to develop the MIP from scratch and demonstrate its scientific value has put a high (too high) burden on SP8, but certainly contributed to deliver a better product which will hopefully satisfy the need of many researchers in the future.