Funded by the European Union

HORIZON-WIDERA-2023-ACCESS-02

ChatMED

Grant ID: 101159214

Bridging Research Institutions to Catalyze Generative AI Adoption by the Health Sector in the Widening Countries

"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Jožef Stefan Institute

Univerzitet u Nišu
MEDICINSKI FAKULTET

TU Graz

Deliverable D2.1

Data Management Plan

Funded by the European Union

Deliverable D2.1 - DATA MANAGEMENT PLAN	
Date:	**February 27, 2025**
Lead Beneficiary:	**FCSE**
Type:	**DMP**
Author(s):	**Kostadin Mishev, Monika Simjanoska Misheva, Katarina Trojacanec, Eftim Zdravevski, Dimitar Kitanovski, Katerina Zdravkova, Riste Stojanov**
Dissemination level:	**PU**
Version:	**7.0**

Abstract: The Data Management Plan (DMP) for the ChatMED project outlines a strategy for managing neurological data to develop Large Language Models (LLMs) for healthcare. Central to the DMP is the TriZ-Flow architecture, a three-zone data processing pipeline (Bronze, Silver, Gold Zones) ensuring data quality, standardization, and GDPR compliance. This decentralized approach minimizes privacy risks while enabling efficient data integration. The DMP emphasizes ethical data handling, including informed consent and data minimization, and adheres to FAIR principles (Findable, Accessible, Interoperable, Reusable) to ensure data is organized, secure, and reusable. It integrates multimodal data (e.g., neuroimaging, patient histories) to enhance AI-driven diagnostics in neurology. The DMP also serves as a training tool for researchers, particularly in Widening countries. It provides infrastructure for the Human Evaluation Framework (HEF), ensuring rigorous evaluation of LLMs in neurology tasks. Designed for scalability, the TriZ-Flow architecture extends beyond ChatMED, ensuring long-term impact. Robust data security measures, including Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC), protect sensitive information. An Ethics Advisor oversees compliance with ethical guidelines.

Disclaimer: The project is funded by the European Union under Horizon Europe Widening participation and spreading excellence programme, grant ID: 101159214. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

HISTORY OF CHANGES

Date	Issue	Author	Summary of main changes
4.11.2024	1.0	Monika S. Misheva	Introduction, Data summary
3.2.2025	2.0	Katerina Zdravkova	Ethics
13.2.2025	3.0	Katarina Trojacanec Dineva	Data summary

Funded by the European Union

20.2.2025	4.0	Kostadin Mishev	FAIR data, Other research outputs, Allocation of resources, Data security
22.2.2025	5.0	Dimitar Kitanovski	Three-zone Data Processing Pipeline
26.2.2025	6.0	Eftim Zdravevski	Three-zone Data Processing Pipeline, FAIR data, Other research outputs, Allocation of resources, Data security
27.2.2025	7.0	Riste Stojanov	Refinement

Table of Contents

List of Figures 4
List of Tables 4
List of Acronyms 4
Introduction 7
ChatMED Through the Prism of Data 7
Position of the DMP within the Project 8
DMP as a Coordination Measure 9
DMP as a Curriculum and Training Tool 9
TriZ-Flow: A Three-Zone Data Processing Pipeline in the Center of the DMP 9
Bronze Zone: Data Ingestion 11
Silver Zone: Data Standardization and Validation 11
Gold Zone: Data Processing and Final Storage 12
TriZ-Flow as a Foundation for the Human Evaluation Framework 12
TriZ-Flow as a Platform for Future Projects 12
Key Messages of the DMP 13
The Data Management Plan 15
Data Summary 15
FAIR Data 19
Making data findable, including provisions for metadata 19
Making data accessible 21
Repository 21

Funded by the European Union

Data 24
Metadata 27
Making data interoperable 28
Increase data re-use 29
Other research outputs 34
Allocation of resources 34
Data security 36
Ethics 36
Other issues 38
Summary 38

List of Figures

Figure 1. WP2 deliverables achieving M1 8
Figure 2. TriZ-Flow zones: Bronze Zone (data ingestion), Silver Zone (data standardization and validation), and Gold Zone (data processing and final storage). 10

List of Tables

Table 1. Alignment among DMP contributions and ChatMED's SOs. 14
Table 2. Repositories to be used in the project. 22
Table 3. Repositories and identifier types. 24

List of Acronyms

ABAC (Attribute-Based Access Control)
AI (Artificial Intelligence)
AMP-AD (Accelerating Medicines Partnership - Alzheimer's Disease)
API (Application Programming Interface)
ATLAS (Anatomical Tracings of Lesions After Stroke)
BAM (Binary Alignment Map)
BBMRI-ERIC (Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium)
CC0 (Creative Commons Zero)
CC BY (Creative Commons Attribution)

Funded by the European Union

CC BY-SA (Creative Commons Attribution-ShareAlike)
CERMEP (Centre d'Exploration et de Recherche Médicale par Émission de Positons)
CSV (Comma-Separated Values)
DAC (Data Access Committee)
DC (Data Controller)
DCAT (Data Catalog Vocabulary)
DICOM (Digital Imaging and Communications in Medicine)
DMP (Data Management Plan)
DPO (Data Protection Officer)
DPs (Data Processors)
EBRAINS (European Brain Research Infrastructure)
EDF (European Data Format)
EGA (European Genome-phenome Archive)
EMG (Electromyography)
EMA (European Medicines Agency)
ENIGMA (Enhancing Neuro Imaging Genetics through Meta Analysis)
EU (European Union)
FAIR (Findable, Accessible, Interoperable, and Reusable)
FASTA (Fast Alignment Search Tool - genomic file format)
FHIR (Fast Healthcare Interoperability Resources)
FMN (Faculty of Medicine, University of Niš)
FOAF (Friend of a Friend - W3C ontology for representing people and relationships)
GDPR (General Data Protection Regulation)
GenAI (Generative Artificial Intelligence)
Git (Version Control System)
GitHub (Online repository hosting service)
GitLab (Version control platform)
H100 (NVIDIA H100 GPU)
HPO (Human Phenotype Ontology)
HL7 (Health Level Seven International)
IB (Image Biomarkers)
IBC (International Brain Connectivity)
i2b2 (Informatics for Integrating Biology and the Bedside)
JSON (JavaScript Object Notation)
JSON-LD (JavaScript Object Notation for Linked Data)
JSI (Jožef Stefan Institute)
LLM (Large Language Model)
LOD (Linked Open Data)
MHA (Medical Health Application)
MFA (Multi-Factor Authentication)
MIMIC (Medical Information Mart for Intensive Care)
MRI (Magnetic Resonance Imaging)
MSB (Most Significant Bit)
NIfTI (Neuroimaging Informatics Technology Initiative)

Funded by
the European Union

NLP (Natural Language Processing)
OASIS (Open Access Series of Imaging Studies)
OBO (Open Biomedical Ontologies)
ODbL (Open Database License)
OMOP (Observational Medical Outcomes Partnership)
ONNX (Open Neural Network Exchange)
OASIS-3 (A Longitudinal Neuroimaging, Clinical, and Cognitive Dataset)
ORCID (Open Researcher and Contributor ID)
PET (Positron Emission Tomography)
PID (Persistent Identifier)
PPMI (Parkinson's Progression Markers Initiative)
PROV-O (Provenance Ontology)
PubMed (Biomedical Literature Database)
PU (Public Use)
RBAC (Role-Based Access Control)
RDF (Resource Description Framework)
REST (Representational State Transfer)
SAML (Security Assertion Markup Language)
SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)
SO (Specific Objective)
SSO (Single Sign-On)
STSE (Short-Term Staff Exchanges)
TF (TensorFlow)
TriZ-Flow (Three-Zone Data Processing Pipeline)
TUGraz (Technical University of Graz)
TXT (Plain Text File Format)
UK Biobank (UK Biomedical Research Database)
VCF (Variant Call Format - Genomics)
WP (Work Package)
XML (Extensible Markup Language)
Zenodo (Open-Access Research Repository)

Funded by
the European Union

Introduction

ChatMED Through the Prism of Data

The ChatMED project represents a groundbreaking initiative aimed at leveraging the power of GenAI to transform the field of neurology, with a particular focus on data-driven solutions. As the healthcare sector increasingly embraces digital transformation, the integration of advanced AI technologies, such as LLMs, offers unprecedented opportunities to enhance diagnostic accuracy, improve patient outcomes, and streamline medical workflows. ChatMED seeks to harness these opportunities by developing a multimodal AI platform tailored to the unique challenges of neurology, a field that inherently relies on the integration of diverse data types, including textual, visual, and auditory information.

At its core, ChatMED is a data-centric project. The initiative recognizes that the success of AI in healthcare, particularly in neurology, hinges on the availability of high-quality, diverse datasets. Neurological disorders, such as Alzheimer's disease, Parkinson's disease, and epilepsy, are complex and multifaceted, requiring a holistic approach to diagnosis and treatment. Traditional AI models, often trained on limited or unidimensional datasets, fall short of capturing the full spectrum of neurological data. ChatMED addresses this gap by focusing on the integration of multimodal data sources, including neuroimaging (MRI, CT scans), electrophysiological data (EEG, EMG), and patient histories, to create a more comprehensive and accurate AI-driven diagnostic tool. The project's data-driven methodology begins with the identification and curation of relevant datasets from trusted repositories. These datasets, which include detailed clinical data, patient histories, and treatment outcomes, serve as the foundation for training and fine-tuning LLMs. By leveraging these rich datasets, ChatMED aims to develop AI pipelines capable of interpreting complex neurological symptoms, suggest differential diagnoses together with tailored treatment recommendations. The integration of multimodal data ensures that the AI models can capture the nuances of neurological disorders, leading to more accurate and reliable outcomes.

The ChatMED project will generate and validate valuable data through structured training activities, STSEs, and summer schools. These initiatives serve not only to enhance capacity building and collaboration but also to produce new datasets that will refine AI models and validate their real-world applicability. All these activities are designed to either generate new data or validate existing datasets, ensuring that the ChatMED platform evolves with high-quality, real-world medical insights.

ChatMED places a strong emphasis on data management and ethical considerations. The project adheres to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, ensuring that all research data is organized, securely stored, and readily available for future use. The consortium is committed to open science practices, making datasets, models, and research outputs accessible to the broader scientific community. This commitment not only fosters collaboration but also accelerates innovation in the field of AI-driven healthcare. The project also addresses the critical issue of data privacy and security, particularly in the context of sensitive

Funded by the European Union

medical information. By implementing robust data protection measures and adhering to GDPR standards, ChatMED ensures that patient data is handled with the utmost care and confidentiality. An Ethics Advisor is embedded within the project to oversee all data-related activities, ensuring compliance with ethical guidelines and mitigating potential risks associated with AI-driven healthcare solutions.

In summary, ChatMED is a pioneering project that places data at the heart of its mission to revolutionize neurology through GenAI. By integrating multimodal datasets, generating new data through training activities, STSEs, and summer schools, and adhering to ethical data practices, the project aims to develop a cutting-edge AI platform that not only enhances diagnostic accuracy, but also paves the way for future innovations in healthcare. Through its data-driven approach, ChatMED seeks to bridge the gap between AI research and clinical practice, ultimately improving patient outcomes and advancing the field of neurology.

Position of the DMP within the Project

The DMP is a critical deliverable in the ChatMED project, playing a pivotal role in achieving the first milestone, M1: Solid coordination measures, curriculums, plans, and HEF established, alongside the other four deliverables from the work package devoted to the establishment of coordinating and supporting measures, WP2: the Plan for Dissemination, Communication, and Exploitation; the Website Establishment; the Detailed Training Curriculum and Talent Mobility Programs, and the Human Evaluation Framework as depicted in Figure 1. This milestone is foundational to the project's success, as it sets the stage for the effective management of data, resources, and research activities throughout the project's lifecycle. The DMP is not only a technical document but also a strategic tool that ensures the project adheres to best practices in data management, open science, and ethical research.

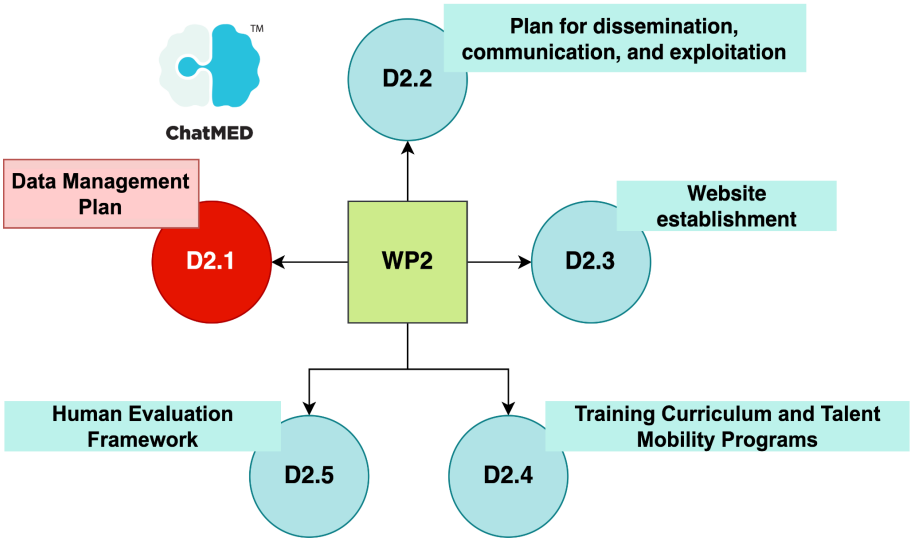

Figure 1. *WP2 deliverables achieving M1.*

Funded by
the European Union

DMP as a Coordination Measure

The DMP is one of the key coordination measures established in the first six months of the project. It serves as a blueprint for how data will be collected, stored, processed, and shared across the consortium. The DMP ensures that all partners are aligned in their approach to data management, which is crucial for maintaining consistency and transparency throughout the project. By defining clear roles and responsibilities for data handling, the DMP facilitates smooth collaboration between the Widening Coordinator (FCSE) and the partners (JSI, FMN, and TU Graz). This alignment is essential for achieving the project's specific objectives, particularly **SO4: Standardize and Disseminate Best Practices in Research and Administration**.

The DMP also supports the establishment of the **Quality Assurance and Ethics Committee**, which oversees data privacy, ethical considerations, and compliance with GDPR. By integrating the DMP into the project's governance structure, the consortium ensures that data management practices are not only efficient but also ethically sound. This is particularly important given the sensitive nature of the medical data used in the project.

DMP as a Curriculum and Training Tool

The DMP is not just a static document, it is a living framework that will be used to train researchers and staff across the consortium. As part of the training activities outlined in WP4, the DMP will serve as a curriculum for teaching best practices in data management, including data standardization, validation, and storage. This training is crucial for building the capacity of researchers in the Widening countries (North Macedonia and Serbia) and ensuring they are equipped with the skills needed to manage large-scale, multimodal datasets.

The DMP also supports the STSE programs, where researchers from the Widening countries will learn advanced data management techniques from the Advanced Partners. By incorporating the DMP into these exchange programs, the project ensures that knowledge transfer is not limited to theoretical concepts but also includes practical, hands-on experience in managing complex data workflows. This aligns with **SO3: Advance Researcher Mobility and Expertise Sharing**, as the DMP becomes a tool for fostering collaboration and skill development across the consortium.

TriZ-Flow: A Three-Zone Data Processing Pipeline in the Center of the DMP

Within the DMP, a need is recognized for building a decentralized and distributed data management platform (TriZ-Flow) as part of the Widening coordinator infrastructure upgrade described in **WP3: Upgrading a Research Management/Administration Unit**, in particular **T3.4: Establishing a knowledge management system**. The initially described robust architecture in Figure 2 ensures data quality, standardization, and transformation at every stage. The TriZ-Flow architecture is divided into three main zones: Bronze Zone, Silver Zone, and Gold Zone. Each zone plays a specific role in the data processing pipeline, ensuring that data is

Funded by the European Union

ingested, standardized, and transformed into its final, high-quality form. To ensure compliance with GDPR and safeguard patient privacy, each zone within the TriZ-Flow architecture - Bronze, Silver, and Gold - can be decentralized and distributed. This decentralization allows for data processing to occur closer to its source, reducing the need for central storage of sensitive data and minimizing privacy risks.

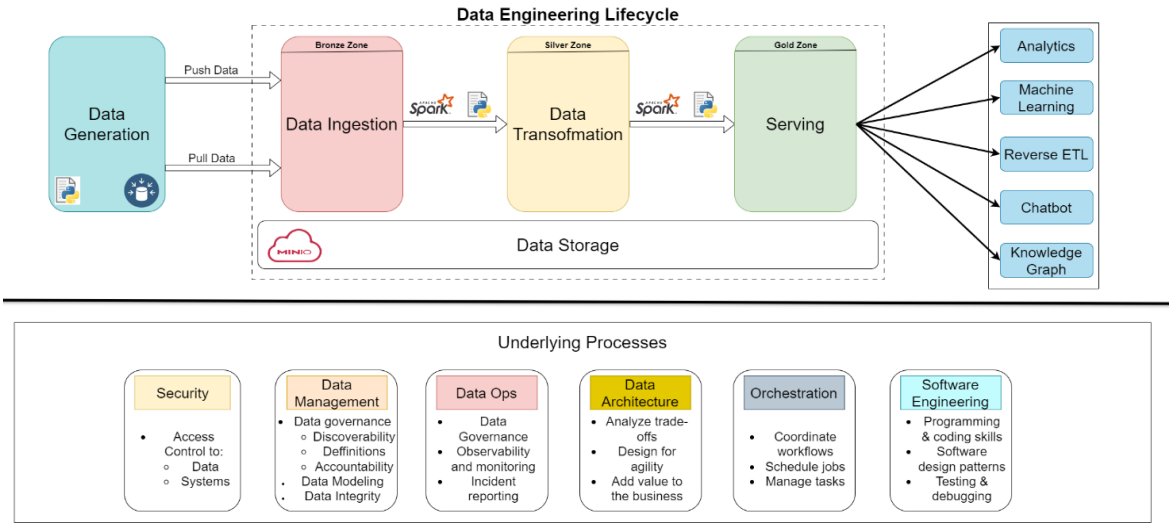

***Figure 2**. TriZ-Flow zones: Bronze Zone (data ingestion), Silver Zone (data standardization and validation), and Gold Zone (data processing and final storage).*

Across all zones, access will be safeguarded through a combination of **Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC)** mechanisms, ensuring that only authorized personnel can interact with specific data types under appropriate conditions. **RBAC** will define baseline access permissions based on predefined roles such as **Data Controller (DC), Data Processor (DP), and Data Protection Officer (DPO)**. These roles will govern broad access rights and responsibilities in line with GDPR requirements. ABAC will provide additional fine-grained control, considering attributes such as data sensitivity, location, processing purpose, and user-specific permissions. This ensures compliance with GDPR's principles, such as data minimization and purpose limitation. The Data Controller will determine access rights, define processing purposes, and establish policies that integrate both RBAC and ABAC rules. The Data Processor will handle data strictly under the controller's instructions, while the DPO will oversee compliance, ensuring that access policies align with GDPR requirements and regulatory obligations. By leveraging both RBAC for structured role-based permissions and ABAC for dynamic, context-aware enforcement, the system will ensure robust security, regulatory compliance, and controlled access to personal data across all zones.

The **process for obtaining access** to any zone will follow a structured, transparent approach. Users seeking access will **(1)** submit a formal request, detailing the purpose and scope of their intended data use. **(2)** This request will be evaluated by the DC in collaboration with the DPO, who will assess whether the proposed use aligns with GDPR principles, such as data minimization, anonymization and de-identification. **(3)** Upon approval, users will be granted access credentials based on their attributes (their role will be one of the attributes), ensuring that

Funded by the European Union

they can only interact with the data necessary for their tasks. **(4)** Continuous auditing and logging mechanisms will track access patterns, ensuring accountability and facilitating compliance reporting.

Through decentralization and robust access protocols, the TriZ-Flow ensures that data privacy is preserved at every stage of the pipeline while enabling meaningful data analysis and knowledge generation.

Bronze Zone: Data Ingestion

The Bronze Zone is the foundational layer where raw data from various sources is ingested and stored in its original format. This zone ensures the complete capture of unaltered data, providing flexibility for reprocessing and future analysis while preserving data lineage and historical tracking. Data is ingested using both push and pull mechanisms, with MINIO serving as the primary storage solution for raw data in formats such as JSON, CSV, and Parquet.

In the **Bronze Zone**, the **Data Controller (DC)** and **Data Protection Officer (DPO)** play critical roles in ensuring GDPR compliance by preventing the collection of unnecessary personal data. The DC sets strict guidelines on what data can be ingested based on its relevance to specific research or operational purposes, while the DPO oversees adherence to data minimization principles and conducts regular audits to verify compliance. During data ingestion, a validation process checks that only authorized data, as defined by these guidelines, is stored in **MINIO**. In addition to storing actual data, also its lineage is stored to ensure that for each data we know its origin, when it was loaded, what kinds of transformations were applied, etc.

Consent management is integrated into the ingestion pipeline to ensure that all personal data collected has explicit, informed consent from data subjects. A consent management system tracks consent terms for each data file (if applicable) or for the agent that is providing the data, enforcing limitations on data use and supporting dynamic consent, allowing subjects to modify or withdraw consent at any time. If consent is withdrawn, the system triggers data deletion or anonymization in the Bronze Zone and notifies downstream zones, ensuring full compliance with GDPR requirements.

Silver Zone: Data Standardization and Validation

The Silver Zone focuses on cleaning, standardizing, and validating the raw data from the Bronze Zone. This stage involves deduplication, format standardization, and the correction of errors or missing values. Data enrichment is also performed here, where additional information is integrated to enhance data context and value. Tools like Apache Spark and Apache Iceberg are used to automate data movement and transformation, ensuring that the data is ready for advanced analytics and machine learning. Depending on the specific data standardization requirements, different datasets - each with potentially distinct processing, compliance, and security considerations - may be handled by different processing frameworks and technologies. These processing pipelines can vary in implementation, leveraging tools optimized for structured, semi-structured, or unstructured data. To ensure seamless integration and governance, these

Funded by the European Union

diverse processing tasks are orchestrated dynamically, aligning with business rules, data quality standards, and regulatory requirements such as GDPR.

Gold Zone: Data Processing and Final Storage

The Gold Zone is the final stage where standardized and validated data is transformed into its final, high-quality form. This processed data is used for advanced applications such as machine learning model training, chatbot development, and knowledge graph construction. The Gold Zone ensures that data is consistent, accurate, and aligned with business objectives, making it ready for a wide range of analytical and operational uses. The processing from the Silver to Gold zone can facilitate different needs, such as, analytical and feature extraction, text standardization, chunking and embedding, time series and signal processing, etc.

The transition from the Silver Zone to the Gold Zone supports a variety of specialized processing needs, depending on the target use case. This includes analytical processing and feature extraction; preparing textual data by chunking and embedding for NLP, vector search, and AI-driven applications; refining sensory data for forecasting and anomaly detection; etc.

To ensure reliability, scalability, and compliance, the Gold Zone leverages specialized processing pipelines and storage solutions tailored to the nature of the data. This includes support for structured, semi-structured, and unstructured formats, with governance mechanisms in place to meet GDPR requirements.

TriZ-Flow as a Foundation for the Human Evaluation Framework

The DMP is closely linked to the establishment of the HEF, which is another critical component of M1. The HEF is designed to assess the performance of the fine-tuned LLMs in neurology-specific tasks. The TriZ-Flow provides the infrastructure and guidelines for managing the data used in the HEF, ensuring that the evaluation process is rigorous, reproducible, and transparent. The TriZ-Flow principles outline how data will be ingested, standardized, validated and safeguard before being used in the HEF. For example, the Bronze Zone of the TriZ-Flow architecture handles raw data ingestion, while the Silver Zone focuses on data standardization and validation. These processes ensure that the data used in the HEF is of high quality and free from biases or inconsistencies. The TriZ-Flow also specifies how data will be stored and accessed, ensuring that the HEF can be easily replicated and scaled across different medical disciplines. Moreover, the TriZ-Flow supports the ethical dimensions of the HEF by ensuring that all data used in the evaluation process complies with GDPR and other relevant regulations. This is particularly important given the project's focus on medical data.

TriZ-Flow as a Platform for Future Projects

The TriZ-Flow pipeline is designed to be a general platform for managing scientific data, not just for the ChatMED project but also for future projects where the Widening Coordinator (FCSE) is involved. The architecture of the TriZ-Flow, which includes the Bronze Zone, Silver Zone, and Gold Zone, provides a scalable and flexible framework for handling diverse datasets. This

Funded by
the European Union

architecture can be adapted to different research contexts, making it a valuable resource for future projects in AI, healthcare, and beyond.

By establishing the TriZ-Flow as a general platform within our DMP, the project ensures that the knowledge and infrastructure developed during ChatMED will have a lasting impact. This aligns with **SO2: Foster Innovation by Promoting the Exchange of Creative Research Methodologies and Strategies**, as the DMP becomes a tool for sharing best practices and fostering innovation across the Widening countries. The DMP also supports the project's commitment to Open Science, as it ensures that data is managed in a way that promotes transparency, reproducibility, and accessibility.

Key Messages of the DMP

This section presents the key messages that highlight the importance of the DMP within the ChatMED project, along with a table that maps the DMP mechanisms to the project's SOs, short-term goals, and long-term goals.

1. **Data as the Foundation of ChatMED:** ChatMED is a data-driven project that leverages GenAI and LLMs to revolutionize neurology. The success of the project hinges on the availability of high-quality, multimodal datasets that capture the complexity of neurological disorders. The DMP ensures that data from diverse sources, such as neuroimaging (MRI, CT scans), electrophysiological data (EEG, EMG), and patient histories, is integrated, standardized, and validated to create a robust foundation for AI-driven diagnostics and treatment recommendations.
2. **TriZ-Flow Architecture:** The DMP introduces a three-zone data processing pipeline (Bronze, Silver, and Gold Zones) that ensures data quality and usability at every stage.
3. **Ethical and FAIR-Compliant Data Practices:** The DMP adheres to the FAIR principles, ensuring that all research data is organized, securely stored, and readily available for future use. The project also prioritizes data privacy and security, particularly for sensitive medical information, by complying with GDPR standards. An Ethics Advisor oversees all data-related activities, ensuring ethical compliance and mitigating risks associated with AI-driven healthcare solutions.
4. **DMP as a Coordination and Training Tool:** The DMP is not just a technical document but also a strategic coordination tool that aligns all partners on data management practices. It supports the establishment of the Quality Assurance and Ethics Committee, ensuring that data handling is both efficient and ethically sound. Additionally, the DMP serves as a training curriculum for researchers, particularly in Widening countries, equipping them with the skills needed to manage large-scale, multimodal datasets. This aligns with **SO3: Advance Researcher Mobility and Expertise Sharing, fostering collaboration and skill development across the consortium.**
5. **Contribution to Milestone M1:** The DMP is one of the five key deliverables in WP2, alongside the Plan for Dissemination, Communication, and Exploitation, the Website Establishment, the Detailed Training Curriculum, and the Human Evaluation Framework. Together, these deliverables contribute to achieving Milestone M1: Solid coordination

Funded by the European Union

measures, curriculums, plans, and HEF established. The DMP ensures that data management practices are standardized, transparent, and aligned with the project's objectives, laying the groundwork for the successful implementation of ChatMED.

6. **A Platform for Future Projects:** TriZ-Flow serves as a general platform for managing scientific data, extending beyond the ChatMED project. Its scalable and flexible architecture can be adapted to future research initiatives, particularly those involving the Widening Coordinator (FCSE). By establishing the platform as a reusable resource, the project ensures that the knowledge and infrastructure developed during ChatMED will have a long, lasting impact.

The following Table 1 provides a clear and structured overview of how the DMP mechanisms align with the SOs, short-term goals, and long-term goals of the ChatMED project. The table highlights the critical role of the DMP in achieving both immediate milestones and long-term impacts.

***Table 1**. Alignment among DMP contributions and ChatMED's SOs.*

Specific Objective	Short-Term Goals	Long-Term Goals	DMP Mechanisms
SO1: Establish a Comprehensive Generative AI Training and Knowledge Exchange Hub	Build a state-of-the-art training facility for GenAI in healthcare.	Create a sustainable hub for AI research and innovation in neurology.	TriZ-Flow pipeline for smooth data flow.
SO2: Foster Innovation by Promoting the Exchange of Creative Research Methodologies and Strategies	Organize training sessions, workshops, and STSEs to share knowledge.	Establish long-term collaborations and innovation pipelines.	DMP provides a curriculum for data management.
SO3: Advance Researcher Mobility and Expertise Sharing	Facilitate staff exchanges and summer schools.	Build a network of skilled researchers in AI and neurology.	DMP aligns data practices across partners.
SO4: Standardize and Disseminate Best Practices in Research and Administration	Develop and implement standardized data management practices.	Ensure long-term adherence to best practices in data management.	DMP teaches FAIR principles and ethical compliance.
SO5: Create a Multimodal Generative ChatMED Platform	Develop a functional AI platform for neurology.	Expand the platform to other medical disciplines and commercial applications.	TriZ-Flow provides high-quality data for AI model training, and enables advanced applications like chatbots and semantic search.

Funded by
the European Union

The Data Management Plan

Data Summary

Will you re-use any existing data and what will you re-use it for?

The ChatMED project employs a rigorous and ethical approach to data management, leveraging well-established and validated neurology-related datasets from reputable sources. These datasets will be used for training and/or fine-tuning LLMs dedicated for neurological applications, enhancing AI-driven clinical decision-making, and developing automated diagnostic and treatment support systems. The integration of these high-quality datasets ensures that AI models are trained on diverse, comprehensive, and representative neurological data, improving both accuracy and generalization abilities.

State the reasons if re-use of any existing data has been considered but discarded.

The decision to reuse these datasets is based on their reliability, widespread availability, and extensive use in prior scientific research. Their structured nature and standardized formats facilitate seamless integration into AI pipelines, accelerating model training and validation. Furthermore, leveraging existing validated data minimizes the need for repetitive data collection efforts, enhancing research efficiency while ensuring responsible data use, promoting efficiency and ethical research practices, while reducing potential biases introduced by small or unrepresentative datasets. Furthermore, leveraging existing data reduces the need for redundant data collection, enhancing research efficiency while ensuring responsible data use. This approach helps avoid potential challenges that may arise from the usage of small or unrepresentative data sets and aligns with the ethical guidelines that prioritize minimizing unnecessary participant involvement.

However, not all existing datasets will be incorporated into the project. Datasets that are proprietary, contain highly sensitive patient information, or have restricted access will be excluded to ensure compliance with ethical standards, licensing regulations, and GDPR requirements. This ensures that the ChatMED project adheres to the highest standards of data privacy, security, and responsible AI development, fostering trust and transparency in AI-driven healthcare solutions.

What types and formats of data will the project generate or re-use?

Publicly available neurological datasets vary widely in their modalities, formats, and applications. These datasets include structural and functional imaging, electrophysiological signals, cognitive assessments, genetic data, pathology segmentations, textual data, etc.

The project will generate and reuse **structured and unstructured medical data** including:

1. **Neuroimaging Data -** high-resolution brain imaging for studying brain structure, connectivity, and pathology.

Funded by the European Union

    a. Formats:
        - DICOM/NIfTI – MRI, PET, fMRI, and DTI scans.
        - MHA – For volumetric segmentation (e.g., BraTS).
        - CSV/XML – Imaging metadata and diagnostic scores.
2. **Pathology-Specific Imaging Data** - labeled pathology segmentations for AI training in detecting neurological abnormalities.
    a. Formats:
        - DICOM/NIfTI – High-resolution neuroimaging.
        - MHA/CSV – Segmentation maps and pathology annotations.
3. **Electrophysiological Data** - EEG, EMG recordings for analyzing epilepsy, neuromuscular disorders, and cognitive function.
    a. Formats:
        - EDF (European Data Format) – EEG, EMG signals.
        - CSV/TXT – Raw time-series data.
4. **Cognitive and Behavioral Data** - neuropsychological assessments, motor function tests, and psychiatric symptom ratings.
    a. Formats:
        - CSV/JSON/XML – Standardized clinical assessment results.
5. **Other Textual Data -** Clinical notes, patient histories, treatment plans, and case studies formatted as structured text (CSV, JSON, XML).
6. **Genomic Data:** Extracted from public genome repositories
    a. Formats:
        - FASTA, VCF – For whole-genome and variant data.
        - BAM, CSV – For transcriptomics and metadata.
        - JSON, TXT – For structured clinical-genomic integration.
7. **Annotated Datasets -** Labeled data for AI model training.
    a. Formats:
        - CSV/JSON – Metadata and clinical labels.
        - NIfTI/MHA – Imaging segmentations.
        - EDF/CSV – Annotated EEG/EMG signals
8. **Pre-trained LLM Models:** Fine-tuned AI models saved in TensorFlow (TF), PyTorch (PT), or ONNX formats.

This data will be curated, preprocessed, and made compatible for fine-tuning of GenAI models.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The primary objectives of the ChatMED project's data management strategy include:

- **Fine-tuning LLMs for clinical decision support.** AI models will be trained to interpret medical text, suggest differential diagnoses, and recommend treatments.
- **Creating a robust knowledge base for Generative AI models.** Integrating multimodal data (text, images, EEG/EMG) to enhance AI comprehension in neurology.

Funded by the European Union

- **Advancing medical research.** AI-powered analysis of neurology-related studies and case studies will improve diagnostic accuracy.
- **Supporting commercialization efforts.** The AI-driven ChatMED platform will serve as a prototype for potential industry adoption.

This aligns with the project's core goal: **bridging the gap between AI research and practical implementation in healthcare across EU and Widening countries.**

What is the expected size of the data that you intend to generate or re-use?

The ChatMED project is expected to handle a large volume of multimodal medical data, requiring high-performance computing resources. The data will be broadly useful across academic, healthcare, technology, and policy sectors, ensuring long-term impact and sustainability beyond the project's duration.

Estimated data sizes include:

- **Structured Text Data:** 10-50GB of medical literature and patient records of which some is annotated with semi-supervised learning approaches.
- **Medical Imaging Data:** 5-10TB of MRI, CT, EEG datasets.
- **Pre-trained AI Models:** Each fine-tuned LLM may require 500 GB-1 TB of storage.
- **Case Study Data:** 50-100GB of labeled patient case studies.

Given these large-scale data requirements, high-performance computing infrastructure with **NVIDIA A100/H100 GPUs** will be used for AI training and inference.

What is the origin/provenance of the data, either generated or re-used?

The data sources for ChatMED come from a mix of existing public repositories and new data collection efforts.

- Existing Data:
    - **Neuroimaging Data**: OASIS, Mind Brain Body, IBC, CERMEP, Forrest Gump fMRI dataset from Hanke et al., ADNI, UK Biobank, SchizConnect, BraTS, ISLES, ATLAS, WMH, PPMI (Parkinson's Progression Markers Initiative), PDBP (Parkinson's Disease Biomarkers Program).
    - **Pathology-Specific Imaging Data**: BraTS – Brain tumor segmentations, ISLES, ATLAS, WMH – Stroke lesion segmentation, Radiopaedia – Expert-curated neuroimaging cases, PPMI, PDBP.
    - **Electrophysiological Data**: PhysioNet – EEG/EMG for epilepsy and sleep disorders, EMGLAB N2001 – Surface EMG for neuromuscular disorders, Mind Brain Body Dataset – Cognitive task-based EEG, PPMI.
    - **Cognitive and Behavioral Data**: ADNI – Alzheimer's disease cognitive assessments, Mind Brain Body Dataset – Linked cognitive and MRI data, SchizConnect – Cognitive tasks and psychiatric scales, PPMI, PDBP - Parkinson's disease-specific motor function tests and cognitive decline.

    - **Other Textual Data**: MIMIC-III, MIMIC-IV, eICU Collaborative Research Database, i2b2 Clinical NLP Challenge Datasets, OpenEHR Clinical Knowledge Manager.
    - **Genomic Data**: ADNI, PPMI, AMP-AD, IGAP, ENIGMA, SchizConnect Genomics Dataset, NeuroLINCS, GAAIN.
    - **Annotated Datasets**: Neuroimaging with Expert Annotations (BraTS, ISLES, ATLAS, WMH, OASIS-3), Electrophysiology (EEG/EMG) with Annotations (TUH EEG Corpus, CHB-MIT Scalp EEG Database, PhysioNet Sleep-EDF Database, EMGLAB N2001), Multimodal Neurological Datasets with Structured Metadata (PPMI, ADNI, MIMIC-III EEG and Clinical Data).
- Newly Generated Data:
    - Neurology-driven scenarios created by medical experts at FMN.
    - Annotated case studies for AI model evaluation.
    - Performance benchmarking datasets.
    - Fine-tuned LLMs that integrate multimodal patient data.

Data management follows FAIR principles and ensures compliance with GDPR and ethical guidelines. In cases, when there is patient or otherwise sensitive data, it would be de-identified and anonymized before being ingested into the ChatMED data platform.

To whom might your data be useful ('data utility'), outside your project?

The data collected and generated by ChatMED will be highly valuable beyond the project's scope, including:

1. Academic and Research Institutions:
    - AI researchers working on medical applications.
    - Neurology and AI-focused research groups.
    - Universities developing AI-powered educational programs.
2. Healthcare Industry & Medical Practitioners:
    - Clinicians seeking AI-assisted diagnostics and decision-making tools.
    - Hospitals looking to integrate AI-based patient management solutions.
    - Pharmaceutical companies exploring AI for drug repurposing in neurological disorders.
3. AI and Technology Companies:
    - Companies developing AI-driven healthcare platforms.
    - Startups working on medical AI solutions.
    - Software developers creating medical AI applications.
4. Public Health Organizations & Policymakers:
    - Institutions working on AI-driven healthcare regulations.
    - EU policymakers monitoring AI applications in medicine.
    - Government bodies investing in AI for national healthcare systems.
5. Patients and Healthcare Advocates:
    - Patient advocacy groups promoting AI-based medical transparency.
    - Open-access tools for patient education on neurological diseases.

Funded by the European Union

The ChatMED platform, datasets, and trained AI models will be made publicly available where feasible, following ethical and legal constraints.

FAIR Data

Making data findable, including provisions for metadata

ChatMED project aims to systematically organize and standardize publicly available neurology-related data, enabling its efficient integration into GenAI pipelines. This includes both fine-tuning LLMs and incorporating data into RAG workflows. To achieve this, the project will develop a specialized data management platform (TriZ-Flow) to support the implementation of the FAIR principles. This platform will serve as the foundation for ensuring that all datasets, data-derived products and publications adhere to the following principles:

- ***Findability***: Each dataset, publication and data-derived product will be assigned a persistent identifier (e.g., DOI) and described using standardized metadata schemas (e.g., Dublin Core, HL7, DCAT) to enhance discoverability. Furthermore the metadata will be assigned a globally unique and persistent identifier.
- ***Accessibility***: Data will be made available for download and review through well-documented APIs and repositories, with transparent access policies that adhere to ethical and regulatory standards. To ensure long-term accessibility, publications and data repositories will implement contingency plans, guaranteeing that metadata remains accessible even if the original resource or repository becomes unavailable.
- ***Interoperability***: Data formats and schemas will adhere to standard vocabularies and carefully selected domain ontologies, ensuring alignment with established health and AI standards such as FHIR, RDF, and JSON for Linking Data (JSON-LD). This will enable efficient integration with AI tools and clinical research platforms. The Data Management Platform will enable open API endpoints using RESTful or GraphQL principles, providing well-documented, secure, and standardized interfaces. These endpoints will allow external computer systems to programmatically access and exchange data, ensuring both interoperability and safety.
- ***Reusability***: Comprehensive metadata, provenance tracking, and licensing information (e.g., Creative Commons or Open Data licenses) will be provided to ensure usability and reproducibility across AI-driven applications and future research. Metadata will be comprehensively documented, including relevant attributes such as data source, creation date, and schema details, to facilitate integration with other compatible datasets and systems.

By adhering to the FAIR principles, ChatMED will deliver high-quality, standardized neurology datasets and data products that empower researchers, developers, and healthcare professionals to advance AI-driven medical insights and decision support systems.

Will data be identified by a persistent identifier?

Funded by the European Union

All datasets, data-derived products, and publications generated and processed within the ChatMED project will be assigned PIDs, such as DOIs or URNs. These PIDs will ensure long-term accessibility, traceability, and unambiguous identification of the data. The DMP developed within the project will maintain a public registry of all PIDs, enabling stakeholders to track data usage, from producers to end-users, and ensuring that data is discoverable by relevant parties.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Yes, rich metadata will be provided to ensure the human and computer discoverability of all datasets, data-derived products, and publications outputted within the ChatMED project. The metadata will include essential attributes such as:

- Descriptive information: Title, description, keywords, and subject categories.
- Technical details: Data format, schema, size, and creation date.
- Provenance: Source, authorship, and version history.
- Licensing: Usage rights and compliance information (e.g., Creative Commons or Open Data licenses).
- PIDs: DOIs or URNs for long-term traceability.

The metadata will adhere to disciplinary and general standards, including:

- Dublin Core for basic resource description.
- HL7 FHIR for health-related data interoperability.
- DCAT for dataset discovery.
- JSON-LD and RDF for structured data representation.

In cases where specific metadata standards do not exist for certain neurology-related data, ChatMED will develop custom metadata schemas based on domain-specific requirements. These schemas will be designed to ensure compatibility with existing standards and will be documented thoroughly to facilitate integration and reuse.
By following these practices, ChatMED will ensure that its data is not only discoverable but also interoperable and reusable across diverse AI-driven and clinical research applications.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, search keywords will be included in the metadata to optimize the discoverability and potential re-use of datasets, data-derived products, and publications within the ChatMED project. These keywords will be carefully selected to reflect the content, context, and domain-specific relevance of the data. Additionally, ChatMED will go beyond and employ a hybrid search approach, combining keyword-based and semantic-based search capabilities for data discovery in the ChatMED's official knowledge repository. This approach will leverage both traditional keyword matching and advanced semantic technologies (e.g., semantic vector search and ontology-based reasoning) to enhance search accuracy and relevance. By integrating domain-specific vocabularies and standard ontologies (e.g., SNOMED CT for clinical terms or MeSH for medical subjects), the hybrid search system will ensure that users can discover resources more effectively, even when using complex or nuanced queries. This dual strategy will maximize the visibility, accessibility, and re-usability of ChatMED's resources in AI-driven applications and clinical research.

Will metadata be offered in such a way that it can be harvested and indexed?

Yes – all metadata will be harvestable through two main routes:

1. *High-level metadata for global discoverability*:
    - High-level metadata (basic information about the datasets, such as titles, descriptions, and keywords) will be made available to globally active resources and federated systems. Examples include platforms like FAIR Data Point, which specialize in making data discoverable across different domains.
    - These systems often provide interfaces that connect to tools like Google Dataset Search, allowing users worldwide to find the data easily. This ensures that the metadata is not only accessible to specialists but also to a broader audience, increasing its visibility and usability.
2. Domain-specific metadata for specialized repositories:
    - For repositories that focus on specific fields (e.g., neurology, healthcare, or AI), the metadata will align with established standards such as HL7 FHIR (for healthcare data), Dublin Core (for general resource description), and DCAT (for dataset catalogs).
    - This alignment allows the metadata to be indexed in two ways:
        - Harvesting: External systems can automatically collect and index the metadata.
        - Submission: Metadata can be manually or programmatically submitted to repositories for indexing.

The specific format and details of the metadata will depend on the capabilities of the repository or system where the metadata is stored. Some repositories may have advanced features for harvesting and indexing, while others may require simpler metadata formats. ChatMED will ensure compatibility with a wide range of systems to maximize the metadata's reach and usability.

Making data accessible

Repository

Will the data be deposited in a trusted repository?

Yes, the data and data-derived products will be deposited in a trusted repository to ensure their long-term preservation, accessibility, and compliance with FAIR principles. Given the importance of transparency and reproducibility in scientific research, all public-domain datasets, along with data-derived products, research papers, and any supplementary materials, will be made publicly available through well-known, trusted repositories that support the modality of the published data.

The selection of repositories will be based on their adherence to best practices in data management, sustainability, and alignment with EU Open Science policies. Preference will be given to repositories that provide persistent identifiers (such as DOIs), support metadata standards, allow for proper version control, and facilitate integration with other research

Funded by the European Union

infrastructures. Potential candidates may include Zenodo[1], OpenNeuro[2], EBRAINS[3], HuggingFace[4] or other domain-specific repositories recommended by the European Open Science Cloud (EOSC).

Additionally, all deposited data will be accompanied by detailed documentation, including metadata, data provenance, and any relevant code or models, ensuring maximum usability for researchers, clinicians, and the wider scientific community. The different types of data and data-derived product modality and their corresponding trusted repositories are described in Table 2.

Table 2. *Repositories to be used in the project.*

1. Raw & Processed Datasets	
Public-domain neurology datasets, preprocessed versions, and any newly derived datasets.	
Zenodo	General-purpose open repository, EU-supported.
OpenNeuro	Specialized in neuroimaging datasets.
EBRAINS	Data Platform (For neuroscience and brain-related datasets).
HuggingFace Datasets	General-purpose repository for accessing and sharing datasets.
2. Data-Derived Products (Models, Annotations, Feature Extracted Data, etc.)	
Trained AI models, extracted features, annotations, or structured datasets derived from public-domain sources.	
Zenodo	AI models and large datasets publishing. https://zenodo.org/communities/chatmed-project/
EBRAINS	Model Repository (For brain-related AI models).
HuggingFace AI Model Hub	Publicly available AI models and easy to access using HuggingFace libraries. Integration with Zenodo will be used for DOI assignment. The following organization space will be used as public repository: https://huggingface.co/ChatMED-Project

[1] https://zenodo.org/
[2] https://openneuro.org/
[3] https://www.ebrains.eu/
[4] https://huggingface.co/

Funded by the European Union

3. Research Papers & Publications	
Peer-reviewed publications, white papers, and technical reports.	
Zenodo	Open Access publications will be achieved at Zenodo. https://zenodo.org/communities/chatmed-project/
arXiv	For preprints of the results before peer-reviewing.
Europe PMC	For biomedical and life sciences papers.
UKIM's central repository	Ss. Cyril and Methodius University of Skopje official repository. https://repository.ukim.edu.mk
4. Software, Code, and Algorithms	
Python/Java/TensorFlow/PyTorch scripts, data processing pipelines, model training code, API endpoints.	
GitHub	Integration with Zenodo for DOI assignment. All the software, code and algorithms will be available at the following site: https://github.com/ChatMED
GitLab	For internal version-controlled collaborative development.
Software Heritage	For long-term archiving of research software.

By depositing the project's outputs in these repositories, we ensure broad accessibility, long-term sustainability, and alignment with Open Science principles, maximizing the impact and reusability of the research.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

At this stage, we have identified several trusted repositories aligned with EU Open Science policies and FAIR principles for depositing different types of data generated by the project, which require no specific arrangement for accessibility and the data will be uploaded via the standard procedures.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Funded by the European Union

The repositories identified for this purpose (Table 3) of the ChatMED project generally provide PIDs such as DOIs or accession numbers, enabling data traceability, long-term accessibility, and compliance with FAIR principles. The repository also resolves these identifiers to digital objects, allowing users to access, cite, and reference the deposited data.

***Table 3**. Repositories and identifier types.*

Repository	Identifier Type	Resolves to Digital Object?	Notes
Zenodo	DOI	Yes	DOI assigned upon upload
EBRAINS	DOI / Handle System	Yes	Neuroscience datasets and models get a persistent identifier
OpenNeuro	DOI	Yes	Neuroimaging datasets get a DOI
Hugging Face	No DOI by default	No (only model URLs)	Zenodo integration (link to Zenodo for DOI integration)
arXiv	arXiv ID	Yes	Persistent identifier for preprints
Europe PMC	PubMed ID / DOI	Yes	Biomedical publications with persistent links
GItHub	No DOI by default	Only repository URL and commit hashes	Zenodo integration - provides a DOI that resolves to specific GitHub repository version

Data

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All data generated in this project will be made openly available whenever possible. However, some datasets may be subject to restricted access due to legal, contractual, or intentional restrictions. For publicly sourced datasets integrated into the system, we will comply with the original data owner's guidelines, licenses, and requirements, such as CC0 (public domain with no restrictions), CC BY (requires attribution), CC BY-SA (requires attribution and share-alike), ODbL (requires attribution and share-alike for databases), and Apache 2.0 (permissive software license with some open data provisions). For datasets generated within the project by consortium partners, we will adhere to all relevant data protection regulations (e.g., GDPR) and contractual agreements.

Funded by the European Union

Where feasible, data will be de-identified and anonymized; otherwise, access will be restricted to authorized users only. Some data may remain restricted due to ethical considerations, security concerns, or commercial interests. In multi-beneficiary projects, specific beneficiaries may choose to keep their data closed if its release would conflict with their legitimate interests, as outlined in the Grant Agreement. If an embargo is necessary, it will allow time for publication or intellectual property protection (e.g., patents). The embargo period will be determined based on the data type and legal requirements but will generally not exceed two years. Once the embargo expires, the data will be made publicly available through appropriate repositories.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

If an embargo is necessary, it will be strictly limited to the time necessary for publication or intellectual property protection (e.g., patents), ensuring compliance with open science principles.. The embargo period will be minimized and will not exceed two years, unless exceptional legal or regulatory requirements necessitate an extension. Once the embargo expires, the data will be made publicly available through FAIR-aligned repositories.

Will the data be accessible through a free and standardized access protocol?

Where possible, data will be accessible through a free and standardized access protocol, ensuring interoperability and ease of use. To maintain security while providing broad accessibility, we will offer API-based access using industry-standard protocols such as REST (over HTTPS) and GraphQL, incorporating authentication mechanisms like OAuth 2.0, API keys, and token-based authentication. Additionally, direct database access will be available through secure protocols, including JDBC and ODBC with SSL/TLS encryption to protect data in transit.

For file-based data transfers, we will support secure protocols such as SFTP (Secure File Transfer Protocol), FTPS (FTP Secure), and HTTPS-based file downloads, ensuring data integrity and confidentiality. Where applicable, authentication methods like SSH keys or multi-factor authentication (MFA) will be enforced to enhance security.

Beyond the access methods provided by the project's data platform, we will utilize open-access repositories that comply with FAIR principles, ensuring data discoverability and long-term availability. These repositories will follow best practices for secure data access, including PIDs (e.g., DOIs), metadata standards, and role-based access control where necessary.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

If data use is restricted due to confidentiality, ethical, or legal considerations, access will be granted based on defined policies and controlled permissions. For sensitive or proprietary data, access may be subject to data-sharing agreements (DSAs), requiring users to formally request access and agree to conditions such as non-disclosure clauses, ethical compliance, and data use limitations. In some cases, data anonymization, de-identification or aggregation may be applied to enable broader sharing while protecting sensitive information.

After the project concludes, data access will be maintained via secure repositories, institutional data archives, or controlled-access databases, ensuring long-term availability. Where applicable, we will deposit data in trusted repositories with managed access controls, allowing continued use

Funded by the European Union

by authorized users. Persistent identifiers (e.g., DOIs) will be assigned to datasets to ensure traceability, and metadata will be published to facilitate discoverability.

For non-public datasets, access requests will be reviewed and managed through a designated data governance body or an automated approval system, ensuring compliance with regulatory and contractual obligations while maximizing responsible data reuse.

How will the identity of the person accessing the data be ascertained?

During the project, access will be managed through authenticated user accounts, institutional agreements, and RBAC and ABAC to ensure that only authorized users can retrieve or modify the data. For external users or collaborators, access will be granted via verified institutional credentials, invitation-based registration, or explicit approval workflows. Data access logs will be maintained to track authentication attempts and activity, supporting auditability and compliance with GDPR, HIPAA, or other relevant data protection frameworks. After the project concludes, access control policies will remain in place, ensuring that only authorized users, based on predefined agreements, retain access to restricted data. Identity verification mechanisms will continue through persistent authentication protocols and periodic revalidation of user permissions, ensuring data security over time.

Authentication will be implemented using secure identity management systems, which may include LDAP (Lightweight Directory Access Protocol), Single Sign-On (SSO) solutions (e.g., OpenID Connect, SAML), or federated identity providers (e.g., ORCID for researchers or institutional login systems), depending on the specific requirements and infrastructure of the project. These mechanisms will be selected to balance security, usability, and interoperability, ensuring seamless yet secure access while mitigating password-related risks. Where appropriate, multi-factor authentication (MFA) will be incorporated to enhance identity verification. This may involve a combination of authentication factors, such as passwords, authentication apps, security tokens, or biometric methods (e.g., fingerprint or facial recognition), based on the level of access required and security policies in place.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

A separate Data Access Committee is not required given the existing governance framework. Access to personal and sensitive data is already managed through a combination of Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC), ensuring that only authorized personnel can interact with specific data types under predefined conditions.

The Data Controller (DC) and Data Protection Officer (DPO) are responsible for evaluating and approving access requests. The structured access process includes:

1. Users submitting a formal request specifying the purpose and scope of data use.
2. The DC and DPO reviewing the request to ensure compliance with GDPR principles, such as data minimization and purpose limitation.
3. Approved users receiving access credentials based on ABAC mechanisms, allowing interaction only with necessary data.
4. Continuous auditing and logging to ensure compliance and accountability.

Since the DC and DPO already oversee access requests and enforce regulatory compliance, an additional Data Access Committee would introduce redundancy. However, if there are highly

Funded by the European Union

complex access requests requiring broader stakeholder input, a committee could be considered. Otherwise, the existing governance structure effectively ensures controlled and compliant data access.

Metadata

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, all metadata generated in the project will be made publicly available and will be licensed under a public domain dedication (CC0), in full alignment with the Grant Agreement and FAIR principles. Additionally, the metadata will include all necessary information to enable users to access the data. This will typically include:

- PIDs (DOI or Accession Numbers) for datasets, ensuring proper citation and long-term discoverability.
- Repository links where the data is stored (e.g., Zenodo, OpenNeuro, EBRAINS, EGA, BBMRI-ERIC).
- Descriptive metadata (e.g., dataset title, version, authors, keywords, methodology).
- Format and standards information, ensuring compatibility with existing research infrastructures.
- Licensing details, confirming that the data itself is openly accessible under an appropriate open license.

By adopting CC0 licensing, the ChatMED project will enable metadata reuse while maintaining compliance with EU Open Science policies.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The data will remain available and findable for a minimum of four years after the project's completion, ensuring continued access for researchers and stakeholders. During this period, the data will be stored in a secure and accessible repository, adhering to best practices for data preservation. Even after the data itself is no longer available, the metadata will be guaranteed to remain accessible, providing essential information about the dataset's context, structure, and origin. This ensures that the research remains traceable and citable, supporting transparency and reproducibility in the long term.

Will documentation or reference about any software needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Yes, comprehensive documentation or references about any software required to access, read, or interpret the data will be included to ensure usability and transparency. This documentation will provide clear instructions on how to use the software, interpret file formats, and process the data. Where possible, the codebases of the software (tied to the release and version used for analysis) will be forked into the ChatMED GitHub repository and archived in the associated Zenodo space to provide full reproducibility. If proprietary software is necessary, detailed guidance on obtaining and using it will be provided, or if possible open source alternatives will be specified.

Funded by
the European Union

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

To ensure data interoperability, we will adhere to community-endorsed best practices, including the FAIR Principles for findability and reusability, Research Data Alliance (RDA) guidelines for metadata harmonization, and W3C Linked Data Principles for structured data integration.

We will adopt Open Data Commons standards for accessible formats like CSV, Parquet, JSON, XML, and NetCDF, and adopt semantic interoperability using RDF and JSON-LD. These formats will enable structured, machine-readable metadata, enabling linked data integration and enhanced discoverability. Furthermore, we will align with domain-specific frameworks such as ISO 19115 for geospatial data, DDI for social sciences, and MIAME for bioinformatics. Metadata will be enriched following OpenAIRE and EOSC guidelines, ensuring persistent identifiers (DOIs, ORCIDs) for long-term traceability. These practices will ensure compliance with global open science initiatives, ensuring efficient data exchange and reuse across disciplines.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

If the use of uncommon or project-specific ontologies and vocabularies is unavoidable, we will ensure their interoperability by providing alignment and mappings to widely recognized and commonly used ontologies (e.g., UMLS, SNOMED CT, MeSH, OMOP, HPO, and FMA for biomedical concepts). These mappings will align project-specific terms with established standards, such as W3C Linked Data Ontologies (e.g., SKOS, FOAF, PROV-O), BioPortal Ontologies (e.g., Gene Ontology, OBI), and ISO metadata frameworks where applicable.

Additionally, all newly generated ontologies and vocabularies will be openly published in machine-readable formats (e.g., RDF, OWL, JSON-LD) in public repositories such as GitHub, Zenodo, or OntoPortal. Furthermore, they will be accompanied by comprehensive documentation detailing the ontology structure, relationships, and intended use cases.This will enable the research community to reuse, refine, and extend the ontologies developed under the ChatMED project, thus improving the domain data interoperability, long-term usability, and alignment with FAIR data principles.

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

To promote transparency, interoperability, and compliance with FAIR data principles, ChatMED will include qualified references to other (meta)data and related research outputs, such as:

- Scientific publications relevant to the project.
- Patents that are linked to the research findings or methodologies.
- External datasets from previous studies, projects or publicly available repositories.
- Internal project datasets to ensure traceability and reproducibility of results.

Funded by the European Union

The quality of these references will largely depend on the extent to which the target data aligns with the 5-star Open Data[5] framework. For restricted or non-public data, the qualified reference will provide detailed descriptions of the dataset, including its access permissions and usage restrictions.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

To ensure the data analysis can be validated and the data can be effectively re-used, detailed documentation will be provided in multiple accessible formats. This will include readme files that outline the methodology, data collection processes, and any relevant contextual information. Codebooks will be included to define variables, explain units of measurement, and describe data structures. Additionally, documentation will cover data cleaning steps, analytical procedures, and any transformations applied to the data. If scripts or code were used for analysis, these will be shared alongside comments or annotations to clarify their functionality. All documentation will be stored in the same repository as the data, ensuring it is easily accessible to users. This comprehensive approach will support transparency, reproducibility, and facilitate data re-use by other researchers.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes, all data generated in the ChatMED project will be freely available in the public domain to maximize accessibility and reuse, fully aligning with the obligations set out in the Grant Agreement and EU Open Science policies.

To ensure the widest possible re-use, the data will be licensed under standard open-access licenses, including:

- CC0 for metadata, ensuring unrestricted use, sharing, and integration with other datasets.
- CC BY 4.0 for datasets and data-derived products, allowing unrestricted reuse while requiring proper attribution.
- MIT/Apache 2.0 for software, ensuring transparency and flexibility for further development and integration.

Additionally, the data will be deposited in trusted repositories (e.g., Zenodo, OpenNeuro, EBRAINS) to ensure long-term availability with PIDs (DOIs or accession numbers) for easy citation and discovery.

Will the data produced in the project be usable by third parties, in particular after the end of the project?

Yes, the data produced in the project will be fully usable by third parties, both during and after the project's completion. To ensure this, the data will be shared in open, non-proprietary formats that

[5] https://5stardata.info/en/

Funded by
the European Union

are widely accessible and interoperable. Comprehensive documentation, including metadata, codebooks, and methodological details, will be provided to enable third parties to understand, interpret, and reuse the data effectively. Additionally, any software or tools developed for data analysis will be made available, preferably as open-source code, to further support reuse. The data will be deposited in a trusted repository with PIDs, ensuring long-term accessibility and citability. These measures will guarantee that the data remains a valuable resource for the research community and other stakeholders beyond the project's lifespan.

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, the provenance of the data will be thoroughly documented using appropriate standards to ensure transparency, traceability, and reproducibility. This documentation will include detailed information about the origin of the data, data collection methods, processing steps, and any transformations or analyses applied. Standards such as Dublin Core, DataCite, or PROV-O will be used to describe the data's lineage, ensuring compliance with best practices in data management. Metadata will capture essential details such as the time and location of data collection, instruments or tools used, personnel involved, and any quality control measures implemented. By adhering to these standards, the provenance documentation will enable users to understand the data's context, assess its reliability, and facilitate its reuse in future research or applications.

Describe all relevant data quality assurance processes.

The ChatMED project implements a comprehensive data quality assurance framework to ensure the integrity, accuracy, and usability of data throughout its lifecycle. This begins with pre-collection planning, where standardized protocols, metadata requirements, and ethical considerations are established, and personnel receive training to ensure consistency in data collection. During data acquisition, real-time validation checks detect outliers, missing values, and format inconsistencies, while cleaning and validation processes, both automated and manual, remove duplicates, enforce standardization, and enrich data with additional contextual information. Robust documentation and provenance tracking guarantee transparency by maintaining metadata and transformation histories, ensuring traceability and reproducibility. Data undergoes peer review and cross-validation against external benchmarks to verify accuracy and compliance with project standards. Version control mechanisms, including change tracking and audit logs, maintain dataset integrity, while secure storage, regular backups, and access controls protect data from loss, corruption, and unauthorized access. Post-collection audits and user feedback loops facilitate continuous quality improvement, complemented by GDPR compliance measures such as informed consent, anonymization, and ethical oversight. Security protocols, including RBAC, ABAC, encryption, and recovery plans, safeguard sensitive data, while adherence to the FAIR principles ensures findability, accessibility, interoperability, and reusability through persistent identifiers, standardized schemas, and transparent access policies. Finally, long-term preservation strategies, including storage in trusted repositories and metadata retention, guarantee sustained availability for future research and AI advancements. Through these rigorous quality assurance measures, ChatMED ensures the reliability of its datasets, ultimately supporting the development of ethical, high-performing AI-driven healthcare solutions.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Funded by the European Union

In addition to managing data in accordance with the FAIR principles, the DMP for the ChatMED project also addresses other research outputs, allocation of resources, data security, and ethical aspects. Below is a detailed discussion of these considerations:

1. Research Outputs Other Than Data

The ChatMED project generates various research outputs beyond raw data, including software, workflows, protocols, AI models, and digital assets. These outputs are managed with the same rigor as data, ensuring they adhere to FAIR principles and are accessible for future use.

Software and code:

- All software, code, and algorithms developed during the project (e.g., Python scripts, TensorFlow/PyTorch models) will be made openly available under open-source licenses (e.g., MIT, Apache 2.0).
- Code will be deposited in repositories like GitHub and Zenodo, with persistent identifiers (DOIs) assigned for long-term accessibility.
- Comprehensive documentation, including README files, user guides, and comments within the code, will be provided to ensure usability and reproducibility.

AI Models:

- Pre-trained and fine-tuned AI models (e.g., Large Language Models for neurology) will be shared via platforms like Hugging Face and EBRAINS, ensuring they are accessible to researchers and clinicians.
- Models will be accompanied by metadata, including training data sources, hyperparameters, and performance metrics, to facilitate reuse and validation.

Workflows and protocols:

- Detailed workflows for data processing, model training, and evaluation will be documented and shared alongside the data and code.
- Protocols for data collection, cleaning, and analysis will be made available to ensure reproducibility and transparency.

Knowledge graphs and annotations:

- Knowledge graphs and annotated datasets will be shared in standardized formats (e.g., RDF, JSON-LD) to support interoperability and reuse in other AI-driven healthcare applications.

2. Allocation of Resources

The DMP outlines the allocation of resources to ensure effective data and research output management throughout the project lifecycle.

Costs for FAIR compliance:

- Storage and archiving: Costs associated with storing data and research outputs in trusted repositories are covered under the project budget.
- Data security: Resources are allocated for implementing robust security measures, including encryption, access control, and backup systems.
- Metadata and documentation: Costs for creating and maintaining comprehensive metadata and documentation are included to ensure data and outputs are FAIR-compliant.

Funded by the European Union

- Training and capacity building: Funds are allocated for training researchers and staff in data management best practices, ensuring they are equipped to handle complex datasets and outputs.

Long-Term preservation:

- Trusted repositories: Data and research outputs will be deposited in trusted repositories that guarantee long-term preservation and accessibility.
- PIDs: Costs for assigning PIDs (e.g., DOIs) to datasets, models, and publications are included to ensure traceability and citability.

Responsibility for data management:

- Data management team: A dedicated team, including a DC, DPO, and DPs will oversee data management activities.
- Ethics and quality assurance committee: This committee will ensure compliance with ethical guidelines, GDPR, and quality assurance processes.

3. Data Security

Data security is a critical aspect of the DMP, particularly given the sensitive nature of medical data. The following measures are implemented to ensure data security:

Secure storage and archiving:

- Encryption: Data is encrypted during storage and transfer to protect it from unauthorized access or breaches.
- Redundant storage: Data is stored in redundant systems (both on-premises and cloud-based) to prevent loss or corruption.
- Trusted repositories: Data and research outputs are deposited in trusted repositories that comply with best practices for secure data storage.

Access Control:

- Role-Based Access Control (RBAC): Access to data and outputs is restricted based on user roles (e.g., Data Controller, Data Processor).
- Attribute-Based Access Control (ABAC): Fine-grained access control is implemented based on attributes such as data sensitivity, user location, and processing purpose.
- Authentication Mechanisms: Secure authentication methods, including multi-factor authentication (MFA) and Single Sign-On (SSO), are used to verify user identities.

Data Recovery and Backup:

- Regular backups: Regular backups are performed to ensure data can be restored in case of loss or corruption.
- Disaster recovery plans: Robust disaster recovery plans are in place to ensure continuity of data access and processing.

Compliance with Regulations:

- GDPR compliance: All data handling processes comply with GDPR regulations, including data minimization, anonymization, and informed consent.
- Ethical oversight: An Ethics Advisor oversees all data-related activities to ensure compliance with ethical guidelines and mitigate potential risks.

Funded by the European Union

4. Ethical Aspects

The ChatMED project places a strong emphasis on ethical considerations, particularly given the sensitive nature of medical data and the use of AI in healthcare.

Informed consent:

- Patient consent: If patients are involved in any of the project activities, they will be provided with detailed information sheets and must give explicit consent if their data is going to be used in the project. This includes consent for data sharing and long-term preservation.
- Hospital consent: Healthcare institutions involved in the project will provide consent for the use of AI technologies for the benefits of the project. Those formal documents will be required by their ethical committee and will be available for review on request by the reviewers.

Data privacy and security:

- Anonymization and de-identification: Sensitive data (if any) will be anonymized or de-identified before being used in the project to protect patient privacy.
- Data minimization: Only the data necessary for the project's objectives is collected and processed, in line with GDPR principles.

Bias and fairness:

- Bias mitigation: Measures are implemented to identify and mitigate biases in datasets and AI models, ensuring fairness and equity in healthcare applications.
- Transparency and explainability: AI models will be designed to be transparent and explainable, enabling clinicians and researchers to understand how decisions are made.

Accountability and liability:

- Ethics committee: Both internal and external Ethics Advisor oversees all project activities to ensure compliance with ethical guidelines and regulatory requirements.
- Liability frameworks: Clear frameworks are established to address accountability and liability issues related to the use of AI in healthcare.

An Ethics Advisor is embedded within the project to ensure all data-related activities adhere to the highest ethical standards. In coordination with the Data Controller (DC) and Data Protection Officer (DPO), the Ethics Advisor reviews data approval requests, verifies that informed consent requirements are met, and confirms adherence to GDPR and other relevant regulations. The Ethics Advisor also provides guidance on anonymization protocols, ensuring that personal identifiers and any sensitive details are properly masked or removed prior to data sharing. Where necessary, they participate in Data Access Committee (DAC) discussions to offer additional insight on complex ethical issues, thus reinforcing transparent, responsible, and ethically sound data management practices throughout the consortium.

5. Long-Term Impact and Sustainability

The DMP ensures that the knowledge and infrastructure developed during the ChatMED project have a lasting impact:

- Reusable platform: The TriZ-Flow architecture (Bronze, Silver, and Gold Zones) is designed to be scalable and adaptable for future projects in AI and healthcare.

Funded by the European Union

- Open science commitment: By adhering to FAIR principles and open science practices, the project ensures that its outputs are accessible and reusable by the broader scientific community.
- Capacity building: Training activities and knowledge exchange programs build the capacity of researchers in Widening countries, enabling long-term collaboration and innovation.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

All digital outputs generated by ChatMED are covered under this DMP, as they are fundamentally composed of data. This includes software, workflows, protocols, AI models and other digital assets, all of which will be managed in accordance with the specified strategies. These outputs will follow the same FAIR principles, storage, licensing, and accessibility measures as specified above, ensuring consistency in their management and long-term availability.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

As mentioned above, all research outputs will be deposited in trusted repositories along with comprehensive metadata, ensuring they are discoverable, accessible, and reusable in alignment with FAIR principles. These repositories feature indexed search functionalities, enabling efficient retrieval and proper attribution of each output.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

The direct costs for making data and research outputs FAIR in the ChatMED project will be minimal, as the project leverages local infrastructure to host its trusted repository and utilizes free EU-supported open science services (e.g., Zenodo, EBRAINS) for data storage and sharing. For indirect costs, data curation and management activities are integrated into the personnel costs for research, as these tasks are performed by the project's researchers and staff as part of their regular project responsibilities. To clarify, during the project lifetime, the computing infrastructure including GPUs and the whole data processing platform will be hosted on-premise in the participating institutions. However, the output of this processing, as well as relevant code developed will be made available through the public data and code repositories mentioned earlier. This approach ensures compliance with FAIR principles while keeping additional expenses low, aligning with the project's budget and Horizon Europe funding guidelines.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Funded by the European Union

The costs related to research data and output management in the ChatMED project will be covered through the personnel costs dedicated to the research component of the Horizon Europe grant. Since data curation, management, and FAIR compliance activities are integral to the research process, these tasks are performed by the project's researchers and staff as part of their regular responsibilities. This approach ensures that no additional budget allocation is required, as the costs are inherently included in the personnel costs for research. By integrating data management into the research workflow, the project remains compliant with Horizon Europe grant conditions while efficiently utilizing available resources.

Who will be responsible for data management in your project?

In the ChatMED project, the scientific committee will play a key oversight role in ensuring that researchers manage data effectively and in compliance with FAIR principles. However, the day-to-day responsibility for data management will be distributed among the following roles:

1. Data Controllers (DCs): Responsible for defining data management policies, ensuring compliance with GDPR, and overseeing data access and usage.
2. Data Processors (DPs): Handle the technical aspects of data processing, including cleaning, standardization, and storage.
3. Data Protection Officer (DPO): Ensures compliance with data protection regulations and oversees ethical considerations.
4. Researchers: Individual researchers will manage the data specific to their tasks, ensuring it is properly curated, documented, and stored according to project guidelines.

The scientific committee will monitor and evaluate how researchers manage different types of data, ensuring alignment with the project's objectives and ethical standards. This collaborative approach ensures robust data management while leveraging the expertise of both researchers and the committee.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Long-term preservation of data in the ChatMED project will be ensured through a combination of local infrastructure and open science platforms, with no additional costs foreseen. The project will utilize the following strategies:

1. Local Infrastructure: The coordinators's local infrastructure, including secure servers and backup systems, will be used to store and preserve data for the long term. This infrastructure is already in place and maintained as part of the institution's resources, ensuring no extra costs are incurred.
2. Open science platforms: Data and research outputs will be deposited in trusted open science platforms such as Zenodo, EBRAINS, and OpenNeuro, which provide free, long-term preservation services. These platforms assign persistent identifiers (e.g., DOIs) to ensure data remains findable and accessible indefinitely.
3. Data selection for preservation: The scientific committee, in collaboration with researchers, will decide which data and outputs are preserved based on their scientific value, relevance, and potential for reuse. This includes datasets, AI models, workflows, and documentation.

Funded by the European Union

4. Duration of preservation: Data will be preserved for a minimum of 5 years after the project's completion, in line with best practices for scientific research. Metadata will remain accessible indefinitely, even if the data itself is no longer available.

By leveraging existing local infrastructure and free open science platforms, the project ensures long-term preservation without additional costs, while maintaining compliance with FAIR principles and Horizon Europe requirements.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

The ChatMED project ensures data security through encryption (at rest and in transit), role-based and attribute-based access control, and secure storage in trusted repositories like Zenodo and EBRAINS. Data transfers use secure protocols (HTTPS, SFTP), and regular backups and disaster recovery plans ensure data recovery. Compliance with GDPR is maintained through anonymization, informed consent, and regular cybersecurity audits. An Ethics Advisor oversees ethical compliance, and researchers receive security training to handle data securely. These measures protect sensitive data while ensuring integrity, confidentiality, and long-term accessibility.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, the data in the ChatMED project will be safely stored in trusted repositories for long-term preservation and curation. The project will use open science platforms such as Zenodo, EBRAINS, and OpenNeuro, which are recognized for their robust security measures, compliance with FAIR principles, and commitment to long-term data preservation. These repositories provide persistent identifiers (e.g., DOIs) to ensure data remains findable and accessible indefinitely. Additionally, sensitive data will be stored in secure local infrastructure with encryption and regular backups. By leveraging these trusted repositories and local systems, the project ensures that data is preserved securely and remains accessible for future research and reuse.

Ethics

Are there, or could there be, any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

The ChatMED project recognizes several ethical and legal considerations that may impact data sharing, particularly in the context of medical data privacy, security, and compliance with regulatory frameworks such as the GDPR[6] and the EU Artificial Intelligence Act[7]. The project ensures that data handling follows strict ethical guidelines, minimizing risks related to patient privacy, informed consent, and bias in AI models. One of the primary concerns is informed

[6] https://gdpr-info.eu/

[7] https://artificialintelligenceact.eu/the-act/

consent, which applies both to healthcare institutions adopting GenAI and to patients whose data may be used. The project follows established best practices from the WHO and the EMA to ensure that patients are fully informed about how their data will be collected, processed, and potentially shared. Special attention is given to data minimization, ensuring that only the necessary information is collected (GDPR, Article 5) and that patients' rights are fully protected.

Additionally, data security and compliance with GDPR are critical. The TriZ-Flow Data Processing Pipeline implements Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) mechanisms, ensuring that only authorized personnel can access specific datasets. Encryption, anonymization, and pseudonymization techniques are applied where necessary to further protect sensitive patient information.

Bias in AI models also poses an ethical risk. ChatMED addresses this by implementing rigorous validation processes, including peer review and benchmarking against external datasets, ensuring that AI-driven insights remain reliable, fair, and free from discriminatory biases. Accountability and transparency are further reinforced by audit trails, version control, and explainability mechanisms, which enable stakeholders to trace how data is used in AI model development.

Finally, long-term data governance is structured to align with FAIR principles and GDPR (Article 5 and 89), ensuring that data remains ethically managed beyond the project's duration. Trusted repositories such as Zenodo, OpenNeuro, and EBRAINS will be used for secure storage, and metadata will be maintained even after datasets are no longer accessible.

All these measures are reviewed by the Ethics Advisor and the Data Protection Officer (DPO) within the ChatMED project, ensuring compliance with both legal and ethical requirements.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

In compliance with GDPR Article 5, the ChatMED project recognizes two types of consent: hospital consent for the use of GenAI in medical applications and patient consent for the use of personal data in collaboration with GenAI. Both consents are crucial to ensuring ethical, legal, and secure data management within the project.

Hospitals participating in ChatMED will provide explicit consent before implementing GenAI-based solutions. This consent is ideally obtained after hospital personnel are familiarized with both the advantages and challenges of GenAI in medical applications. While the consent process may be brief, it must ensure that the institution acknowledges its responsibilities regarding data governance, security, and regulatory compliance.

Potential patient consent in ChatMED will consist of two essential parts if patients are involved in any scenario:

1. An information sheet explaining how personal data will be used, stored, shared, and potentially preserved for research and AI model training.

2. A confirmation of consent, where patients explicitly agree to data processing, anonymization measures, and the possibility of long-term preservation under GDPR regulations.

Patients will be provided with a transparent and voluntary consent process, which ensures that they:

- Fully understand the purpose and scope of data usage in AI-driven medical research.
- Are aware of their right to withdraw consent at any time without affecting their medical care.
- Are informed about encryption, anonymization, and controlled access mechanisms that protect their data.

By obtaining both hospital and patient consent, ChatMED ensures that data sharing, security, and ethical considerations align with legal regulations and best practices, fostering trust among patients, healthcare providers, and AI researchers.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Yes, in addition to the trusted open science platforms (e.g., Zenodo, EBRAINS, OpenNeuro), the ChatMED project will also utilize the www.repository.ukim.mk portal, which is the official institutional repository of the Ss. Cyril and Methodius University of Skopje. This portal will be used to store and share research outputs, including datasets, publications, and software, ensuring compliance with institutional and national data management policies. The repository supports long-term preservation, persistent identifiers, and metadata standards, aligning with FAIR principles. By leveraging this portal alongside international repositories, the project ensures broad accessibility and adherence to both local and global data management best practices.

Summary

This **DMP** outlines a comprehensive strategy for managing the data within the ChatMED project and beyond, distinguishing itself with key non-generic contributions:

1. **TriZ-Flow Architecture**: The DMP introduces a three-zone data processing pipeline (Bronze, Silver, and Gold Zones) designed to ensure data quality, standardization, and compliance with GDPR. This decentralized and distributed architecture allows for secure data processing closer to its source, minimizing privacy risks.
2. **Ethical and FAIR-Compliant Data Practices**: The DMP emphasizes ethical data handling, including informed consent, data minimization, and compliance with GDPR. It also adheres to FAIR principles to ensure data is organized, securely stored, and available for future use.

Funded by
the European Union

3. **Integration of Multimodal Data**: The project focuses on integrating diverse data types (e.g., neuroimaging, electrophysiological data, patient histories) to create a comprehensive AI-driven diagnostic tool for neurology. This multimodal approach enhances the accuracy and reliability of AI models.
4. **Training and Capacity Building**: The DMP serves as a training curriculum for researchers, particularly in Widening countries, equipping them with skills to manage large-scale, multimodal datasets.
5. **Foundation to the HEF**: The DMP provides the infrastructure and guidelines for managing data which will be used in the HEF, ensuring rigorous, reproducible, and transparent evaluation of LLMs in neurology-specific tasks.
6. **Long-Term Impact and Sustainability**: The TriZ-Flow architecture is designed to be a scalable platform for future projects, extending beyond ChatMED. The DMP ensures that the knowledge and infrastructure developed during the project will have a lasting impact, aligning with Open Science practices.
7. **Data Security and Access Control**: The DMP implements Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) mechanisms to ensure secure data access. It also includes robust data recovery and backup plans to protect sensitive medical information.
8. **Ethical Oversight**: An Ethics Advisor is embedded within the project to oversee all data-related activities, ensuring compliance with ethical guidelines and mitigating risks associated with AI-driven healthcare solutions.

Overall, the DMP provides a robust framework for managing data responsibly, ensuring ethical compliance, and enabling innovation in AI-driven healthcare, particularly in the field of neurology.