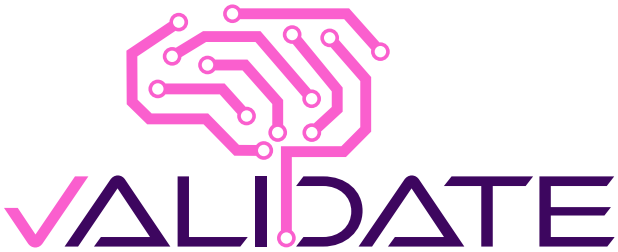

DELIVERABLE

Project Acronym: **VALIDATE**

Grant Agreement number: **101057263**
Project Title: **Validation of a Trustworthy AI-based Clinical Decision Support System for Improving Patient Outcome in Acute Stroke Treatment**

D3.1 – Research data management plan (DMP)

Revision: 1.0

Authors and Contributors	Julian Aengenheister (IBM iX); Adam Hilbert (Charité – Universitätsmedizin Berlin); Giorgio Colangelo (NORA Health S.L.)			
Responsible Author	Julian Aengenheister		**Email**	Julian.aengenheister@ibmix.de
	Beneficiary	IBM iX	**Phone**	+491627836508

Project co-funded by the European Commission within **HORIZON-HLTH-2021-DISEASE-04-04**		
Dissemination Level		
PU	Public, fully open	x
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101057263

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
0.1	13.10.2022	Julian Aengenheister	IBM iX	Initial creation
1.0	20.10.2022	Julian Aengenheister	IBM iX	First release version

Date of delivery	Contractual:	28.10.2022	Actual:	31.10.2022
Status	final ☒ /draft ☐			

Abstract (for dissemination)	This report describes how project VALIDATE'S data are being re-used, created, collected, processed, interpreted, preserved, and accessed. The intention of using the guideline provided by the European Commission is to proof our process of research data management adheres to the common understanding of the research community in EU funded projects. To create a standardised result, we use the Data Stewardship Wizard with its Common DSW Knowledge Model (ID: dsw:root:2.4.4) knowledge model, a project that created a collaboration tool to create data management plans according to Horizon 2020 standards for research data management and scores to measure whether the project follows FAIR and open data principles. With a scale from 0.0 to 1.0 (lowest to highest) and scores of 0.8 to 1.0 for Findability, Accessibility, Interoperability, Reusability and Openness, our results show this project generally follows the FAIR and open data standards. The creation of this Data Management Plan (DMP) also revealed that certain practices related to findability and reusability are not following all the guidelines yet and suggests that they should be addressed during the delivery of the project. Other results of project VALIDATE's ongoing work packages may influence details of this research data management plan which will be updated regularly.
Keywords	Data Management Plan (DMP), data management

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

Table of Content

Executive Summary

For the VALIDATE project, data is central to provide information to both, machine-learning models, and clinicians. As we, ourselves, rely on FAIR data, that enables systems to collaborate with each other, we also ensure FAIR data principles across all beneficiaries of the project, allowing for better data governance during the project and for further development and recreation after the project has ended. By providing a data management framework, these described principles guide us in managing our data and facilitate knowledge discovery standards.

Data sets are findable and accessible if descriptive metadata, that is understandable to humans and machines, is added. This includes data sets to be stored in trusted repositories. Using a formal and broadly accepted language for knowledge representation, metadata also enables the data sets to be interoperable. Finally, accurate information on usage licence and origin makes data sets reusable.

This data management plan (DMP) is based on the Horizon 2020 FAIR DMP template[1] which has been digitized by ds-wizard.org[2]. It guides us with the right questions at the right time and will be a living document that will be reviewed and updated regularly and be released in its final version at the end of the project as it is expected to be influenced by subsequent deliverables, e.g. "*D3.2 integrated requirements report covering technical and user requirements"*.

The major point of this report is, for the data being created and collected, that formats and processes are already standardised and established according to FAIR data principles. This is particularly important as the created data is sensitive and includes personal patient information.

With well-established formats (DICOM, FHIR, JSON, XML) we agreed on widely used standards allowing us to reuse open-source toolsets and frameworks that can already collaborate. It is expected that this list of formats is nonexhaustive and may be extended throughout the project.

The required data sets we can already foresee to train the machine-learning model and to show inside the mobile app, include electronic patient records, medical imaging data and laboratory results. Access to these data sets is dependent on the decision of every institution's ethical committee and yet to be described in a consortium agreement. A detailed list of variables will be described during the project and an updated version of the research data management plan is to be expected.

To determine their condition, patients who agreed to the informed consent form will be interviewed using standard patient-reported outcome measures (PROM) survey 90 days after treatment. Following EU GDPR, informed consent can be withdrawn at any time. By complying with the regulations described in GDPR, we maintain a strict level of data privacy and security standards to protect the patients' data. Given that we process personal data of EU citizens or residents, and offer services to such people, the GDPR applies to us. By following the GDPR, we ensure data integrity, security, and privacy, collect only minimal data, and store it on certified servers. We also anonymise personal data as soon as possible.

[1] European Commission. (2016). Data management – H2020 Online Manual. URL: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

[2] Pergl, Robert & Hooft, Rob & Suchánek, Marek & Knaisl, Vojtěch & Slifka, Jan. (2019). "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. Data Science Journal. 18. 10.5334/dsj-2019-059.

1 Data Summary

This chapter summarizes the survey results of chapters “II. Re-Using Data”, “III. Creating and Collecting Data”, and “IV. Processing Data” (see Appendix A) related to the purpose of the collected data and its relation to the objectives of the project as well as expected types, formats, origin, size, and utility of the data sets.

1.1 Datasets

We collect data from questionnaires and electronic patient records. The datasets are:

- **Patient-Reported Outcome Measures (PROM)** – This data set will collect information about the patient's condition 90 days after treatment. The PROM will be collected using a tool from one of our partners NORA Health S.L. and only if the patient has given consent to the informed consent form.
- **Clinical Data** – This data set contains the electronic patient record. It will be used to train the machine-learning model. Individual records of this data set will be shown to the clinician in the mobile application.
- **Medical Imaging Data** – This data set contains CT and MIR scans. It will be used to train the machine-learning model. Individual records of this data set will be shown to the clinician in the mobile application.
- **Clinical Laboratory Data** – This data set contains the patients' laboratory data. It will be used to train the machine-learning model. Individual records of this data set will be shown to the clinician in the mobile application.

1.2 Data formats and types

We will be using the following data formats and types:

- Digital Imaging and Communications in Medicine standard (DICOM;.dcm)
  It is a standardized format. This is a suitable format for long-term archiving.
- Fast Healthcare Interoperability Resources (FHIR)
  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.
- JavaScript Object Notation (JSON)
  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.
- Extensible Markup Language (XML)
  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2 FAIR Data

According to the article "The FAIR Guiding Principles for scientific data management and stewardship"[3] digital assets shall be findable, accessible, interoperable, and reusable. This article advocates specifically for researchers to focus on adhering to these principles to support the increasing use of computational resources needing access to data.

The following paragraphs provide an overview of the survey results in chapter "III. Creating and Collecting Data" (see Appendix A) describing how our data will be findable, accessible, interoperable, and reusable.

2.1 Making data findable, including provisions for metadata

- Patient-Reported Outcome Measures (PROM)
- Clinical Data
- Medical Imaging Data
- Clinical Laboratory Data

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret data in the future. We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2 Making data openly accessible

We will be working with the philosophy *as open as possible* for our data.

All our data can become completely open immediately.

Data that is not legally restrained will be released after a fixed period (The data will be deposited at the end of the project. The results by the consortium the data will be embargoed until 1 year after project end.), unconditionally.

Metadata will be openly available. Metadata will be available in a form that can be harvested and indexed (managed by the used repository / repositories).

We have a consortium agreement that arranges Intellectual Property.

For our produced data, conditions are as follows:

- Patient-Reported Outcome Measures (PROM)
- Clinical Data
- Medical Imaging Data
- Clinical Laboratory Data

2.3 Making data interoperable

We will be using the following data formats and types:

- Digital Imaging and Communications in Medicine standard (DICOM;.dcm)
  It is a standardized format.

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

- Fast Healthcare Interoperability Resources (FHIR)
  It is a standardized format.
- JavaScript Object Notation (JSON)
  It is a standardized format.
- Extensible Markup Language (XML)
  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- Informed Consent Ontology (ICO) (https://doi.org/10.25504/FAIRsharing.b9znd5)

2.4 Increase data re-use (through clarifying licenses)

The metadata for our produced data will be kept as follows:

- Patient-Reported Outcome Measures (PROM)
- Clinical Data
- Medical Imaging Data
- Clinical Laboratory Data

As stated already in Section 2.2, all our data can become completely open immediately.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3 Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

Julian Aengenheister is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Marc Ribo and Adam Hilbert are responsible for finding, gathering, and collecting data.

To execute the DMP, no additional specialist expertise is required.

We require the following hardware or software in addition to what is usually available in the institute: To determine their condition, patients who agreed to the informed consent form will be interviewed using a standard patient-reported outcome measures (PROM) survey 90 days after treatment. Following EU GDPR, informed consent can be withdrawn at any time. PROM surveys will be managed using software of one of the beneficiaries, NORA Health S.L. An additional tool to manage informed consent will be researched during the project.

4 Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...).

All personal data will be anonymized as early as possible.

We are running the project in a collaboration between different groups and institutes. However, there is no collaboration agreement in the project that describes who can have access to what data, yet. The creation of a collaboration agreement is still ongoing and will describe conditional access for each data set.

5 Ethical aspects

As project VALIDATE not only creates and collects data but also aims to provide medical recommendations to clinicians a particular focus is on ethical usage and processing of data. At this stage of the project an overview of the ethical aspects of the data can be found below.

During this project work packages “WP1 Trustworthy AI” and “WP2 Clinical Model Development” pay specific attention to the ethical use of data in machine-learning aided AI models and will require this data management plan to be updated as soon as the following deliverables are due:

- D1.1 VALIDATE trustworthy AI framework and manual
- D1.2 Z-inspection process result report
- D1.6 SOP guideline on trustworthy AI development of an AI-based clinical decision support system
- D2.5 SOP guideline on model development, validation, and lifecycle management of AI models for prognostic tools

Overview of ethical aspects:

- Patient-Reported Outcome Measures (PROM)
  - It contains personal data.
  - It contains sensitive data.
- Clinical Data
  - It contains personal data.
  - It contains sensitive data.
- Medical Imaging Data
  - It contains personal data.
  - It contains sensitive data.
- Clinical Laboratory Data
  - It contains personal data.
  - It contains sensitive data.

Data we collect

We will collect data connected to a person, i.e. "personal data". We ask the data subjects for their consent. We will collect consent for our use as well as reuse of the data. The consent form will be available for re-users. An ethical committee will make an ethical review on the project. We need to conduct a data protection impact assessment (DPIA).

6 References

European Commission. (2016). Data management – H2020 Online Manual. URL: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

Pergl, Robert & Hooft, Rob & Suchánek, Marek & Knaisl, Vojtěch & Slifka, Jan. (2019). "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. Data Science Journal. 18. 10.5334/dsj-2019-059.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

7 Appendix A: Data Stewardship Wizard Survey & Results

Due to the complexity of the data management plan survey the raw result of the survey can be accessed in a structured format via the following URL: https://dsw.ibmix.de/projects/f82f2ef8-c156-427d-bf84-8ac6d0025eb4. For access to this tool, please contact Julian Aengenheister, julian.aengenheister@ibmix.de.