In-depth assessment of potential new data integration into the MIP (relevance, clinical benefits, integration feasibility, risks, costs)

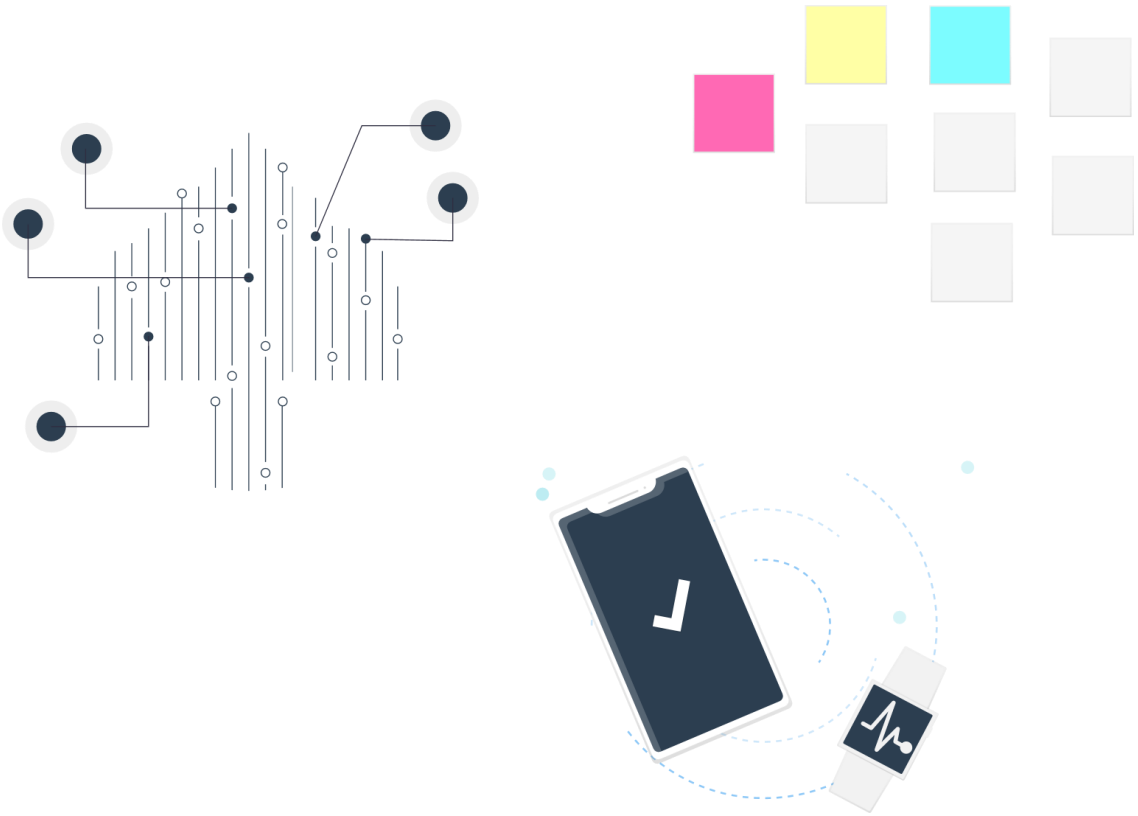

Figure 1: Novel types of data that will potentially be integrated in the MIP (Omics, Brain Connectivity Information, and Wearable Data).

Project Number:	785907	Project Title:	Human Brain Project SGA2

Document Title:	In-depth assessment of potential new data integration into the MIP (relevance, clinical benefits, integration feasibility, risks, costs)
Document Filename:	c2975 *(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)
Contributor(s):	Olivier DAVID, SHORT PARTNER NAME (P#), Section 3.1 Philippe RYVLIN, CHUV (P27), Section 3.3
SciTechCoord Review:	Jacek MANTHEY, CHUV (P27)
Description in GA:	A report describing the systematic assessment of the potential relevance, technical solutions, level of risk and cost/benefit ratio associated with the integration of new types of data into the MIP including -omics, data captured by wearable devices and complex neuroimaging data.
Abstract:	This document summarizes from the 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. The specific recommendations for the future development of the MIP are collected in the accompanying document "Recommendations for the MIP technical development during SGA3", with component ID C2976.
Keywords:	MIP evolution, omics, fMRI, iEEG, wearables, new types of data, new biomarkers.
Target Users/Readers:	MIP Developers and Users.

Table of Contents

Table of Tables

Table of Figures

List of acronyms

BOLD: Blood Oxygen Level-Dependent

CDE: Common Data Elements

CSV: Comma-Separated Values

DICOM: Digital Imaging and Communication On Medicine

DTF: Directed Transfer Function

DNA: Deoxyribonucleic acid

EDA: Electrodermal Activity

EEG: Electroencephalogram

EHR: Electronic Health Record

EMG: Electromyogram

fMRI: Functional-MRI

HRV: Heart Rate Variability

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

PDC: Partial Directed Coherence

PPG: Photoplethysmogram

SDK: Software Development Kit

SNP: Single Nucleotide Polymorphism

VPN: Virtual Private Network

XML: Extensible Markup Language

1. Highlights

- The current MIP is already almost prepared for the integration of fMRI and EEG/iEEG data. The BIDS specification [1] and open-source pipelines such as those published in the BIDS-Apps [2] repository are the most cost-effective way of including more advanced neuroimaging techniques and features at the MIP Federated level.
- The objective of integrating multiple types of Omics data has to be narrowed according to the target diseases and the biomarkers that are known to be meaningful from the clinical perspective. From the technical point of view, we recommend to adopt the workflows provided by the Galaxy project [3] and integrate them in the MIP, in line with what has already been done during SGA2.
- Regarding data from wearable devices, the temporal properties of this type of information makes it difficult to adapt it to the current MIP scheme. Furthermore, important challenges arise regarding data acquisition, transmission and storage, and also from the data privacy and security perspectives. Our recommendation in this sense is the adoption of existing solutions for the management of data from wearable devices, such as the open-source RADAR [4] platform, its installation as a parallel software stack, and integrating into the MIP just the interesting long-term features that can be adapted to the MIP model.

2. Summary

During SGA1 and SGA2, the MIP has been primarily oriented to the federated analysis of variables obtained from MRI images in the context of patients with dementia. The final objective was to find reproducible dependencies between these variables, diagnostic outputs, and other variables from the EHR including demographics, genetic biomarkers or neuropsychology measures obtained from different tests. For this, the MIP supports the acquisition of basically two types of data: 1) MRI images stored in a hospital PACS database in the DICOM or NIFTI formats; and 2) semi-structured data extracted from the hospital EHR database in a standard data representation format (CSV, JSON or XML). For a more detailed overview of the current MIP architecture and underlying technologies, we refer the reader to the section 2 in the accompanying report C2976 [5].

Then, with the view to prepare SGA3, different novel types of data have been proposed to expand the type of analysis that can be performed at the platform, and to benefit not only from the Big Data paradigm, but also from an Information Fusion approach. The proposed types of data, which will guide the structure of this document, are the following:

1. **Brain connectivity information:** This type of data refers to much more complex neuroimaging data than the current T1 MRI, and includes structural and functional connectivity maps extracted from fMRI, EEG or iEEG. The main property of these data is its dynamic nature, and the ability to move from a region-of-interest-based analysis to voxel-based analysis.
2. **Omics:** This is a quite generic term that refers to a broad spectrum of biological information. Probably the primary representative is genomics, but it also includes proteomics or metabolomics, that may be as relevant as genomics in the context of the diseases currently targeted by the MIP such as dementia [6]. The main property of this type of data is the large variability among data types, formats, and sizes, and the challenge posed to hospitals to acquire it in a standardized way and compatible for the federated analysis.
3. **Data from wearable devices:** The recent explosion of wearable devices capable of obtaining long-term physiological information of a very different nature opens the door to a more cost-effective, personalized and comprehensive monitoring of the health status and evolution of a patient. These data are also dynamic in nature, but the time scale is completely different from the brain connectivity data sources. For example, while a fMRI spans in the range of minutes [7], the wearable monitoring can be pseudo-continuous for weeks, months or even years.

In the following sections we will discuss the potential integration into the MIP of these types of data, specifying the potential features that could be relevant for the targeted conditions (dementia, epilepsy and Parkinson's disease), how they could be implemented, and the ethical, risks, and cost aspects that should be considered. This assessment will be done according to the protocol defined at the beginning of SGA2 [8], and the technical details about the changes that will have to be implemented in the MIP architecture are described in the accompanying report [5].

3. Potential new biomarkers and features

From a data analysis perspective, the specification of the features to capture would be inexorably bounded to the target problem. Regarding the future of the MIP, this is (and should be) an open question, so by no means we will try to enumerate all the possible biomarkers and features that could be extracted from the new types of data. Nevertheless, we will point out to a series of features that cover the spectrum of data types and operations that should be supported by the platform in order to allow state-of-the-art analyses on the current target diseases.

3.1 Brain Connectivity Information

The secondary data elements that are obtained from the raw dynamic neuroimaging data (fMRI and iEEG) by means of different data processing algorithms are commonly called *"derivatives"* [9]. These *"derivatives"* may be used directly as MIP variables, or they could be treated as an intermediate abstraction susceptible to further analysis and aggregation before entering the *Data Factory*. For example, in the case of epilepsy, the main objective is usually to quantify spatial (iEEG) and temporal (fMRI) aggregations that behave as distinct networks regarding the origin and propagation of seizures. Based on the premises assumed for the definition of these networks, the following classification has been proposed [10]:

- **Connectivity features:** This approach tries to describe the brain function through mathematical estimates of the links between two signals originating from different brain regions, and therefore reflecting how the activity is coordinated between them [11]. The signals could be different iEEG channels or BOLD signals from fMRI. To quantify the links, many different methods have been proposed in the literature, including linear coherence, linear regression analysis, mutual information, nonlinear regression analysis, similarity indexes, or correlation coefficients, amongst others [12].
- **Causality features:** This category groups together those methods that, beyond analysing the link between brain areas, attempt to estimate the direction of coupling, associating it with the notion of causality. One of this measures is the Directed Transfer Function (DTF) [13], that has been used to localize the source region of epileptic seizures [14]. Another feature implementing this idea of causality is the Partial Directed Coherence (PDC) [15], that works in the frequency domain. Both the DTF and PDC assume a linear relation between the signals, but this constraint has been relaxed for example in the work by Wendling et al. [16], which is based on a nonlinear regression measure that has been augmented by a "direction index" that captures the asymmetry of the nonlinear correlation coefficient.
- **Graph theory-based features:** This approach is based on the representation of the brain connections as graph structures at different scales (local, global), and the subsequent mathematical manipulation of these structures. For example, in The Virtual Brain [17] the connectivity is represented by long-range and local connectivity matrices, from which a variety of quantitative metrics can be extracted. Two particularly relevant parameters that can be evaluated from the graph representations of the brain are the clustering coefficient and the characteristic path length. In the specific case of epilepsy, these parameters have been found to be very relevant for the characterization of the ictal activity [18]. Another important feature that allows to summarize potentially very complex graphs is the degree, that represents the number of connections linking each node to other nodes [19].

3.2 Omics

Conceptually, the features that can be extracted from omics data are the most similar ones to the current MIP variables. They are generally static, easily quantifiable and well suited to the types of statistical analysis currently supported by the MIP. However, deciding which biomarkers will be finally included for federated analysis will have to be based on the capabilities of the hospitals and health centres in the federation to obtain that information. There are literally hundreds of omics-based biomarkers reported in the literature that claim a relation with the current target diseases, but for many of them the experimental evidence is still too weak [6]. This is an excellent use case for MIP, as it makes it possible to study the statistical incidence of these biomarkers on a much larger scale than most of the published studies. In general, we can classify the omics biomarkers according to the molecules under study, that will determine the techniques used to extract the information. We can highlight the following categories:

- **Genetic biomarkers:** Basically, they consist of genetic mutations that are identified by DNA analysis. This information is becoming increasingly available in hospitals, enabling the extraction of different SNPs that are associated to conditions such as Alzheimer's disease [6] or epilepsy [20].
- **Epigenomics:** Among epigenetic biomarkers, microRNAs have raised particular interest in recent years for the diagnosis of neurodegenerative diseases [21]. These RNAs regulate gene expression at post-transcriptional level, and are differently expressed in the brain under pathological conditions. Many of the related biomarkers can be obtained directly from blood or serum samples [6,20].
- **Metabolomics:** Most diseases cause alterations in the metabolic function that lead to variations in the concentration of different metabolites. These variations are susceptible to detection in a non-invasive way from simple blood or plasma samples [22]. In the case of Alzheimer's disease, most of the metabolites identified as potential biomarkers are related to lipid and amino acid metabolisms [23], while for epilepsy the levels of myo-inositol and glutathione have shown a good discriminative power in several studies [24].
- **Proteomics:** Proteomic studies are particularly promising in the study of neurological diseases, since the brain uses signalling proteins, found in blood, to control bodily functions (e.g. peripheral and central inflammatory, and immune mechanisms). Therefore, changes in these signalling proteins will likely modify the phenotype of different diseases in blood samples [25,26]. A detailed list of proteomics-based biomarkers for Alzheimer's disease can be found in [6].

3.3 Wearable Data

Wearable devices can potentially provide large amounts of dynamic data that could allow longitudinal analyses with different temporal granularity. However, a limitation on the number of sensors that a patient can wear simultaneously has to be considered, and therefore an individual monitoring profile will be required for each patient (or group of patients) based on their disease and the target analysis. Below we describe the main parameters that have shown relation with the diseases currently targeted by the MIP, and that can be obtained through wearable devices.

- **Tachycardia**: As a potential extracerebral indicator of changes in the autonomous response, this is an event of primary interest since it is common during epileptic seizures. It may be induced not only by seizure activity per se through its effects on the central autonomic and limbic networks, but also indirectly via seizure-induced catecholamine release [27,28]. Indeed, tachycardia is related to increased sympathetic activity and has been reported in up to 80% of all epileptic seizures [27]. It occurs consistently in generalized tonic-clonic seizures but also in many focal seizures [29]. Importantly, tachycardia might start before any other detectable clinical sign or symptom [30]. While being sensitive, tachycardia is not a specific marker because of inevitable changes of heart rate with daily life activities like walking, climbing stairs, or emotional states.

- **Heart Rate Variability (HRV)**: HRV is another biomarker of interest related to the heart rhythm. HRV is controlled by reciprocal sympathetic and parasympathetic influences and serves to adapt the cardiovascular function to external and internal demands [31]. Rhythmic components of HRV that are of interest in the seizure-detection context belong to the high frequency and low frequency bands. A recent paper reported that a low-frequency/high-frequency ratio increase could be successfully used for predicting focal epileptic seizures (88.3% sensitivity and 86.2 % specificity) [32].
- **Electrodermal activity (EDA)**: This variable refers to the dynamics of skin electrical conductance, including slow changes in basal conductance level and transient skin conductance responses. In contrast to cardiac regulation, EDA depends solely on sympathetic control of the function of sweat glands [33]. It is closely linked to emotional and mental arousal. Poh et al. [34] reported the first long-term recording of EDA in the patients with epileptic seizures. The authors showed that seizures are associated with a clear increase in EDA, though much greater in generalized tonic-clonic seizures than in focal seizures without secondary generalization [34]. Thus, EDA has become one of the reliable ways to detect generalized tonic-clonic seizures, but does not appear capable to reliably detect other seizure types in isolation. Also, EDA offers the possibility to assess the patient's stress level in general, and might be able to detect stress patterns that could trigger seizures and help their forecasting.
- **Respiratory changes:** in the context of epilepsy, alterations in the respiratory activity result from abnormal activation of the respiratory centres in the brainstem and can be manifested by tachypnea, bradypnea, apnea, hypoventilation, and hypercapnia [35]. These changes occur both in generalized and focal seizures and are especially common in seizures originating from the mesial temporal structures [36]. It is reflected by hypoxemia measured through pulse oxymetry (SpO2). In several studies, hypoxemia below 90% was observed in about one third of all seizures, both focal and generalized [37]. Tachypnea associated with epileptic seizures is also of interest owing to its specific pattern that differs from increases in ventilation during activities of daily living [38]. Post-ictal respiratory abnormalities also play a major role in sudden death in epilepsy, suggesting that their appropriate monitoring might help preventing these sudden deaths.
- **Motor activity:** There are a number of biomarkers that can be obtained from the analysis of the motor activity of the subject, and that have proven to be related with different neurological diseases. The motor activity can be well characterized by 3D-accelerometers located at different body positions. In dementia patients, gait analysis has shown a good classification performance by using a reduced number of gait parameters such as gait speed, average kinetic energy, compensation motions or gait harmony [39]. Similar results were obtained for patients with Parkinson's Disease, for the assessment of mobility [40] and freezing of gait [41]. In the case of patients with epilepsy, apart from autonomous changes, a large variety of involuntary body movements can occur during seizures and be detected by 3D-accelerometry. While varying between patients, they are often very reproducible from one seizure to the other in the same patient. The generalized tonic-clonic seizures are now well characterized both by 3D-accelerometry and electromyography, with specific patterns that clearly differ from physiological activity and from psychogenic non-epileptic seizures [42]. Their detection with wrist accelerometers is very high [43]. 3D-accelerometer have also been used to characterize and detect other seizure-related movements during focal seizures, such as hypermotor seizures [44].

4. Data types and processing pipelines

This section deals with the specific implementation of the three considered types of data, by means of data formats, storage structure and management, and processing pipelines. It is not an exhaustive list of all the available standards and specifications, but a description narrowed to the context of the MIP, considering its current state and the viable options in the nearest future.

4.1 Brain Connectivity Information

The source data to obtain brain connectivity information will be fMRI and iEEG sessions. Both types of data are fully targeted by the BIDS [1] specification, which is clearly a safe bet for the implementation. BIDS basically describes how to organize a dataset of neuroimaging sessions and its corresponding metadata, and there are extensions to cover EEG signals. The underlying imaging format for fMRI in BIDS is NIFTI [45], and therefore it is usually required a conversion from DiCOM, which is a common format in hospital PACS. However, there are open source tools such as dcm2nii [46] that make this process straightforward.

Regarding BIDS for EEG (and iEEG) [47], four different formats are accepted for signal storage:

- European data format (.edf)
- BrainVision Core Data Format (.vhdr, .vmrk, .eeg) by Brain Products GmbH
- The format used by the MATLAB toolbox EEGLAB (.set and .fdt files)
- Biosemi data format (.bdf)

While the specification is clear regarding the organization of the raw data, there are no official lists of *derivatives* variables that could be extracted to infer the brain connectivity information. Since this is an open question, the MIP can be a leading actor, proposing specific implementations for the target features.

As explained in the accompanying report [5], the integration of this type of data would require the *Data Capture* and *Data Factory* components of the MIP to be able to read and process this variety of formats to be fully compliant with the standard. However, we consider that supporting just edf files would be a good compromise.

Regarding the processing pipelines for feature extraction, some of the most relevant and general-purpose procedures in the literature are available in the BIDS-Apps [2] repository as open-source containerized implementations. Their integration in the *Data Factory* would provide a wide range of significant features with minimum cost. These features could then be included as new variables in the CDE-Normalized database, and they would be ready for use in the available catalogue of experiments. An example is the rsHRF application, that is able to work with individual fMRI sessions and characterize the hemodynamic response function (HRF) of the brain [48], in a format that is compatible with the MIP CDE database.

A very different challenge would be the integration of novel processing algorithms to treat this type of data in the pipelines. For example, we can think of Deep Learning-based methods, which are now the dominant machine learning tool in big data environments [49]. In this sense, we may distinguish two use cases:

1. Deep Learning for image processing and feature extraction: In this case, a trained neural network model would be used as a preprocessing algorithm for obtaining a set of variables of interest (image segmentation, voxel classification, etc.), and therefore it should be deployed in the *Data Factory*. The availability of containerized versions of the most popular Deep Learning frameworks such as Tensorflow [50] or Pytorch [51] makes this option fully feasible without affecting the basic design of the platform. An important aspect to be evaluated in this case is a probable increase in the computing requirements of the MIP Local facilities.
2. Deep Learning for model design and evaluation: This would imply the integration of Deep Learning training and inference methods in the *Algorithm Factory*, supporting a distributed scheme to be used at the Federated level. Distributed training of neural networks is a very relevant topic, and most of the frameworks support it to some extent, but probably the algorithm orchestrator in the current MIP architecture cannot provide the required performance to consider this a feasible feature to implement in the near future. As discussed in the accompanying report [5], the redesign of the architecture including the new *MIP Central* component is the most cost and risk-effective approach to this end.

Another type of processing that gets a lot of attention from the scientific community is the longitudinal analysis of fMRI [52] and iEEG [53]. However, this type of analysis has not been considered

in the MIP design, and adapting the architecture would require major changes, mainly in the anonymization module of the *Data Capture* and in the user front-end.

In summary, using the BIDS specification for implementing the fMRI (NIFTI) and iEEG (edf) data types in the MIP is the less risky and most cost-effective solution. Since BIDS is already supported as the organization scheme for T1 MRI, this means that the changes required for the integration will be minimal. It will be necessary, however, to implement the specific connectors for these new types of data with the hospitals' data storage systems, but the scheme remains the same. For preprocessing and feature extraction pipelines, BIDS-Apps provide a wide range of ready-to-deploy popular methods, and other existing open-source solutions could be adapted to a containerized implementation within a reasonable time and effort.

From the ethical point of view, the current considerations applied for T1 MRI images can be extended to these new types of data without additional constraints if every session is managed individually. However, if there is a need to include longitudinal analyses, the anonymization scheme of the MIP has to be revised to keep the consistency on the patients' identification over different sessions.

4.2 Omics

In the case of omics, there is an enormous diversity of data formats, data types and computational methods [54] associated on the one hand to the variety of information to which the term can refer, and on the other hand to the presence of multiple vendors and acquisition devices with proprietary solutions. In order to enable a technical discussion that will facilitate us to reach the implementation level, here we will focus on genetic information, which is in the short term the most interesting to be included in the MIP. However, the conclusions drawn can be extended to other types of omics, such as proteomics or metabolomics.

A clear example of a type of omics data that could be included in the current MIP up to the federated levels is the value of different SNPs [55] associated with a target condition, such as for example Alzheimer's Disease [56]. Each SNP could become a new variable in the CDE catalogue, and their value could be encoded with three different categorical values, typically 0 (homozygous), 1 (heterozygous), and 2 (other homozygous).

The main technical difficulties rely on data acquisition and transformation. In this case, due to the variety of equipment for obtaining the genetic information, no prior constraints should be imposed at the hospital level regarding data formats. Fortunately, there are open-source initiatives that deal with these problems and provide a standardized interface to define acquisition and transformation workflows for many different types of biological data. In particular, we believe that the Galaxy project [3] should become the basis for all the pipelines involving omics, and this is in line with the development that has been carried out during SGA2. It is important to note that Galaxy is a very large community project that expands over many different fields in biomedical research, and in this sense, it is essential to determine the specific requirements for the MIP that can be solved by the Galaxy stack, and restrict the integration span to keep the costs under control.

Another related and large-scale project from which the MIP can benefit is *ENIGMA* [57], an international consortium for collaboration between researches in imaging genomics, neurology and psychiatry. *ENIGMA* proposes multi-parameter meta-analyses based on imaging and genetic information, and one of the core concepts is the one of "protocols": These are standardized automatic or semi-automatic procedures that will ensure that the data from different sources are compatible, and that the extracted information is relevant to a particular end. It is worth exploring the potential synergies between the MIP and *ENIGMA*, by integrating some of the *ENIGMA* protocols in the MIP Local. This would, on one hand, make it easier to obtain relevant variables from genetic data to be included in the CDE-Normalized database, and on the other hand, facilitate to many of the hospitals in the MIP network to participate in the *ENIGMA* consortium and in some of the specific working groups, such as ENIGMA-Epilepsy [58] or ENIGMA-Parkinson's [59].

On the other hand, and even if the interest is not so clear than for fMRI and iEEG data, we have to consider the potential integration of Deep Learning-based techniques for the processing of the omics data [60]. The technical challenges are equivalent than in the previous case, with the additional difficulty that the data cannot be moved to the *MIP Central* component, as stated below.

Finally, and although it is not strictly within the scope of the MIP, we consider highly desirable to achieve some coordination between hospitals in obtaining and locally storing the raw data related to omics. The objective is to avoid excessive efforts later in data harmonization when the variables of interest will enter the MIP, and in this sense our recommendation is the definition of some simple common protocols for data acquisition and storage.

From the ethical and legal perspectives, omics data is the most challenging of the three new targeted types of data. In principle, all omics should be excluded from the *MIP Central* component, so the considered processing pipelines have to be compatible with the current MIP architecture.

4.3 Wearable Data

We can summarize the properties of the data obtained from wearable devices with the following three words: personal, longitudinal and long-term. This is in clear contrast with the current MIP vision, in which each sample in the federated data storage is assumed to be independent from the others, and time is not explicitly considered in the analysis. While a full convergence between these two visions is at this moment unrealistic, we will discuss some strategies that can help to leverage the wearable data to some extent without completely breaking the current design.

First, regarding data acquisition, representation and management, a wearable context has to be oriented to data streams, that in principle could be continuous for long periods. This is a technically solved issue, but the current MIP architecture is not yet adapted. Also, since a large volume of data will be produced outside the Hospital domain and will be transmitted through regular Internet connections, a specific effort has to be made in the design of a security and privacy-aware system for remote patient register, authentication and wearable data communication. In this sense, it is of primary importance to assess the legal terms of the specific wearable devices that will be used. For example, for the Empatica E4 device [61], that is one of the considered devices to integrate in the future, it is not clear what data is transferred to the private cloud storage owned by the company, and the provided SDK requires that all connections to the devices to be logged in the Empatica cloud.

Also, another important property of wearable data is its poor quality and lack of robustness, derived from the acquisition in completely uncontrolled scenarios. This makes it essential to adopt multi-parameter approaches to data analysis, in which the redundancy between different biosignals helps to increase the robustness of the extracted features. But the current MIP is not adapted for multi-parameter processing, and this would require significant changes on the *Data Capture* and *Data Factory* components.

In summary, to properly support the integration of wearable data we consider it necessary the development of an additional non-trivial architecture for data management that involves both the patient and the hospital endpoints. There are vendors offering infrastructures and services to build this type of architectures in a cloud-based fashion [62], but taking into account the Open-Source nature of the MIP, we consider that the potentially most fruitful and cost-effective solution is the integration with the RADAR-base m-health architecture [4]. This is also an Open-Source architecture based on Apache Kafka [63] that solves the full pipeline for medical wearable data acquisition, transmission, storage and processing. It also provides specific tools for data analysis and visualization, so the full stack could be deployed in parallel to the MIP, and then the connection could involve just the features selected to incorporate to the federated analysis. We identify no major risks in adopting this solution, but the development efforts would be certainly far from negligible, and much higher than for the other two considered types of data.

From the ethical and data privacy perspectives, this type of data opens new dimensions to consider, insofar as sensitive data will exist outside hospital facilities and will be transmitted totally or partially via Internet. Besides the potential issues that can arise from the terms and policies of the manufacturers of the wearable devices, such as the Empatica case described above, our recommendation is to provide special protection to the data stored locally in the patient home (in a base station, PC, or typically the patient's smartphone) and to communications. As described in the accompanying document, this could be achieved by: 1) encrypting the locally stored data, and 2) setting a VPN network between the hospital and all the devices with Internet connectivity involved in the acquisition of wearable data.

5. Risks and costs

In this section we summarize the risk and cost assessment of different implementation alternatives for the three target novel types of data. Risks are qualitatively evaluated in three levels (low, medium and high), while costs are estimated in Persons-Month, considering a scenario in which one high-qualified professional would be in charge of the full technical work. Also, it has to be noted that some costs could be shared between related tasks (for example, the total cost of the generation of synthetic data of different types would be less than the sum of the individually estimated costs).

Table 1: Risks and costs for the integration of brain connectivity information

Description	Risk	Cost (PM)
Integration of fMRI data in the current *Data Capture*	Low	1
Integration of EEG/iEEG data in the current *Data Capture*	Low	2
Integration of *BIDS-Apps* for fMRI feature extraction in the current *Data Factory*	Low	2
Tailored feature extraction for EEG/iEEG in the current *Data Factory*	Medium	6
Tailored feature extraction for fMRI in the current *Data Factory*	Medium	9
Deep Learning algorithms for feature extraction in the *Data Factory*	Medium	6
Deep Learning algorithms in the *Algorithm Factory*	Medium	9
Deep Learning algorithms in the *MIP Central*	Medium	6
Deep Learning algorithms at the Federation Level	High	24
New anonymization scheme for longitudinal fMRI/EEG/iEEG sessions	High	6
Longitudinal analysis of fMRI/iEEG data	High	12
Multi-parameter feature extraction (MRI/fMRI/iEEG) in the current *Data Factory*	High	9
Multi-parameter feature extraction in the *MIP Central*	High	6
Generation of synthetic fMRI data	High	36
Generation of synthetic EEG/iEEG data	High	24
Generation of synthetic features from fMRI/EEG data	High	15

Table 2: Risks and costs for the integration of omics data

Description	Risk	Cost (PM)
Integration of acquisition of genetic data (FASTQ format) in the *Data Capture*	Low	2
Execution of *Galaxy* workflows from the *Data Factory*	Low	3
Execution of workflows for genetic data inside the *Data Factory*	Medium	6
Implementation of *ENIGMA* protocols in the *Data Factory*	Medium	9
Federated analysis of genetic data from different sources	Medium	12
Deep Learning algorithms in the *Algorithm Factory*	Medium	9
Deep Learning algorithms at the Federation Level	High	24
Execution of *Galaxy* workflows in the *MIP Central*	High	6
Implementation of *ENIGMA* protocols in the *MIP Central*	High	9
Generation of synthetic features from genetic data	High	15

Table 3: Risks and costs for the integration of wearable data

Description	Risk	Cost (PM)
Deployment of a parallel *RADAR* platform for managing wearable data	Low	6
Integration of wearable features from *RADAR* in the current *Data Factory*	Low	2
Integration of wearable features from *RADAR* in the *MIP Central*	Low	6
Longitudinal analysis on a parallel *RADAR* installation	Low	6
Deep Learning algorithms in the *MIP Central*	Medium	6
Deep Learning algorithms at the Federation Level	High	24
Longitudinal analysis on the MIP	High	12
Multi-parameter feature extraction from wearable data in the current *Data Factory*	High	9
Multi-parameter feature extraction from wearable data in the *MIP Central*	High	6
Generation of synthetic wearable data	High	24
Generation of synthetic features from wearable data	High	12

6. Conclusions

In this report we have analysed the potential integration of three new types of data in a future version of the MIP. Each of these types poses different challenges and development requirements, but in general all of them are viable and highly desirable. The current MIP is already almost prepared for the integration of fMRI and EEG/iEEG data, and the BIDS specification and open-source pipelines such as those published in the BIDS-Apps repository are the most cost-effective way of including more advanced neuroimaging techniques and features at the MIP Federated level.

Regarding the integration of omics data, this objective has to be narrowed according to the target diseases and the biomarkers that are known to be meaningful from the clinical perspective. From the technical point of view, we recommend to adopt the workflows provided by the Galaxy project and integrate them into the MIP, in line with what has already been done during SGA2.

Finally, the temporal properties of the data produced by wearable devices makes it difficult to adapt it to the current MIP scheme. Furthermore, important challenges arise regarding data acquisition, transmission and storage, and also from the data privacy and security perspectives. Our recommendation in this sense is the adoption of existing solutions from the RADAR platform, its installation as a parallel software stack, and integrating into the MIP just the interesting long-term features that can be adapted to the MIP model.

7. References

1. 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).

2. 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).

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

4. 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).

5. Teijeiro, T. & Atienza, D. Recommendations for the MIP Technical Development During SGA3 (C2976). (2020).

6. Peña-Bautista, C., Baquero, M., Vento, M. & Cháfer-Pericás, C. Omics-based Biomarkers for the Early Alzheimer Disease Diagnosis and Reliable Therapeutic Targets Development. *Curr Neuropharmacol* **17**, 630–647 (2019).

7. Murphy, K., Bodurka, J. & Bandettini, P. A. How long to scan? The relationship between fMRI temporal signal to noise and necessary scan duration. *Neuroimage* **34**, 565–574 (2007).

8. Teijeiro, T. & Atienza, D. Protocol for the Systematic Assessment of Novel Types of Data Integration into the MIP (MS 8.3.1). (2018).

9. Holdgraf, C. *et al*. iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Sci Data* **6**, 102 (2019).

10. Bartolomei, F. *et al*. Defining epileptogenic networks: Contribution of SEEG and signal analysis. *Epilepsia* **58**, 1131–1147 (2017).

11. Stam, C. J. *et al*. The relation between structural and functional connectivity patterns in complex brain networks. *International Journal of Psychophysiology* **103**, 149–160 (2016).

12. Wendling, F., Ansari-Asl, K., Bartolomei, F. & Senhadji, L. From EEG signals to brain connectivity: A model-based evaluation of interdependence measures. *Journal of Neuroscience Methods* **183**, 9–18 (2009).

13. Kaminski, M. J. & Blinowska, K. J. A new method of the description of the information flow in the brain structures. *Biol. Cybern.* **65**, 203–210 (1991).

14. Dai, Y., Zhang, W., Dickens, D. L. & He, B. Source Connectivity Analysis from MEG and its Application to Epilepsy Source Localization. *Brain Topogr* **25**, 157–166 (2012).

15. Baccalá, L. A. & Sameshima, K. Partial directed coherence: a new concept in neural structure determination. *Biol Cybern* **84**, 463–474 (2001).

16. Wendling, F. & Bartolomei, F. Modeling EEG signals and interpreting measures of relationship during temporal-lobe seizures: an approach to the study of epileptogenic networks. *Epileptic Disord* **Spec Issue**, 67–78 (2001).

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

18. Ponten, S. C., Bartolomei, F. & Stam, C. J. Small-world networks and epilepsy: Graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clinical Neurophysiology* **118**, 918–927 (2007).

19. Courtens, S. *et al*. Graph Measures of Node Strength for Characterizing Preictal Synchrony in Partial Epilepsy. *Brain Connectivity* **6**, 530–539 (2016).

20. Pitkänen, A. *et al*. Advances in the development of biomarkers for epilepsy. *The Lancet Neurology* **15**, 843–856 (2016).

21. García-Giménez, J. L. *et al*. Epigenetic biomarkers: A new perspective in laboratory diagnostics. *Clinica Chimica Acta* **413**, 1576–1582 (2012).

22. Han, X. *et al*. Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS ONE* **6**, e21643 (2011).

23. Oberacher, H. *et al*. Targeted Metabolomic Analysis of Soluble Lysates from Platelets of Patients with Mild Cognitive Impairment and Alzheimer's Disease Compared to Healthy Controls: Is PC aeC40:4 a Promising Diagnostic Tool? *JAD* **57**, 493–504 (2017).

24. Filibian, M. *et al*. In vivo imaging of glia activation using ^{1}H-magnetic resonance spectroscopy to detect putative biomarkers of tissue epileptogenicity: *Glia-Related Biomarkers of Epilepsy*. *Epilepsia* **53**, 1907–1916 (2012).

25. Shi, M., Caudle, W. M. & Zhang, J. Biomarker discovery in neurodegenerative diseases: A proteomic approach. *Neurobiology of Disease* **35**, 157–164 (2009).

26. Jiang, W. *et al*. Preliminary explorations of the role of mitochondrial proteins in refractory epilepsy: Some Findings From Comparative Proteomics. *J. Neurosci. Res.* **85**, 3160–3170 (2007).

27. Sevcencu, C. & Struijk, J. J. Autonomic alterations and cardiac changes in epilepsy. *Epilepsia* **51**, 725–37 (2010).

28. Shmuely, S., van der Lende, M., Lamberts, R. J., Sander, J. W. & Thijs, R. D. The heart of epilepsy: Current views and future concepts. *Seizure* **44**, 176–183 (2017).

29. Leutmezer, F., Schernthaner, C., Lurger, S., Potzelberger, K. & Baumgartner, C. Electrocardiographic changes at the onset of epileptic seizures. *Epilepsia* **44**, 348–54 (2003).

30. Nilsen, K. B., Haram, M., Tangedal, S., Sand, T. & Brodtkorb, E. Is elevated pre-ictal heart rate associated with secondary generalization in partial epilepsy? *Seizure* **19**, 291–5 (2010).

31. Thayer, J. F. & Lane, R. D. Claude Bernard and the heart-brain connection: further elaboration of a model of neurovisceral integration. *Neurosci Biobehav Rev* **33**, 81–8 (2009).

32. Moridani, M. K. & Farhadi, H. Heart rate variability as a biomarker for epilepsy seizure prediction. *BLL* **118**, 3–8 (2017).

33. Critchley, H. D. Electrodermal responses: what happens in the brain. *Neuroscientist* **8**, 132–42 (2002).

34. Poh, M. Z. *et al*. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. *Conf Proc IEEE Eng Med Biol Soc* **2010**, 4415–8 (2010).

35. Blum, A. S. Respiratory physiology of seizures. *J Clin Neurophysiol* **26**, 309–15 (2009).

36. Kothare, S. V. & Singh, K. Cardiorespiratory abnormalities during epileptic seizures. *Sleep Med* **15**, 1433–9 (2014).

37. Ryvlin, P. *et al*. Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (MORTEMUS): a retrospective study. *Lancet Neurol* **12**, 966–77 (2013).

38. Osorio, I. & Schachter, S. Extracerebral detection of seizures: a new era in epileptology? *Epilepsy Behav* **22 Suppl 1**, S82-7 (2011).

39. Gietzelt, M., Wolf, K.-H., Kohlmann, M., Marschollek, M. & Haux, R. Measurement of Accelerometry-based Gait Parameters in People with and without Dementia in the Field: A Technical Feasibility Study. *Methods Inf Med* **52**, 319–325 (2013).

40. Weiss, A. *et al*. Toward Automated, At-Home Assessment of Mobility Among Patients With Parkinson Disease, Using a Body-Worn Accelerometer. *Neurorehabil Neural Repair* **25**, 810–818 (2011).

41. Moore, S. T. *et al*. Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry. *J NeuroEngineering Rehabil* **10**, 19 (2013).

42. Beniczky, S., Conradsen, I., Pressler, R. & Wolf, P. Quantitative analysis of surface electromyography: Biomarkers for convulsive seizures. *Clin Neurophysiol* **127**, 2900–7 (2016).

43. Joo, H. S. *et al*. Spectral Analysis of Acceleration Data for Detection of Generalized Tonic-Clonic Seizures. *Sensors (Basel)* **17**, (2017).

44. Ulate-Campos, A. *et al*. Automated seizure detection systems and their effectiveness for each type of seizure. *Seizure* **40**, 88–101 (2016).

45. Larobina, M. & Murino, L. Medical Image File Formats. *J Digit Imaging* **27**, 200–206 (2014).

46. Rorden, C. *dcm2nii DICOM to NIfTI conversion*. (2007).

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

48. Wu, G.-R. *et al*. A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data. *Medical Image Analysis* **17**, 365–374 (2013).

49. Litjens, G. *et al*. A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017).

50. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. https://arxiv.org/abs/1603.04467.

51. Paszke, A. *et al*. PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 8024–8035 (Curran Associates, Inc., 2019).

52. Telzer, E. H. *et al*. Methodological considerations for developmental longitudinal fMRI research. *Developmental Cognitive Neuroscience* **33**, 149–160 (2018).

53. Kiral-Kornek, I. *et al*. Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System. *EBioMedicine* **27**, 103–111 (2018).

54. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat Rev Genet* **14**, 333–346 (2013).

55. Single Nucleotide Polymorphism - SNPedia. https://www.snpedia.com/index.php/Single_Nucleotide_Polymorphism.

56. Giri, M., Zhang, M. & Lü, Y. Genes associated with Alzheimer's disease: an overview and current status. *Clin Interv Aging* **11**, 665–681 (2016).

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

58. ENIGMA-Epilepsy « ENIGMA. http://enigma.ini.usc.edu/ongoing/enigma-epilepsy/.

59. ENIGMA-Parkinson's « ENIGMA. http://enigma.ini.usc.edu/ongoing/enigma-parkinsons/.

60. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief Bioinform* **18**, 851–869 (2017).

61. Real-time physiological signals | E4 EDA/GSR sensor. *Empatica* https://www.empatica.com/research/e4.

62. Oracle Internet of Things Cloud Service. *Oracle Help Center* https://docs.oracle.com/en/cloud/paas/iot-cloud/index.html.

63. Apache Kafka. *Apache Kafka* https://kafka.apache.org/.