Recommendations for the MIP Technical Development During SGA3

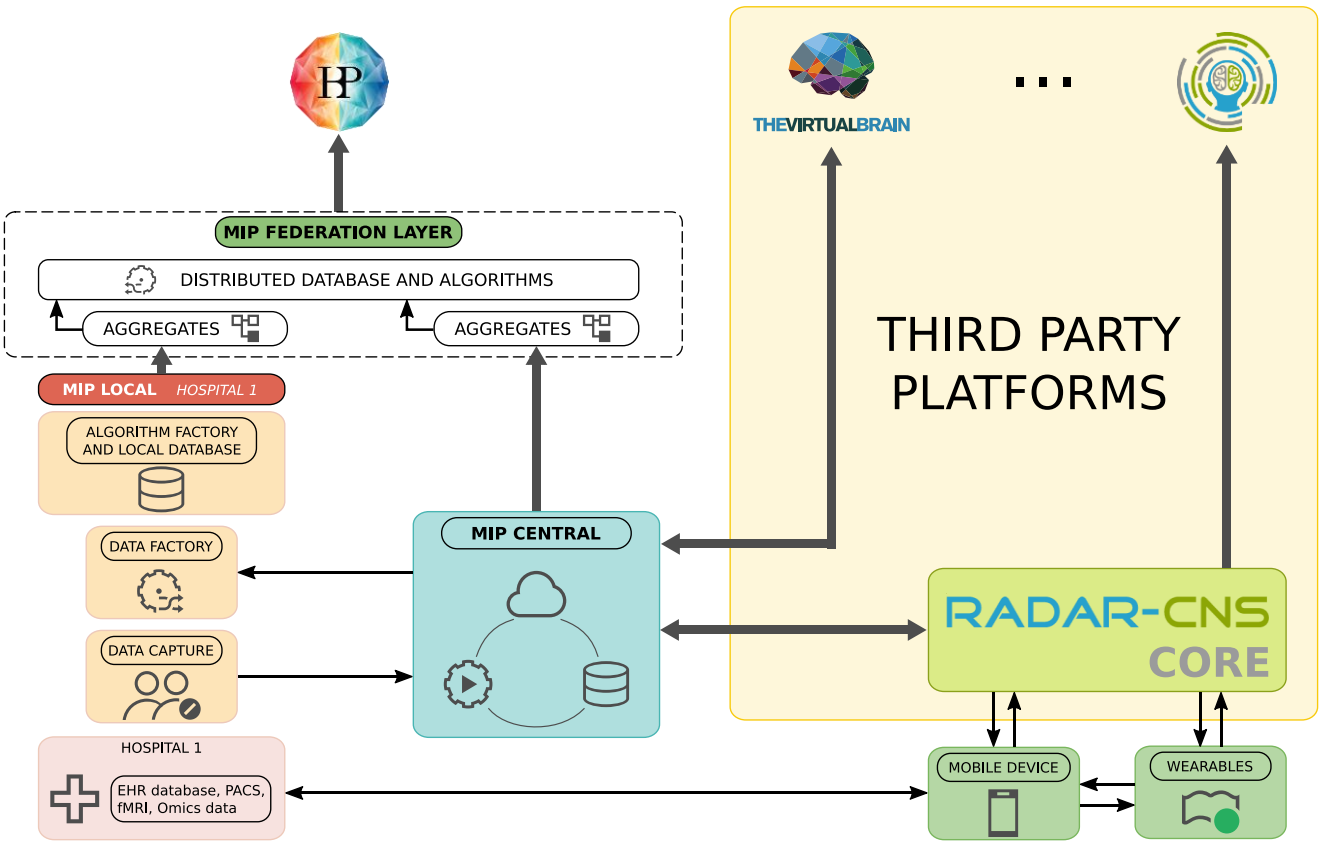

Figure 1: Recommended evolution of the MIP architecture and its possible integration with third party projects.

Project Number:	785907	Project Title:	Human Brain Project SGA2

Document Title:	Recommendations for the MIP technical development during SGA3
Document Filename:	c2976 *(right-click, select "update field" to update)*
Deliverable Type:	Report
Work Packages:	WP50 – 8.3
Dissemination Level:	PU
Planned Delivery Date:	SGA2 M24 / 31 03 2020
Actual Delivery Date:	SGA2 M24 / 31 03 2020
Author(s):	Tomas TEIJEIRO, EPFL (P1)
SciTechCoord Review:	Jacek MANTHEY, CHUV (P27)
Description in GA:	A document summarising the recommendations for future MIP technological developments during SGA3.
Abstract:	This document 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. This integration is discussed in detail in the accompanying document "In-depth assessment of potential new data integration into the MIP", with component ID C2975. Here, we begin with a high-level description of the MIP architecture and of its different software components, and then we analyse the evolution of each of them from a technical perspective.
Keywords:	MIP, software architecture, technical development, recommendations, SGA3.
Target Users/Readers:	MIP Architects and Developers.

Table of Contents

Table of Figures

List of acronyms

CDE: Common Data Elements

CSV: Comma-Separated Values

DICOM: Digital Imaging and Communication On Medicine

EEG: Electroencephalogram

EHR: Electronic Health Record

EMG: Electromyogram

fMRI: Functional-MRI

GUI: Graphical User Interaction

iEEG: Intracranial-EEG

JSON: Javascript Simple Object Notation

MIP: Medical Informatics Platform

MRI: Magnetic Resonance Image

NIFTI: Neuroimaging Informatics Technology Initiative

PACS: Picture Archiving and Communication Systems

PC: Personal Computer

PPG: Photoplethysmogram

SDK: Software Development Kit

SNP: Single Nucleotide Polymorphism

TUI: Text-based User Interaction

VPN: Virtual Private Network

XML: Extensible Markup Language

1. Highlights

- The MIP implementation should continue to be based on a microservices architecture, and rely on well-maintained open-source projects for all tasks that have already been addressed in other related contexts.
- To make it feasible to integrate new types of data, novel analysis algorithms and third-party tools, the MIP architecture should be revised to include a new main component, called *MIP Central* in this document, that centralizes the data and algorithms for which it is not viable to build a distributed version in reasonable time and costs.
- For a proper exploitation of the proposed new types of data (brain connectivity, omics and wearable data), the following analysis methods have found to have a major interest: multi-parameter analysis, longitudinal analysis and deep-learning-based methods.
- From an experimental point of view, the possibility of generating realistic synthetic data for virtual cohorts of patients should be explored as a high risk – high reward priority.
- The main development efforts on the current MIP components should be oriented to improve the user experience and the robustness of the platform.
- The current codebase should be simplified and organized with clearer objectives.

2. The MIP at a glance

From a simplistic perspective, the MIP can be summarized as a software bundle that allows the user to perform wide-scope statistical analyses on scalar and categorical features obtained from brain MRI images and EHR records.

However, there are a set of challenging design constraints that make the project tremendously complex in its implementation, deployment and maintenance. The most important of these constraints, which in turn are the main strong points of the MIP over other systems with similar aims, are:

- The ability to perform global analyses simultaneously over data coming from different hospitals (Federated analysis).
- No individual data leaves the hospital where it was acquired.

The second point guarantees a data privacy level that goes far beyond simple data anonymization, and this should facilitate the engagement of new hospitals and health institutions in the MIP. This would lead to a much easier and less burdensome recruitment of new cohorts of subjects for the data analysis experiments, and in the end, to make possible studies with a volume of data several orders of magnitude greater than the state-of-the-art.

These two constraints make the MIP distributed in nature, and this affects the full data processing pipeline: from data capture and storage to the data analysis algorithms.

In this report we follow a top-down analysis approach: we first describe the overall architecture of the platform, each component and its role, and how this architecture should evolve to support the functional changes that are expected from the integration of new tools and different types of data. Then, we will focus on each individual component and discuss how it has to be adapted to the new architecture.

Figure 2 shows the current software architecture of the MIP. First, we may distinguish two main layers: The MIP Local, that runs exclusively in one Hospital or medical facility; and the MIP Federation Layer, that performs large-scale federated analyses by using the interfaces provided by several MIP Local instances.

The MIP Local, in turn, comprises three main layers: 1) Data Capture, that is coupled with the specific information systems of the hospital, and that is in charge of reading the data from the EHR and MIR image databases and perform full anonymization; 2) Data Factory, that performs the preprocessing,

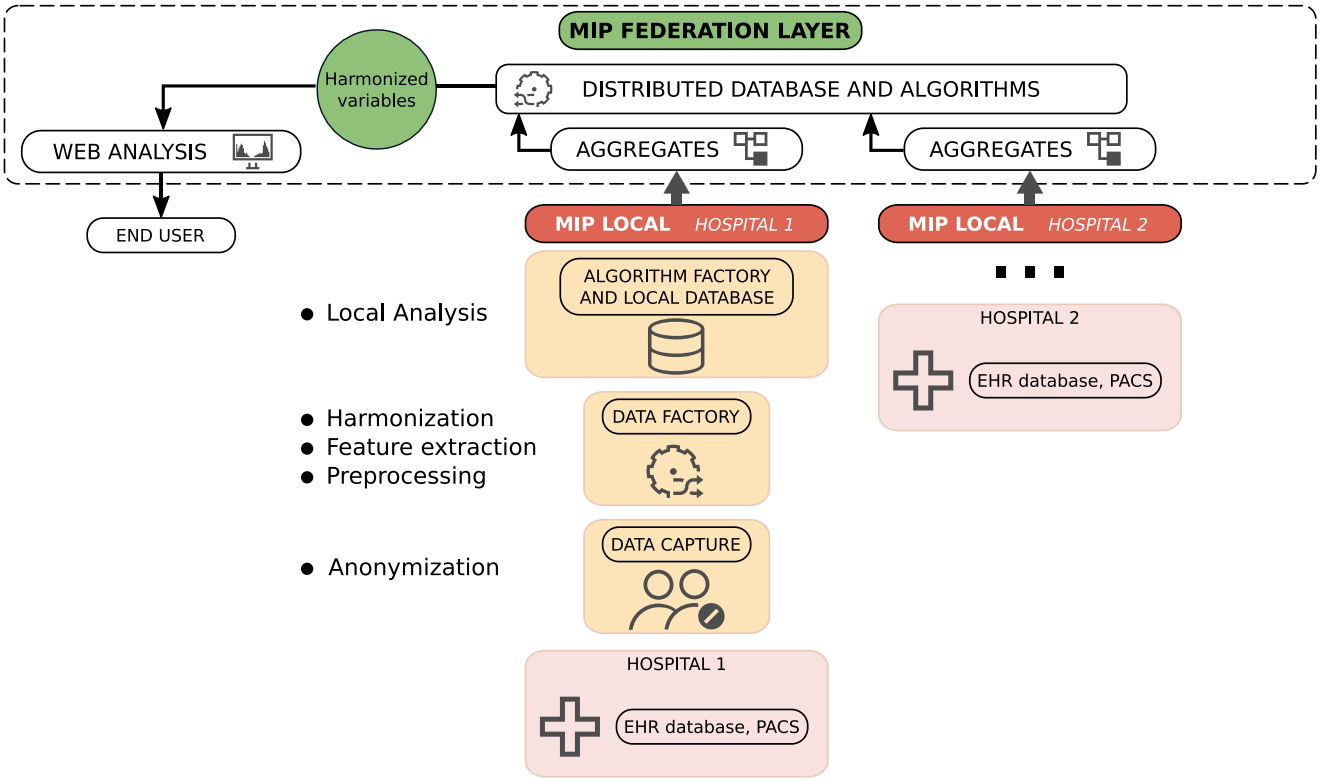

Figure 2: Current MIP architecture

feature extraction, and harmonizes the features according to a common scheme for all MIP Local instances; and 3) Algorithm Factory, that allows to perform local analyses of the Hospital data, and runs the distributed part of the algorithms at the Federated level.

The main technical challenges posed by this architecture derive from its highly distributed nature, and can be seen from two different perspectives:

1. **Data management:** In order to support federated analysis, all the hospitals have to provide at least a subset of the same set of features, and most importantly, the same feature needs to have the same semantics in all the hospitals and form a coherent global distribution. This is not a difficult problem in theory, but in practice it is extremely hard to manage, and very expensive in terms of time and human resources.
2. **Data analysis:** Given the data privacy constraints, all the analysis algorithms at the federated level have to be distributed in nature. This is undoubtedly the most significant restriction for including new algorithms in the MIP, due to the practical and even theoretical limitations for adapting many of the state-of-the-art statistical and Machine Learning methods.

From the implementation perspective, the MIP is based on microservices running on individual containers. This is undoubtedly an appropriate design choice, and the paradigm should be kept in the evolution of the platform. Regarding the underlying technologies and software dependencies, all components have their roots on well-established Open Source software, which guarantees the independence of the project and the possibility of maintenance in the long term.

3. Architecture evolution

Looking into the future of the MIP, there are two main perspectives that need to be addressed: the inclusion of new types of data, and the development and integration of new analysis techniques. In both cases we have very specific examples of what will be pursued in the medium term, so we will take them as a reference to design the changes that the current architecture will require.

Regarding the new types of data, in the SGA2 grant agreement [1] three main novel data sources have been identified to be of major interest for the future of the platform: 1) brain connectivity

information (fMRI, intracranial EEG), 2) multiple types of omics data, and 3) sensory data coming from wearable devices (EEG, PPG, EMG, etc.). While the specific challenges and requirements imposed by each of these data types are discussed in details in the accompanying report [2], here we will just summarize the influence they will have in the global MIP architecture:

- **Brain connectivity information:** In this case we main distinguish two main types of data: fMRI and iEEG. In the first case, given the standardization efforts of the neuroimaging research community, crystalized in initiatives such as BIDS [3] that is currently supported by the MIP, makes it possible to directly integrate this type of data without any change. Of course, specific preprocessing and feature extraction pipelines have to be included, but many of the most used ones, such as fMRIPrep [4] or C-PAC [5], provide open-source implementations and Docker containers for fast deployment, perfectly fitting the current MIP design. In this sense, many open source workflows are being published via the BIDS-Apps initiative [6]. Regarding iEEG, there is a crucial importance on the length of the records, and more specifically in the type of analysis that is pursued. If we consider an iEEG as a single test for which we can extract a collection of features, then it can be processed in the same way as MRI and fMRI images, and the current BIDS data structure can be used. However, if the desired analysis is a long-term longitudinal study, then significant changes would be required at all levels in the MIP. This is discussed below in more detail.
- **omics data:** Conceptually, the integration of omics data into the MIP does not require any architectural change. Of course, the *Data capture* and *Data factory* components will need specific workflows and pipelines for obtaining the required biomarkers and transform them into numeric features compatible with the CDE database. This may raise scalability issues due to high computational requirements, but the current design is adequate.
- **Sensory data coming from wearable devices:** Unlike the other types of data, data from wearable devices is intrinsically oriented to a long-term, personalized analysis, and it is usually meaningless to try to reduce the information to a set of static, discrete features. For this reason, their fit into the current MIP architecture is not natural, and it would require significant changes in all components: in the *Data capture*, to be able to acquire streams of pseudo-continuous data; in the *Data factory*, to support multimodal preprocessing pipelines that simultaneously analyse different parameters: in the *Algorithm factory*, to support longitudinal analysis; and in the *Federation layer*, to allow an effective interaction with the user.

But most of the evolution difficulties are related to the other perspective: data analysis. We have identified three main use cases that have been referred as important by different stakeholders and that would require major changes in the platform:

1. **Longitudinal analysis:** At this moment, it is quite tricky to include temporal information in the features that can be analysed in the MIP, in order to assess the dynamic behaviour of the target diseases. Also, at the Federated level it is not possible to identify a patient after their data entered the *Data Factory*, so if new data from the same patient comes later, this information cannot be used in the analysis.
2. **Multimodal analysis:** With the integration of new types of data in the platform, it arises the need to combine the information provided by each of them. This can be already done at the feature level in the *Algorithm Factory* and the *MIP Federation*, but so far it is not considered in the preprocessing stage. This could benefit from approaches such as sensor fusion, and particularly in the case of non-robust data such as the those obtained from wearable devices.
3. **Advanced machine learning:** The inclusion of new machine learning methods in the MIP is an extremely time-consuming task in terms of human resources. The need to design a distributed version of the method, and a subsequent implementation from scratch in most of the cases makes it unfeasible to keep the path with all the libraries and tools that are constantly appearing in this field. Also, the current workflow-oriented execution of the analysis algorithms at the federated level is simply too expensive to execute computationally demanding methods, such as deep neural networks. Finally, many of the state-of-the-art methods require access to the raw data, and this is not allowed at the federated level.

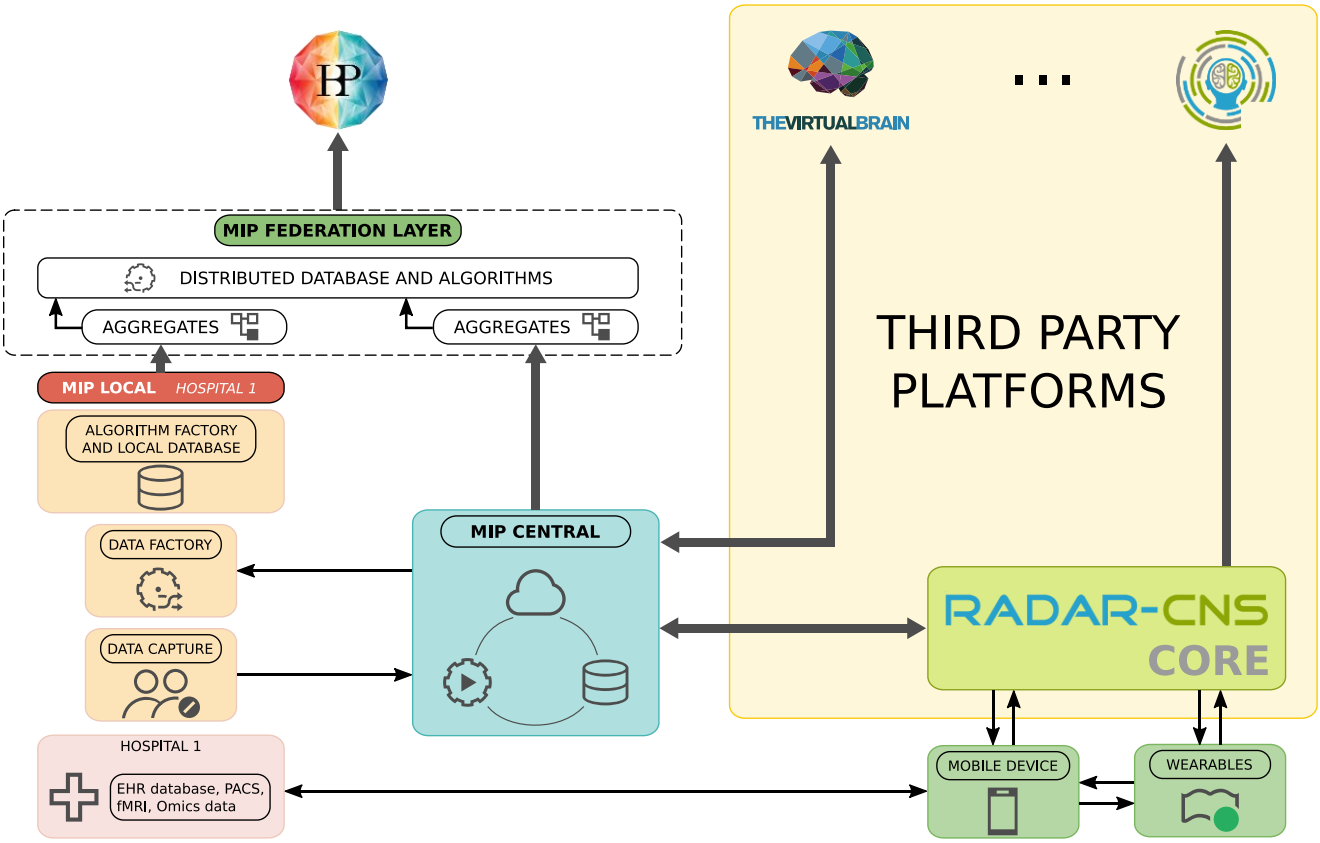

Figure 3: Proposal for the MIP Architecture Evolution

On the other hand, the tools provided by design at the MIP federated level are rather limited for the usual workflows of data analysts, specifically regarding data exploration and visualization. In this sense, the data privacy restrictions do not allow to go much further, but an idea worth exploring is the provision of "Virtual Cohorts", that is, synthetically generated datasets that mimic the distribution of a selected population in the real data, and which can be used with full access for model training and tuning. Then, the results would be validated with the real, inaccessible data, in order to confirm the experimental reproducibility.

A fundamental part of the development of the MIP during SGA2 has been the integration with third-party neuroscience projects that share common interests. This interaction has been the basis for identifying and characterizing most of the weaknesses discussed above, and also for proposing possible solutions. Figure 3 illustrates the envisioned architecture for future versions of the MIP. The most relevant change with respect to the current one is the inclusion of the *MIP Central* component. This would be implemented as a cloud service, and its purpose is to remove specific limitations regarding data access and analysis methods. A broad range of raw data from the hospitals can be published here from the *Data Capture*, that is, with proper anonymization, and this data may be used either by analysis algorithms deployed at the *MIP Central* as microservices (for examples, deep neural networks), or by third party software with specific authorization. An example of this would be the Virtual Brain project [7]. The *MIP Central* will also be connected with the *Data Factory* component of the different hospitals, in order to publish new features that are compatible with the CDE database, and therefore include them in the current and future federated analysis tools.

However, there is a high risk in this architecture that has to be managed very carefully. This risk is that for short-term convenience, all data and algorithms would be included in the *MIP Central*, turning it into a non-scalable and unmanageable hotchpotch of many different types of data. Thus, the protocols for including data and analysis methods have to be very clear and motivated, with a thorough ethical inspection and detailed technical constraints.

As examples of third-party platforms we have included two projects that are presently collaborating with the HBP SP8: The Virtual Brain [7] and RADAR-CNS [8]. The Virtual Brain is a perfect example of an application that has a data processing flow and user interaction completely independent from the

MIP, but that can be used to generate high-level features that would then be integrated in the federated analysis.

In the case of RADAR-CNS, the link can be potentially much deeper. Since this is an open-source platform implementing a full stack for the management of longitudinal monitoring data obtained through wearable devices, most of the solutions should be directly adopted by the MIP for the integration of wearable data, using the RADAR CORE platform as a middleware. In this way, most of the problems related to data acquisition, storage, querying and analytics would be already solved, or at least partially.

In the following sections we describe how the specific components and sub-components of the architecture have to evolve to be adapted to the proposed new global architecture of the MIP. To increase the utility and readability of the document, we have posed it as a sort of roadmap list associated to high-level development tasks. In order to understand the motivation for each of these tasks, we refer to the discussions in the accompanying report [2].

3.1 Data Capture

- Update the anonymization scheme in order to support longitudinal analyses and version control of the data. This means that the identifiers have to be consistent along time, while at the same time the untraceability to the original patients has to be preserved.
- Use the current BIDS implementation for acquiring fMRI data, and adopt BIDS-EEG [9] for the ingestion of EEG/iEEG data.
- Design an ingestion system for the data from wearable devices. Since this data comes from outside the hospital facilities, the acquisition has to take this into consideration. In the scenario of adopting the RADAR platform as a middleware for all the functionality related to the wearable devices, this means that a connector with the RADAR storage system has to be developed. Regarding communications security, we recommend the implementation of a VPN between the Data Capture, the RADAR middleware and all the remote devices gathering wearable data.

3.2 Data Factory

- Integrate the stack of the Galaxy project [10] for the processing and feature extraction of the variables related to omics data.
- Open the possibility to implement the protocols from the *ENIGMA* project [11]. Since most of the protocols are semi-automatic, this will probably require some kind of GUI or TUI for direct interaction between an expert user and the Data Factory.
- The MIP Central should be able to push normalized variables in the CDE-Database like another local processing pipeline. The most reasonable entry point for doing this is the Data Factory.
- A version control system for the data should be implemented to guarantee reproducibility of the experimental results. Both at the MIP-Local, MIP-Central and MIP-Federated levels. The DataLad [12] initiative can be inspiring and a good initial reference in this sense. This can have a relevant impact on the anonymization module.
- The computing capabilities and requirements of the MIP Local will have to be revised according to the new functionalities of the Data Factory. In particular, GPU computing may be required if Deep Learning methods are adopted for preprocessing and feature extraction on fMRI or EEG data. This may also be necessary for data generation methods.
- The generation of synthetic data for building Virtual Cohorts of patients that can be shared without any ethical restrictions is more interesting in the Data Factory, with access to the full collection of raw data. However, the complexity is also much higher, and the cohort would be restricted to the distributions that can be observed in the population of a particular hospital.

3.3 Algorithm Factory

- In the latest version of the MIP, the Algorithm Factory has been replaced by the Exareme engine. At this point, we don't envision any major changes in this design, since more complex analysis methods that do not fit the current scheme will be implemented in the MIP Central. The recommendations for this component are focused on improving the robustness of the current algorithms and on the orchestration at the federation level.

3.4 MIP Central

- For the publication of shared data at the MIP Central level, a common API should be defined.
- The MIP Central should be able to act as another MIP Local instance, in order to support all the types of analysis currently implemented in the MIP Federation layer.
- There is a need to define the computing environment that the integrated third-party applications will have access to. In this regard, we recommend to give the highest possible flexibility to make it easier the engagement of new partners and enrich the analysis features of the platform.

3.5 MIP Federation

- The CDE-Normalized database should be the first target for the generation of synthetic data and Virtual Cohorts of patients. Given the tabular nature of the variables this is a much easier problem than in the Data Factory, but on the other hand it adds the complexity derived from its distributed nature.
- Following the path of SGA2, significant efforts should be devoted to increase the robustness and performance of the Federation layer. From a user perspective, this is probably at this point the most discouraging aspect of the MIP.

3.6 User Interface

- The management of errors and unexpected situations has to be much more explicit and informative. In general, the feeling from the user perspective is that the platform needs more testing effort. This should have a particularly relevant position in the next software development project plan.
- Another main issue identified from the user perspective is the responsiveness of the interface. Even if the distributed execution of analysis algorithms on multiple datasets is as slow as the slowest of the nodes, and assuming a significative speed improvement is not possible without a major redesign of the platform components, the user experience can greatly benefit from a better feedback and a real-time update of the status and evolution of the experiments.

4. Codebase evolution

An overview of the MIP open source codebase (https://github.com/HBPMedical) can provide some insights about design and code management decisions that should be revised in the future. From the 119 non-archived code repositories, 60 have not been updated in the last 2 years. This suggests that the MIP software ecosystem may be oversized and that a significant number of the sub-components may not be necessary, so there is a potential simplification of the platform design. While the high-level architecture is clearly described in several documents and in the knowledge base [13] we miss a more comprehensive description of all the sub-components that comprise the actual implementation

of the MIP. These sub-components should be related with one or more code repositories, and their inter-dependencies and dependencies with third-party projects should be clearly described.

From a technical perspective, we recommend that efforts should be devoted to remove the following dependencies:

- Matlab: Requiring a Matlab instance in production usually causes a significant memory and computing overload, and constitutes a scalability bottleneck. Also, the license management can pose an additional complication.
- Python 2.7: The branch 2.x of Python has been officially discontinued from January 1st, 2020. This means that no security updates will be provided, so it is compelling to update all software relying on this version and move to Python 3. This includes very important dependencies such as exareme or Galaxy.

5. Conclusions

In this report we have analysed the current and the immediate future development of the MIP platform from the technical perspective. We have given a particular emphasis to the integration of new types of data beyond MRI images, and the technical challenges that arise from that. As a conclusion, we suggest the following six recommendations for the upcoming development of the MIP:

1. The MIP implementation should continue to be based on a microservices architecture, and rely on well-maintained open-source projects for all tasks that have already been addressed in other related contexts.
2. To make it feasible to integrate new types of data, novel analysis algorithms and third-party tools, the MIP architecture should be revised to include a new main component, called *MIP Central* in this document, that centralizes the data and algorithms for which it is not viable to build a distributed version in reasonable time and costs.
3. For a proper exploitation of the proposed new types of data (brain connectivity, omics and wearable data), the following analysis methods have found to have a major interest: multi-parameter analysis, longitudinal analysis and deep-learning-based methods.
4. From an experimental point of view, the possibility of generating realistic synthetic data for virtual cohorts of patients should be explored as a high risk – high reward priority.
5. The main development efforts on the current MIP components should be oriented to improve the user experience and the robustness of the platform.
6. The current codebase should be simplified and organized with clearer objectives.

6. References

1. Human Brain Project Specific Grant Agreement 2 | HBP SGA2 Project | H2020 | CORDIS | European Commission. https://cordis.europa.eu/project/rcn/220793/factsheet/en.

2. Teijeiro, T. & Atienza, D. In-depth assessment of potential new data integration into the MIP (relevance, clinical benefits, integration feasibility, risks, costs) (C2975). (2020).

3. Gorgolewski, K. J. *et al*. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 1–9 (2016).

4. Esteban, O. *et al*. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods* **16**, 111–116 (2019).

5. Craddock, C. *et al*. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Front Neuroinform* **42**, (2013).

6. Gorgolewski, K. J. *et al*. BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS Computational Biology* **13**, e1005209 (2017).

7. Sanz Leon, P. *et al*. The Virtual Brain: a simulator of primate brain network dynamics. *Front. Neuroinform.* **7**, (2013).

8. Ranjan, Y. *et al*. RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR Mhealth Uhealth* **7**, e11734 (2019).

9. Pernet, C. R. *et al*. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data* **6**, 103 (2019).

10. Afgan, E. *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**, W537–W544 (2018).

11. ENIGMA - Enhancing Neuro Imaging Genetics by Meta-Analysis. http://enigma.ini.usc.edu/.

12. DataLad. https://www.datalad.org/.

13. HBP Medical Knowledge Base. https://hbpmedical.github.io/.