

ITU-T Focus Group Deliverable

(09/2023)

Focus Group on Artificial Intelligence for Health
(FG-AI4H)

**DEL10.8 – FG-AI4H Topic Description
Document for the Topic Group on AI for
neurological disorders (TG-Neuro)**

ITU-T FG-AI4H Deliverable

DEL10.8 – FG-AI4H Topic Description Document for the Topic Group on AI for neurological disorders (TG-Neuro)

Summary

This topic description document (TDD) specifies a standardized benchmarking for AI in neurological diseases. It covers scientific, technical and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence, benchmarking, data audit, clinical relevance, data quality, ethics, health, neurological diseases, overview, regulations, topic description, topic groups.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1 of the Deliverable DEL10.8 on "*FG-AI4H Topic Description Document for the Topic Group on AI for neurological disorders (TG-Neuro)*" approved on 16 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Ferath Kherif
CHUV
Switzerland

Marc Lecoultre
ML Labs,
Switzerland

E-mail: ferath.kherif@chuv.ch

E-mail: ml@mlab.ai

© ITU 2025

Some rights reserved. This publication is available under the Creative Commons Attribution-Non Commercial-Share Alike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; <https://creativecommons.org/licenses/by-nc-sa/3.0/igo>). For any uses of this publication that are not included in this licence, please seek permission from ITU by contacting TSBmail@itu.int.

If you wish to reuse material from this publication that is attributed to a third party, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder.

Table of Contents

	Page
1 Introduction.....	1
2 About the FG-AI4H topic group on neurological disorders	2
2.1 Documentation	2
2.2 Status of this topic group	3
2.3 Topic Group participation	5
3 Topic description	6
3.1 Subtopic Neurocognitive diseases	6
4 Ethical considerations	8
5 Existing work on benchmarking	10
5.1 Subtopic neurocognitive disorders	10
6 Benchmarking by the topic group.....	11
6.1 Subtopic on neurocognitive disorders	12
7 Overall discussion of the benchmarking.....	27
8 Regulatory considerations	27
8.1 Existing applicable regulatory frameworks.....	27
8.2 Regulatory features to be reported by benchmarking participants	28
8.3 Regulatory requirements for the benchmarking systems	28
8.4 Regulatory approach for the topic group	28
Annex A – Glossary	31
Annex B – Declaration of conflict of interests	33

DEL10.8 – FG-AI4H Topic Description Document for the Topic Group on AI for neurological disorders (TG-Neuro)

1 Introduction

This topic description document specifies the standardized benchmarking for TG-Neuro systems. It serves as Deliverable 10.8 (DEL10.8) of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

This topic group is dedicated to AI being used against neurocognitive diseases. It provides an empirical basis for testing the clinical validity of machine-learning-based diagnostics for Alzheimer's disease (AD) and related dementia syndromes (defined by DSM V as 'neurocognitive disorders') using real-world brain imaging and genetic data. With increased life expectancy in modern society, the number of individuals who will potentially develop dementia is growing proportionally. Current estimates are that over 48 million people worldwide are suffering from dementia, bringing the social cost of care to 1% of the world's gross domestic product (GDP). These numbers led the World Health Organization to classify neurocognitive disorders as a global public health priority.

Compared with visual assessment, automated diagnostic methods based on brain imaging are more reproducible and have demonstrated a high accuracy in separating AD from healthy ageing, but also in the clinically more challenging separations between different types of neurocognitive disorders. Similarly, although ApoE genotypes carrying higher risk for AD are easily obtainable, this information is rarely integrated in machine-learning-based diagnostics for AD. Although encouraging, implementations into clinical routine have been challenging.

The goal of this topic group is to create a standardized benchmark for evaluating AI systems for the diagnosis and treatment of neurocognitive disorders. By using real-world data it aims to evaluate the clinical validity and reproducibility of these systems and ultimately contribute to the integration of AI-based diagnostics into clinical practice. It is expected that this benchmarking will not only improve the accuracy and efficiency of diagnostics, but also reduce the social and economic burden associated with neurocognitive disorders.

A large representative sample will be created and will be used for the creation of the models. The models will be then validated (see benchmarking methods below) on the real-world undisclosed patient data.

The benchmarking process will be based on the most modern methods used by the ML community, but also on the recommended methodology for clinical trials.

Two subtopic groups were established:

- Neurocognitive diseases, led by Kherif Ferah (CHUV, Switzerland) and Marc Lecoultré (MLLab, Switzerland).
- AI based Parkinson's disease screening and management, led by Khondaker Abdullah Al Mamun (mamun@cse.uiu.ac.bd; AIMS Lab, United International University, Bangladesh).

The first subtopic progressed, and the results are documented in this deliverable, while the second subtopic group did not progress significantly and remains for further study; results will be reported in a future opportunity. Hence, "neurocognitive diseases" is used across this deliverable.

2 About the FG-AI4H topic group on neurological disorders

The introduction highlights the potential of a standardized benchmarking of AI systems for neurological disorders to help solving important health issues and to provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Neuro at Meeting B, New York, 15-16 November 2018.

FG-AI4H assigns a topic driver to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During the FG-AI4H meeting B, New York, 15–16 November 2018, Ferath Kherif from the University Hospital of Lausanne (CHUV) was nominated as topic driver for the TG-Neuro.

Current members of the topic group on AI against neurocognitive diseases are shown in Table 1.

Table 1 – Topic group members

Name	Bio
Kherif Ferah, Vice-director LREN, CHUV, Switzerland	Senior Lecturer at the University of Lausanne and vice-director of the Laboratoire de Recherche en Neuroimagerie (LREN) of Département des Neurosciences Cliniques (DNC) at the University Hospital of Lausanne (CHUV). He obtained his PhD in neuroscience at Pierre and Marie Curie University, Paris. He was research fellow at MRC-CBSU in Cambridge and then at the Wellcome Trust Centre for Neuroimaging in London before his arrival in Lausanne in 2010. He used functional imaging to probe cognitive function and used my mathematical background to test new hypotheses pertaining the explanation of individual differences.
Marc Lecoultré, MLLab.ai, Switzerland	Expert in AI & Data Science, strong entrepreneurship professional with a master's degree from the Swiss Federal Institute of Technology, a Graduate Certificate from Stanford and multiple certifications in Lean Management and AI domains. He has founded several companies in these fields. He has practised AI and machine learning for over 15 years. He has worked on dozens of projects in various companies and industries. He is an editor and actively participates in the WHO/ITU focus group on AI for health.

The topic group would benefit from further expertise of the medical and AI communities and from additional data.

2.1 Documentation

This document is the TDD for the TG-Neuro. It introduces the health topic, including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for neurocognitive disorders. It describes the existing approaches for assessing the quality of neurocognitive disorder systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subclauses for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as Deliverable DEL10.8 "Neurocognitive disorders (TG-Neuro)". The topic group is expected to submit input documents reflecting updates to the work on this deliverable (Table 2) to each FG-AI4H meeting.

Table 2 – Topic group output documents

Number	Title
FGAI4H-C-020-R1	Status report for AD' use case
FGAI4H-B-013-R1	Proposal: Using machine learning and AI for validation of AD' biomarkers for use in the clinical practice
FGAI4H-R-016-A01	TDD update (TG-Neuro) [same as Meeting L]
FGAI4H-R-016-A02	CfTGP Update (TG-Neuro) [same as Meeting E]
FGAI4H-S-016-A03	Status update on TG-Neuro

2.2 Status of this topic group

With the publication of the "call for participation" of the current topic group members, it is expected to be shared within their respective networks of field experts.

2.2.1 Status update for meeting D

The following is an update of activities since meeting D:

- The updated Call for topic group participation for TG-Neuro was published on the ITU website and can be [downloaded here](#).
- Several e-mail exchanges with the topic group members to request inputs and updates to the TDD.
- Networks reached out to via e-mail and social media (LinkedIn, Twitter), sharing the call for topic group participation and to spread the word.
- Preliminary interest expressed by several groups and individuals interested in contributing to the topic group and are following up with them individually.

2.2.2 Status update for meeting E

The following is an update of activities since meeting D:

- A new submission regarding Standardization of MRI Brain Imaging for Parkinson Disease made by Biran Haacke, Prof. Mark Haacke, Mark Messow from The MRI Institute for BMR in Canada.
- a300 patients' datasets added to the Alzheimer's data that will be available for AI solutions. New quantitative and semiquantitative methods for assessing image quality included.
- Several discussions with clinical research groups and hospitals that will be interested to join the neurocognitive disease. The discussion is ongoing and still at a preliminary stage; new groups from Italy and Bulgaria are likely to be integrated.
- Onboarding Prof. Alexander Tsiskaridze (neurologist) from Ivane Javakhishvili Tbilisi State University's Faculty of Medicine, Georgia. He may provide data, new topics and AI solutions.
- Two meetings with the Norwegian Ministry of Health and Care Services to include stakeholders from northern Europe in the FG.
- A discussion with an EU official on the topic of defining cloud/computing infrastructure needs for health research. A meeting/workshop is planned for October, final date to be determined. Ferath Kherif will be presenting the neurocognitive disease group.

2.2.3 Status update for meeting F

There are no activity updates from meeting F.

2.2.4 Status update for meeting G

The following is an update of activities since meeting F:

- Identification of new cohorts to be included.
- Create a catalogue of potential studies that can be included in the future (potential large datasets, challenge: harmonization, only in Europe).

2.2.5 Status update for meeting H

The following is an update of activities since meeting G:

- Rename TG-Cogni (neurocognitive diseases) to TG-Neuro "Neurological disorders". The neurocognitive diseases use case becomes a subtopic group within TG-Neuro.
- Cover the AI-based Parkinson's disease screening and management use case as a subtopic group within the TG-Neuro (ex TG-Cogni). The subtopic is led by Khondaker Abdullah Al Mamun (mamun@cse.uiu.ac.bd), AIMS Lab, United International University, (Bangladesh).
- Requests: From a few startups (3).
- Data: Improved feature extraction from data and quality measures.
- Metadata registry.
- Develop generic tools for data curation, quality control and provenance.
- Develop, implement and deploy tools to extract brain morphology, genomic, proteomic behavioural and cognitive features from clinical and research databases.
- Contribution DASH.
 - Data capture:
 - Distributed sites;
 - Data quality;
 - Curation;
 - Standards;
 - Formats etc.
 - Algorithms:
 - Decentralized, locally hosted data sets federated platform.

2.2.6 Status update for meeting K

The following is an update of activities since meeting H:

- Best practices for data sharing, sourcing data access, quality and curation.
 - Data confidentiality;
 - Data security and privacy;
 - Data anonymization;
 - Data de-identification;
 - Data minimization;
 - DAQCOR Indicators: descriptive system for planning and reporting observational studies.
- First Metadata Model: Dementia use case.

2.2.7 Status update for meeting M

The following is an update of activities since meeting K:

- Created a minimal viable product for data sourcing with:

- Data catalogue;
- Data curation;
- Data exploration.
- Link to EU and worldwide initiative;
- Established research contract with Brazil/Fundação Cruz (Fiocruz);
- List and connect to other partners/platforms (e.g., AD Workbench/Gates);
- Complete the data catalogue;
- MVP data catalogue;
- Federated algorithms and hybrid cloud infrastructure;
- Audit trial.

2.2.8 Status update for meeting N

The following is an update of activities since meeting M:

- Attracting data providers who are also data users;
- Increasing the diversity and representativeness of the data, along multiple dimensions such as:
 - Psychiatric health;
 - Gender;
 - Geographical locations (environment factors, social, culture, etc.)

2.3 Topic Group participation

The participation in both, the Focus Group on AI for Health and in a TG was generally open to anyone (with a free ITU account). For this TG, the corresponding 'Call for TG participation' (CfTGP) can be found here:

- <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-R-016-A02.docx>

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

- <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Neuro.aspx>

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG zoom link:

- <https://itu.zoom.us/my/fgai4h>

All relevant administrative information about FG-AI4H – such as upcoming meetings or document deadlines – will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

- <https://itu.int/go/fgai4h/join>

The TG-Neuro does not have a specific mailing list in addition to the general FG-AI4H mailing list.

Regular FG-AI4H workshops and meetings take place about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

- <https://itu.int/go/fgai4h>

3 Topic description

This clause contains a detailed description and background information of the specific health subtopic for the benchmarking of AI in neurocognitive disorders and how this can help to solve a relevant real-world problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies and streamline FG-AI4H meetings. However, in some cases, different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. This TDD currently covers the subtopic on AI against neurocognitive diseases. Subtopics for dementia or Parkinson could be covered in future versions.

This topic group is dedicated to AI against neurocognitive diseases. It provides an empirical basis for testing the clinical validity of machine-learning-based diagnostics for neurological disease (AD' or Parkinson's disease) and related dementia syndromes (defined by DSM V as neurological disorders) using real-world brain imaging and genetic data.

Additional conditions that are relevant to this topic group may be added in the future.

3.1 Subtopic Neurocognitive diseases

3.1.1 Definition of the AI task

This clause provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is not about the benchmarking process (this will be discussed more detailed in clause 4). This clause corresponds to [DEL3](#) "*AI4H requirement specifications*", which describes the functional, behavioural and operational aspects of an AI system.

With increased life expectancy in modern society, the number of individuals who will potentially develop dementia is growing proportionally. Current estimates are that worldwide over 48 million people suffering from dementia, bringing the social cost of care to 1% of world's gross domestic product (GDP). These numbers led the World Health Organization to classify neurocognitive disorders as a global public health priority. The topic systematically addresses previous limitations by using real-world imaging and genetic data obtained in the clinical routine that are analysed with predictive machine-learning algorithms, including benchmarking and cross-validation of the learned models. The intended integrative framework will assign a level of probability to each of several possible diagnoses to provide an output that is readily usable and interpretable by clinicians. Beyond this immediate impact on clinical decision-making and patient care, a flexible strategy allows for scaling the framework by integrating further clinical variables – neuropsychological tests, imaging and CSF biomarkers, to name but a few that will lead to new areas of research developments.

3.1.2 Current gold standard

This clause provides a description of the established gold standard of the addressed health topic.

Compared with visual assessment, automated diagnostic methods based on brain imaging are more reproducible and have demonstrated a high accuracy in separating AD from healthy ageing, and also in the clinically more challenging separations between different types of neurocognitive disorders. Similarly, although ApoE genotypes carrying higher risk for AD are easily obtainable, this information is rarely integrated in machine-learning-based diagnostics for AD. Although encouraging, implementations into clinical routine have been challenging.

The group's own and others' studies on structural imaging already considered more than two diagnostic options or used probabilistic rather than categorical diagnostic labels. These pattern recognition machine-learning-based approaches run on a standard PC and rely on a set of labelled training data – for example, structural magnetic resonance imaging (MRI) and reliably established diagnostic label for each subject – to diagnose new cases in the absence of expert radiologists. They

also permit a fully automated detection and quantification of specific pathologies (e.g., white matter hyperintensities or microbleeds).

3.1.3 Relevance and impact of an AI solution

This clause addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

The proposal is novel, has translational importance and is potentially applicable to epidemiological, pharmacological and therapeutic studies in all clinical domains seeking to explore various aspects of health-related big data and validate their accuracy as biomarkers. It will not only advance the scientific understanding of ageing-associated cognitive decline and neurocognitive disorders. It will also provide a model for infrastructure and technology for the creation of large-scale projects in different fields of research for the benefit of patients, clinical and basic science researchers.

3.1.4 Existing AI solutions

This clause provides an overview of existing AI solutions for the same health topic that are already in operation. It contains details of the operations, limitations, robustness and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in clause 6.

Supervised classification methods for predicting clinical outcome and analysing variance in data have been used successfully. Previously, SVM classifiers have been applied to anatomical data for diagnosing different forms of dementia. However, multivariate pattern recognition approaches have typically been applied to unimodal data, motivating the development of a methodological approach to accommodate multiple-modal data. Recently this methodology has been applied in order to develop predictive models for healthy ageing and found that the mean prediction error was significantly reduced when all measurements were combined. Table 3 provides summaries of other AI solutions.

Table 3 – AI and machine-learning in neurodegenerative disease care

Reference	Supporting System	Domain	Features	Methodology	Target Users
Bruun et al. 2019 [1]	Clinical Decision Support System, PredictND tool	Dementia: vascular, frontotemporal, Alzheimer's, subjective cognitive decline.	<ul style="list-style-type: none"> – Clinical test – MRI visual – Data analytics 	Objective comparison of data	Clinicians, neurologist
Rao et al. 2017b; 2017a; 2020 [2–4]	CDS-CPL: Clinical Decision Support and Care Planning Tool	Alzheimer's disease and related dementia: ADRD	<ul style="list-style-type: none"> – Online questionnaire – Evidence-based recommendations – Physical exam techniques – Referrals, medications 	Differential diagnosis, individualized care plans	Caregivers, NPs and Pas
Mitchell et al. 2018 [5]	An advance care planning video decision support tool	Promote goal-directed care for advanced dementia patient	<ul style="list-style-type: none"> – Medical records – Bedford Alzheimer Nursing Severity-Subscale 	Providing care after viewing the video	Nursing home residents
Tolonen et al. 2018 [6]	Clinical Decision Support System, PredictND tool	Designed for differential diagnosis of different types of dementia	<ul style="list-style-type: none"> – Multiple diagnostic tests such as neuropsychological tests, MRI and cerebrospinal fluid samples 	Multiclass disease State index classifier, visualization of its decision-making	Support physician

Table 3 – AI and machine-learning in neurodegenerative disease care

Reference	Supporting System	Domain	Features	Methodology	Target Users
Vashistha et al. 2019 [7]	AI-based clinical decision systems (CDSs) along with point-of-care diagnosis	Neurodegenerative disorders such as Parkinson's disease, amyotrophic lateral sclerosis (ALS), AD, epilepsy	<ul style="list-style-type: none"> Machine learning and wearables based therapeutics A combinatorial intelligent system for the prediction of PD development by machine learning 	Markov decision processes (MDP) and dynamic decision networks	Neurodegenerative disorder specialist

4 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal and social concerns that have to be considered in this context. They are discussed in Deliverable [DEL01](#) "*Ethics and governance of artificial intelligence for health*", which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This clause refers to DEL01 and should reflect the ethical considerations of the TG-Neuro.

- What are the ethical implications of applying the AI model in real-world scenarios?
 - Ethical implications include issues related to patient consent, privacy and data retention, as well as data bias and concerns related to AI models, including trust and the potential risk of misdiagnosis.
- What are the ethical implications of introducing benchmarking? (Having the benchmarking in place itself has some ethical risks: e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk.) The following points should be considered:
 - There is a risk in relying on metrics that cannot comprehensively capture the nuances of clinical decision-making. In addition, the cost of health care may be a factor that consciously or unconsciously influences final decisions.
 - The ethical implications of benchmarking in the context of neurodegenerative diseases are particularly challenging because of the lack of definitive diagnostic tests and the high rate of misdiagnosis by clinicians.
 - With clinicians being wrong 30% of the time, it is very risky to base AI on learning from potentially incorrect labels, which could amplify and spread existing inaccuracies and biases in diagnosis.
 - This could lead to the development of AI models that systematically reproduce the same errors made by clinicians, resulting in incorrect or delayed diagnoses and suboptimal treatment decisions.
- What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
 - Ethical considerations include the potential hesitation of patients and their families to participate in data collection, especially when there is no definitive diagnosis or treatment.
 - Risk of stigmatization and discrimination against individuals diagnosed with neurodegenerative diseases.

- What risks face individuals and society if the benchmarking is wrong, biased or inconsistent with reality on the ground?
 - This could result inappropriate treatment recommendations or delayed treatment, irreversible damage caused by ineffective medications and increased health care costs (unnecessary tests, treatments and hospitalizations).
 - Loss of confidence in AI models and ethical issues due to bias (discrimination against certain demographic groups).
- How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
 - Informed consent from patients and their families that clearly explain the purpose, use of the data and potential risks.
 - Data minimization: Collect only the information necessary for the intended purpose.
 - Withdraw consent and record deletion: Provide patients and their families with the opportunity to withdraw consent and request deletion of their records at any time.
 - Data storage is in secure and encrypted databases with strict access controls to limit access to only authorized personnel.
 - Access logs are reviewed to determine who accessed the data, when and for what reason.
 - Data sharing: data is shared only with authorized individuals and organizations, under a data transfer agreement. Obtain additional consent from patients and their families or ethic committee before sharing data with third parties or for secondary purposes.
 - Data security: Firewalls are implemented to prevent unauthorized access, regular security audits are conducted.
 - Data security training is provided to medical and non-medical staff.
 - Compliance with laws and regulations: Comply with all data privacy laws and regulations (e.g., GDPR).
 - Privacy impact assessment (PIA) is conducted with a data privacy officer (DPO).
- How is it ensured that benchmarking data are representative and that an AI offers the same performance and fairness? (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of 'inclusion and exclusion criteria' for test data?)
 - Fairness across different demographic groups and contexts is critical for TG-neuro, especially considering the role of socioeconomic status (SES), gender, education and lifestyle on the incidence, severity and prevalence of neurological disease.
 - TG-Neuro proposes to conduct a broader and representative data collection to obtain data from high-, low- and middle-income countries and that takes into account various factors such as age, gender, ethnicity, SES, education level and lifestyle.
 - Identify and eliminate potential biases, such as over- or under-representation of certain demographic groups or disease subtypes (define clear and comprehensive inclusion and exclusion criteria).
 - Benchmarking is conducted in different demographic groups and contexts independently (via collaboration with external organizations and institutions) to assess the performance and fairness of the model.

- What are your experiences and learnings from addressing ethics in your TG?
 - Ethical considerations in data collection: an important aspect of the work was to address ethical concerns related to data collection. These included ensuring informed consent, protecting privacy, clarifying data ownership and management issues and obtaining ethics committee approvals when necessary. Some important lessons were learned from experience working on ethical issues in TG-Neuro. Participation in the WHO-ITU focus group facilitated access to global collaboration and allowed us to share knowledge and best practices and address ethical challenges together.
 - Obtaining data continues to be a major challenge. There are ethical and logistical barriers (privacy concerns, consent management, data ownership) to accessing and sharing data, making the development of fair and robust AI models important. Potential of federated learning: As a potential solution to some of the data challenges, federated learning has emerged as a viable option because it allows models to be trained on decentralized data, eliminating the need for data sharing and reducing privacy concerns. Implementing federated learning requires significant local effort and the availability of adequate human and technological resources, which is not possible in all areas, particularly in low- and middle-income countries.

5 Existing work on benchmarking

This clause focuses on the existing benchmarking processes in the context of AI and neurocognitive disorders for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics and clinical evaluation attempts). The goal is to collect all relevant learning from previous benchmarking that could help to implement the benchmarking process in this topic group.

5.1 Subtopic neurocognitive disorders

5.1.1 Publications on benchmarking systems

While a representative comparable benchmarking for neurocognitive disorders does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This clause summarizes insights from the most relevant publications on this topic. It covers parts of the Deliverable [DEL7](#) "*AI for health evaluation considerations*", [DEL7.1](#) "*AI4H evaluation process description*", [DEL7.2](#) "*AI technical test specification*", [DEL7.3](#) "*Data and artificial intelligence assessment methods (DAISAM)*", and [DEL7.4](#) "*Clinical Evaluation of AI for health*".

The following items are for further study:

- What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
- What are the most relevant approaches used in literature?
- Which scores and metrics have been used?
- How were test data collected?
- How did the AI system perform and how did it compare to the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
- How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?
- Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?

- What are the most relevant gaps in the literature (what is missing concerning AI benchmarking)?

5.1.2 Benchmarking by AI developers

All developers of AI solutions for neurocognitive disorders implemented internal benchmarking systems for assessing the performance. This clause will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

The primary data are already available and growing in volume. Data will include both real-world patient data and data collected from research cohorts. The data will include clinical scores, diagnostic, cognitive measures and biological measures (PET, MRI, fMRI, lab results).

The data include patients on more than 6 000 patients on dementia (one of the largest patient' cohorts) different stages of the disease (subjective complains, mild impairments or demented).

5.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This clause reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by Deliverable [DEL7.5](#) "*FG-AI4H assessment platform*" (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

With the advent of electronic health records (which) and picture archiving and communication systems (PACS), clinical researchers have the ability to access information pertaining to groups of patients in their hospital, provided they have the informed consent of each individual patient.

Because of privacy regulations related to patient privacy and security, both twichEHR and PACS systems were designed to collect data on patients in a single hospital. Patient medical records remained scattered across a large number of hospitals, clinics and private practices around the world, as was the case with paper-based medical records prior to the introduction of their electronic form.

Today, integrating scattered electronic patient records and PACS is a major challenge, not only because of patient data protection, but also because of incompatible ICT solutions. As a result, clinical researchers can only access data stored in their own hos'itals' systems.

Global leaders in medical informatics have addressed this challenge by developing solutions for two distinct purposes:

- Solutions for managing content and processing research data (e.g., LORIS and CBwhichn);
- EHR systems for sharing patient data between clinicians (e.g., Cerner and Epic Systems);
- Data catalogues (e.g., EMIF and GAAIN).

None of the three different groups of solutions support data analytics use cases.

6 Benchmarking by the topic group

This clause describes all technical and operational details regarding the benchmarking process for the TG-Neuro, subtopic on neurocognitive disorders AI task including subclauses for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: [DEL5](#) "*Data specification*" (introduction to Deliverables 5.1–5.6), [DEL5.1](#) "*Data requirements*" (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL5.2](#) "*Data acquisition*", "*Data annotation specification*", "*Training and test data specification*" (which provides a systematic

way of preparing technical requirement specifications for datasets used in training and testing of AI models), "*Data handling*" (which outlines how data will be handled once they are accepted), [DEL5.6](#) "*Data sharing practices*" (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), [DEL06](#) "*AI training best practices specification*" (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL7](#) "*AI for health evaluation considerations*" (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), [DEL7.1](#) "*AI4H evaluation process description*" (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), "*AI technical test specification*" (which specifies how an AI can and should be tested *in silico*), [DEL7.3](#) "*Data and artificial intelligence assessment methods (DAISAM)*" (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), [DEL7.4](#) "*Clinical evaluation of AI for health*" (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL7.5](#) "*FG-AI4H assessment platform*" (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL9](#) "*AI for health applications and platforms*" (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL9.1](#) "*Mobile based AI applications*", and [DEL9.2](#) "*Cloud-based AI applications*" (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 Subtopic on neurocognitive disorders

The benchmarking of the TG-Neuro subtopic on neurocognitive disorders is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This clause outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent clauses, where the actual benchmarking methodology for each version will be described.

6.1.1 Benchmarking version 1

A large representative sample will be created and will be use for the creation of the models. The models will be then validated (see benchmarking methods below) on the real-world undisclosed patient data.

The benchmarking process will be based on the most modern methods used by the machine learning community, but also on the recommended methodology for clinical trials.

6.1.1.1 Overview

This clause provides an overview of the key aspects of this benchmarking iteration, version [Y], for the TG-Neuro group, which is focused on neurodegenerative diseases. The overall scope of this benchmarking iteration is to specify the functionality planned for benchmarking by TG-Neuro. This iteration aims to perform initial benchmarking focusing on the technical and operational dimensions, as well as the scientific and clinical potential of AI tools applied to neurodegenerative diseases using data such as MRI, PET scans and clinical memory scores. Features added to benchmarking in this iteration include:

Benchmarking selection criteria: Specific criteria for selecting use cases that demonstrate the clinical value added by AI tools for neurodegenerative diseases. These criteria will help ensure that the selected use cases are relevant, meaningful and representative of the challenges faced in neurology departments and memory clinics.

Specifications for use of the benchmarking system: Detailed specifications for the use of AI methods in various platform scenarios, including the goals and needs of clinicians and researchers in neurology departments and memory clinics, end-to-end scenarios for the use of the platform in each of the participating hospitals and the benefits of using the platform in a clinical context.

Clinical use case specifications: Detailed specifications for the clinical use cases targeted by the benchmarking, including the specific neurodegenerative conditions, required input data (e.g., MRI, PET scans, clinical memory scores), expected outcomes and potential impact on patient outcomes.

Use cases from each of the participating data providers (e.g., hospitals) where benchmarking will be applied, as well as user testimonials, will be detailed and analysed in a future version of this document. The planned benchmarking includes the goals and needs of clinicians and researchers in neurology departments and memory clinics, the end-to-end scenario for using the platform in each of the hospitals, and the benefits of using the platform in a clinical context.

6.1.1.2 Benchmarking methods

This clause provides details about the methods of the benchmarking version 1. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources and legalities).

6.1.1.2.1 Benchmarking system architecture

This clause covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

The platform is built by TG-Neuro by combining federated learning techniques with robust privacy-preserving mechanisms with the aim to facilitate data sourcing from different institutions while ensuring data security and privacy.

The architecture focused on data management and privacy-preserving machine learning. We leverage the principles of Data Mesh, an emerging data platform architecture that promotes decentralized, domain-oriented data ownership and ensures efficient data integration and sharing among various stakeholders. This architecture will incorporate data cleansing and harmonized pre-processing to ensure data consistency and compatibility across multiple sources.

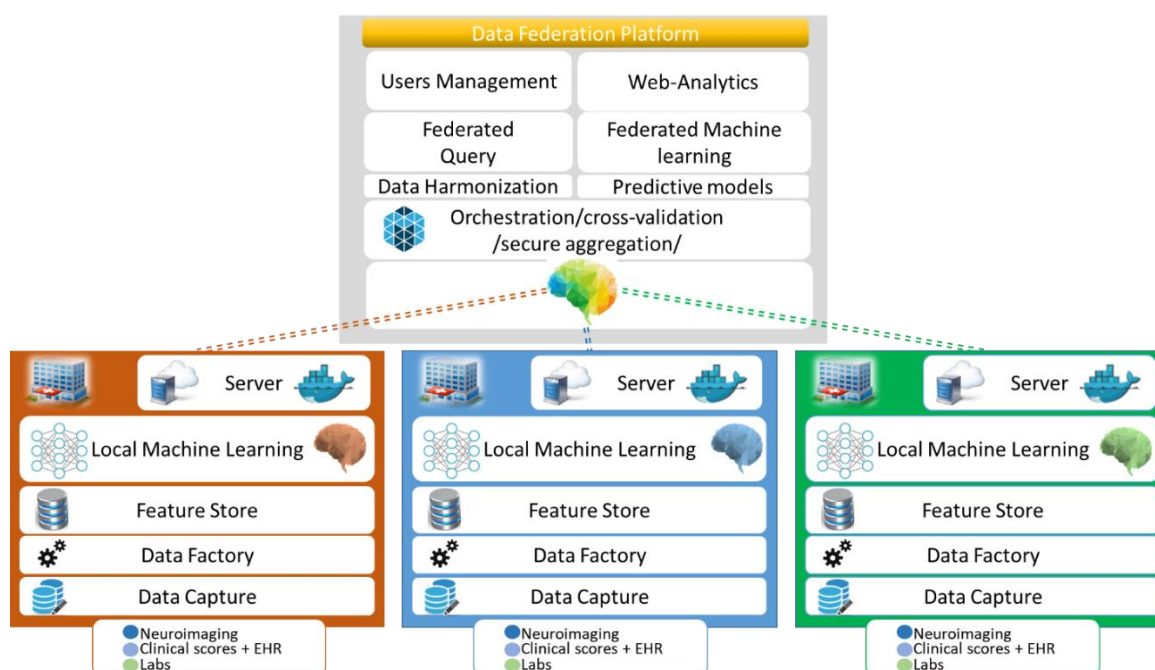


Figure 1 – Federated benchmarking platform

6.1.1.2.2 Benchmarking system dataflow

This clause describes the dataflow throughout the benchmarking architecture. The TG-Neuro Benchmarking Platform (Figure 1) is a distributed information system that:

- collects de-identified health-related and privacy-sensitive patient data from hospital information systems (EHR systems and PACS) and research datasets (ADNI, ESDS, PPMI, etc.) related to neurodegenerative diseases – data capture components;
- processes captured neuroimaging and other patient biomedical and demographic data to extract patient health-related characteristics – data factory components;
- harmonizes and normalizes feature data types across datasets captured from different hospitals and research databases – data factory components;
- provides permanent patient feature data in each participating hospital – characteristic feature store components;
- provides a set of pre-integrated statistical methods and predictive machine-learning algorithms, including benchmarking and cross-validation of learned models.

6.1.1.2.3 Safe and secure system operation and hosting

This clause addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage or data loss.

In the case of a manufactured data source (versus self-generated data), it is possible to refer to the 'manufacturer's prescriptions:

Data De-identifier replaces the following personally identifiable information with pseudonyms:

- Information exported from EHR systems in CSV format;
- Information from neuroimages stored in the headers of DICOM files.

Data De-identifier saves the files with de-identified data to storage; see Figure 2.

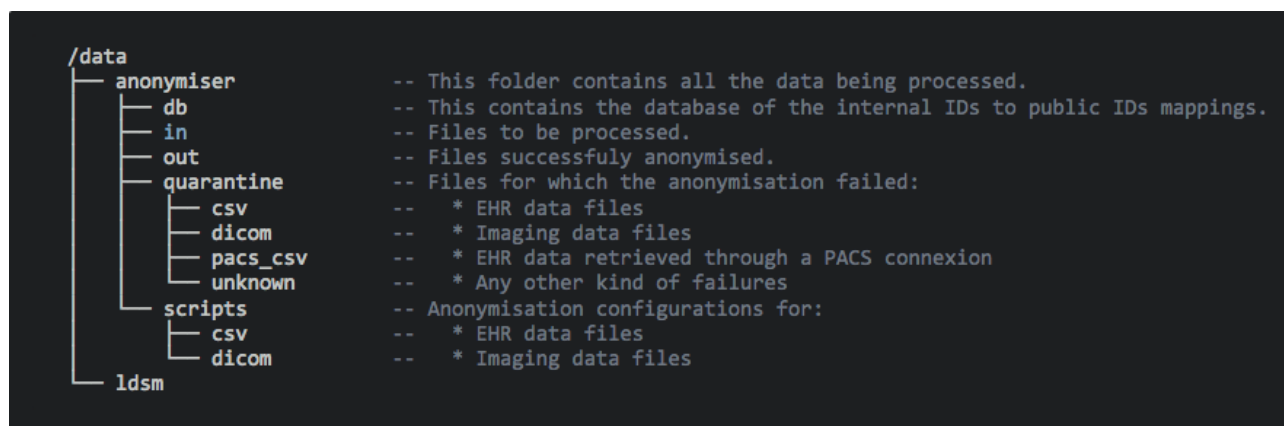


Figure 2 – Data folder organization for the de-identification processing

Data used for both training and test datasets are stored in the features database, in the features data store subsystem.

The model benchmark & cross-validation component performs data split using k-fold cross-validation. This method of data sampling divides the complete dataset into K disjoint parts of roughly the same size. K different models are trained on $K - 1$ parts each while being tested on the remaining one part of the data. That is done on all K parts exactly once to ensure that every data row is used equally often for training and exactly once for testing. The resulting K test errors are then averaged to get the final error estimate of the model, which was built on the complete dataset.

The following items are for further study:

- How are the data protected against data loss (e.g., what is the backup strategy)?
- What mechanisms are in place to ensure that proprietary AI models, algorithms and trade secrets of benchmarking participants are fully protected?
- How is it ensured that the correct version of the benchmarking software and the AIs are tested?
- How are automatic updates conducted (e.g., of the operating system)?
- How and where is the benchmarking hosted and who has access to the system and the data (e.g., virtual machines, storage and computing resources, configurational settings)?
- How is the system's stability monitored during benchmarking and how are attacks or issues detected?
- How are issues (e.g., with a certain AI) documented or logged?
- In case of offline benchmarking, how are the submitted AIs protected against leakage of intellectual property?

6.1.1.2.4 Benchmarking process

This clause describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

The following items are for further study:

- How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
- How do possible participants learn about an upcoming benchmarking?
- How can one apply for participation?
- What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine-learning technology used, company size, company location)?
- Are there any contracts or legal documents to be signed?
- Are there inclusion or exclusion criteria to be considered?
- How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
- How can participants test their interface (e.g., is there a test dataset in a case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
- Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
- If there are problems with an AI, how are problems resolved? (E.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)
- How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
- In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

6.1.1.3 AI input data structure for the benchmarking

This clause describes the input data provided to the AI solutions as part of the benchmarking of neurocognitive disorders. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking.

Whole brain images from MRI, PET or CT scans

- Image file format: DICOM or NIFTI format.
- Image file names: Image names will be anonymized to exclude any patient identifying information.
- Image resolution: Images will be supplied in their original resolution as captured from the MRI scanner.

Neuroimaging-derived features

The neuromorphometric processing component (SPM12) uses NIFTI data for computational neuroanatomical data extraction using voxel-based statistical parametric mapping of brain image data sequences:

- Each T1-weighted image is normalized to MNI (Montreal Neurological Institute) space using non-linear image registration SPM12 Shoot toolbox.
- The individual images are segmented into three different brain tissue classes (grey matter, white matter and CSF).
- Each grey matter voxel is labelled based on a neuromorphometrics atlas (constructed by manual segmentation for a group of subjects) and the transformation matrix obtained in the previous step. Maximum probability tissue labels were derived from the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labelling". These data were released under a Creative Commons Attribution-Non-Commercial (CC BY-NC) licence. The MRI scans originate from the OASIS project, and the labelled data was provided by Neuromorphometrics, Inc. under an academic subscription.

Additional information for the medical systems will be provided in text-delimited format:

- Count of vascular lesions;
- History;
- Genetic;
- Memory score;
- Executive functioning scores;
- Comorbidity symptoms;
- Verbal fluency;
- Delayed memory scores;
- Motor scores;
- Psychiatric questionnaires;
- Alcohol use;
- Temperature.

6.1.1.4 AI output data structure

This clause describes the output data the AI systems are expected to generate in response to the input data. It is similar to the input data structure for the benchmarking. It covers the details of the data format, coding and error handling at the level of detail needed for an AI to participate in the benchmarking.

The output of the algorithm should be a CSV file in text format with the following columns:

- ID of the data set processed;
- The algorithm parameters, e.g., variables used e.g., demographic, brains, etc.;
- The diagnosis of cognitive disorders and disease severity:
 - AD;
 - Mild cognitive impairment (MCI);
 - Cognitively normal (CN);
 - Other mixed dementia (MD).

6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous clauses, the benchmarking system needs to know the expected correct answer (sometimes called 'labels') for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following clause.

A separate CSV file in text format will be provided containing the following columns:

- ID of the records;
- Label or annotation of the MRI scans;
- Label and annotation of other biological data.

6.1.1.6 Scores and metrics

Scores and metrics are at the core of the benchmarking. This clause describes the scores and metrics used to measure the performance, robustness and general characteristics of the submitted AI systems for neurodegenerative diseases.

The primary stakeholders for the benchmarking scores and metrics include clinicians, researchers and health care data providers.

Selected scores and metrics: The origins of these scores and metrics are standard in machine learning and clinical research. These metrics were chosen for their relevance to clinical practice, ease of interpretation and ability to compare across AI solutions. The exact definition/formula of the scores and metrics is based on the labels and AI output data structures defined in the previous clauses and aggregated across the entire datasets from different locations.

General criteria for selecting scores and metrics include relevance to clinical practice, ease of interpretation, comparability of AI solutions and ability to correct for bias in the data set. The selected scores and metrics do not explicitly correct for bias in the dataset. However, this can be accounted for by ensuring that the test dataset is representative of the real distribution of the condition.

In addition, it is proposed to integrate clinician feedback by measuring the clinical utility. This measure assesses the impact of the automated decision in term of impact on the clinical path of the patients, impact on the treatment and impact on the relatives.

For robustness

Test accuracy: F1 score

$$F1 = 2*TP/(2TP + FP + FN)$$

Where:

TP is the number of true positives: number of positive cases (e.g., patients with a disease) that were correctly identified as positive by the model.

FP is the number of false positives: number of negative cases (e.g., patients without the disease) that were incorrectly identified as positive by the model.

FN is the number of false negatives: number of positive cases that were incorrectly identified as negative by the model.

TN (true negative) is the number of negative cases that were correctly identified as negative by the model.

For medical performance

Clinical sensitivity (true positive rate or recall): This is calculated as $TP/(TP + FN)$. This is the probability that a patient with a neurodegenerative disease is correctly identified as having the disease by the AI model. It indicates the ability of the model to correctly identify individuals with neurodegenerative disease among all actual positive cases.

Clinical specificity (true negative rate): It is calculated as $TN/(TN + FP)$. This is the probability that a patient without neurodegenerative disease is correctly identified as having no disease by the AI model. It indicates the ability of the model to correctly identify individuals without neurodegenerative disease among all actual negative cases.

Clinical precision (positive predictive value): This is calculated as $TP/(TP + FP)$. This is the probability that a patient identified by the AI model as having a neurodegenerative disease actually has that disease. It indicates the effectiveness of the model in accurately predicting the presence of neurodegenerative disease in patients identified as positive by the model.

The following items are for further study:

- How does this consider the general guidance of WG-DAISAM in [DEL7.3](#) "Data and artificial intelligence assessment methods (DAISAM)"?
- Detailed advantages and disadvantages of each chosen metric.
- Addressing the reproducibility of results across different benchmarking iterations.

6.1.1.7 Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources and storage.

Quality control of the dataset: To assess the quality of the data, the DACORD framework for the design, documentation, and reporting of data-curation methods (Ercole et al., 2020 [8]) was applied. This framework provides a comprehensive approach to ensure data quality and reliability. One of the top drivers of TG-Neuro, F. Kherif, was a member of the DACORD collaborators that created the framework and the indicators, as listed in Table 4. This involvement ensures a deep understanding and rigorous application of the framework in the data quality control process.

Table 4 – DACQORD indicators

Study phase	Dimension	Indicator
Design time	Correctness	1. The case report form (CRF) has been designed by a team with a range of expertise.
	Completeness	2. There is a robust process for choosing and designing the dataset to be collected that involves appropriate stakeholders, including a data-curation team with appropriate skill mix.
	Concordance	3. The data ontology is consistent with published standards (common data elements) to the greatest extent possible.
	Concordance	4. Data types are specified for each variable.
	Correctness	5. Variables are named and encoded in a way that is easy to understand.
	Representation	6. Relational databases have been appropriately normalized: steps have been taken to eliminate redundant data and remove potentially inconsistent or overly complex data dependencies.
	Representation	7. Each individual has a unique identifier.
	Representation	8. There is no duplication in the data set: data has not been entered twice for the same participant.
	Completeness	9. Data that is mandatory for the study is enforced by rules at data entry and user reasons for overriding the error checks (queries) are documented in the database.
	Completeness	10. Missingness is defined and is distinguished from 'not available', 'not applicable', 'not collected' or 'unknown.' For optional data, 'not entered' is differentiated from 'not clinically available' depending on research context.
Design time	Plausibility	11. Range and logic checks are in place for CRF response fields that require free entry of numeric values. Permissible values and units of measurement are specified at data entry.
	Correctness	12. Free text avoided unless clear scientific justification and (e.g., qualitative) analysis plan specified and feasible.
	Concordance	13. Database rule checks are in place to identify conflicts in data entries for related or dependent data collected in different CRFs or sources.
	Representation	14. There are mechanisms in place to enforce or ensure that time-sensitive data is entered within allotted time windows.
	Completeness	15. There is clear documentation of interdependence of CRF fields, including data entry skip logic.
Design time	Correctness	16. Data collection includes fields for documenting that participants meet inclusion/exclusion criteria.
	Representation	17. The data entry tool does not perform rounding or truncation of entries that might result in precision-loss.
	Plausibility	18. Extract/transform/load software for batch upload of data from other sources such as assay results should flag impossible and implausible values.

Table 4 – DACQORD indicators

Study phase	Dimension	Indicator
	Representation	19. Internationalization is undertaken in a robust manner, and translation and cultural adaption of concepts (e.g., assessment tools) follows best practice.
	Concordance	20. Data-collection methods are documented in study manuals that are sufficiently detailed to ensure the same procedures are followed each time.
	Correctness	21. All personnel responsible for entering data receive training and testing on how to complete the CRF.
	Correctness	22. The CRF / eCRF are easy to use and include a detailed description of the data-collection guidelines and how to complete each field in the form. They are pilot tested in a rigorous pre-specified and documented process until reliability and validity are demonstrated.
Design time	Concordance	23. Data collectors are tested and provided with feedback regarding the accuracy of their performance across all relevant study domains.
	Correctness	24. Data collection that requires specific content expertise is carried out by trained and/or certified investigators.
	Correctness	25. Assessors are blinded to treatment allocation or predictor variables where appropriate and such blinding is explicitly recorded.
	Correctness	26. There is a clear audit chain for any data processing that takes place after entry, and this should have a mechanism for version control if it changes.
	Representation	27. Data are provided in a form that is unambiguous to researchers.
	Concordance	28. For physiological data the methods of measurement and units are defined for all sites.
	Correctness	29. Imaging acquisition techniques are standardized (e.g., magnetic resonance imaging).
	Correctness	30. Biospecimen preparation techniques are standardized.
	Correctness	31. Biospecimen assay accuracy, precision, repeatability, detection limits, quantitation limits, linearity and range are defined. Normal ranges are determined for each assay.
	Correctness	32. There is automated entry of the results of biospecimen samples.
Training and testing	Completeness	33. A team of data-curation experts are involved with pre-specified initial and ongoing testing for quality assurance.
Run-time	Completeness	34. Proxy responses for factual questions (such as employment status) are allowed in order to maximize completeness.
	Representation	35. Automated variable transformations are documented and tested before implementation and if modified.
	Completeness	36. There is centralized monitoring of the completeness and consistency of information during data collection.

Table 4 – DACQORD indicators

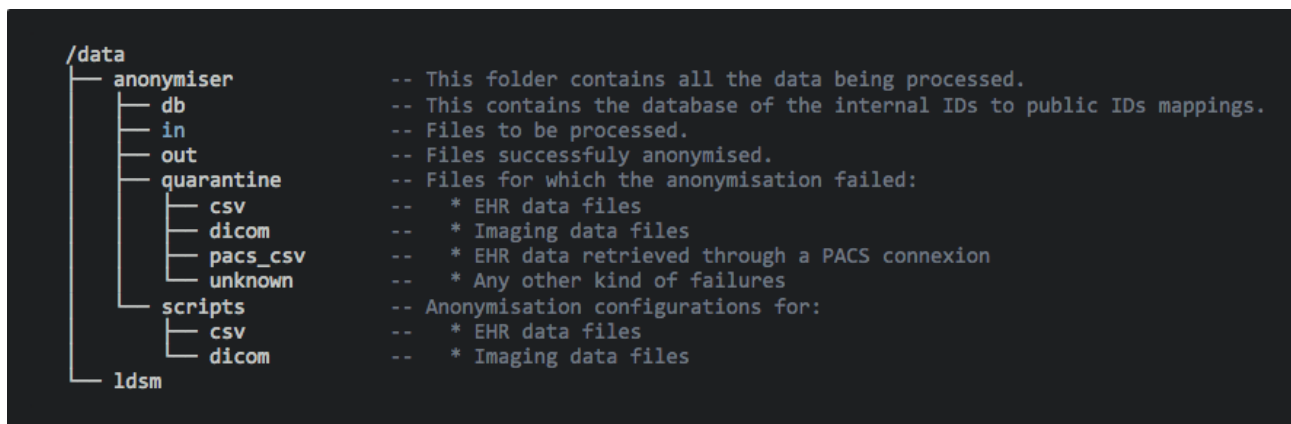
Study phase	Dimension	Indicator
	Plausibility	37. Individual data elements should be checked for missingness. This should be done against pre-specified skip-logic/missingness masks. This should be performed throughout the study data acquisition period to give accurate 'real time' feedback on completion status.
Run-time	Plausibility	38. Systematic and timely measures are in place to assure ongoing data accuracy.
	Correctness	39. Source data validation procedures are in place to check for agreement between the original data and the information recorded in the database.
	Plausibility	40. Reliability checks have been performed on variables that are critical to research hypotheses, to ensure that information from multiple sources is consistent.
	Correctness	41. Scoring of tests is checked. Scoring is performed automatically where possible.
	Correctness	42. Data irregularities are reported back to data collectors in a systematic and timely process. There is a standard operating procedure for data irregularities to be reported back to the data collectors and for documentation of the resolution of the issue.
	Representation	43. Known/emergent issues with the data dictionary are documented and reported in an accessible manner.
Post-collection	Representation	44. The version lock-down of the database for data entry is clearly specified.
	Correctness	45. A plan for ongoing curation and version control is specified.
	Representation	46. A comprehensive data dictionary is available for end users.

Data privacy

Data De-identifier replaces the following personally identifiable information with pseudonyms:

- Information exported from EHR systems in CSV format;
- Information from neuroimages stored in the headers of DICOM files.

Data De-identifier saves the files with de-identified data to storage; see Figure 3.

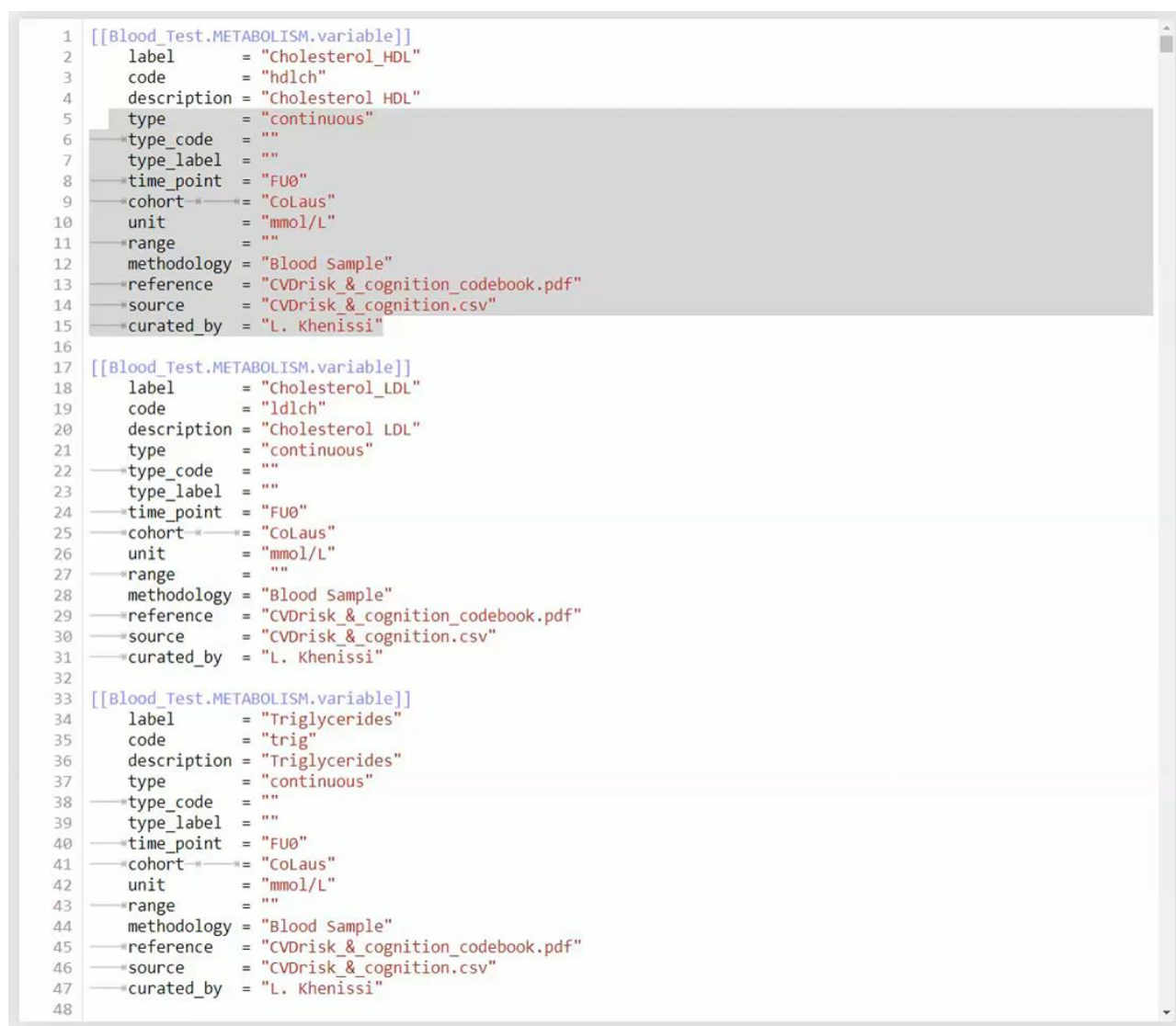
**Figure 3 – Data file storage organization**

Data annotation and metadata

The variables in the datasets were annotated using the Cde common data element principle and metadata description was encoded in the TOML file.

The available data (Appendix I) are described using the concept of a common data element, which was enriched with new hierarchical definition for biological data.

Data catalogue format is a TOML file. Clinicians (neurologists, neuropsychologists, ...) complemented the variable descriptions with attributes according to FDA standards for clinical trial (see the example in Figure 4).



```
1  [[Blood_Test.METABOLISM.variable]]
2      label      = "Cholesterol_HDL"
3      code       = "hdlch"
4      description = "Cholesterol HDL"
5      type       = "continuous"
6      type_code  = ""
7      type_label = ""
8      time_point = "FU0"
9      cohort     = "CoLauS"
10     unit       = "mmol/L"
11     range      = ""
12     methodology = "Blood Sample"
13     reference   = "CVDrisk_&_cognition_codebook.pdf"
14     source      = "CVDrisk_&_cognition.csv"
15     curated_by  = "L. Khenissi"
16
17  [[Blood_Test.METABOLISM.variable]]
18      label      = "Cholesterol_LDL"
19      code       = "ldlch"
20      description = "Cholesterol LDL"
21      type       = "continuous"
22      type_code  = ""
23      type_label = ""
24      time_point = "FU0"
25      cohort     = "CoLauS"
26      unit       = "mmol/L"
27      range      = ""
28      methodology = "Blood Sample"
29      reference   = "CVDrisk_&_cognition_codebook.pdf"
30      source      = "CVDrisk_&_cognition.csv"
31      curated_by  = "L. Khenissi"
32
33  [[Blood_Test.METABOLISM.variable]]
34      label      = "Triglycerides"
35      code       = "trig"
36      description = "Triglycerides"
37      type       = "continuous"
38      type_code  = ""
39      type_label = ""
40      time_point = "FU0"
41      cohort     = "CoLauS"
42      unit       = "mmol/L"
43      range      = ""
44      methodology = "Blood Sample"
45      reference   = "CVDrisk_&_cognition_codebook.pdf"
46      source      = "CVDrisk_&_cognition.csv"
47      curated_by  = "L. Khenissi"
48
```

Figure 4 – Example of data catalogue model

6.1.1.8 Data-sharing policies

This clause provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data-sharing period, patient consent, and update procedure (see also [DEL5.5](#) on *Data handling* and [DEL5.6](#) on *data-sharing practices*). In addition, international standards for data sharing, such as the General Data Protection Regulation (GDPR) and relevant data protection laws, apply as a legal framework for data sharing. A data-sharing contract was signed to set the terms for sharing data for AI research. The content of the contract included the following key elements:

- Purpose and intended use of data;

- Period of agreement;
- Description of data;
- Metadata registry;
- Data harmonization;
- Data update procedure;
- Data-sharing scenarios:
 - Ability of data to be shared in public repositories;
 - Data being stored in local private databases (e.g., hospitals).
- Rules and regulation for patients' consent;
- Data anonymization and de-identification procedure;
- Roles and responsibilities:
 - Data provider;
 - Data protection officer;
 - Data controllers;
 - Data processors;
 - Data receivers.

The following items are for further study:

- Which legal framework was used for sharing the AI?
- Was a contract signed and what was the content?

6.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

The following items are for further study:

- Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
- Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group and ethnicity)?
- Was a baseline assessed as part of the benchmarking?
- How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous clause.
- What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

6.1.1.10 Reporting methodology

This clause discusses how the results of the benchmarking runs will be shared with the participants, stakeholders and general public.

The following items are for further study:

- What is the general approach for reporting results (e.g., Leader board versus drill-down)?
- How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
- How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
- What additional metadata describing the AI models have been selected for reporting?
- How is the relationship between AI results, baselines, previous benchmarking iterations and/or other benchmarking iterations communicated?
- What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
- What is the publication strategy for the results (e.g., website, paper and conferences)?
- Is there an online version of the results?
- Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
- Are there any known limitations to the value, expressiveness or interpretability of the reports?

6.1.1.11 Result

The aim is to provide a machine-learning model to automatically detect dementia, as depicted in Figure 5. The outcome model with the requirement of having reasonable performances in terms of the different losses and metrics defined and must be able to explain its predictions. The approach used a three-dimensional scan of the brain as input, namely the raw T1-weighted magnetic resonance images (MRI) of the patient brain.

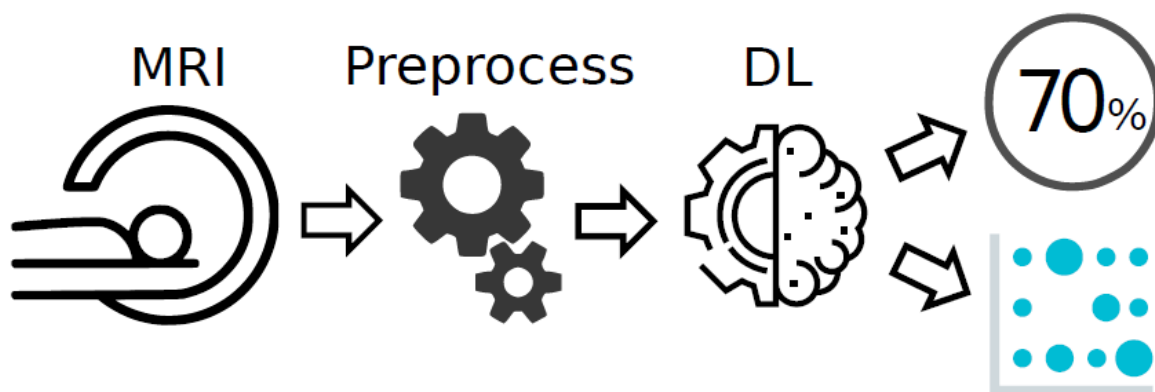


Figure 5 – AI model for automatic detection of dementia cases

To evaluate and compare the performance of the different models, the metrics described previously were used. Using the confusion matrix shown in Figure 6, an accuracy of 81.87% was calculated. Although at first glance this metric suggests good performance, it is important to note that random guessing would yield an accuracy of 70.76%. Therefore, it is more informative to consider the precision, which is 74%, and the recall, which is 67.27%.

Figure 7 shows the curves receiver operating characteristic (ROC) and precision-recall (PR), which provide a more nuanced understanding of the performance of the model. Since the data set is unbalanced, it is advisable to focus on the PR curve rather than the ROC curve. Additionally, Figure 8 shows selected slices in which the hippocampus is visible. While the Shap explanation is difficult to

interpret, both the GradCam and FullGrad methods focus on the hippocampus. This is particularly evident in the output of the FullGrad algorithm, where the area of maximal attention on the selected slice is within the hippocampus.

- Acc: 81.87%
 - Random: 70.76%
- Precision: 74%
- Recall: 67.27%

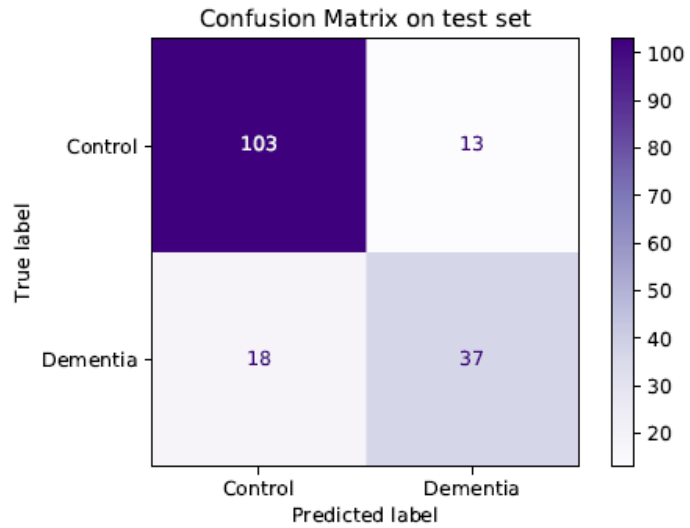
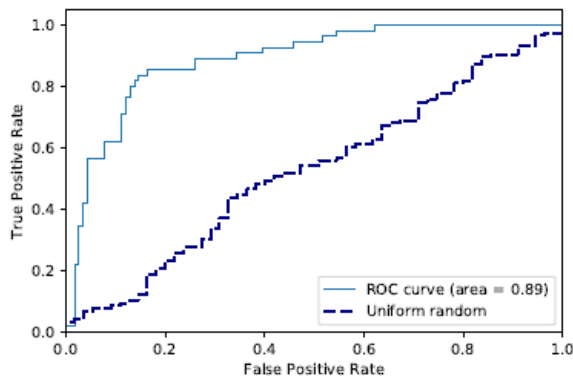
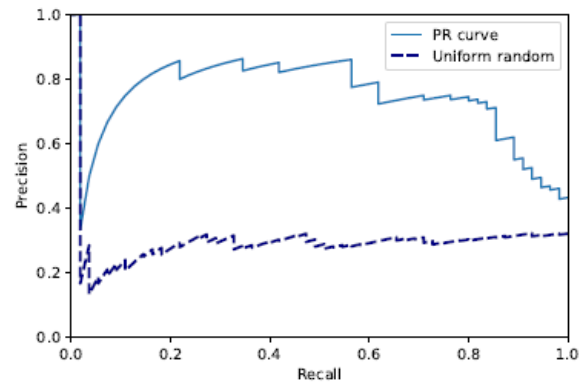


Figure 6 – AI model's result on the test dataset



(a) ROC curve.



(b) Precision and Recall curve.

Figure 7 – Evaluation curves

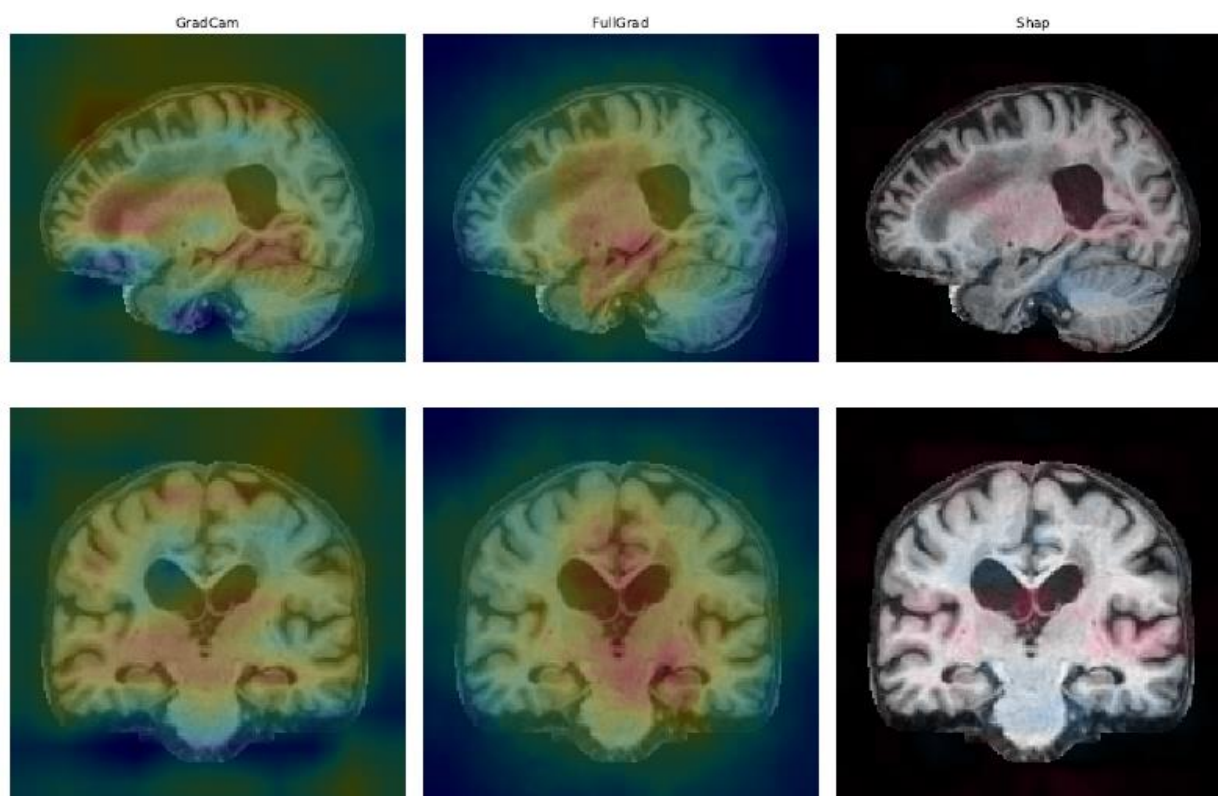


Figure 8 – Outputs of the explainer algorithm on one patient with dementia

6.1.1.12 Discussion of the benchmarking

This clause discusses insights of this benchmarking iterations and provides details about the "outcome" of the benchmarking process (e.g., giving an overview of the benchmark results and process).

Although the results obtained and the insights gained through visualization are promising, this benchmarking iteration in TG-Neuro was conducted with only a limited amount of labelled data due to difficulties accessing a custom dataset. In order to obtain comprehensive results, it is necessary to apply the entire pipeline to a broader dataset with labels provided.

Further, TG-Neuro is interested in evaluating the pipeline in conjunction with other imaging modalities, rather than relying solely on T1 images. For instance, studies have demonstrated a link between brain iron density and dementia. Due to this, T1-weighted images may not fully reveal the cause of dementia, and that only its consequences may be observable at this time.

Currently, the results have been presented to only a few clinicians, and the next important step is to share the results with more clinicians so that they may provide feedback on the pipeline and enable it to be further refined based on their requirements.

6.1.1.13 Retirement

This clause addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this clause [DEL04](#) "*AI software lifecycle specification*" (identification of standards and best practices that are relevant for the AI for health software life cycle).

The following items are for further study:

- What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency or published)?

- What happens to the submitted AI models after the benchmarking?
- Could the results be reproduced?
- Are there legal or compliance requirements to respond to data deletion requests?

7 Overall discussion of the benchmarking

This clause discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in clause 6.1.1.12).

TG-Neuro benchmarking has provided valuable insights and provided a solid foundation for future research. Next steps involve analysing larger data sets, exploring different modalities, and collaborating with clinicians to improve the pipeline and contribute to the understanding and early detection of dementia.

The following items are for further study:

- Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
- Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
 - Did the AI system perform as predicted relative to the baselines?
 - Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust and quality of service)?
- Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?

8 Regulatory considerations

For AI-based technologies in health care, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on "*Regulatory considerations on AI for health*" (WG-RC) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL2](#) "*Overview of Regulatory Concepts on Artificial Intelligence for Health*" (which provides an educational overview of some key regulatory considerations), [DEL2.1](#) "*Mapping of IMDRF essential principles to AI for health software*", and [DEL2.2](#) "*Good practices for health applications of machine learning: Considerations for manufacturers and regulators*" (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). [DEL04](#) identifies standards and best practices that are relevant for the "*AI software lifecycle specification*". The following clauses discuss how the different regulatory aspects relate to the TG-Neuro.

8.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following clause summarizes the most

important aspects that AI manufacturers need to address if they are developing AI systems for neurocognitive disorders.

The following items are for further study:

- What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR and ISO; maybe the systems in this topic group always require at least "MDR class 2b" or maybe they are not considered a medical device)?
- Are there any aspects to this AI system that require additional specific regulatory considerations?

8.2 Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This clause outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

The following items are for further study:

- Which certifications and regulatory framework components of the previous clause should be part of the metadata (e.g., as a table with structured selection of the points described in the previous clause)?

8.3 Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This clause outlines the regulatory requirements for software used for benchmarking in this topic group.

The following items are for further study:

- Which regulatory frameworks apply to the benchmarking system itself?
- Are viable solutions with the necessary certifications already available?
- Could the TG implement such a solution?

8.4 Regulatory approach for the topic group

Building on the outlined regulatory requirements, this clause describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the [DEL2](#) "*Overview of Regulatory Concepts on Artificial Intelligence for Health*".

The following items are for further study:

- Documentation and transparency:
 - How will the development process of the benchmarking be documented in an effective, transparent and traceable way?
- Risk management and lifecycle approach:
 - How will the risk management be implemented?
 - How is a life cycle approach throughout development and deployment of the benchmarking system structured?

- Data quality:
 - How is the test data quality ensured (e.g., could the process of harmonizing data of different sources, standards and formats into a single dataset cause bias, missing values, outliers and errors)?
 - How are the corresponding processes document?
- Intended use and analytical and clinical validation:
 - How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
- Data protection and information privacy:
 - How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis and destruction)? This is especially relevant if real patient data is used for the benchmarking.
- Engagement and collaboration:
 - How is stakeholder (regulators, developers, health care policymakers) feedback on the benchmarking collected, documented and implemented?

References

- [1] Bruun, Marie, Kristian S. Frederiksen, Hanneke F. M. Rhodius-Meester, Marta Baroni, Le Gjerum, Juha Koikkalainen, Timo Urhema, et al. 2019. 'Impact of a Clinical Decision Support Tool on Dementia Diagnostics in Memory Clinics: The PredictND Validation Study'. *Current Alzheimer Research* 16 (2): 91–101. <https://doi.org/10.2174/1567205016666190103152425>
- [2] Rao, Anitha, Marguerite Manteau-Rao, Heather Petkunas, Elizabeth Deck, and Neelum T. Aggarwal. 2017a. '[O5–02–02]: Using Neuroscience-Based Technology to Generate Clinical Decision Support and Customized Care Plans for G0505'. *Alzheimer's & Dementia* 13 (7S_Part_30). <https://doi.org/10.1016/j.jalz.2017.07.513>
- [3] ———. 2017b. '[P3–485]: using an Alzheimer's Disease and Related Dementia Clinical Decision Support and Care Planning Software System in an Outpatient Nurse Practitioner Clinic'. *Alzheimer's & Dementia* 13 (7S_Part_24). <https://doi.org/10.1016/j.jalz.2017.06.1704>
- [4] Rao, Anitha, Benjamin Miller, Tanzila Kulman, Paulo Pinho, and Neelum T Aggarwal. 2020. 'Novel Application of Digital Dementia Phenotyping and Risk Classification for Insurance and Longevity Risk Modeling: Health Services Research / Cost of Care'. *Alzheimer's & Dementia* 16 (S10). <https://doi.org/10.1002/alz.044372>
- [5] Mitchell, Susan L., Michele L. Shaffer, Simon Cohen, Laura C. Hanson, Daniel Habtemariam, and Angelo E. Volandes. 2018. 'An Advance Care Planning Video Decision Support Tool for Nursing Home Residents With Advanced Dementia: A Cluster Randomized Clinical Trial'. *JAMA Internal Medicine* 178 (7): 961–69. <https://doi.org/10.1001/jamainternmed.2018.1506>
- [6] Tolonen, Antti, Hanneke F. M. Rhodius-Meester, Marie Bruun, Juha Koikkalainen, Frederik Barkhof, Afina W. Lemstra, Teddy Koene, et al. 2018. 'Data-Driven Differential Diagnosis of Dementia Using Multiclass Disease State Index Classifier'. *Frontiers in Aging Neuroscience* 10 (April): 111. <https://doi.org/10.3389/fnagi.2018.00111>
- [7] Vashistha, Rajat, Dinesh Yadav, Deepak Chhabra, and Pratyosh Shukla. 2019. 'Artificial Intelligence Integration for Neurodegenerative Disorders'. In *Leveraging Biomedical and Healthcare Data*, 77–89. Elsevier. <https://doi.org/10.1016/B978-0-12-809556-0.00005-8>
- [8] Ercole, Ari, Vibeke Brinck, Pradeep George, Ramona Hicks, Jilske Huijben, Michael Jarrett, Mary Vassar, Lindsay Wilson and DAQCOR collaborators. 2020. 'Guidelines for Data Acquisition, Quality and Curation for Observational Research Designs (DAQCOR)'. *Journal of Clinical and Translational Science* 4 (4): 354–59. <https://doi.org/10.1017/cts.2020.24>

Annex A

Glossary

This clause lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AD	Alzheimer's Disease	A progressive neurologic disorder that causes the brain to shrink and brain cells to die.
AI	Artificial Intelligence	
AI4H	Artificial Intelligence For Health	
AI-MD	AI-based Medical Device	
CfTGP	Call for Topic Group Participation	
CN	Cognitively Normal	A term used to describe an individual who is not experiencing noticeable memory problems or other cognitive impairments.
DAQCORD	Data Acquisition, Quality and Curation for Observational Research Designs	
DEL	Deliverable	
DICOM	Digital IMAGING and Communications in Medicine	A standard for the communication and management of medical imaging information and related data.
DPO	Data Protection Officer	An enterprise security leadership role required by the General Data Protection Regulation (GDPR).
EHR	Electronic Health Records	Digital version of a patient's paper chart.
EMIF	European Medical Information Framework	A platform that allows the secure discovery, access and analysis of harmonized and federated data from diverse real-world and clinical trial sources.
FDA	Food and Drug Administration	
FGAI4H	Focus Group on AI for Health	
fMRI	functional MRI	An MRI that measures brain activity based on associated blood flow changes.
FN	False Negative	A result that incorrectly indicates the absence of a condition.
FP	False Positive	A result that incorrectly indicates the presence of a condition.
GAAIN	Global Alzheimer's Association Interactive Network	A platform that provides access to a vast collection of Alzheimer's disease research data, tools and resources.
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	

Acronym/Term	Expansion	Comment
ISO	International Organization for Standardization	
ITU	International Telecommunication Union	
LORIS	Longitudinal Online Research and Imaging System	A web-based data and project management software for neuroimaging research studies.
MCI	Mild Cognitive Impairment	A stage between the expected cognitive decline of normal ageing and the more severe decline of dementia.
MD	Mixed Dementia	A condition in which abnormalities characteristic of more than one type of dementia occur simultaneously.;
MDR	Medical Device Regulation	
MRI	Magnetic Resonance Imaging	A medical imaging technique used to visualize internal structures of the body in detail.
PET	Positron Emission Tomography	A functional imaging technique that helps to show how your tissues and organs are functioning.
PIA	Privacy Impact Assessment	A tool for identifying and assessing privacy risks throughout the development life cycle.
SaMD	Software as a Medical Device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H topic group works. This document is the TDD for the TG-Neuro.
TG	Topic Group	
TN	True Negative	A result that accurately indicates the absence of a condition.
TP	True Positive	A result that accurately indicates the presence of a condition.
WG	Working Group	
WHO	World Health Organization	

Annex B

Declaration of conflict of interests

None provided.
