AI Stroke Neurorehabilitation

Public report

Data Management Plan

Action number: **101057655**
Action Acronym: **AISN**
Action title: **Integrating AI in Stroke Neurorehabilitation**
Author(s): **Petra Ritter - Charité, Ciprian Ciulpan - EBRAINS**

Submission date: 02 May 2024
Public project details can be found on www.AI-SN.eu.

Funded by the European Union

Table of Content

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	31.Mar.2023	Initial version
1.1	09.Aug.2023	Inclusion of Personal Data Protection Section
1.2	14.Aug.2023	FAIR sensitive data
1.3	03.Mar.2024	Partner input changes

Introduction

This document constitutes the Data Management Plan for the AISN Project.
The data management plan (DMP) shall outline how data will be collected, processed, stored, shared, and preserved throughout the AISN research project. The DMP scope includes all aspects of data management, from data collection to data sharing and preservation. It describes the types of data that will be collected, the methods and tools used for data collection and processing, as well as the mechanisms for data storage, backup, and security. Additionally, the DMP outlines how the data will be shared with other researchers, including any restrictions on access and use, and the measures taken to ensure data privacy and confidentiality. Finally, the DMP will cover aspects regarding the long-term preservation and accessibility of the data, including the data archiving and curation processes that will be used.

The AISN platform for clinical decision-support and intervention delivery is built from the integration of three validated TRL8 or higher platforms for data acquisition and access (EBRAINS' Knowledge Graph, TRL8), clinical interpretation, and whole-brain simulation (CHARITE's Virtual Research Environment, Virtual Brain Cloud, TRL4/5), and intervention planning, delivery and optimization (EODYNE's Rehabilitation Gaming System, TRL6) with advanced AI tools for the extraction of predictive information from complex high-dimensional and heterogeneous medical data (SADDLE's Bayesian Inference Engine, TRL5). All these constituent platforms have been tested in realistic environments of direct relevance to the current use case: a pathway for establishing standard operating procedures and technologies for the integration of AI in stroke care supporting clinical decision-making and treatment delivery. Data is obtained from individual patients and cohorts and integrated into patient-specific models. These are used to predict treatment outcomes and inform interventions delivered online using mobile/desktop apps and wearables closing the loop with clinicians and patients. The AISN pipeline implements a continuous care pathway and provides a concrete example of how the Action Plan for Stroke Europe of the European Stroke Organization can be realized. It thus provides a realistic and necessary context in which the fundamental question on guidelines of AI-enhanced healthcare can be addressed in the domain of stroke and beyond.

1. Data summary

2.1. Data sources

To the possible extent from a modelling point of view, the project relies on pathological image data and their derived models of The Virtual Brain (TVB) which have partially been created and registered outside of the project already.
Otherwise, except for generic, shared metadata constructs such as “controlled terms” of the openMINDS metadata structures, the project relies on data generated and collected within the project itself.

The project aims to process these types of data:

- Datasets from consortium partners and publicly available
- Pseudonymised clinical data extracted from medical information systems
- Derived data related to the usage and processing of the data
- Experimental data collected during the clinical trial

Additionally, log files of the services involved in the overall process will record information about the quantity and regularity of data exchange and interaction as well as general information required for maintenance.

1.1.1. Pseudonymised clinical data extracted from medical information systems

Structural and functional neuroimaging data: volumetric data (T1, weighted T1, T2), single-photon emission CT, fluid attenuation inverse recovery (FLAIR), positron emission tomography (PET), diffusion-weighted imaging (DWI) and functional MRI (fMRI), neuropsychological evaluation, stereo encephalography (SEEG), neuropsychological evaluations, locomotion and kinematics. Such data is collected and stored in the Hospital Information System of the hospital(s).

1.1.2. Derived data related to the usage and the processing of the data: data profiles, pre-processing and medical data flows, analytics, biomedical and statistical simulation models, etc.

1.1.3. Experimental data collected during the clinical trial

Patient outcome measurements will be collected, covering ICF categories at baseline, after the end of the therapy and at 6 months follow-up, mainly text and numerical values. The AISN interventions, based on the RGS system will collect data from different platforms: RGSClinic, RGSWear and RGSWeb. Each platform records Kinematic data (movement) of the patients during the sessions (and continuously in the case of RGSWear) as well as performance indicators and task complexity modulators. The estimated number of patients to be recruited during the trial is 115 and it is expected that each patient performs daily training sessions during the period of the trial.
Whilst sensor data (of wearables) and RGS session information will be provided by the partner Eodyne, the pathological image data as well as the TVB models and simulations are provided by Charité. The AI models are generated by Saddlepoint. The size of the data varies between the several types of data due to their different representation and resolution (in the case of

images). We expect the size of the individual files in ranges from a few hundred KB to several MBs in the case of textual, structured information (e.g., JSON / XML) while image data could easily take hundreds or even thousands of MBs. To ensure the suitability for the establishment of a solid AI model, the collected data must originate at least from several 100eds of different subjects.
The estimated volume of data is 10GB per patient, so in total AISN could collect 1.1TB of data. In general, each patient will use RGS for about 4 weeks.

2.2. Data registration

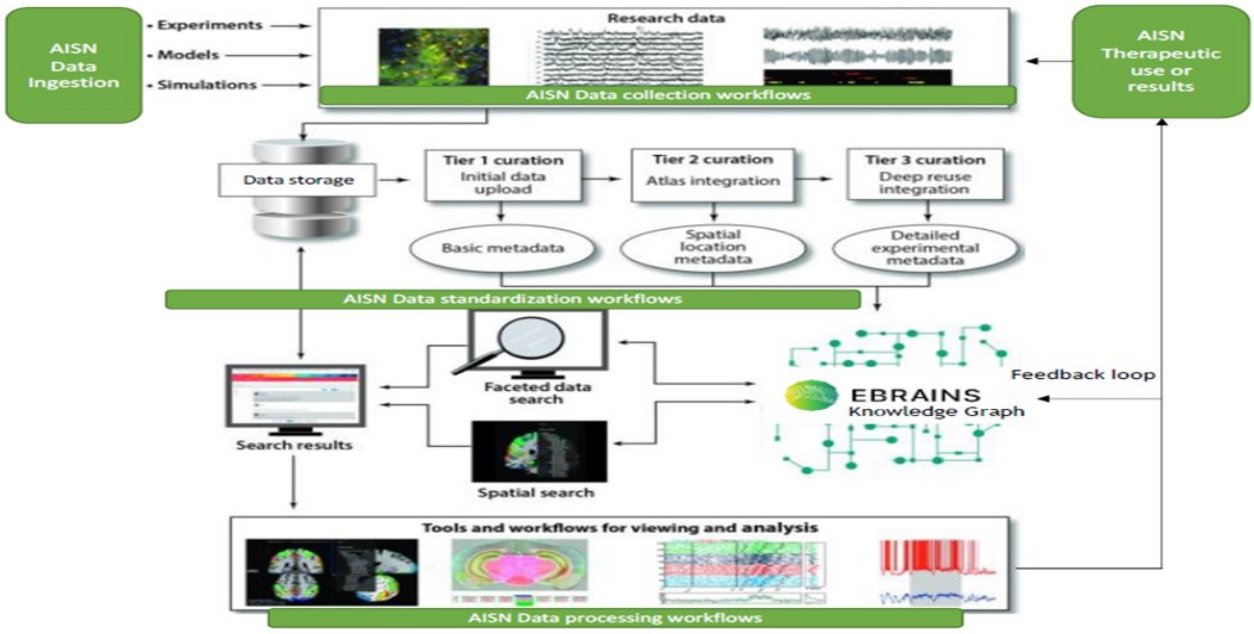
The data will be registered and made available for the refinement of the AI models which can lead to reuse as part of reapplication in the model generation process. Generated generic AI models are meant for reuse e.g., inside RGS (Rehabilitation Gaming System) to improve the level of personalization, recovery predictions, and for giving indications to potential optimizations in the treatment process.
For the above-mentioned types of data, appropriate metadata is generated and recorded in the Knowledge Graph following the openMINDS[1] metadata structures developed by the Human Brain Project and used by EBRAINS Research Infrastructure to allow categorisation and criteria-based selection of the data structures for further processing.

2.3. AI training data

By training AI models with the collected data, those models are going to be one of the generated derivatives of the project.
The data generated and gathered by the means of wearables and activity logs of the use of the RGS system as well as the use of clinical information shall allow for improvement in the simulation and modelling of the long-term development of stroke patients. By closing the data loop between the various components involved, improved feedback cycles shall lead to a continuously improved and individualized therapeutic application and treatment of individual patients.

2.4. Data Flow and Processing

[1] https://github.com/HumanBrainProject/openMINDS



Schematic view of the data flow of the AISN project and the integrations with the EBRAINS KG. The main components are Data ingestion to link with the AI pipeline and link with visualizations; Data standardization workflows including NLP extraction intended to support clinical report standardization; Data collection workflows within the protected BRAINS Health Data Cloud (including the clinical trial requirements) ensure that legal and ethical standards for data generation, use, storage, and sharing are applied; Knowledge Graph further developments: to ensure maximal usability of data the consortium will also work together to build on existing terminology (e.g., ontologies, controlled vocabularies, and terminologies) to ensure high quality metadata annotation of all data produced and used by the project.

The RGS applications record user and session data and stores them in an RGS-specific data repository/database. Automated processes ensure the selection of valid data (e.g., the removal of incomplete sessions) and applies de-identification mechanisms, like those that are part of the clinical data. The resulting data is maintained in a separate file repository to ensure that only valid data is registered and to avoid data contamination.

2. FAIR data

2.1. Making data findable, including provisions for metadata

The project will make the data discoverable and reusable by their metadata registration in the EBRAINS Knowledge Graph. Information about provenance and usage conditions can be registered as part of the overall metadata provisioning.

The data and models meant for external reuse and/or with the requirement for citability will be provided with a persistent identifier (Digital Object Identifiers) whilst the data meant for internal processing will be referenceable by a KG (Knowledge Graph) system-specific identifier as well as potentially pre-existing external identifiers. For temporary data which is not going to be annotated in the EBRAINS Knowledge Graph, no such identifier will be made available.

Whilst file name conventions and directory structures of the underlying data are highly dependent on the capturing devices and follow the logic the manufacturer assigned, the data (to the granularity of the individual file) will be annotated by following the openMINDS metadata structures and ensure the appropriate and rich description of the involved entities which will go beyond the restrictions provided by hierarchical file structures. Alongside the possibility to assign free text keywords, the assignment of predefined "controlled terms" in semantically explicitly specified properties allow contextualized and well-defined annotation as well as an automated linkage to external resources such as ontological terms and their hierarchies. Whenever possible, existing community schemas and ontologies will be used. If any of those resources need to be extended, it will be done so via an extension of the OpenMINDS metadata schema used by EBRAINS RI (Research Infrastructure).

Additionally, we strive to apply the brain imaging data structure (BIDS) community standard to all data. This standard is continuously extended for new data types by so called BIDS extension proposals (BEP)[3] one of which presently under discussion being the BEP034 Computational Modelling[4] led by Charité.

The registration of the metadata in the EBRAINS Knowledge Graph allows a clean versioning of the metadata which can be made accessible via the publicly available Knowledge Graph6[5] Search as well as via the EBRAINS Knowledge Graph Query API[6]. Therefore, it can be made accessible for both, humans as well as computer programs/scripts.

2.2. Making data accessible

The data produced will be persisted – according to their sensitivity level and use – in appropriate, trusted long-term (>10 years) and potentially access-protected repository. For non-sensitive data, meant for public sharing, this would be the EBRAINS data services. More sensitive data (e.g., brain image data) will be in a dedicated certified infrastructure complying with the required regulations such as the Virtual Research Environment and EBRAINS Health

[3] https://bids.neuroimaging.io/get_involved.html#extending-the-bids-specification
[4] https://zenodo.org/record/7962032
[5] https://search.kg.ebrains.eu
[6] https://query.kg.ebrains.eu

Data Cloud[7].

Regardless of their location, the registration of the metadata in the EBRAINS Knowledge Graph will be done centrally for all involved data (except for temporary data as part of the processing pipelines which is not registered at all). This allows the generation of persistent identifiers regardless of their location. Whenever possible by legal restrictions (e.g., GDPR), the metadata is meant to be made available publicly and open. The metadata stored alongside the persistent identifier (usually a DOI) is therefore going to be available even if the data or the KG would be no longer available. Also, because the metadata in the KG is kept separately from the underlying data, these two entities are going to have different lifecycles and therefore metadata can be kept separately from the data.

The data will be provided – depending on the underlying repository - as downloadable resources over HTTPS by default (information about how to access data as well as license information will be provided by the metadata in a machine-readable format). Additional services such as ZIP services / other protocols will be discussed as part of the implementation phase.

Sensitive data will be accessible within the dedicated infrastructures for sensitive data such as the Virtual Research Environment and Health Data Cloud. These infrastructures provide functionality and compute resources that facilitate even large-scale processing activities in compliance with GDPR.

For access-protected data, EBRAINS Health Data Cloud provides access control, encryption and network isolation mechanisms as well as organizational measures (e.g. a roles concept) to provide data protection and information security. Within Health Data Cloud (that is based on the VRE architecture), the data can be accessed via web browser as well as by scripts / other services by well-known standards and no specific software or toolset is required and any specific documentation (other than the API documentation) is required.

As foreseen by GDPR, the legal data controllers of specific data sets will remain in control of their data. It is solely under their responsibility to exert control over who gains access to their data for which processing purposes. Consequently, no additional data-access committees are initially foreseen.

The project information classification scheme requires information assets to be protectively marked into one of 3 classifications (excluding public information which does not need to be marked). The way the information is handled, published, moved, and stored will be dependent on this scheme.
The classes of information are:

- Level 0: Public (or unclassified): Much of the information held by the organization is freely available to the public via established publication methods. Such items of information have no classification and will not be assigned a formal owner or inventoried.

[7] https://www.ebrains.eu/health-research-platforms/health-platforms/work-with-health-data-2

- Level 1: Protected: For information that is not published freely by the organization, some of this may be classified as Protected. This is typically information, which is relatively private in nature, either to an individual or to the organization and, whilst its loss or disclosure is unlikely to result in significant consequences, it would be undesirable. The criteria for assessing whether the information would be classified as Protected include whether its unauthorised disclosure would:
    - Cause distress to individuals
    - Breach of proper undertakings to maintain the confidence of information provided by third parties
    - Breach of statutory restrictions on the disclosure of information
    - Cause financial loss or loss of earning potential, or facilitate improper gain
    - Give an unfair advantage to individuals or companies
    - Prejudice in the investigation or facilitating the commission of a crime
    - Disadvantage the organization in commercial or policy negotiations with others
- Level 2: Restricted: The level above Protected is that of Restricted. This information would be more serious if it were disclosed to unauthorised persons and result in significant embarrassment to the organization and possibly legal consequences. The criteria for assessing whether the information would be classified as Restricted include whether its unauthorised disclosure would:
    - Affect relations with other organizations adversely
    - Cause substantial distress to individuals
    - Cause financial loss or loss of earning potential or facilitate improper gain or advantage for individuals or companies
    - Prejudice in the investigation or facilitating the commission of a crime
    - Breach of proper undertakings to maintain the confidence of information provided by third parties
    - Impede the effective development or operation of organizational policies
    - Breach of statutory restrictions on disclosure of information
    - Disadvantage the organization in commercial or policy negotiations with others
    - Undermine the proper management of the organization and its operations

  Information falling into the classification of “Restricted” will typically be handled by Work package leaders and above, with some employees of lower clearance being given access only in specific circumstances.
- Level 3 Confidential: The highest level of classification is that of Confidential. This is reserved for information which is highly sensitive and would cause major reputation and financial loss if it were lost or wrongly disclosed. The criteria for assessing whether the information would be classified as Confidential include whether its unauthorised disclosure would:
    - Materially damage relations with other organizations (i.e., cause formal protest or other sanction);
    - Prejudice of individual security or liberty;
    - Cause damage to the operational effectiveness or security of the AISN partners
    - Work substantially against organizational finances or economic and commercial interests of the AISN partners;

- Substantially undermine the financial viability of the AISN partners and major organizations;
- Impede the investigation or facilitate the commission of a serious crime;
- Impede seriously the development or operation of organizational policies;
- Shut down or otherwise substantially disrupt significant business operations.

Access to information assets defined as "Confidential" will be tightly controlled by the General Assembly of AISN and in many cases, numbered copies of documents will be distributed according to specific procedures. In case of confidential personal data the above mentioned technical and organizational measures do apply (implemented in the Health Data Cloud/Virtual Research Environment).

2.3. Making data interoperable

The project will register metadata about the generated and collected data by application of the openMINDS standard. Alongside the possibility to describe the research context in a fine-granular way, this standard also allows the annotation of individual files with so-called "ContentTypes" which are an extension of IANA MediaTypes[8]). This allows to explicitly specify which software can produce and/or consume a specific file and therefore be key for interoperability.

To ensure maximal usability of data the consortium will also work together to build on existing terminology (e.g., ontologies, controlled vocabularies, and terminologies) to ensure high quality metadata annotation of all data produced and used by the project. To support this EBRAINS will provide professional curation support, which will facilitate metadata annotation. The metadata will be accessed depending on the users' level. This will be based on adapting the EBRAINS Knowledge Graph leveraging its existing searchable and discoverable functionalities.
OpenMINDS integrates and harmonizes various vocabularies and ontological terms, whilst keeping the reference to external identifiers (e.g., Interlex, Uberon, …). The vocabulary to be used when applying openMINDS is standardised by so-called "controlled terms" which allow inter-disciplinary interoperability by providing term mapping and integration of external identifiers for mapping purposes. Since those controlled terms are a community effort, not-yet specified terms can and will be contributed to the openMINDS project incl. their mappings to external references.
As part of the metadata, openMINDS also allows to register provenance information allowing to link e.g., which data and/or model derived from which other data structures.

2.4. Increase data re-use

We will annotate the data identified to be potentially shareable with the required information for interpretation as part of the metadata registration and will aim for an as-complete representation as possible. One part of this contains the possibility to link data descriptors, example code and other external resources providing additional information which can be

[8] https://www.iana.org/assignments/media-types/media-types.xhtml

specified whenever it is needed. The published data will remain available after the project as part of the EBRAINS data and knowledge services. The provenance will be registered by the openMINDS metadata structures and before publication, the data will pass the curation processes of EBRAINS which include automated and manual quality as well as ethical checks.

3. Other research outputs

Alongside the pure data, the project will generate AI models. The models will be registered the same way as the data by being represented in the EBRAINS Knowledge Graph, described with openMINDS structures, and registered with persistent identifiers (DOIs). Accordingly, the defined management plan for data also applies for those models.

4. Allocation of resources

The annotation processes as well as the persistence should be automated to the max and be made part of the general data flow. Accordingly, the costs shall be kept as low as possible. The cost distributes across:

- Operating the metadata management system
- Operating the long-term storage system incl., the costs “per GB”
- Registration of DOIs (Digital Object Identifiers)
- Establishing the integration of the data processing pipelines with the metadata management system and the long-term storage solution.
- Maintenance of the integration between the data processing pipeline with the metadata management system and the long-term storage solution
- Contribution of added terms / customizations / extensions of openMINDS and its impacts on the metadata management system

For the duration of the AISN project, data pipeline operational costs will be covered by the project work plan.

5. Data security

Data is backed according to the backup plans of the infrastructures where they are stored, for instance VRE has a backup schedule that can be adapted to the needs of the project.

Data is only persisted on sites proven to be technically approved to keep the data of their sensitivity level and distributed only according to a strictly specified data flow within the internal components. It must be ensured that data required for processing is either kept in the components only temporarily (ideally in memory only) and is removed after the processing immediately or the processing is executed on the sites of the data itself. Only data stored on the long-term storage is meant for persistence after the processing is complete. All data (not only the one categorized as sensitive) is always transferred via industry-standard encrypted transport channels (e.g., HTTPS).

Similarly, for metadata, the according metadata management system will ensure the recovery of data with an acceptable maximum data loss of 1 day (24 hours). The metadata - independent of its sensitivity level - is always transferred via industry-standard encrypted transport channels (HTTPS) between systems. The metadata management system provides a fine-grained permission management control system allowing restricted access to individual parts of the metadata.
For authentication, industry-standard authentication systems are in place (oAuth 2.0).

The Virtual Research Environment and EBRAINS Health Data Cloud provide health data solutions on certified so called “critical” infrastructure following ISO 27001 information security standards. This comprises auditing and certification renewal every two years. Amongst many other measures in place, one is the continuous scanning for novel vulnerabilities using state of the art technologies and mechanisms in place to remove any new vulnerabilities immediately after detection. On top of this information security standard, the VRE/HDC have been audited for EU data protection compliance. Thus, users of these infrastructures can demonstrate compliance with GDPR when processing sensitive data.

6. Personal Data Protection and Ethics

This section of the deliverable focuses specifically on certain aspects of management and processing of personal data in the AISN project. This section outlines the ethical considerations pertaining to data management.

Personal data

Article 4(1) of the GDPR defines personal data as "any information related to an identified or identifiable natural person". An "identifiable" natural person is someone who can be identified through direct identifiers such as a name, or indirect identifiers such as an identification number, location data or a combination of health data. During the AISN project, the following personal data will be processed:

- Pseudonymised clinical data extracted from medical information systems
- Data related to health and experimental data collected during the clinical trials
- All other personal data collected from external natural persons who participate in the clinical trial or other project related activities. This may include the names and contact details of external participants.

Article 1(5) of the GDPR defines pseudonymisation as "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information". Through the process of pseudonymisation, the elements which can be used to identify the data subject are eliminated and kept separately. However, with the use of additional information, the data subject could still be identified therefore, pseudonymised personal data is still regulated under the GDPR as personal data.

Regarding data anonymization, the level of de-identification is restricted to the removal of person related tags and annotations. Due to the nature of data potentially containing individual markers such as biometric information, the level of de-identification cannot be seen to be neither pseudonymization nor anonymization from a GDPR perspective. Accordingly, the resulting data has still to be treated as personal, sensitive data.
Partners EBRAINS, CHARITE, SRU and EODYNE have broad experience in data management. When sharing data, models and software on AISN KG, VRE and RGS, researchers must comply with European Union and national legal and ethics standards. For raw and/or derived personal data obtained from living human subjects, compliance with the EU General Data Protection Regulation (GDPR) must also be assured. AISN/EBRAINS Ethics Compliance Management ensures that all accepted data meet ethical and legal requirements. In addition, users that want access to GDPR-sensitive data in AISN need to sign and comply with the applicable Data Processing Agreements that is the EC Standard Contractual Clauses[9].

The Virtual Research Environment at Charité is a certified and audited infrastructure where users can demonstrate compliance with GDPR when processing personal data. GDPR and other legal and ethical aspects for the data stored and processed in the scope of this project, see D5.2 Legal and Ethical Framework Report.

[9] https://commission.europa.eu/publications/standard-contractual-clauses-controllers-and-processors-eueea_en

AISN maintains a separate Data Management Committee with regulatory expertise from the University of Vienna and ethical expertise from the University of Oxford, ensuring the highest standards of privacy, confidentiality and patient participation.

In the context of the AISN project, the informed consent of the data subject forms the legal basis for the processing of non-sensitive personal data. Where special categories personal data is processed, two legal grounds under Article 9(2) are applicable. These grounds are the explicit consent of the data subject (Article 9(2)(a)) and the research exemption (Article 9(2)(j)).

Further details on personal data processing and GDPR compliance in the AISN project can be found in D5.2 Legal and Ethical Framework V1 as well as in D5.1 Data Protection Impact Assessment.

7. Other issues

We will not make use of other national/funder/sectorial/departmental procedures for data management.