Project No. 964220

Intelligent digital tools for screening of brain connectivity and dementia risk estimation in people affected by mild cognitive impairment

Deliverable 1.4

AI-Mind data governance and data management protection framework

Part 1: AI-Mind Data Governance Framework p11-39

Part 2: Managing Data in AI/ML Development Projects in Healthcare p41-58

This document is for research purposes only and its contents are not prescriptive. Accordance with this document may not be sufficient to ensure compliance with and does not replace regulatory requirements from: MDR/IVDR, the EU act on high-risk AI, or other statutory requirements.

WP 1 – Concept governance and requirements of the AI-Mind Connector and AI-Mind Predictor

Authors	DNV, Lurtis, UCM, OUS, AALTO, IRCCS, HUS, OsloMet, Brainsymph
Lead participant	DNV
Delivery date	25th February 2022
Dissemination level	Public
Type	Report

Version 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964220

Revision History

Author(s)	Description	Date
Part 1		
Atle Stein Kvalheim, Harry Hallock, Serena Marshall and Abdillah Suyuthi (DNV)	Draft deliverable	01/05/2021
Victor Ayllón and José María Peña Sánchez (Lurtis)	Revision 1st draft	17/08/2021
Haneef Awan (TSD)	Revision 2nd draft	01/10/2021
Christoffer Hatlestad-Hall (Brainsymph)	Revision 2nd draft	05/10/2021
Federico Ramírez Toraño (UCM)	Revision 2nd draft	06/10/2021
Francesca Miraglia (IRCCS)	Revision 2nd draft	14/10/2021
Susanne Merz and Enrico Glerean (AALTO)	Revision 2nd draft	14/10/2021
Hanna Renvall (HUS)	Revision 2nd draft	17/10/2021
Anis Yazidi (OsloMET)	Revision 2nd draft	18/10/2021
Part 2		
Abdillah Suyuthi (DNV)	Draft deliverable	23/12/2021
Abdillah Suyuthi, Harry Hallock, Serena Marshall (DNV)	Revision 1st draft	25/01/2022
D1.4 Combined		
Abdillah Suyuthi, Harry Hallock, Serena Marshall (DNV)	Draft deliverable (v1)	27/01/2022
Victor Ayllón (Lurtis)	Revision 1st draft	31/01/2022
Federico Ramírez Toraño (UCM)	Revision 1st draft	03/02/2022
Ira Haraldsen (OUS)	Revision 1st draft	03/02/2022
Susanne Merz (AALTO)	Revision 1st draft	04/02/2022
Francesca Miraglia (IRCCS)	Revision 1st draft	04/02/2022
Hanna Renvall (HUS)	Revision 1st draft	05/02/2022
Anis Yazidi (OsloMET)	Revision 1st draft	07/02/2022
Abdillah Suyuthi, Harry Hallock, Serena Marshall (DNV)	Revision 2nd draft (v2)	07/02/2022
José María Peña Sánchez (Lurtis)	Revision 2nd draft	09/02/2022
Abdillah Suyuthi, Harry Hallock, Serena Marshall (DNV)	Final Version	15/02/2022
Andreia Cruz (accelCH)	Final Revision	21/02/2022
Harry Hallock, Serena Marshall (DNV)	Final Version	23/02/2022

Abbreviations

AI	Artificial Intelligence
BIDS	Brain Imaging Data Structure
CA	Consortium Agreement
CANTAB	Cambridge Neuropsychological Test Automated Battery
CD	Continuous Delivery
CDR	Central Data Repository
CD4ML	Continuous Delivery for Machine Learning
CI	Continuous Integration
CM	Continuous Monitoring
CT	Continuous Training
DL	Deep Learning
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DQ	Data Quality
DTA	Data Transfer Agreement
EC	European Commission
EEG	Electroencephalography
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GDPR	General Data Protection Regulation
H2020	Horizon 2020
HBP	Human Brain Project
ISO	International Organization for Standardization
IVDR	In Vitro Diagnostic Regulation
MCI	Mild Cognitive Impairment
MDR	Medical Device Regulations
MEG	Magnetoencephalography
ML	Machine Learning
MLOps	Machine Learning Model Operationalisation Management
NLP	Natural Language Processing
SQL	Structured Query Language
TSD	Services for Sensitive Data (Tjenester for Sensitive Data)
UI	User Interface
UiO	University of Oslo
WP	Work Package

Definitions

Actor	A person, organisation or system that has one or more roles that initiates or interacts with activities [1].
AI-Mind Model	A program, algorithm or mathematical model derived from AI-Mind data using either classical machine learning or deep learning techniques. The AI model allows it to reach a conclusion or make a prediction when provided with sufficient information e.g., AI-Mind Predictor and Connector models.
AI-Mind Tool	Final user application in the format of a web-accessible TSD-hosted service to which users (initially partners within the project) connect to get access to the AI models and to obtain predictions on specific participant data, i.e., AI-Mind Predictor and Connector tools.
Algorithm	A process or set of rules to be followed in calculations or other problem-solving operations.
Anonymisation	The process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party [2].
Application	Packaging of the model to enable its deployment to its production/use environment (e.g., software application)
Architecture	(System) fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution [3].
Data	A representation of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automatic means. A collection of values assigned to base measures, derived measures and/or indicators [4]. In AI-Mind, the following are defined as data: participants' clinical data (e.g., EEG/MEG measurements, blood samples, results from neuropsychological testing, etc), processed data, detail of the problem framing, featured data, training, validation, and test datasets, the trained model (binary file), and the model evaluation result. For more details see Part 1 Section 5.1.
Data governance	The strategy, policies, processes, roles, and responsibilities that are required for data management and continuous monitoring and improvement in data quality. This is implemented by development, execution, and supervision of plans, policies, programs and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles [5].
Data management	Processes to ensure that important data assets are formally managed throughout the enterprise and that data governance goals are achieved. It establishes and utilises processes, controls, and technologies that operate on data and enable compliance with policies and governance directions of data maintenance and data value chain in the enterprise [6].
Data processing	Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction [7].
Data quality	A measurement to which the data conforms to syntactic, semantic, and pragmatic quality definitions. Syntactic quality is the degree to which data conforms to requirements stated by the metadata. Semantic quality is the degree to which data corresponds with that which it represents. Pragmatic quality is the degree to which data is appropriate and useful for a particular purpose [6].
Data-driven	Decisions are made according to data analysis and data interpretation.

Entity	Concrete or abstract thing (element e.g., physical artefact, concept, actor) in the domain under consideration [8].
Evaluation score	Piece of information (usually a number) which represents the performance measure of a ML model.
Featured data	A measurable property of a phenomenon under consideration, determined by problem framing for model development.
Hyperparameters	Any configuration which is used to control the learning process, external to the model.
Metadata	Data about data or data elements, possibly including data descriptions; information about data ownership, access paths, access rights and data volatility [9]. For example, in AI-Mind metadata describing the participant clinical data, i.e., description of the content of EEG data file, or names and ranges of values of the different neuropsychological tests.
Parameters	A configuration that determines how input data is transformed into the desired output, internal to the model.
Personal data	Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [10].
Problem framing	A systematic effort to understand, define, structure, and decompose the AI/ML problem, which allow the formulation of hypotheses and/or approaches to solve the problem.
Processed data	Transformed data that has undergone required cleansing, imputation, organisation and structuring etc.
Prospective Data	Data collected from participants enrolling into the AI-Mind research study. It will consist of data from 1000 MCI patients who will undergo multiple (four time points over two years) MEG, EEG and CANTAB testing, and genetic (APOE4) and P-tau 181 tests once during the project.
Pseudonymisation	The 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, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person [10].
Raw data	Unprocessed data that is available for the development team to produce AI/ML solution(s).
Requirement	A statement that translates or expresses a need and its associated constraints and conditions [11].
Retrospective Data	data originating from multiple data controllers, which can be shared beyond the scope of the original project they were collected for, based on the subjects' informed consent. In AI-Mind, retrospective data will be shared by four of the five clinical partners (excluding HUS).
Role	A defined function to be performed by a project team member, such as testing, filing, inspecting, coding [12].
Stakeholder	An individual or organisation having a right, share, claim, or interest in a system or in its possession of characteristics that meet their needs and expectations [13].
Source data	Retrospective and prospective data that has been transferred to the TSD Platform by clinical partners but has yet to be processed or transformed within the TSD Platform.
Trained model	The culmination of featured data input run through an algorithm.

TSD Platform	Centralised secure database provisioned in TSD (located at the University of Oslo, Norway), where AI-Mind data will be collated and processed.
Unprocessed data	Data not yet processed by the AI/ML development team. Some pre-processing may have been carried out by the data provider.
WP5 Platform	The platform, including UI, used by AI-Mind's clinical partners for the clinical trial, for data analysis and visualisation. It will be hosted within TSD.

Partner Short Names

OUS	Oslo University Hospital
AALTO	Aalto University
accelCH	accelopment Schweiz AG
Brainsymph	BrainSymph AS
DNV	Det Norske Veritas
HUS	Helsinki University Hospital
IRCCS	Scientific Institute for Research, Hospitalization and Healthcare, San Raffaele Roma
Lurtis	Lurtis Rules S.L
OsloMET	Oslo Metropolitan University
UCM	Complutense University of Madrid

Table of Contents

Executive Summary

In AI-Mind, data initially takes the form of collected data such as participants' clinical data (e.g., EEG/MEG measurements, neuropsychological testing, etc). This data is then processed and used for developing ML/DL algorithms, thus data later also takes the form of featured data, training, validation, and test datasets and the trained model, plus more. Due to DNV's designation as a Notified Body under the EU medical device regulation (MDR), and thus DNV's need to maintain impartiality, DNV will not be part of the latter processes nor work with those types of data. As such, D1.4 has been split into two parts:

- D1.4 Part 1: AI-Mind Data Governance Framework – specific for AI-Mind; this details relevant regulations, policies and standards related to data governance, the AI-Mind Data Strategy and Data Governance Policy, DTA and DPIA. Based on these, a plan for AI-Mind Data Management and the types and roles of different Data, Processes and Actors are outlined.
- D1.4 Part 2: Managing Data in AI/ML Development Projects in Healthcare – generic for healthcare projects; this details the different types of data used in AI/ML projects, and general considerations on how to handle these types of data.

The splitting of this deliverable into two parts enables D1.4 to cover all aspects related to governance and management of the different types of data in AI-Mind, but at two different levels, where Part 1 is specific for AI-Mind, and Part 2 is generic for healthcare projects.

D1.4 Part 1 provides a framework which will enable AI-Mind partners to exercise authority and control over the management of the AI-Mind data. This will enable data to be managed and processed in a compliant and secure manner by partners which ensures quality, and security.

D1.4 Part 2 introduces the complexities associated with AI/ML development projects in healthcare, outlines the types of data in such projects, and provides general considerations about data quality and management for those data.

This document is for research purposes only and its contents are not prescriptive. Accordance with this document may not be sufficient to ensure compliance with and does not replace regulatory requirements from: MDR/IVDR, the EU act on high-risk AI, or other statutory requirements.

D1.4 Part 1: AI-Mind data governance framework

Organisation of D1.4 Part 1

Figure 1 presents the overview of how Part 1 is organised. The background and motivation, objective, scope, main references, interdependencies with other AI-Mind deliverables, and AI-Mind stakeholders are described in the Introduction Chapter 1. The starting point on developing a framework for AI-Mind data governance and data management is the regulations, standards, and best practices, which are presented in Chapter 2. Chapter 3 outlines the AI-Mind data strategy, data governance policy, relevant parts of the data transfer agreement, and the data protection impact assessment as per regulations. AI-Mind data management, which describes the requirements on how to implement the data strategy and data governance policy is presented in Chapter 4. Chapter 5 outlines the specificity of the AI-Mind project related to data and processes in the project, as well as actors involved in the processes. The solid numbered circle in Figure 1 represents the chapter number. Each chapter from Chapters 2-5 are drawn as a rectangle inside another rectangle and this signifies the relationship of those chapters as described previously.

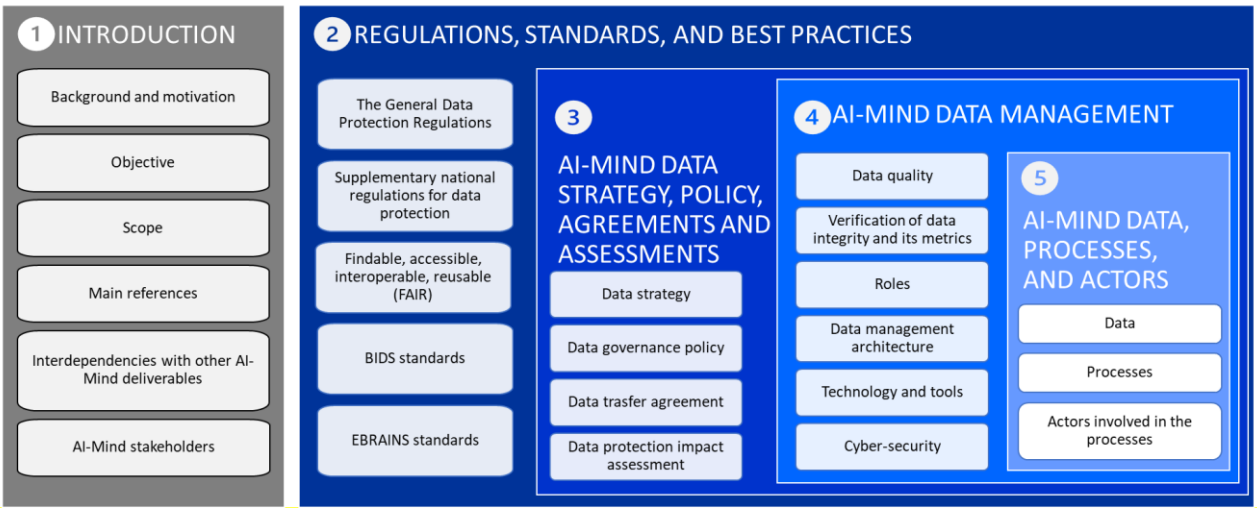


Figure 1. Organisation of part 1 of this deliverable (i.e., D1.4 Part 1).

1 Introduction

1.1 Background and motivation

The overarching aim of AI-Mind is to create an AI-supported cloud-based diagnostic system that integrates a portfolio of new and existing data analysis methods that are non-invasive, widely available and low-cost.

An important component of the work to achieve this goal is **data governance**, i.e., the exercise of authority and control over the management of the data assets. Data governance is often viewed as high-level executive data stewardship.

The DAMA-DMBOK Guide describes data management as "the planning and execution and oversight of policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data and information assets."[5]

The present data governance framework has thus been developed to describe how data shall be managed and processed within the defined scope to ensure compliance, quality, and security.

1.2 Objective

The main objectives of D1.4 Part 1 are to:

- Define the scope of the data governance and management framework.
- Align the governance of personal data within the project with national legislations, best practices, and recommendations.
- Analyse the requirements arising from the data transfer agreement (DTA) and Consortium Agreement (CA).
- Identify the roles and needs of AI-Mind stakeholders.
- Describe the data governance policies and, on a high level, the governance roles, data/metadata, quality, processes, management, architecture, technology/tools, and cybersecurity.

Importantly, D1.4 Part 1 specifies the framework, not a detailed implementation.

1.3 Scope

D1.4 Part 1 i.e., the AI-Mind data governance framework, covers aspects and considerations related to:

- Prospective and retrospective participant data.
- Prospective and retrospective participant metadata.
- Data about the transfer of prospective and retrospective participant data and metadata.
- The transfer, processing, and storing of the above-mentioned data and metadata to the central data repository (CDR) in TSD.

Due to DNV's designation as a Notified Body under the EU medical device regulation (MDR), and thus DNV's need to maintain impartiality, there are certain restrictions on what can be covered in D1.4. D1.4 Part 1 does not cover:

- Feature extracted data *(covered at a high, non-AI-Mind specific level in D1.4 Part 2)*
- AI/ML models *(covered at a high, non-AI-Mind specific level in D1.4 Part 2)*
- Processed data used by the AI tools in the WP5 Platform.
- Data produced by the AI tools.
- Any processes beyond the CDR in TSD Platform.

- The clinical platform.

This framework has been developed based on the following assumptions:

- All prospective data will undergo pseudonymisation at the clinical partners' site, prior to being transferred to the TSD Platform. Refer to the Clinical study dossier (deliverable 5.1) for more information about the initial ID (local inclusion) and AI-Mind participant ID (generated once inclusion criteria is fulfilled).
- All retrospective data will be anonymised at the clinical partners' site, prior to being transferred to the TSD Platform. The anonymisation techniques are described in deliverable 9.4.
- Cambridge Neuropsychological Test Automated Battery (CANTAB) data will be transferred to TSD Platform from Cambridge Cognition Ltd. cloud storage via an API solution implemented on an intermediate server hosted and maintained by TSD.

1.4 Main references

Part 1 uses the following ISO standards and technical reports:

- ISO 8000-8:2015 Data quality — Part 8: Information and data quality: concepts and measuring [8]
- ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes [13]
- ISO 25237:2017 Health informatics — pseudonymisation [2]
- ISO/IEC/IEEE 42010:2011 Systems and software engineering — Architecture description [3]
- [ISO/IEC/IEEE 15939:2017] Systems and software engineering — Measurement process [4]
- [ISO/IEC 2382:2015] Information technology — Vocabulary [9]
- [ISO/IEC/IEEE 29148:2011] Systems and software engineering — Life cycle processes — Requirements [11]

1.5 Interdependencies with other AI-Mind deliverables

AI-Mind is a project that requires a breadth of disciplines and expertise to deliver its goals. Thus, this deliverable will either provide input or benefit from the output of other deliverables:
The following have provided input for D1.4 Part 1:

- D8.3 Project handbook (delivered M3)
- D2.1 Standardisation of available and prospective data collection (delivered M8)
- D1.3 Guidelines for data management and sharing (delivered M9)
- D5.1 Clinical study dossier (Clinical study protocol, Informed Consent Form) (delivered M9)
- D4.1 Software requirements & specifications (SRS) (delivered M10)

D1.4 Part 1 may provide input for:

- D2.2 Implementation of data governance and data sharing frameworks and database structure at the central database (delivered M12)
- D2.3 Standardisation protocols for data collection and pre-processing (to be delivered in M16)

- D2.4 Standardisation of data integration and data repository management (to be delivered in M37)
- D2.5 Updated standardisation protocols for data collection and pre-processing (to be delivered in M38)
- D1.6 Best clinical guidelines for implementation of AI-Mind (to be delivered in M60)

1.6 AI-Mind stakeholders

Through collaboration, the AI-Mind project aims to develop innovative solutions to diagnose mild cognitive impairment (MCI) more readily and predict dementia risk (see Figure 2).

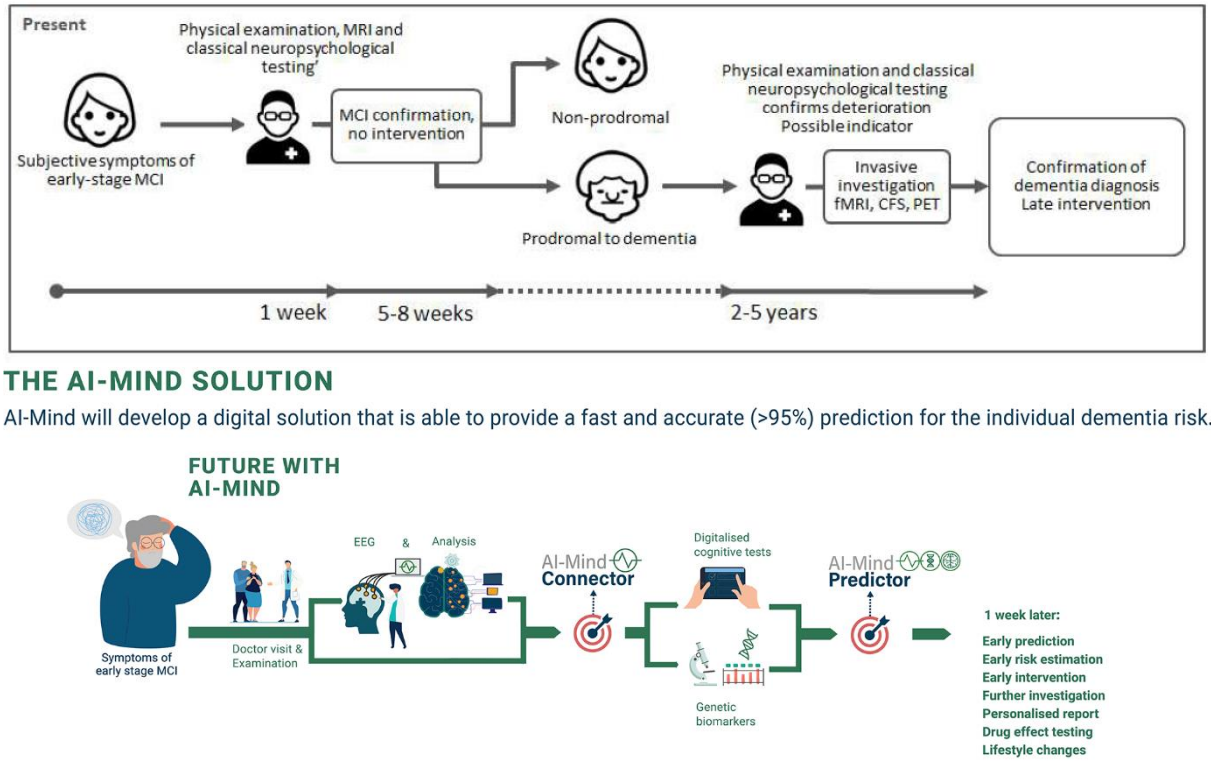


Figure 2. The present and future AI-Mind patient journey[1]

Many AI-Mind project partners have a history of collaborations, which has established a framework for working together, knowledge sharing and trust. Ways to innovate and collaborate with new and future stakeholders must be considered, as all these different actors make up the stakeholders of the AI-Mind project.

To provide the guidelines for data governance, security, protection, and management of the AI-Mind data, the role and needs of current and future key stakeholders have been identified and are depicted in Table 1. All AI-Mind project partners are responsible for results dissemination and where applicable, preparation and dissemination of scientific manuscripts.

[1] 2021 © accelopment Schweiz CH

Table 1. AI-Mind stakeholders, their roles, and needs

Stakeholder		**Role**	**Needs**
Participants		Provide personal information at multiple time points.	Personal data protection and high-quality care.
AI-Mind Project Partners	Clinical partners (data owner and data collector)	Engage with, collect, quality assure and transfer participant data; validate AI-Mind predictions.	Secure and efficient procedures relating to data collection, transfer, processing, and storage. Being able to cast prediction on the existing participants and validate the results.
	AI researchers	Access and build the AI-Mind connector and predictor models.	High-quality complete data sets.
	IT professionals and software developers	Develop the necessary infrastructure to enable collaboration between clinical partners, AI researchers, systems architect, data stewards and innovation manager. Embedding the AI models into the final AI-Mind tools (Connector and Predictor).	Complete overview of data, legal/ethical requirements, relevant stakeholders' needs.
	Data steward	Manage and maintain the retrospective and prospective source data in the TSD platform, ensure that the data is available whenever needed, maintenance of the data management process, harmonise the data format, especially for retrospective data.	Data management procedures and tools that help to monitor and manage the flow of data.
	Systems architect	Design, manage and maintain the architecture descriptions developed in the project.	Information on the requirements and specifications of the project and feedback from other consortium partners to design/modify the data pipeline and manage the data flow.
	Platform administrator	Responsible for the technical and administrative IT procedures within TSD.	Data management procedures and tools that help to monitor and manage the IT platform and the flow of data.
	Innovation manager	Ensure the commercialisation of AI-Mind tools.	Complete overview of legal requirements for the developed tools, and overview of grant, consortium and data transfer agreements.

	General Assembly	Guide AI-Mind project and tool development.	Transparency of information flow from AI-Mind.
Advocacy	Policy makers	Inform and promote change to existing health laws and procedures.	Notification of changes required to benefit patients, healthcare, industries, and society.
	Patient groups	Ensure project(s) align and deliver according to patient's needs.	Promising applications implemented into the clinic.
Future users / Post project	Clinics	User of the commercialised AI-Mind tools and platform.	Reliable, affordable, accurate and interoperable tools.
	AI-Mind product owner (i.e., commercial partners)	Provision of the commercialised AI-Mind tools and platform.	Ongoing maintenance and development of platform, route to market.
	Patients	Beneficiary of improved prediction tools.	Improved quality of life outcomes.
	Data sharing initiatives e.g., EBRAINS	Platform for sustainable use of personal data generated.	Access to participant data for new and alternative needs.
	Researchers	Creation of new tools and technologies.	Access to data.
Funder	European commission	Provision of funds and guidelines.	Strengthened economies and societies.

This data governance framework (i.e., D1.4 Part 1) accounts for the needs of all current stakeholders for the duration of the AI-Mind project and sets the foundation for a framework that should be developed for the needs of future stakeholders.

2 Regulations, standards, and best practices

To ensure the protection of personal data, as well as the interoperability and sustainability of the AI-Mind data, AI-Mind will comply with various regulations, standards, and best practices. They are listed in the following sub-sections.

2.1 The General Data Protection Regulations

The (EU) 2016/679 General Data Protection Regulation (GDPR) entered into force from May 25th, 2018, applies to all member states (EU/EEA) and acts to harmonise data protection and privacy laws across Europe [10]. The GDPR lists the following rights for individuals: the right to be informed, the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object, and also rights around automated decision making and profiling.

Specific data processing laws relevant to AI-Mind are described in deliverable 1.3 - *Guidelines for data management and sharing*. The GDPR requires minimal requirements to be met and documentation supporting evidence of compliance relating to:

- **Lawfulness, fairness and transparency –** the basis for legal processing of data, and creation of a privacy and information policy to share with participants, process for editing of participant details. AI-Mind will ensure that participants will receive information about how AI-Mind will process their data, and they must provide informed consent to participate before joining the study.
- **Purpose limitation** – determination of what information is processed. AI-Mind has ensured that an outline of what participant data will be collected and processed has been described in deliverable D5.1.
- **Data minimisation –** ensuring that only data required for fulfilling AI-Mind objectives is collected and processed. AI-Mind has ensured that prior to commencement of the project, the plan for necessary data to be included for achieving AI-Mind objectives was made. An outline of the refined data collection plan has been described in deliverable D5.1.
- **Accuracy –** reasonable steps must be taken to ensure that inaccurate or incomplete data is rectified or erased. AI-Mind has ensured that tools to estimate data accuracy are described in Section 4.5 Technology and tools, and they will be developed in WP2, WP3 and WP4.
- **Storage limitation –** development of a robust retention policy to ensure that personal data is not kept for longer than required. AI-Mind has ensured that the DTA signed between data owners (UCM, HUS, IRCCS, UCSC) and the recipient (OUS) defines the storage retention time. This is summarised in Section 3.3.8 Data retention and disposal clause and more details can be found in Clause 2.1 viii in the DTA.
- **Integrity and confidentiality (security) –** determination of who has or should have access, consider data protection throughout the data lifecycle. AI-Mind has ensured that the DTA defines an access and protection policy (Section 3.3.1 Practical terms for transfer, use and storage of the data).**Data** and derivatives of the data *shall be processed and / or used only within the recipient's computational infrastructure* [TSD Platform] *and handled in accordance with the General Data Protection Regulations (GDPR)*(1.2.iii.).
- **Identification -** Methods to ensure confidentially of the data will be developed. AI-Mind has ensured this, see Section 3.3.4 Confidentiality and 3.3.5 Access revocation. Section 4.1.2 of D2.1 describes the pseudonymisation process. Furthermore, TSD is compliant with Norwegian data protection laws.
- **Accountability –** assigning responsibility for data protection (e.g., data protection officer (DPO)), creation of data processing agreements. AI-Mind has ensured that each clinical partner has its own DPO, and the OUS DPO acts as the overarching DPO for the project. A DTA has been signed by all clinical partners and allows the use of data by all AI-Mind partners. Information about accountability regarding data governance and partner's roles can be found in Section 4.3.

2.2 Supplementary national regulations for data protection

All countries involved within AI-Mind operate within the EU-EEA and therefore must comply with the GDPR. In addition to this, countries may have their own rules and legislation that supplements the

GDPR. Details about these can be found in AI-Mind deliverable 1.3 -*Guidelines for data management and sharing*.

2.3 Findable, accessible, interoperable, reusable (FAIR)

AI-Mind will follow the FAIR principles [14], thus the data within reason as dictated by the DTA, will be:

- Findable:
  - F1. (meta)data are assigned a globally unique and persistent identifier.
  - F2. data are described with rich metadata (defined by R1 below).
  - F3. metadata clearly and explicitly include the identifier of the data it describes.
  - F4. (meta)data are registered or indexed in a searchable resource.
- Accessible:
  - A1. (meta)data are retrievable by their identifier using a standardised communications protocol.
    - A1.1 the protocol is open, free, and universally implementable.
    - A1.2 the protocol allows for an authentication and authorisation procedure, where necessary.
  - A2. metadata are accessible, even when the data are no longer available.
- Interoperable:
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles.
  - I3. (meta)data include qualified references to other (meta)data.
- Reusable:
  - R1. meta(data) are richly described with a plurality of accurate and relevant attributes.
    - R1.1. (meta)data are released with a clear and accessible data usage licence.
    - R1.2. (meta)data are associated with detailed provenance.
    - R1.3. (meta)data meet domain-relevant community standards.

How AI-Mind will achieve FAIR data is captured in Section 4 - AI-Mind Data Management.

2.4 BIDS standard

The electroencephalography (EEG) and magnetoencephalography (MEG) data collection procedures in the AI-Mind project have been developed to ensure compliance with the Brain Imaging Data Structure (BIDS) standard (https://bids.neuroimaging.io). The BIDS standard follows FAIR principles, merges data from different brain imaging modalities into the same general framework, and additionally specifies naming conventions for structuring brain imaging data and metadata.

2.5 EBRAINS standards

EBRAINS, a new digital research infrastructure continuing from the Human Brain Project (HBP)[15], [16], has awarded the AI-Mind project a service category 3 voucher for Brain modelling and simulation workflows: integrated tools to create and investigate models of the brain. AI-Mind has established a collaboration to exchange (meta)data with EBRAINS and/or use EBRAINS services and will also align with an associated project called HealthDataCloud. Should (meta)data exchange occur, AI-Mind will align our own data schema with the openMINDS schema and ensure compliance with EBRAINS Knowledge Graph.

3 AI-Mind Data Strategy, Policy, Agreements and Assessments

3.1 Data strategy

The governance of data and plans to ensure that this goal is addressed adequately within a project should always be an integral part of any organisation or consortium's data strategy. A data strategy should provide goals and objectives to ensure data is used effectively and efficiently in the current and future tasks. A data strategy requires a supporting data management plan.

AI-Mind's data strategy is to ensure that AI-Mind-generated data provides value to partners and participants on a sustainable basis. Data creation, transfer, sharing and processing must be handled safely and securely within the AI-Mind project. After or externally to the AI-Mind project, data must be shared securely within the legal confines, as agreed upon by the data providers (i.e. clinical partners) for continued value creation (see alignment through 2.5 EBRAINS standards).

3.2 Data governance policy

Policy statements:

- Within TSD, data will be processed in a GDPR compliant manner.
- Data can only be accessed by personnel authorised by the:
    - Respective principal investigator for data located at each of the five clinical partner sites,
    - TSD Platform co-ordinator (project coordinator, Ira Haraldsen) for data located on TSD.
- Participant data shall be pseudonymised, and where possible, anonymised.
- Quality of participant data (e.g., minimum common requirements across sites for EEG/MEG raw data) shall be evaluated and logged. Requirements for data quality, standardisation and pre-processing will be established in D2.3 and participant inclusion/exclusion criteria in D5.1.
- International standards and best practices for data governance shall be used where possible.
- Data, metadata and processes shall be controlled and managed. See Section 5 AI-Mind Data, Processes and Actors.

3.3 Data Transfer Agreement

Data transfer agreements (DTAs) were signed between OUS (recipient) and (legal) representatives from the four clinics (provider; UCM, UCSC, IRCCS and HUS) that will transfer their data to the TSD

Platform at OUS/UiO. Details relating to data sharing and minor differences between the DTAs were identified and described in D1.3.

The DTA (version 1 amendment 1; signed 28/04/21) establishes conditions for the transfer and use of data within the AI-Mind consortium and specifies requirements for confidentiality, publications, protection and ownership. These are detailed accordingly below, and where the Consortium Agreement (version 1; signed 27/04/21) is deferred to, then additional information is referred to from there. The DTA covers the transfer of generated data (material and data) that is stated for inclusion as described within Annex 1 of the AI-Mind grant agreement (964220), and additional existing EEG data see section 5 AI-Mind Data, Processes and Actors, for details.

Within the following sub-sections, *italicised* text represents information quoted directly from the DTA.

3.3.1 Practical terms for transfer, use and storage of the data

Practical terms for use and transfer of the data (clause 1) are agreed upon, namely that:

- Data is made available on *a royalty free and non-exclusive basis*, for the specified research only, in accordance with the subsequent conditions (1.1.).
- Data will only be used by individuals working within the recipient's institution for the purpose of conducting AI-Mind research, *to the exclusion of any commercial application* not contractually mutually agreed with the provider (1.2.i.).
- Data and derivatives of the data *shall be processed and / or used only within the recipient's computational infrastructure* [TSD Platform] *and handled in accordance with the General Data Protection Regulations (GDPR)*(1.2.iii.).

3.3.2 Identification

Data will only be made available in pseudonymised form (personal identifiers replaced with code), and there are no circumstances in which the recipient should attempt to re-identify the participants (1.1.). Additionally, data, and derivatives of the data, *will not be used, whether alone or in conjunction with other information, in any effort whatsoever to establish the individual identities* of any data subject (1.6.).

3.3.3 Consent

One legal basis for using the data collected and processed within AI-Mind is Article 6 (1) of the GDPR (a) consent. The DTA states that data, and derivatives of the data*, shall not be duplicated, transferred, distributed, or supplied to any third party..., for any purpose or use without the prior written consent of the provider* (1.2.ii.)

More details relating to the different legal bases for data sharing and processing within AI-Mind can be found in Section 2.3.5 (Legal basis for data processing by AI-Mind's clinical partners) of D1.3 - *Guidelines for data management and sharing*.

Within AI-Mind some partners have based their processing on Article 6 - Lawfulness of processing (1) (a) consent (for one or more specified purposes), (1) (c) legal obligation and (1) (e) protection of public interest and/or some partners have based their processing on Article 9 – Processing of special categories of data (2) (j) necessary for scientific research purposes in accordance with Article 89 (*which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject*), (2) (a) *the participant has given explicit consent for processing* to the specified purpose(s), and (2) (i) necessary for reasons of public interest in the area of public health.

3.3.4 Confidentiality

Confidentiality (clause 2) provisions state that unless the DTA explicitly permits, data, and derivatives of the data, is disclosed in the strictest confidence and therefore:

- (2.1.i.) cannot be disclosed by the recipient,
- (2.1.ii.) must be handled with at least the same degree of care and security (*but always at least a reasonable degree of care*) as the recipient uses with their own confidential data
- (2.1.iii.) sole use of the data is for conducting the stated research, and not *for any other purposes without the prior written consent of provider or until further agreement*s can be made.
- (2.1.iv.) data use is restricted to employees within the institution who require such access in order to carry out the research, in accordance with confidentiality and non-disclosure obligations. As such (2.1.v.) data must be *protected from unauthorised access*, any deviations must be reported in writing to the provider and reasonable steps must be taken to regain control and confidentiality.
- (2.1.vii.) legally binding requests for data disclosure must be conveyed to the provider promptly.

3.3.5 Access revocation

Access authorisations must be revoked as soon as they are no longer required (2.1.vi.)

3.3.6 Duration and termination

The DTA comes into force from its effective date (April 2021) and *will remain in force for the duration of the research* (7.1.), although termination of the DTA may be initiated by either recipient or provider with one month's written notice at any time (7.2.), without prejudice to rights and liabilities of either Party accrued prior to expiry or termination (7.3.), and with continuation of provisions which have an effect either expressly or impliedly, after expiry or termination (7.4).

3.3.7 Contract breach liability clause

If contract breach, or inability to fulfil obligations stated in appendix 1 [of the DTA], the DTA may be terminated, the degree of success of the research shall be communicated (in writing) and data use will be discontinued. (1.3., 1.4., 1.5.)

The DTA defers to the Consortium Agreement for AI-Mind with respect to liability (clause 5). The Grant Agreement, chapter 6 - section 2, article 46 details the liability for damages of the commission (46.1) and the beneficiaries (46.2). *The commission cannot be held liable for any damage caused by or to the beneficiaries or to third parties as a consequence of implementing* the AI-Mind project. *The beneficiaries (except in the case of force majeure) must compensate the commission for any damage it sustains* as a result of non-compliance or implementation action.

3.3.8 Data retention and disposal clause

When the research is completed, or upon other termination of this DTA, (see DTA 1.5), *followed by request and instructions from the provider*, the data, and data copies, must be deleted. *The recipient may keep a copy to the extent it is required* to archive data for compliance with applicable laws (2.1.viii).

3.3.9 Governing law and court of choice for resolving disputes

Three of the providers (HUS, IRCCS, UCSC) defer to the Consortium Agreement with respect to governing law and jurisdiction, where mediation of disputes shall be carried out under the jurisdiction of the courts of Brussels (clause 10 of the DTA). One provider (UCM) requires their DTA to be *governed by Spanish law and jurisdictions*, with tribunals in Madrid having *exclusive right to deal with any dispute which may arise in connection with this agreement.*

The grant agreement of AI-Mind (Article 57.2) details the process for dispute settlement if the *interpretation, application, or validity of the Agreement cannot be settled amicably*. *The General Court — or, on appeal, the Court of Justice of the European Union — has sole jurisdiction*.[2] However, if such a dispute is between the Commission and OUS, accelCH, BrainSymph, DNV, and/or OsloMet; *the competent Belgian courts have sole jurisdict*ion.

3.3.10 Limitations of the DTA

The DTA *imposes no obligation on the recipient with respect to any information which*:

- *at the time of the disclosure is generally available to the public* or becomes so later, *otherwise than through the fault or negligence of the recipient* (2.2.i.).
- *can be shown by written records to have been in the recipient's possession prior to the time of the disclosure* (2.2.ii.), *is rightfully given by to the recipient by a third party under no* confidentiality obligation(2.2.iii.), *is independently developed* (2.2.iv.), *has been identified as no longer confidential by the provider* (2.2.v.) or has been requisitioned by a court or administrative order (2.2.vi.).

[2] Actions must be brought under Article 272 of the Treaty on the Functioning of the EU (TFEU).

However, the data *shall not be deemed to be available to the public or be in the recipient's possession merely because they can be reconstructed* from multiple public sources if none of these sources *actually teaches or suggests the entire combination*, along with its meaning and importance (2.3.).

3.3.11 General provisions, auditing, and notices

The DTA makes general provisions (clause 6) relating to superseding prior communications and agreements (6.1.), delivering notices under this agreement in writing to the address detailed (6.2.), non-liability for loss, damage, or delay as a result of *force majeure* (6.3) and no failure on the part of either party to exercise any right or remedy the DTA shall be construed as a waiver thereof.

A duty to confirm compliance and allow audits (clause 8) states that upon request the recipient shall provide sufficient information to satisfy that the terms of the DTA are met (8.1.) and that requests relating to processing will be dealt with promptly, with submission of auditing for use of data and processing activities. The AI-Mind grant agreement states that *the European Court of Auditors (ECA) may — at any moment during implementation of the action or afterwards* —have the right of access for the purpose of checks and audits.[3]

Notices (and the like) that are necessary in accordance with the DTA (clause 11) *shall be in writing* (11.1.) and *notifications may be made by fax or* email as long as source and destination can be evidenced.

Two providers acknowledge that data protection requirements (clause 12) will defer to the Consortium Agreement of AI-Mind (HUS and UCM), the other two (IRCCS and UCSC) do not mention clause 12. The Grant Agreement states (39.2) that *the beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements).*

3.3.12 Third parties

Rights and obligations covered by the DTA cannot be transferred from the recipient to third parties without prior consent of the provider (9.1) and either party is entitled to demand renegotiation (9.2).

3.3.13 Intellectual property

The DTA defers to the Consortium Agreement for AI-Mind with respect to intellectual property (clause 3 of the DTA). Agreements on background (data, know-how or information) held by beneficiaries which is needed for implementation of AI-Mind (Article 24 of the GA), and ownership of results (tangible or intangible) that will be generated (Article 26 of the GA) were made in writing and are built upon principles described in the project proposal.

[3] Under Article 287 of the Treaty on the Functioning of the European Union (TFEU) and Article 161 of the Financial Regulation No 966/201218

3.3.14 Publications

The DTA defers to the CA for AI-Mind with respect to publication (clause 4 of the DTA). UCM additionally specify that published data must be *irreversibly anonymised* to preserve participant confidentiality. Article 29 of the GA states the obligations of beneficiaries to disseminate results, *unless it goes against their legitimate interests*, as soon as possible. Advance notice (45 days – unless otherwise agreed) of intended results that will be disseminated must be disclosed to other beneficiaries. Any other beneficiary may object within 30 days of receiving notification if it can show that its legitimate interests in relation to the results or background would be significantly harmed. In such cases, the dissemination may not take place unless appropriate steps are taken to safeguard these legitimate interests. In the Consortium Agreement, specific reference to *any planned publication* is made, where 30 days notice prior to publication must be made to the other parties, and objections must be made within 21 days after receipt of the notice.

3.4 Data Protection Impact Assessment

To support the further processing of AI-Mind data into EBRAINS, a data protection impact assessment (DPIA) has been carried out by AI-Minds' host institution OUS. For more information about DPIAs refer to deliverable 1.3. The processing of data into EBRAINS can be considered as compatible with the original purposes (of scientific research) of AI-Mind (the GDPR Article 5 (1) (b)). The precondition for such further processing is that the guarantees arising from Article 89 are complied with. Article 6 (1) (e) and Article 9 (2) (j) will form the bases for this data processing, as mentioned in 3.3.3 Consent. It is likely that each of the informed consent documents prepared by each of the clinics (in a participant's native language) will require updates to reflect this information.

4 AI-Mind Data Management

The general purpose of data management is to generate, obtain, confirm, transform, retain, retrieve, disseminate, and dispose data to designated stakeholders. Data management plans, execute and control the provision of data to designated stakeholders in a manner that is unambiguous, complete, verifiable, consistent, modifiable, traceable, and presentable.
Preparations for data management in AI-Mind include:

- Defining the data strategy (section 3.1 - Data strategy).
- Defining the data that shall be managed. This document mainly applies to the source data received from the clinical participants. (Section 5.1 - Data).
- Designating authorities and responsibilities for data management, including legislation, privacy, and security (sections 3.3 - Data Transfer Agreement, 4.3 - Roles, and 4.6 - Cyber-security).
- Defining the content, formats, and structure of data. Sections 5.1 - Data and 5.2 - Processes describe this on a conceptual/overview level, the implementation shall be prepared in WP2.

- Defining the data processes, including quality controls (sections 4.1 - Data quality 4.2 - Verification of data integrity and its metrics and 5.2 - Processes).

4.1 Data quality

Requirements are or will be developed with respect to:

- the participant clinical data that will be transferred.
- the transfer itself, e.g., at the right time, at the right volume, in a secured manner.

The quality of the source data shall be controlled to determine whether the data meets the specified requirements. The ISO 8000-8 concepts shall be applied when measuring data quality [8]:

- Syntactic quality - to what degree do the data conform to stored meta-data (e.g., data model).
- Semantic quality - to what degree do the data correspond to represented external phenomena.
- Pragmatic quality - to what degree is the data suitable and worthwhile for a given use.

The **syntactic quality** measurements shall be based on the data model that is developed in WP2. The data model will be based on the conceptual model and responses from the clinical partners to requests about the metadata. In addition, a description of file structures that deal with data to be stored in the database shall be used for measuring syntactic quality.

Semantic quality will not be measured, as it is neither appropriate nor possible for the project to compare the transferred data with the individual participants' actual characteristics.

Pragmatic quality will not be measured directly. One of the core tasks of the project is to find out if it is the right (usable) data that is transferred and used for the development of the AI-Mind tools. There will be an internal measurement of how suitable the AI models are for their purpose (e.g., diagnosis and prediction), which in essence is an indirect measure of the pragmatic qualities of the data being collected.

In addition, the project must assess whether the data transmissions have the agreed frequency, size, etc. The requirements should be reassessed at regular intervals during the project's lifetime.

4.2 Verification of data integrity and its metrics

Quantification of the outcomes following adherence to the requirements for ensuring data quality can be collected and measured. Example metrics to collect could be accuracy, completeness, consistency and integrity, for example, through measurement of a number of non-complying instances/total number of instances.

The participant data received from the clinics will, for all practical purposes, be verified in accordance with syntactic quality rules:

- **Entity integrity**: every entity shall have a unique identifier.

  e.g., Two different participants having the exact same identifier (Site ID + Participant ID) is a violation of Entity integrity.[4]

- **Referential integrity**: every entity that is referenced shall exist as an entity in its own right.

  e.g., When the data for a specific Participant (identified by Site ID + participant ID) is referring to a specific Clinic (identified by Site ID) this Clinic also has to exist in the data (in the domain of discourse), if not it is a violation of Referential integrity.[4]

- **Domain integrity**: all attribute values shall be within the specified domain.

  e.g., If the participant's birthdate is mandatory and required to be less than today's date – 20 years, then lack of a date and the date 2100-04-01 are both violations of Domain integrity. Other examples are that *number_electrodes* must be greater than 19 or that *duration_eyes_closed* must be greater than 4 minutes.

- **User-defined integrity**: all user-defined constraints shall be complied with.

  e.g., A User defined integrity rule could be "for each Visit ID in the Session ID, the time interval between two different data sets shall be at least 6 months". Another example could be "for *Number_EEG_Channels* in the metadata, the value must be 128".

The Entity and Reference integrity rules are general rules that the data set must comply with for the data to be coherent. Acceptance of data sets that do not comply 100% with the Entity and Reference integrity rules should be considered with respect to how it will affect the use of the data.

All syntactic rules shall be reflected in and derived from the data model. Each individual quality rule is therefore realised on the basis of this. The metric may be developed by registering for each individual quality rule the number of checks that fail in relation to the total number of checks for the relevant rule.

Specific syntactic rules i.e., data quality parameters will be specified in D2.1, D2.3, D2.5 and D4.1 and implemented in D2.2 and D2.4.

4.3 Roles

AI-Mind is a project set within a limited timeframe; thus, roles mainly related to a permanent enterprise may not be included. In addition, roles that in larger companies would have been described separately may be merged. One person can take on several roles. Data governance roles required within an AI-Mind partner's organisation are their own responsibility. The roles listed below are for the AI-Mind project, and include:

- **Data owner (each clinical partner)**

[4]The participant numbering system will be based on the EAN-8 system, enabling generation of barcodes associated with specific participant numbers. The whole 8-digit number is named the Session ID, which is a unique combination of three sub-IDs: the Site ID (2 digits), the Participant ID (3 digits), and the Visit ID (2 digits). The final digit is a check digit, used to verify the structure of the Session ID stem. See section 4.2 in D2.1 for more details.

  - Responsible for the source data (independent data controllers). These are also responsible for the data until uploaded to the TSD Platform.
  - For other data (e.g., computational models) and tools produced within the project, their ownership is defined according to the AI-Mind Consortium Agreement.
  - The data owners are also responsible for deciding how data shall be handled post project.
  - Considered independent data controllers, in accordance with the CA under the GDPR, for their own processing of personal data. See Section 2.3.2 in D1.3 for more details.
- **Data steward (representative from each clinical partner)**
  - Responsible for managing and maintaining the retrospective and prospective source data in the TSD Platform. Liaises with the platform administrator.
  - Included is both the responsibility to ensure that the data is available whenever needed, and the maintenance of the data management process (including monitoring of data quality metrics) on behalf of the project.
  - Responsible for harmonising the data format, especially for retrospective data.
- **Systems architect (at least one representative for the overall project)**
  - Responsible for designing, managing and maintaining the software and data architecture descriptions developed in the project.
- **Platform administrator (at least one representative for the overall project)**
  - Responsible for the technical and administrative IT procedures within TSD Platform.
  - Liaises with TSD staff who are responsible for the hardware and network layer for the TSD Platform.
  - Supervise data import, data processing and authorising data access and/or export to EBRAINS.

4.4 Data management architecture

Architecture descriptions aim to assist the understanding of the system's essence and key properties pertaining to its composition, behaviour, and evolution.

Architecture descriptions are used by the parties that create, utilise, and manage the system to improve communication and co-operation, enabling them to work in an integrated, coherent way.

The AI-Mind project will develop and manage its own architecture descriptions – both for data and processes. A common project vocabulary shall also be included. The total set of descriptions shall be consistent, e.g., the conceptual data model found in this document shall be consistent with the logical data models and implemented system described elsewhere in the project. The consistency must be handled throughout the life of the project.

Developing an agreed architecture description is an essential activity in the initial phases of this project. Management and maintenance of the architecture shall be executed by the Systems architect role.

The high-level architecture descriptions relevant here are found in sections 5.1 - Data and 5.2 - Processes. This architecture descriptions shall comply with the more detailed views prepared in D4.1 by Lurtis (Section 9.1, Appendix 1: Platform Architecture).

4.5 Technology and tools

In general, the technology and tools described in this sub-section are utilised to achieve the main objective of AI-Mind data governance, ensuring the integrity, quality, and security of all data and metadata during its collection, transfer, processing, storing, and export. The technology and tools described here shall be able to track the flow of the data and its transformation along the way. For example, records of every single actor (who is logging in to TSD and accessing the data) shall be collected and retained. Any data transformation shall be recorded as well as the actor who performed it and the date and time when the transformation was performed. The technology and tools described here shall be able to assess the quality of the data and display its quality status in an informative way.

Some of the required tools, which are described here, are standalone tools, others are part of larger IT solutions. A high-level overview of the tools, covering tool-objective, how they are likely to be used, and other specific features are described in the subsequent paragraphs.

AI-Mind data will be stored in a specific purpose database engine, such as PostgreSQL. In addition, a database file system will be utilised with only its metadata stored at the database engine. The technology and tools described here shall be able to manage both methods of storing data. The mapping between the metadata stored in the database engine and the path of where the file is stored shall be kept up to date. There shall be no pathless metadata and conversely there shall be no parentless file. AI-Mind database shall be backed up regularly. A procedure to restore the database with minimal AI-Mind process disruption whenever needed shall be established.

If a type of data is expected to undergo various transformations (either manual or automatic process), its transformation status shall be *true* or *successful* (not failed) at all times. If there is any failure during the transformation process, the actor involved in the process shall be noted and the date and time when it happened shall be recorded. Roll back procedure shall be planned to accommodate such failure and any other failures, including the ones which are not foreseeable. The summary of the status of various data and processes shall be made available for corresponding actors.

Most of the time, any process a user performs on AI-Mind platform involves retrieving data from the database and/or file from the database file system, processing it, and then storing the results within a complimentary database and/or database file system. To guarantee data integrity, a user (depending on their role and right) shall utilise a special purpose API to perform CRUD (create, read, update, and delete) operations on the data. Manual data access (i.e., any CRUD operation using e.g., Windows Explorer or any command prompt) shall not be allowed. The CRUD tool is required to automatically record the user ID and corresponding date and time when any operation to the database is done. The tool performs some data quality checks prior to admission of the data to the database and/or file system. The data quality items to be checked and the corresponding algorithms are subjected to the agreement with the actors who are responsible for the current process and the subsequent processes. This data quality check process is illustrated in Figure 3.

According to the best practice of data storing, here is the list of AI-Mind databases and their corresponding data:

1. Data lake (DL) - Provides storage for the raw data sets (in their original format) uploaded by each partner.

2. Data warehouse (DW) – the main storage and work area for the data pipeline (outlined in D4.1). Provides storage for the raw, standardised, and featured data sets, as well as the trained and deployed models.
3. Code Repository - Provides storage for software code repositories and ML/DL model's serialised files.
4. Platform Management System Repository - Provides storage for all the data necessary to manage the internal AI-Mind platform, i.e., platform administration files (e.g., configuration parameters, log files, diagnostic files etc.)

According to their purpose, we may categorise the necessary tools for AI-Mind data governance and management as follows:

1. Read only dashboard presenting the status and statistics of AI-Mind data quality and data management at various stages along the data value chain.
2. APIs to perform and record CRUD operations at various stages along the data value chain.
3. APIs to check data quality at various stages along the data value chain.
4. Feature to check whether the user is authorised to perform any CRUD operation at a specific stage.
5. Optional notification generator for the actors responsible in the next stage.
6. A tool to manage errors and troubleshoot reports when operating the above tools.

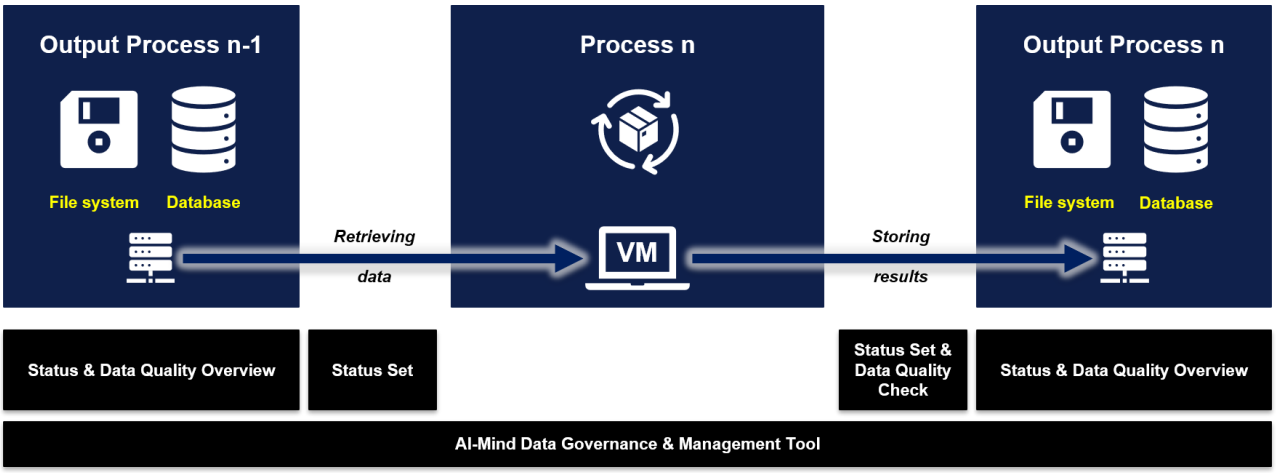


Figure 3. General principle of AI-Mind data quality check during data transformation of Process n. Note: n {1, 2, …, N}, where N is the total number of processes in the AI-Mind data value chain.

These aspects are mapped into the requirements in D4.1-*Software requirements & specifications (SRS)*.

4.6 Cyber-security

Security and protection measures for the sensitive clinical data generated within AI-Mind can be considered at three data staging areas.

1. At the point of creation at clinical partner locations – within their secure servers. This step is the responsibility of the clinical partners and will follow participant data protection procedure operations.
2. During transfer and upload to the central secure database (TSD Platform) – Data will be safely uploaded, and the transfer will be quality checked for its' integrity.
3. Within the TSD Platform. TSD offers a secure environment for storing and processing sensitive data in compliance with the Norwegian Personal Data Act (thus indirectly the GDPR) and Health Research Act. AI-Mind data will operate within this environment, and a summary of the important cybersecurity details is provided below.

Specific cybersecurity requirements for AI-Mind partners using the TSD Platform include:

- Access and log in to TSD requires two-factor authentication (2FA). For Windows machines these utilise PCoIP and Kerberos for password verification, and for Linux machines these utilise Thinlinc and nginx proxy for password verification. The FreeRadius server is used for the second factor.
- For each project in which a participant is involved in, a separate username and password is required on TSD.
- To prevent data duplication and transfer outside of TSD, functionalities that enable uncontrolled transfer of data (such as cut-and-paste functionalities, the mapping of local drives, USB forwarding) are disabled. Only project admins can transfer data out (or give permission to other users to do likewise). Data transfer can be achieved using two different methods:
    - Filelock: Access is obtained through a STFP network to a physical machine outside the TSD network following 2FA, which is synchronised to a virtual machine within the TSD network. Thus, files put on physical machine are automatically copied to virtual machine. This allows for strong control of user import/export privileges and logging of files name and eventually checksum data.
    - HTTPS API: Authentication using 2FA, BankID (Norwegian system for secure electronic ID verification) or basic authentication (only allows for importing data and can only be run on specific machines). The API is integrated with TSD's internal Identity Provider and authentication and authorisation system.
- Two physical gateway machines (FreeBSD servers) will manage all incoming traffic; directing it for inspection through firewall servers. The provisioning system (master database) defines the rules for the jumphosts, which in turn automatically generate rules to ensure projects remain separate. These jumphosts allow communications between internal TSD and trusted external computers.
- The backup system is based on commvault and is encrypted. The encryption key is only available on dedicated terminal servers and stored in safes.
- Monitoring is based on the USIT log system (Nivlheim) and Zabbix
- Standard UiO virus and malware check are run on all Windows Servers

More details about the gateways, provisioning system, network, storage, databases, directories and file structure and servers can be found in Section 9.2 Appendix 2: TSD security details.

5 AI-Mind Data, Processes and Actors

5.1 Data

This project shall manage both data and metadata.

Participant clinical data contain results from various tests (EEG, MEG, CANTAB, blood sampling, textual data, technical and clinical logs) performed at the clinical sites, and which are transferred to the AI-Mind project together with corresponding metadata. Also included are co-registration data, quality characteristics related to the clinical data, and data that apply to the actual transfer (for a full overview of prospective data see Table 1 of D2.1).

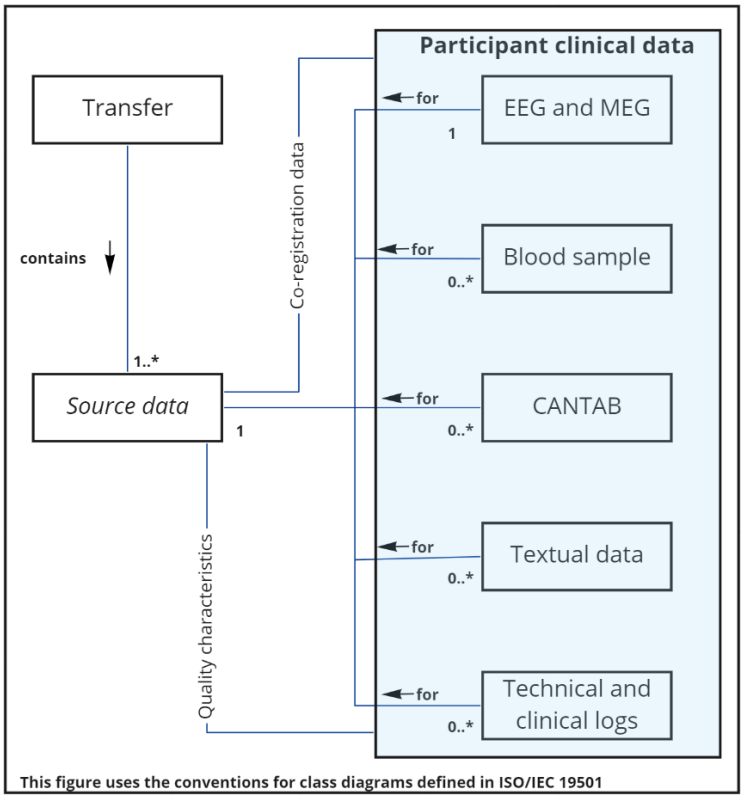


Figure 4. Conceptual model intended to illustrate the source data composition: transfer and participant clinical data, as well as the data flow (arrows) and relationships between them. * Indicates unlimited quantity.

The source data (see Figure 4) contains:

- **Transfer** - data concerning the actual transfer (e.g., size, date, sender etc).
- **Participant clinical data** – the data derived from assessments and measurements with the participant. This participant clinical data shall be pseudonymised, and where possible, anonymised.

This **Participant clinical data** consists of:

- **EEG** - data set containing the results of EEG examinations conducted by the clinics using 128 EEG electrodes. EEG files are exported in their native format as .cnt (64-bit) to ensure no metadata is

lost, and in Brainvision.eeg format to comply with the BIDS standard. Data and metadata will be stored in native format and converted to a BIDS-compliant format in the TSD server.

- **MEG** - data set containing the results of MEG examinations conducted by HUS and UCM. MEG files are exported in their native format (.fif) to comply with the BIDS standard. Data and metadata will be stored in native format and converted to a BIDS-compliant format in the TSD server.
- **CANTAB** - data sets containing the results of various cognitive examinations conducted by the clinics, such as CANTAB (Cambridge Neuropsychological Test Automated Battery), MMSE, MoCA, etc.
- **Blood sampling** - data set containing the results of genetic analysis of APOE (Apolipoprotein E genotypes) and plasma biomarker (e.g., P-tau181, or other phosphorylated forms of tau protein) based on samples collected at each clinical site at a participants first visit and analysed at OUS.
- **Textual data and technical and clinical logs** - Questionnaire on the participant's demographics including age, sex, education, medication, possible comorbidities such as cardiovascular disease, diabetes, tobacco, alcohol consumption, depression, dyslipidaemia, diet.

This **Participant clinical data will also have corresponding metadata.**

Both retrospective and prospective participant datasets are utilised in the project:

- **Retrospective** data - originating from multiple data controllers; data which can be shared beyond the scope of the original project they were collected for, based on the subjects' informed consent. In AI-Mind, retrospective data will be shared by four of the five clinical partners (excluding HUS).
- **Prospective** data – collected from participants enrolling into the AI-Mind study. It will consist of data from 1000 MCI participants who will undergo multiple (four appointments over two years) MEG, EEG and CANTAB testing, and genetic (APOE4, P-tau181, or other phosphorylated forms of tau protein) tests once during the project.

More information about the retrospective and prospective participant data can be found in table 1 of D2.1 and table 3 of D5.1.

DNV is designated as a Notified Body under the EU MDR and is thus required to maintain impartiality. Below are additional data types not shown in Figure 4, and out of scope for D1.4 Part 1. However, these data types, and the raw data, are discussed on a high level in Part 2 of this deliverable, which covers managing data for AI/ML development projects:

- Processed data
- Details of the problem framing
- Featured data
- Trained model
- Model evaluation result

Other types of data collected or used (e.g., project management data, development documentation, communications, dissemination info) in the project are excluded from this deliverable. They will be governed by the Data Management Plan (D8.3).

5.2 Processes

An overview of the workflow for the AI-Mind project is given in Figure 5, which at a general level suggests what needs to be done, not how. These details are explained in the views developed in D4.1 by Lurtis, see Section 9.1, Appendix 1: Platform Architecture.

The roles described in these processes may be generalisations of more specified actors (certain stakeholders listed in 1.6 - AI-Mind stakeholders) in the processes.

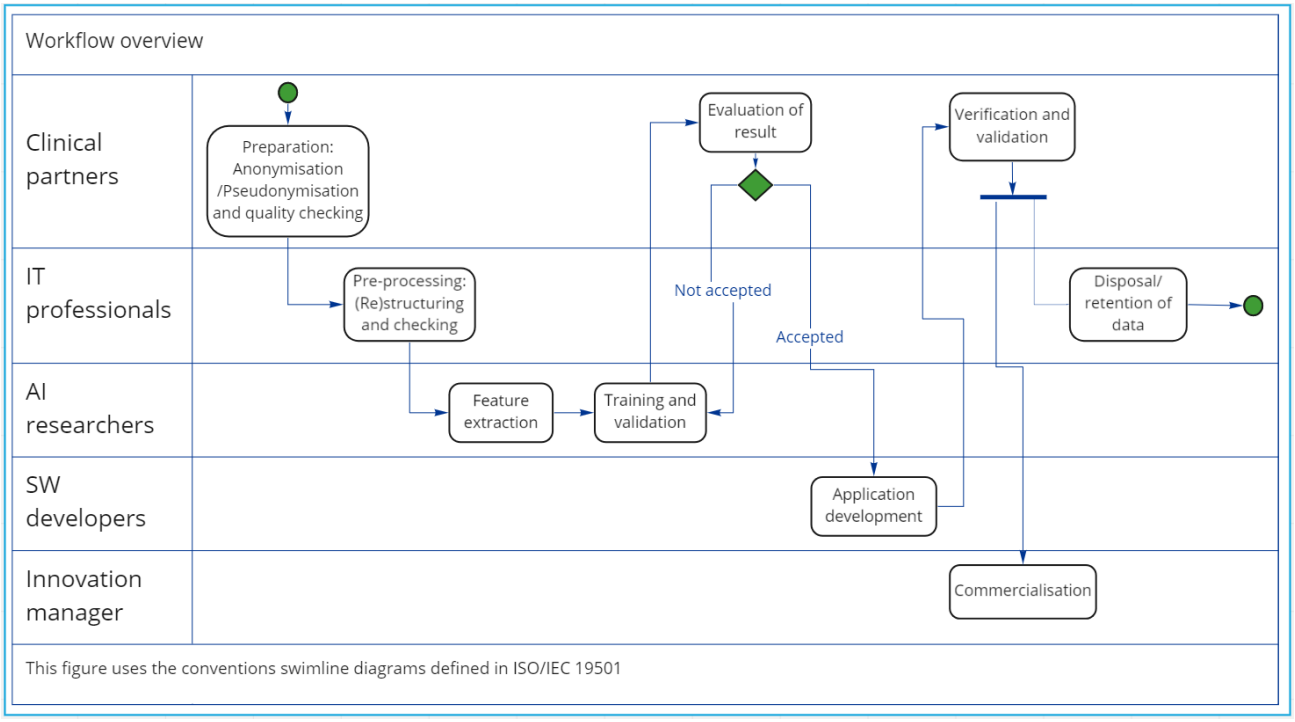


Figure 5. Overview of the general process - A workflow of stepwise activities and actions can be represented graphically by an activity diagram. The actions may be organised into swim lanes, which are used to organise responsibility for the actions. Arrows run from the start towards the end and represent the order in which activities happen. Ellipses represent the actions. The diamond represents decisions. Bars represent the start (split) or end (join) of concurrent activities. The green circle represents the start and end of the workflow.

For the processes within the scope of D1.4 Part 1, the processes and roles are described in some detail. Processes and roles outside the scope are only listed.

Preparation

The Clinical Partner shall:

- Assemble the complete set of all participant data for each participant visit as outlined in Section 4.1 of D2.1.
    - Prospective data - EEG, MEG, CANTAB, participant demographics data.
        - Blood tests (only collected once - on a participants' first visit) will be sent to OUS AI-Mind biobank for centralised genetic analysis.
    - Retrospective data – as agreed for each clinical partner.
- Pseudonymise or anonymise the data as outlined in Section 4.2 of D2.1

- Ensure the data set complies with the agreed quality and minimum compliance requirements as outlined in Section 4.3 of D2.1:
    - Collate transfer data: truly raw data and the pre-processed data
- Upload the data using the TSD data portal data transfer tool, using a site-unique URL to their designated TSD target directory (as outlined in Section 4.4 of D2.1).

Further details are in:

- Section 2.1 Data Collection Process in D4.1-*Software requirements & specifications (SRS).*
- Section 4 Data Collection Procedure in D2.1-*Standardisation of available and prospective data collection*.

Potential risks related to the data collection and preparation procedures are outlined in Section 7 of D2.1.

Pre-processing within TSD

The IT Professional shall be responsible for the following tasks performed (as automatically as possible):

- Receive and log the transfer from the clinical partner.
- Decrypt the data sets.
- Assemble the complete set per participant, per time point.
- After verifying that the complete data set complies with the agreed requirements:
    - Prepare the dataset to be used in the development of the AI-Mind tools. This includes posting the data to the correct storage locations, such as database tables and file structures.

Further details are in:

- Section 2.2 Extraction, Transformation & Loading Process (ETL) in D4.1-*Software requirements & specifications (SRS)*
- Section 4 Data Collection Procedure in D2.1-*Standardisation of available and prospective data collection*
- D2.3-*Standardisation of prospective data and pre-processing procedures* (due for delivery M16).

Due to DNV's designation as a Notified Body under the EU medical device regulation (MDR), and thus DNV's need to maintain impartiality, there are certain restrictions on what can be covered in D1.4.

Processes outside D1.4 Part 1 scope:

- **Feature extraction**. A stage to select and/or combine features from the raw data for reducing the amount of the data to be trained and at the same time meaningfully representing the information in the data. *See D1.4 Part 2*
- **Training & validation**. A process to train ML algorithm with training dataset in order to produce ML model. The trained model is then to be validated and tested using validation and

test datasets. Here a final ML model is selected based on its performance against the test dataset. *See D1.4 Part 2*

- **Evaluation of result.** The selected ML model is tested (if possible) in a real environment using the actual dataset. *See D1.4 Part 2*
- **Application development**. The selected ML model is deployed into the real infrastructure. This may involve the development of data pipeline and inference engine.
- **Verification & Validation**. The end solution is to be verified and validated, which is usually in the form of User Acceptance Test.
- **Disposal/Retention of data**.
- **Commercialisation**. The solution is ready to be operated and commercialised.

5.3 Actors involved in the processes

For a general description of these roles, please see sections 1.6 AI-Mind stakeholders and 4.3 Roles.
Actors involved in the processes within the scope of D1.4 Part 1:

- **Clinical Partners** – are the owners, independent data controllers, and transferers of the source data. They deliver data to the project at an agreed quality. As owners, the clinical partners are crucial in deciding what to do with the data after it has been used in the project or after the project has ended.
- **IT professionals** – shall manage and maintain the systems where the data is stored and support the Data Stewards in ensuring that the data is available whenever needed.

Actors involved in the processes outside the scope of D1.4 Part 1:

- **AI researchers**
- **SW developers**
- **Innovation Manager**

6 Conclusion

Effective identification of strategies within AI-Mind have been carried out from the onset of the project to ensure that:

- Regulations, standards and best practices are incorporated.
- Minimal requirements for data quality are met.
- Responsibilities for roles are assigned.
- A data management architecture is defined.
- Technology and tools are fit-for-purpose.
- Robust cyber-security procedures are in place.

Efficient data management and governance reduce potential risks that can arise during the data processing lifecycle. Implementation and continuous monitoring of this framework will be carried out

through other deliverables within the AI-Mind project and may be used to guide data governance for data that is outside the scope of this document.

7 Acknowledgements

This work is a collaboration with valuable input from partners in the project, especially Lurtis.

8 References for D1.4 Part 1

1. The Open Group. TOGAF® Standard [Internet]. 9.2. 2018. 537 p. Available from: https://www.opengroup.org/togaf
2. The International Organization for Standardization. ISO 25237:2017 Health informatics — Pseudonymization [Internet]. 1st ed. 2017. 62 p. Available from: https://www.iso.org/standard/63553.html
3. The International Organization for Standardization; the International Electrotechnical Commission; IEEE. ISO/IEC/IEEE 42010:2011 Systems and software engineering — Architecture description [Internet]. 1st ed. 37 p. Available from: https://www.iso.org/standard/50508.html
4. The International Organization for Standardization; the International Electrotechnical Commission; IEEE. ISO/IEC/IEEE 15939:2017 Systems and software engineering — Measurement process [Internet]. 1st ed. 2017. 39 p. Available from: https://www.iso.org/standard/71197.html
5. DAMA. Data Management Body of Knowledge (2nd Edition). Technics Publications; 2017. 590 p.
6. DNV. Data quality assessment framework DNVGL-RP-0497 [Internet]. 2017. Available from: https://rules.dnv.com/docs/pdf/DNV/rp/2017-01/dnvgl-rp-0497.pdf
7. The European Parliament and the Council of the European Union. General Data Protection Regulation - article 4 [Internet]. via Intersoft Consulting. Available from: https://gdpr-info.eu/art-4-gdpr/
8. The International Organization for Standardization. ISO 8000-8:2015 Data quality — Part 8: Information and data quality: Concepts and measuring [Internet]. 1st ed. 2015. 15 p. Available from: https://www.iso.org/standard/60805.html
9. The International Organization for Standardization; the International Electrotechnical Commission. ISO/IEC 2382:2015 Information technology — Vocabulary [Internet]. 1st ed. 2015. 4 p. Available from: https://www.iso.org/standard/63598.html
10. The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da [Internet]. European Union and EEA; 2016. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj
11. The International Organization for Standardization; the International Electrotechnical Commission; IEEE. ISO/IEC/IEEE 29148:2011 Systems and software engineering — Life cycle

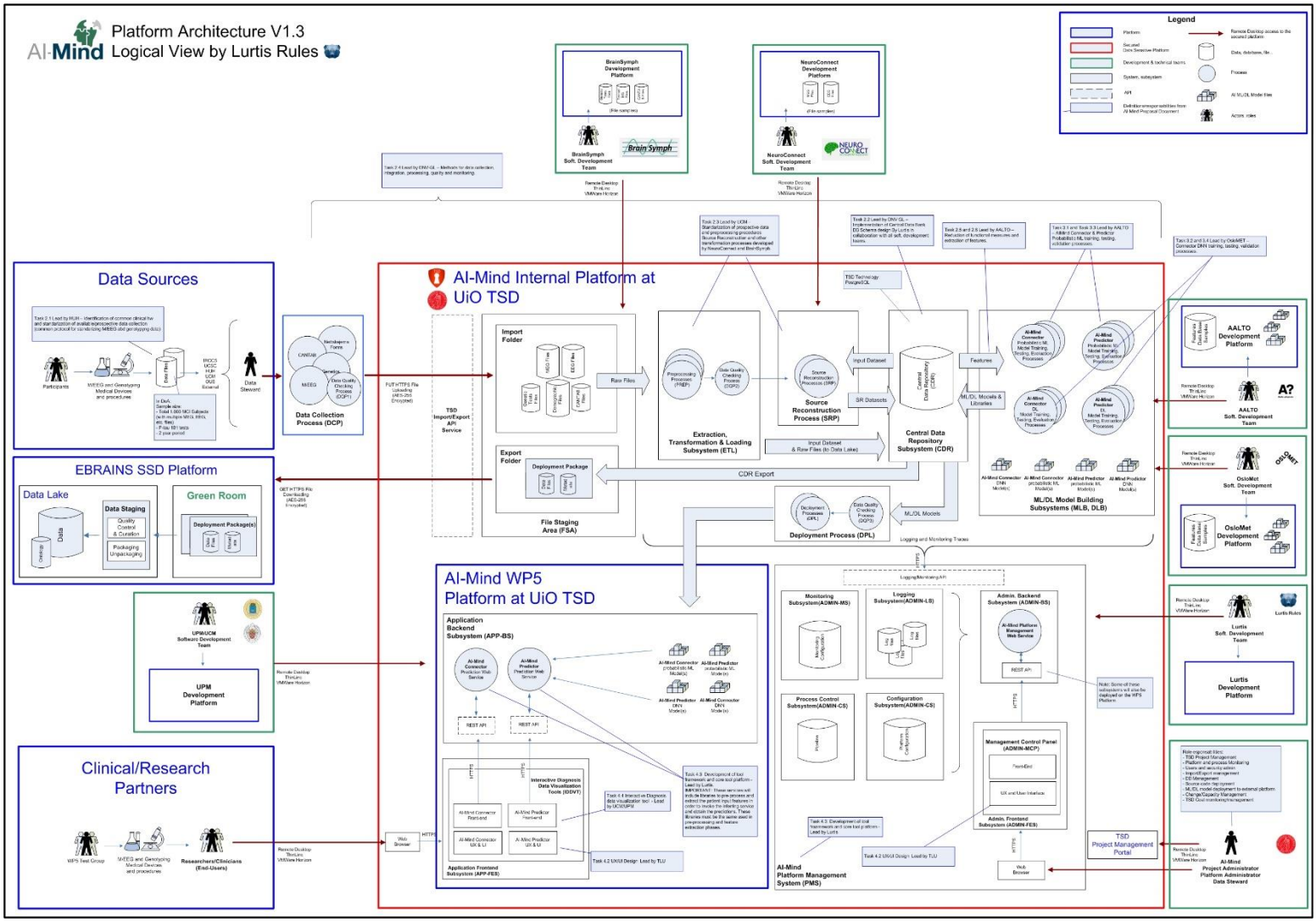
processes — Requirements engineering [Internet]. 1st ed. 83 p. Available from: https://www.iso.org/standard/45171.html

12. Project Management Institute. A Guide to the Project Management Body of Knowledge (PMBOK® Guide) [Internet]. 5th ed. Project Management Institute, Inc.; 2013. 616 p. Available from: https://repository.dinus.ac.id/docs/ajar/PMBOKGuide_5th_Ed.pdf
13. The International Organization for Standardization; the International Electrotechnical Commission; IEEE. ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes [Internet]. 1st ed. 2015. 108 p. Available from: https://www.iso.org/standard/63711.html
14. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018.
15. Discover EBRAINS [Internet]. Available from: https://ebrains.eu/discover
16. Swiss Federal Institute of Technology in Lausanne. Human Brain Project [Internet]. Available from: https://www.humanbrainproject.eu/en/

9 Appendix

9.1 Appendix 1: Platform Architecture

Logical View of platform architecture – Figure 3 from D4.1



9.2 Appendix 2: TSD security details

The following are links to documentation regarding the use of TSD and its cybersecurity:

- https://www.uio.no/english/services/it/research/sensitive-data/about/description-of-the-system.html
- https://www.uio.no/english/services/it/research/sensitive-data/help/
- https://www.uio.no/english/services/it/research/sensitive-data/help/hpc/index.html
- https://www.uio.no/english/services/it/research/sensitive-data/help/hpc/dragen.html
- https://www.uio.no/english/services/it/research/sensitive-data/help/import-export.html
- https://www.uio.no/english/services/it/research/sensitive-data/help/software/containers.html
- Whitepaper Feb 2020 (contact TSD tsd-contact@usit.uio.no for a copy)
- Risk Analysis (ROS) Feb 2020 (contact TSD tsd-contact@usit.uio.no for a copy)

----------This page intentionally left blank----------

D1.4 Part 2: Managing Data in AI/ML Development Projects in Healthcare

Organisation of D1.4 Part 2

Figure 1 presents the overview on how Part 2 is organised. An introduction to AI, AI/ML development projects and complexity on data handling in AI/ML development projects are described in Chapter 1. An overview of the data involved in an AI/ML development project is presented in Chapter 2. Chapter 3 outlines data handling considerations in AI/ML development projects with reference to the proposed EU AI-Act and DNVGL-RP-0510. The solid numbered circle in Figure 1 represents the chapter number. Each chapter in Figure 1 is drawn as a rectangle inside another rectangle to signify the relationship of those chapters as described previously.

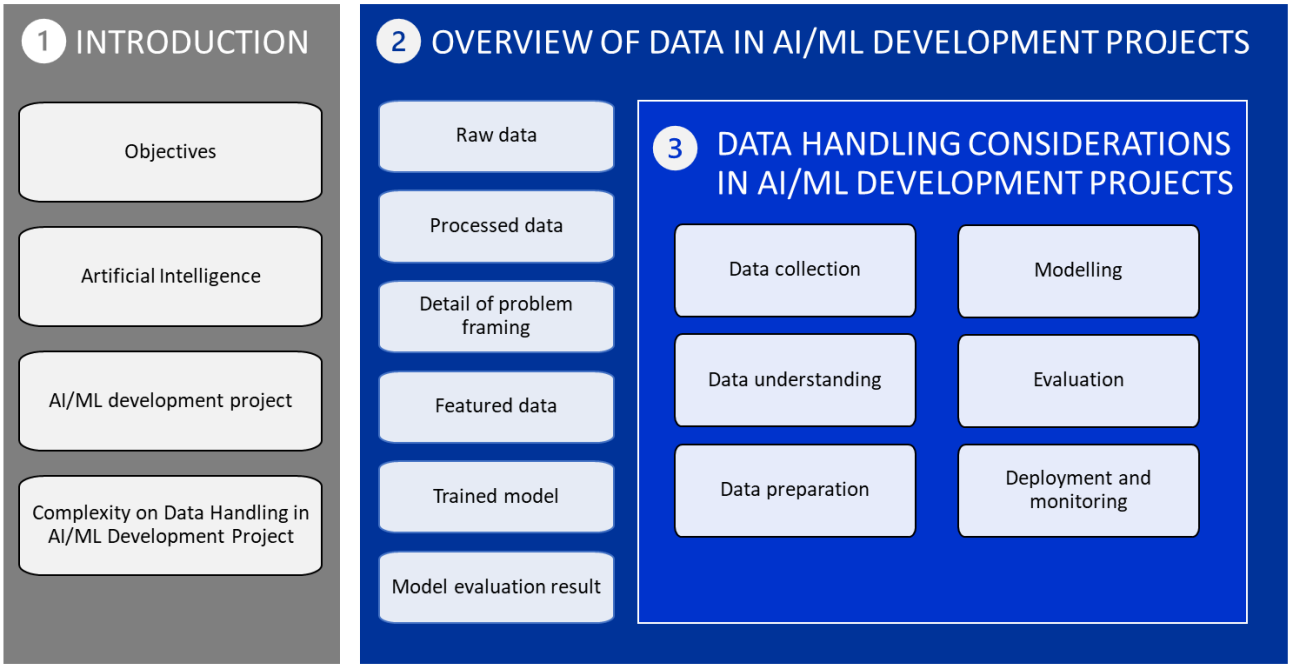


Figure 1. Organisation of part 2 of this deliverable.

1 Introduction

1.1 Objectives

The overall objectives of part 2 of this deliverable are:

1. To introduce the complexities associated with AI/ML development projects and their associated data
2. To provide an overview of data in a typical AI/ML development project,

3. To outline general approaches to data quality and management for those data, based on best practices.

1.2 Artificial Intelligence

The proposed European Artificial Intelligence (AI) Act published on the 21st April 2021 aims to implement the development of an ecosystem of trust by proposing a legal framework for trustworthy AI [1]. According to this Act, an AI system is a system that is developed with one or more of the techniques and approaches listed below and can, for a given set of human-defined objectives using learning, reasoning or modelling, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. AI techniques and approaches according to the same proposed Act (with minor modifications) include the following:

1. **Machine learning approaches**, including supervised, unsupervised and reinforcement learning, using a wide variety of methods, including deep learning.
2. **Logic- and knowledge-based approaches**, including knowledge representation, inductive (logic) programming, knowledge bases, inference, and deductive engines, (symbolic) reasoning and expert systems. The rules and knowledge bases may be formed based on statistical approaches, Bayesian estimation, search, and optimisation methods.

1.3 AI/ML Development Project

Machine learning (ML), as a subset of AI, is 100 % data driven. However, not all AI can be considered data-driven. Logic and knowledge-based approaches are examples of AI which are not always data-driven but knowledge/expert-driven, and there are other areas such as heuristic optimisation and agent-based simulation that are problem-driven. In D1.4 Part 2, only AI and ML which employs data as its main ingredient is included.

The main objective of an AI/ML development project is to deliver AI/ML solutions. The project typically includes processes such as business/problem understanding, data collection, data understanding, data preparation, modelling, evaluation, deployment, and monitoring. Managing the quality of the processes, data, transformed data, and the trained models are paramount for the success of an AI/ML development project. Reference is made to DNVGL-RP-0510 [2] which provides a framework to perform the assessment on the quality of data-driven models and algorithms. DNVGL-RP-0510 acts as the main reference in D1.4 Part 2. The recommended practice was established based on de-facto standard CRISP-DM (Cross-Industry Standard in Data Mining) [3]. It covers the assessment on the organisation and development environment, processes, and the models.

1.4 Complexity on Data Handling in AI/ML Development Project

A typical AI/ML development project involves data as its main ingredient to successfully deliver the intended AI/ML solution. The data undergoes various transformations along the process of the development project. The nature of the process of typical AI/ML development projects is more science than engineering, hence the term *data science*. See Figure 2 for a schematic concept on how a machine learning problem evolves during a development project in order to produce the final model.

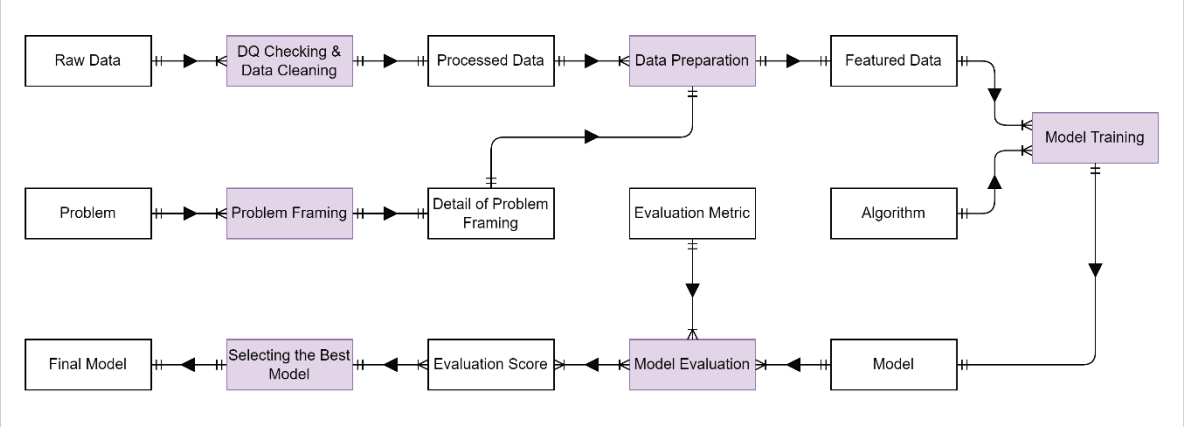


Figure 2. A machine learning problem evolves during a development project in order to produce the final model. This diagram uses Crow's foot notation [4]. The arrow suggests the workflow direction. Note: for each process (purple boxes) there may be a need to return to previous processes in the workflow. Arrows to indicate this have been omitted for the sake of clarity.

A project typically starts with proposing several options to frame the problem at hand. As indicated in Figure 2, a single machine learning problem may have one or more problem framings (one-to-many relationship). Problem framing is a systematic effort to understand, define, structure, and decompose the problem, which allows the formulation of hypotheses and/or approaches to solve the problem. For each option of problem framing (see Section 2.3), a corresponding hypothesis is proposed, which eventually will be tested by training algorithms on the data to create a model.

The raw data, as described in Section 2.1, is quality checked and cleaned for detectable artefacts.

The processed data, as described in Section 2.2, is then transformed to extract the intended features which satisfy each approach identified during the problem framing. It is possible to have more than a single set of extracted features for each option of problem framing from the same raw data. Figure 2 indicates this with the one-to-many relationship.

Once the featured data, as described in Section 2.4, is extracted, it can be used to train a specific machine learning algorithm to produce a trained model. The featured data is split into training, validation, and test datasets, see Figure 3. The model is trained using the training set and then confirmed with the validation dataset. The test dataset should not be utilised until the end of the training process when final model performance is evaluated. For each algorithm, there are parameters and hyperparameters that can be selected in order to produce the best performing model. Either the

final (best) model for each algorithm or all of the models may be kept. For the latter case, the selected hyperparameters should be recorded.

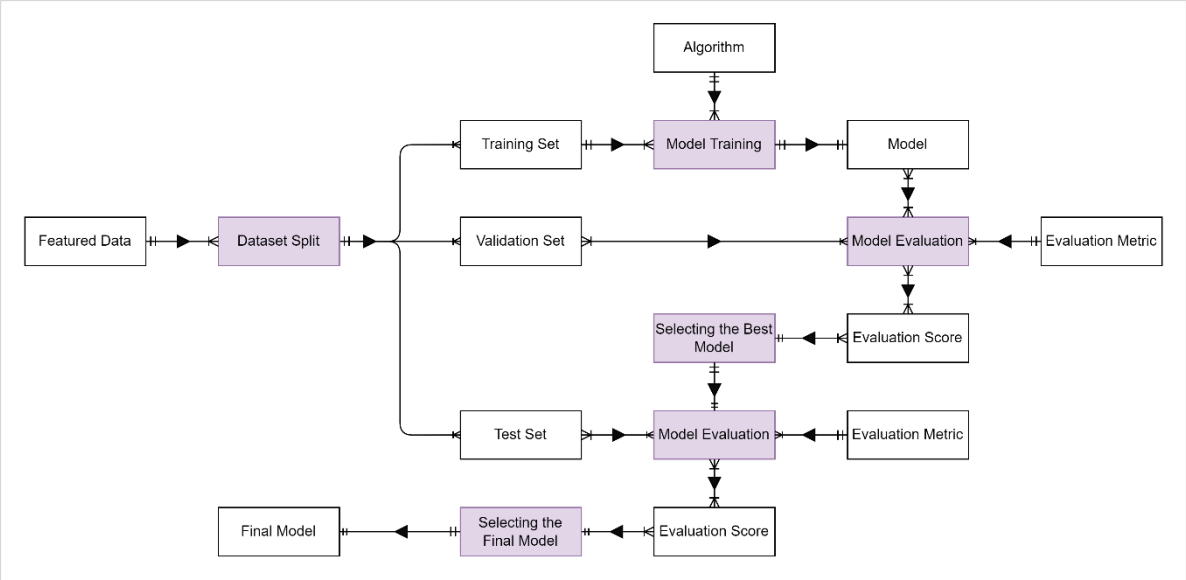


Figure 3. A set of featured data is usually split into training, validation and test sets in various different ways. The diagram shows the process in which each dataset is used. This diagram uses Crow's foot notation [4]. The arrow suggests the workflow direction. Note: for each process (purple boxes) there may be a need to return to previous processes in the workflow. Arrows to indicate this have been omitted for the sake of clarity.

A typical machine learning task, such as a regression problem or classification problem, can be solved with many different algorithms. Therefore, it is best practice to train the featured data for all possible algorithms within the domain of the machine learning task at hand. This situation is indicated in Figure 2 as one-to-many relationship between the Featured Data and the Model Training. Figure 3 indicates this situation by presenting one-to-many relationship between the Training Set and Model Training.

AI/ML development projects are commonly done in an iterative manner and never through a linear process. The 'solving' of a single problem with the same set of raw data may end up with a large number of models. Thus, the relationship between the trained model and the problem framing, as well as the data and all of its chains of transformation, should be documented. The results of previous steps may need revisiting and reanalysing multiple times in order to learn and obtain clues about how to improve the model further and to avoid repetition of previous approaches which yield ineffective models.

During this iterative process, multiple sets of featured data may be combined. The proposed problem may be split into more specific problems, which might end up with a set of models instead of a single model. Those models may be used in isolation, or combined in series or parallel fashion, depending on the circumstances, in order to form the final model.

See Figure 4 for an example scenario on how a machine learning problem evolves during a development project.

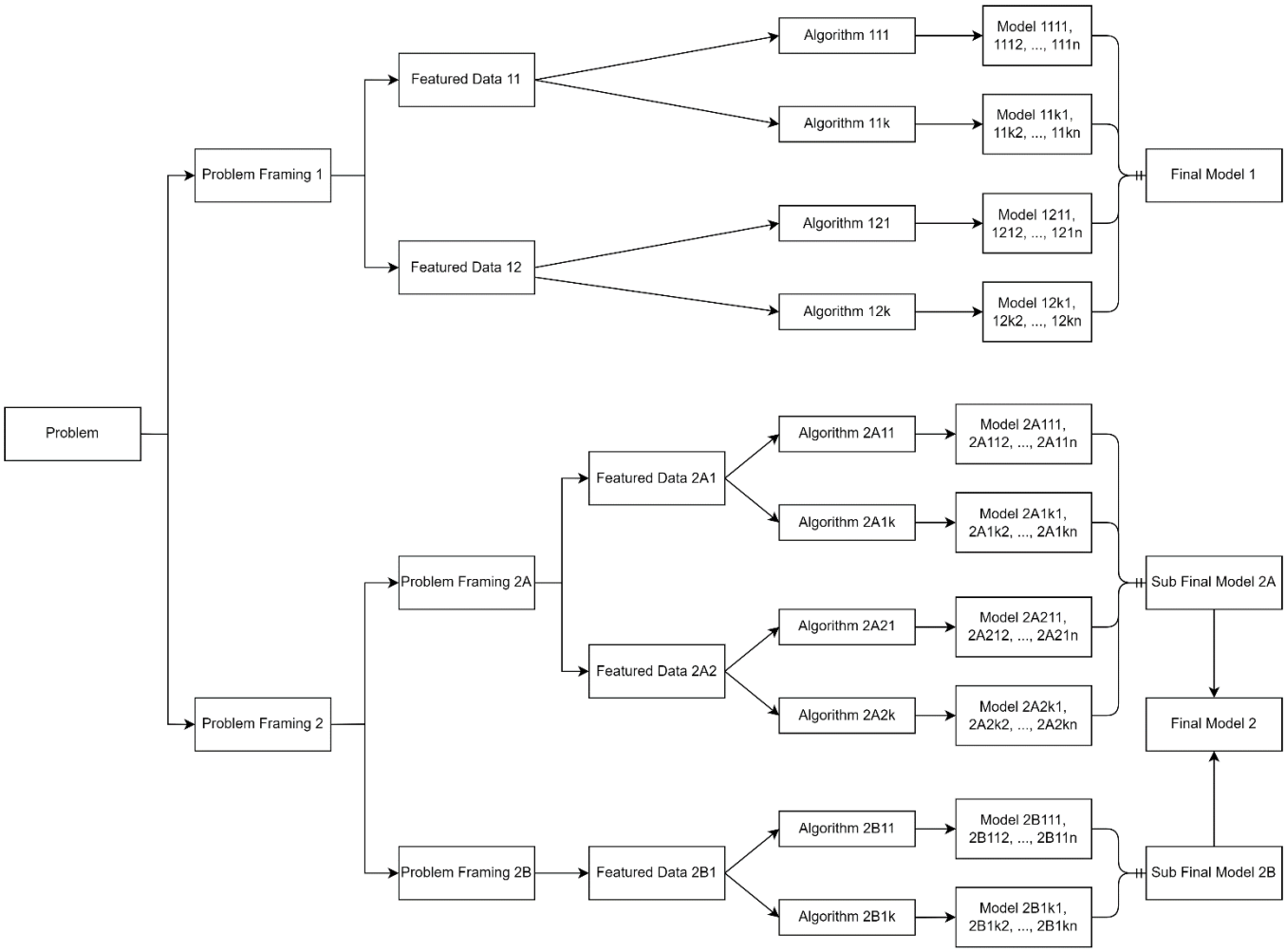


Figure 4. An example scenario on how a machine learning problem evolves during a development project. "k" and "n" signify that k and n items are available but are not shown due to limited space.

In Figure 4, for Problem Framing 1, two types of featured data (11 and 12) are utilised for model development with different algorithms (111, 11k and 121, 12k) to create the corresponding models, from which the best model is chosen as Final Model 1. For Problem Framing 2, this is split further (2A, 2B), where different types of featured data (2A1, 2A2 and 2B1) are utilised for model development with different algorithms to create the corresponding models, from which the best model is chosen as sub-final models 2A and 2B. These sub-final models are then merged into Final Model 2. During deployment, the operator may decide to run only the best performing model out of the two final models (i.e., Final Model 1 or Final Model 2). Alternatively, both final models may be run sequentially, and the operator may use the prediction which maximises the benefit or which minimises the loss/risk.

2 Overview of Data in AI/ML Development Projects

According to traditional data science terminology, the term **data** in an AI/ML project generally includes raw data, processed data, featured data, as well as training, validation, and test datasets. As such, the end result of model training, i.e., the trained model, the metadata, the model evaluation result and the detail of the problem framing are not usually included. However, D1.4 Part 2 considers the trained AI/ML model(s) from a data management and quality control perspective, thus all of these are considered data and are discussed in more detail within this section.

2.1 Raw Data

Raw data is here defined as unprocessed data that is available for the development team to produce AI/ML solution(s). The term "unprocessed" indicates that processes are not yet done by the development team, however, it is possible that the data has been pre-processed, for example pseudonymised, by the data provider prior to sending it to the development team.

2.1.1 Multiple Data, Formats and Sources

Almost all AI/ML development projects for both real-world data problems (e.g., clinical, marketing, sentiment analysis) and those with an industrial focus (e.g., sensors data from machinery in operation) involve more than a single set of data. These datasets may come in various formats and may not be available from a single data source, as such having multiple data providers is usually common.

For example, to identify the underlying cause of a 'rare-disease', diagnostic data may come from a wide spectrum of sources including observational tests, blood tests (for e.g., genetic biomarkers), X-rays, CT scans, ultrasounds and MRI. These data may come in over a long time period, as future tests will be based on results from previous tests. Data may be provided in tabulated format, binary data from a CSV, free text from a medical record, a digital image, etc. It is likely that this data will also come from different sources e.g., GP, geneticist, other patients, etc.

Unless there is a standardisation (or agreement) on data format, more data sources mean more work is required to transform the incoming data to an appropriate format for downstream use and/or processes.

For the purpose of predicting the risk of dementia, we may need data from EEG tests, demographic information, cognitive test scores, blood sample, etc. The EEG test result is basically time series data recorded with multiple electrodes/channels (for more details, see example below in 2.1.2). Demographic information consists of multiple data fields, such as age, sex, education, etc. The cognitive test scores may be obtained from traditional or computerised neuropsychological testing (e.g., CANTAB, MMSE etc), which often results with output in a CSV file format.

2.1.2 Storage: Database versus File System

Data may be stored in complex databases and/or alternatively stored more simply as singular or sets of files in a file system. Databases provide increased data consistency, ease of data access, lower

redundancy and improved back-up/recovery processes compared to file systems. However, file systems offer higher efficiency retrieval, back-up, compactivity and editing of data held within, with lower cost and design complexity. It is feasible within an AI/ML project that both storage systems are used.

Consider data in a tabulated format, which consists of a number of data records and data fields. A data record is a basic data structure, and when present in a spreadsheet, they are usually called rows. A data record or row is a collection of fields, and these fields may have different data types, typically in a fixed number and sequence. As a rule of thumb, if a single data record involves a large number of data fields, then it is better to store the record as a file.

For example, in an image recognition problem, there are m x n data fields for a single grayscale image, where m and n are the number of pixels in the X and Y directions, respectively. For a colour image, the number of data fields is 3 x m x n in order to accommodate the red, green, and blue colour ranges. A proper image recognition problem using an Artificial Neural Network (ANN) might require thousands if not millions of images. Therefore, it is usually best to store image datasets in a file system rather than to compile the images into a database.

Another example is in the field of neurophysiological and psychophysiological research and clinical applications. EEG testing can be used to detect abnormalities in the electrical activity of the brain. An EEG cap records sensor data through multiple electrodes, and the result of the test is stored in an EEG file. The size of the file depends on the number of electrodes, which can be a single channel or as many as 256 channels [5], and the duration of the EEG test. The file contains time-series data of the voltage values of the EEG, which usually has a sampling frequency of 250 to 2000 Hz. The file size is often in the order of 500 MB or greater. Therefore, it is better to store the data in a file system.

2.2 Processed Data

Upon collection, the raw data may not meet the predetermined data quality requirements for its intended use, and thus cleaning and pre-processing is often required. Syntactic data quality, commonly the most important at this stage, is the degree to which data conforms to their specified syntax, i.e., requirements stated by the metadata. Examples of syntactic data quality issues include missing values, wrong formatting, anomalous values, and replicated values (See D1.4 Part 1 Section 4.1 Data quality for more information).

In order to tackle data quality problems, we need to perform data cleaning for artefacts and data imputation. Data imputation means replacing selected entries in the dataset with alternative values.

Organising the data is usually necessary to standardise the format and structure of the data if it originates from non-standardised sources with varying format and schema.

The result is a set of processed data that is stored in an organised way to allow easy access during the latter processes.

2.3 Detail of Problem Framing

Problem framing can be achieved only after both the problem and the data are collated and sufficiently understood. The detail of problem framing should be noted and managed and it may include:

1. The decision to treat the problem either as binary classification, unidimensional regression, multi-class single-label classification, multi-class multi-label classification, multidimensional regression, or clustering,
2. The required feature,
3. The target output, and
4. The model design/structure, i.e., whether it is a single model, or multiple models working together, or a composite model which works with non-ML model/algorithm, etc.

2.4 Featured Data

In general, to train a machine learning model, the optimum solution of the following equation must be estimated:

Equation 2-1

$$f(\boldsymbol{X}) = \boldsymbol{y}$$

Where $f()$ is a machine learning function which maps $\boldsymbol{X}$ to $\boldsymbol{y}$, $\boldsymbol{X}$ is a matrix of feature (or input) variables, and $\boldsymbol{y}$ is a vector (for a single) or matrix (for multiple) of target (or output) variable(s). Each row in $\boldsymbol{X}$ is a data record. Each column of $\boldsymbol{X}$ represents a feature. A feature is a measurable property of a phenomenon under consideration.

Most of the time, one cannot simply feed-in the processed data into the $\boldsymbol{X}$ matrix. Instead, careful selection of measures from the processed data and/or additional processing may be required to produce featured data. In the case of time-series data, once the features to be extracted are decided, the timestamp must be aligned, especially if the data comes from different sources and is available in various sampling frequencies.

Implicitly, there are unlimited numbers of featured data that may be produced from a set of processed data. The featured data must be managed, their origin, methods and processes on how they were produced must be properly documented.

For example, to predict the risk of cardiovascular disease, data from tests measuring some of the following may be needed: blood pressure, molecules in the blood (e.g., cholesterol), genetic biomarkers, heart function (via electrocardiogram) and heart structure (via computed tomography scan). Some of these data may be a one-off measurement at a single point in time, numerical (260 mg/dl blood cholesterol) or binary (yes/no for the presence of troponin biomarker). Other data may be sampled over a period of time, such as the electrocardiogram or the CT scan. If the patient has historical records of, for example, a previous electrocardiogram, it might not be provided with the same sampling frequency. Therefore, timestamp alignment might be necessary. Furthermore, for other reasons, one might not want to use the instantaneous value for each timestamp the data is provided, as features. For example, rolling or running of aggregated values over a certain period of

interest as features, may provide more value. The aggregated values can be average value, standard deviation, minimum value, maximum value, sum value, etc. The end result is featured data, which is ready to be fed into the training of a machine learning model.

2.5 Trained Model

The machine learning function $f()$, i.e., the algorithm, in Equation 2-1 can be parametric or non-parametric. Parametric algorithms simplify the function by summarising the data with a set of parameters of fixed size. Some examples of parametric machine learning algorithms include linear regression, logistic regression, linear discriminant analysis, Naive Bayes, and simple neural networks. Non-parametric algorithms do not make strong assumptions about the form of the learning function. These algorithms have the ability to learn any function from the training data. Examples of non-parametric algorithms are k-nearest neighbour, decision tree, and support vector machine.

The broad range of algorithms and their respective hyperparameters and setups make it difficult to store the trained model in a relational database with a fixed schema. Therefore, the trained model is best stored as a file (for example, pickle file), and only the metadata of the trained model should be stored in a relational database.

Metadata related to a trained model may include the following (non-exhaustive):

- **Model structure and sub-models (if any).** See Section 1.4 and Figure 4 for more detail on this.
- **Pre-processing techniques**. Pre-processing techniques refer to the processes of extracting features from the data. These processes may be quite complex. Therefore, it is preferable to provide a combination of short descriptions in free text format and a pointer to a particular class or function, where the feature extraction is coded and implemented.
- **Algorithm for each model and/or sub-model**. Sub-models might be necessary if a single generic model is not sufficient to cover the whole problem space. The root cause of this is usually the lack of available data, which barely covers the whole range of the problem dimensions. Sometimes, it is necessary to segregate the problem into several sub-problems and build models separately. For example, a machine learning model to predict life expectancy may consist of several sub-models, where each sub-model represents a different health domain or risk factor that together contribute to the life expectancy prediction.
- **Hyperparameters**. Hyperparameters refer to any configuration which controls the learning process. A hyperparameter is often set by the data scientist and is independent of the training data. The setup of a hyperparameter is usually necessary to help estimate model parameters. Different machine learning algorithms have different sets of hyperparameters which need to be tuned accordingly. Examples of hyperparameters for SVR (Support Vector Regression) algorithm are kernel type to go to higher dimensions and regularisation parameter to reduce the error by avoiding overfitting. An example hyperparameter for K-NN classifier algorithms is the number of nearest neighbours. Examples of hyperparameter for decision tree algorithms are the type of function to measure the quality of a split, the maximum depth of the tree, and the minimum number of samples required to be at a leaf node. Hyperparameters must be

decided prior to the training. There exist, nonetheless, techniques to perform hyperparameter tuning (grid search, Bayesian optimisation, or heuristic [evolutionary] search) to optimise the configuration of the hyperparameters and a given target dataset to obtain the best training performance.

- **Parameters**. A set of parameters of a particular machine learning algorithm determines how input data is transformed into the desired output. Parameters are estimated from the data. A trained ML model is simply described by its algorithm and its parameters (which were trained from the data).
- Decisions on which **cross-validation technique** to use when the available data is scarce/limited. Cross-validation is a statistical method used to estimate the performance of machine learning models using only the training data.
- Selected **model performance metrics** and their corresponding results.
- **Output-processing routine**. The output of a machine learning model at the end of its inference process might not be readily readable by humans. For example, the classification problem might output binary number, e.g., 0 or 1, or probability of being class A or B. A routine to convert the output of the ML model to human readable format should be defined and its reference (to a particular function or class) may be part of the trained model metadata.

Additionally, the software version and corresponding library versions utilised when building the model could also be part of the metadata.

2.6 Model Evaluation Result

Model testing or model evaluation is an integral part of ML model development. The model is evaluated on the test dataset, which should not have been exposed (used) in the model training. Data leakage, the accidental use of sharing data/information between the training and test dataset, must be avoided due to its tendency to overestimate the performance of the tested model. For example, one cannot use data to train the algorithm and then use that algorithm to make a prediction on the same data.

At minimum, the documented model evaluation result may consist of:

- The details of the evaluation metrics,
- The description of the test dataset,
- The connection to the description of the model and the problem framing.

Different machine learning problems have different ways (or metrics) of measuring the model performance, for example:

1. Regression problem - Typical metrics are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Root Mean Squared Log Error (RMSLE), R Squared or Coefficient of determination (R2), and Adjusted R Squared.
2. Classification problem – Typical metrics usually have an association with the confusion matrix, i.e., True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN),

Accuracy, Precision, Recall/Sensitivity, Specificity, F1 Score, or other aggregative metrics (e.g., Brier score).

3. Clustering problem - Typical metrics are Silhouette Coefficient and Dunn's Index.

3 Data handling considerations in AI/ML development projects

This section provides some topics to consider when handling data in an AI/ML development project. The topics were derived from the proposed EU AI-Act 2021 [1] (see Section 6 Appendix 6.1) and adapted from best practices as documented in DNVGL-RP-0510 Framework for assurance of data driven algorithms and models [2] which is based on de-facto standard CRISP-DM [3]. In the subsequent sections, the topics are arranged in the following order: data collection, data understanding, data preparation, modelling, evaluation, and deployment and monitoring.

3.1 Data Collection

Data collection is not necessarily part of the data science framework but represents the process of gathering and measuring information. This process applies especially when data was collected for a certain purpose, which then at a later stage is re-purposed for building AI/ML models. Lately, data collection is increasingly purposed to provide data specifically for building AI/ML models.

3.1.1 Data Handling Considerations

1. Prior to collecting any data, it is important to focus on the data that is actually needed to reach the objectives of building the AI/ML models.
2. Data to be collected should be decided, defined, and documented.
3. Deciding what data to collect includes decisions about:
    a. What type of information is needed?
    b. How should the data be collected?
4. Each data element should be defined in a way that leaves no ambiguity for humans or computers, both during the collection and during the processing or modelling.
5. Metadata may be outlined as a way to define the data more precisely. Here, metadata is defined as information that describes the data element. As a minimum, it should include data type. Other details to include could be unit of measure, semantic and pragmatic validity range, unique identifier, short name, long name, information about whether it is a derived value or measured value, expected or pre-set precision, information about whether it is compressed and how it is compressed (as well as related parameters, e.g., dead band, deviation range), etc.
6. Data may originate from sensors, which is usually numeric or Boolean, or manually input by a human operator. Manually input data can be numeric, Boolean, or free text.
7. The design of data collection should consider the possible necessity to control the quality of the collected data. Data quality may consist of semantic, pragmatic, and syntactic qualities. It

is optimal to handle and fix those three data qualities at the data source. At later stages, this can be more difficult.

8. Careful design of data collection may save a lot of time and resources during data preparation.

3.2 Data Understanding

Data understanding focuses on describing and exploring the data to assess its quality and its suitability for the intended modelling. At the end of this stage, it will be decided whether the data is sufficient for the modelling phase, whether and which action should be taken to obtain more or better-quality data, or whether the project goals should be adjusted.

3.2.1 Data Handling Considerations

1. Data engineering competence is required to collect and prepare the data.
2. Data science competence is required to develop an understanding of the properties and possible shortcomings of the data.
3. The following issues related to data quality could be checked:
    a. Is the data complete?
    b. Are there data syntactic quality issues?
    c. Are there data pragmatic quality issues?
    d. Are there data semantic issues? These might be difficult to assess at this stage. This concern should be handled during Data Collection (see Section 3.1).
4. If data quality issues exist, is there any chance to fix it at the source and have it re-sent?
5. The following issues related to data representativeness could be checked:
    a. Is the modelling related data size/volume sufficient?
    b. Is the data representative for the defined problem?
    c. Is the data unbiased?
6. If data was found to be inadequate for the defined problem, then:
    a. Consider obtaining more and better data.
    b. Consider adjusting the project goal.
7. A data quality report should be made readily available, demonstrating the suitability of the data with respect to requirements.

3.3 Data Preparation

Data preparation focuses on selecting the data, fixing data quality issues, constructing, integrating, re-formatting, and harmonising the dataset to make them ready for input to modelling. Feature extraction (according to problem framing, see 2.3 Detail of Problem Framing), encoding and embedding occur as part of this step.

Data harmonisation refers to efforts to combine data from different sources and produce a comparable dataset. This process usually involves data mapping from the original format to a standardised format.

Feature extraction is a process to transform the raw data into numerical features, which is suitable as input for the training of machine learning models. During the transformation, one or more of the following processes might be executed:

1. The number of features might be reduced,
2. Some features might be combined into, or represented by, a new feature,
3. A collection of more granular data might be summarised using simple statistics (mean, standard deviation, variance, skewness, kurtosis, etc.) or other means,
4. Categorical data might be encoded,
5. Data in text format is transformed into numeric,

Embedding and encoding are parts of feature extraction process, and they are mainly implemented in the Natural Language Processing (NLP) domain but has gained popularity in different machine learning application domains. NLP is concerned with programming computers to process and analyse large amounts of natural language data. Processing free text sections of a patient's medical journal, or processing voice recordings from the medical experts, to build AI/ML models are two examples of NLP applications. Encoding is a process to transform words, phrases, or sentences into a numeric system by means of vocabulary mapping. Embedding is a more efficient way of transforming text materials to a numeric system [6].

3.3.1 Data Handling Considerations

1. Data quality issues should be addressed, including those related to:
    a. Missing data records,
    b. Missing data field values, and
    c. Other data quality issues.
2. The resulting (integrated and harmonised) data set should be sufficiently consistent. Data consistency refers to reflexive, symmetric, and transitive attributes [7]. Reflexive means consistency with itself. Symmetric refers to situations where data record A is consistent with B, B is also consistent with A. Transitive refers to situations where data unit X is consistent with Y and Y is consistent with Z, and data unit X is consistent with Z.
3. The format of all relevant data fields should be systematically analysed and adapted where necessary. All data formats should be adequate for the subsequent analysis. The data format is a formalised description of how data and information of a particular type can be structured so that a user can read and process it. Data formats are often standardised and are used across the system and applications. There are 3 categories of data format:
    a. File-based data format. Data is stored in one or more files, which are usually placed in an arbitrary folder. Examples of file-based data formats include but are not limited to comma-separated values (CSV), tab-separated values (TSV), Extensible Markup Language (XML), JavaScript Object Notation (JSON), shapefile format (SHP), the compiled machine code version of an SHP ASCII-based shape entities file (SHX) etc.
    b. Directory-based (file system) data format. In this type of data format, whether there are one or more files, they are all stored in the parent folder in a particular manner. Example of the use of such data format is images, EEG files, etc.

c. Database connection. This is an organised collection of data, which is typically stored electronically in a computer system. For example, relational database (e.g., MySQL, PostgreSQL, MariaDB), NoSQL database (e.g., MongoDB, CouchDB), graph database (e.g., ArangoDB, Neoj, OrientDB), etc.

3.4 Modelling

Modelling includes iterative processes of training, assessment, and adjustment, until satisfactory results are obtained. In each iteration, adjustments can be made to the data partitioning (train/validation), feature extraction and training strategy (cross-validation, hyperparameter tuning).

3.4.1 Data Handling Considerations

1. Parameters are listed, and hyperparameters are specified, together with any relevant rationale.
2. Model parameters and hyperparameters are recorded and managed properly.
3. Partitioning of data is done without introducing bias or error (e.g., data leakage, the test dataset has different statistics to the training dataset, etc)
4. Thresholds on data quality for training/validation/testing are documented.
5. The model is appropriately documented, see Section 2.5 Trained Model.
6. The model's expected sensitivity to data quality problems is documented.
7. The data used to validate/test the model is representative.
8. The data used to validate/test the model is independent of the training data, therefore there is no data leakage.
9. The data used to validate/test the model has similar statistical properties to the training data.

3.5 Evaluation

The evaluation stage of the model assesses whether goals, success criteria and requirements are satisfied. Even if the success criteria are not satisfied, evaluating the final model with respect to the projects' objectives may provide a useful insight that can be used in future iterations and/or other projects.

3.5.1 Data Handling Considerations

1. The model's sensitivity to expected data quality problems is understood and quantified.
2. Models in high-risk systems, such as clinical investigations, are tested with extended test sets based on context-specific objectives.

3.6 Deployment and Monitoring

Deployment of the model (including the pre-processing codes to prepare the input data) to its production/use environment (application) and subsequent monitoring of the performance of the model is carried out during its operation.

To facilitate this, disciplines known as Machine Learning Model Operationalisation Management (MLOps) and Continuous Delivery for Machine Learning (CD4ML) have been developed to meet the needs of ML model deployment and monitoring.

MLOps is the establishment of processes around the designing, building and deployment of models into the production environment. It couples the model to considerations relating to the incoming data quality, model decay and model locality, to improve quality, transparency and agility. MLOps includes continuous integration (CI) using additional tests and validation of data and models, continuous delivery (CD) through an automatic deployment of the ML model within an ML training pipeline, continuous training (CT) and re-training of ML models for redeployment, and continuous monitoring (CM) of production data and model performance metrics [8].

CD4ML can be considered a subsidiary of MLOps, through which multi-disciplinary teams of data scientists, DevOps team members, data engineers and business representatives, and CI/CD enables the implementation and automation of ML models into agile, user-focussed production pipelines [9].

3.6.1 Data Handling Requirements

1. A deployment plan has been issued.
2. A monitoring and maintenance plan for the application has been issued.
3. The assumptions underlying the application’s operation are sufficiently documented.
4. The assumptions underlying the application's operation are valid.
5. Any contingency/redundancy/fallback mechanisms used in operations to handle wrong predictions or other application failure modes are sufficiently documented.
6. The validity of the assumptions underlying the application’s operation are continuously/periodically monitored.
7. The data quality of input data is continuously/periodically monitored.
8. The performance of the application is continuously/periodically monitored.

4 Conclusion

In the context of data governance, data no longer simply refers to the raw data being collected during a project or the input data being fed into an algorithm. Data encompasses the raw data, processed data, featured data, the training, validation, and test datasets as well as the end result of model training, i.e., the trained model (binary file), the metadata, the model evaluation result and the detail of the problem framing.

As a result, data handling in AI/ML projects is becoming more complex and multi-faceted, with governance and management required at multiple stages of the development process. As the propensity of these projects increase, more robust processes and documentation, such as the consideration discussed throughout D1.4 Part 2, will be required to ensure adequate and acceptable data handling.

5 References for D1.4 Part 2

[1] The commission to the European Parliament, "Proposal for a Regulation of the European Parliament and of the Council "Laying down harmonised rules on Artificial intelligence (ARTIFICIAL INTELLIGENCE ACT) and amending certail union legislative acts"," EUR-Lex, 2021.

[2] DNV GL, "DNVGL-RP-0510: Framework for assurance of data-driven algorithms and models," DNV GL, 2020.

[3] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing,* vol. 5, no. 4, 2000.

[4] G. C. Everest, "Basic Data Structure Models Explained with a Common Example," in *Proceedings Fifth Texas Conference on Computing Systems*, Austin, TX, 1976.

[5] T. Lau, J. Gwin and D. Ferris, "How Many Electrodes Are Really Needed for EEG-Based Mobile Brain Imaging?," *Journal of Behavioral and Brain Science,* vol. 2, pp. 387-393, 2012.

[6] B. Bengfort, R. Bilbro and T. Ojeda, Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning, O'Reilly Media, 2018.

[7] P. Shi, Y. Cui, K. Xu, M. Zhang and L. Ding, "Data Consistency Theory and Case Study for Scientific Big Data.," *Information,* vol. 10, no. 4, p. 15, 2019.

[8] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz and M. Plöd, "MLOps Principles," 2022. [Online]. Available: https://ml-ops.org/.

[9] A. Miller, "CD4ML Continuous Delivery Machine Learning," BMC Blogs, 2021. [Online]. Available: https://www.bmc.com/blogs/cd4ml-continuous-delivery-machine-learning/.

6 Appendix

6.1 Appendix 1: Proposed EU AI-Act

The table below has paragraphs taken from Article 10 (Data and data governance) of the proposed EU AI-Act and where they are addressed in this document (i.e., D1.4 Part 2).

Paragraphs	**Discussed in Section/s**
1. High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5.	NA
2. Training, validation and testing data sets shall be subject to appropriate data governance and management practices. Those practices shall concern in particular,	NA
(a) the relevant design choices;	3.1.1
(b) data collection;	3.1.1
(c) relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment, and aggregation;	3.2.1, 3.3.1
(d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent;	3.1.1, 3.2.1, 3.3
(e) a prior assessment of the availability, quantity and suitability of the data sets that are needed;	3.1.1, 3.2.1
(f) examination in view of possible biases;	3.2.1, 3.4.1
(g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.	3.2.1, 3.3.1, 3.4.1
3. Training, validation, and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.	3.1.1, 3.4.1, 3.5.1
4. Training, validation, and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural, or functional setting within which the high-risk AI system is intended to be used.	3.1.1
5. To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in Article 9(1) of Regulation (EU) 2016/679, Article 10 of Directive (EU) 2016/680 and Article 10(1) of Regulation (EU) 2018/1725, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.	Not mentioned.
6. Appropriate data governance and management practices shall apply for the development of high-risk AI systems other than those which make use of techniques	3.5.1, 3.6.1

involving the training of models in order to ensure that those high-risk AI systems comply with paragraph 2.	