INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FG-AI4H-P-016-A01

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plenary	Helsinki, 20-22 September 2022

DOCUMENT

Source:	TG-Neuro Topic Driver
Title:	Att.1 – TDD update (TG-Neuro) [same as Meeting L]
Purpose:	Discussion

Contact:	Ferath Kherif CHUV Switzerland	E-Mail: ferath.kherif@chuv.ch
Contact:	Marc Lecoultre ML Labs, Switzerland	E-Mail: ml@mllab.ai

Abstract: Calling on members of the medical and artificial intelligence communities with a vested interest in AI against neuro-cognitive diseases! Become engaged in the group dedicated to establishing a standardized benchmarking platform for AI against neuro-cognitive diseases within the International Telecommunication Union (ITU)/World Health Organization (WHO) Focus Group on "Artificial Intelligence for Health" (FG-AI4H).

This version of TDD is the same as seen in Meeting L, reproduced for easier reference as a Meeting N document.

Change notes: *Topic Driver: Please list the changes of the current TDD version in comparison to earlier versions. This can include content updates in specific sections, additional or completed sections, updates on subtopics, etc.*

Version 7 (submitted as FGAI4H-X-# to meeting K in location E-meeting)

- Updated to new template
- Updated xyz

..

Contributors

Name Some Company/Institute Country	Tel: +4x xx xxxxxxxxx Email: some.name@somecompany.com
Name Some Company/Institute Country	Tel: +4x xx xxxxxxxxx Email: some.name@somecompany.com

CONTENTS

Page

Page

List of Tables

Page

List of Figures

Page

FG-AI4H Topic Description Document

Topic group-[TG-GROUP NAME]

1 Introduction

Topic Driver: Add a short (half page) introduction to the topic. The introduction should provide a general overview of the addressed health topic and basic information about the AI task, including the input data and the output of the AI. The objective and expected impact of the benchmarking should also be described. More detailed information about the topic will appear in section 1.3.

This topic description document specifies the standardized benchmarking for TG-Neuro systems. It serves as deliverable No. [YOUR DEL ID] of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

This topic group is dedicated to AI against neuro-cognitive diseases. We provide 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 become demented is growing proportionally. Current estimates count world-wide 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 Organisation to classify neurocognitive disorders as a global public health priority.

Compared to visual assessment, automated diagnostic methods based on brain imaging are more reproducible and have demonstrated a high accuracy in separating AD from healthy aging, but also 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.

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's 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. Thus, assessment of clinical validity involves measurement of the following metrics derived from the confusion matrix:

- Test accuracy: F1 score

- Clinical sensitivity: ability to identify those who have or will get the disease = TP/(TP+FN)

- Clinical specificity ability to identify those who do not have or will not get the disease
  =TN/(FP+FN)

- Clinical precision the probability that the disease is present when the test is positive
  = sensitivity x prevalence / (sensitivity x prevalence + (1-specificity) x (1-sensitivity))

In addition, we propose 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 …).

The primary data are already available and growing in volume. Data will include both real world patient's 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 patients' cohort) different stages of the disease (subjective complains, mild impairments or demented)

2 About the FG-AI4H topic group on Neuro-Cognitive disorders

The introduction highlights the potential of a standardized benchmarking of AI systems for neuro-cognitve disorders to help solving important health issues and 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 the meeting [MEETING NO.] in [LOCATION AND DATE OF YOUR FOUNDING MEETING].

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 FG-AI4H meeting [MEETING NO.] [LOCATION AND DATE OF YOUR FOUNDING MEETING], [TOPIC DRIVER] from [TOPIC DRIVER INSTITUTION] was nominated as topic driver for the TG-NEURO.

Current members of the topic group on AI against neuro-cognitive diseases include:

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 Departement 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 Lecoultre, 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 founded several companies in these fields. He practiced 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 to 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

Topic Driver: As the structure of the TDD document is the same for all topic groups, you only need to fill in the green placeholders [].

This document is the TDD for the TG-[YOUR TOPIC]. 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 [YOUR TOPIC]. It describes the existing approaches for assessing the quality of Neuro-Cognitive disorders 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 subsections 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 "DEL 10.[YOUR DEL ID] Neuro-Cognitive disorders (TG-NEURO)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic group output documents

Number	Title
FGAI4H-C-020-R1	FGAI4H-C-020-R1: Status report for Alzheimer's disease use case
FGAI4H-B-013-R1: Proposal	FGAI4H-B-013-R1: Proposal: Using machine learning and AI for validation of Alzheimer's disease biomarkers for use in the clinical practice

The working version of this document can be found in the official topic group SharePoint directory.

- [INSERT THE **LINK** TO YOUR **TOPIC GROUP SHAREPOINT FOLDER HERE** AND UPLOAD THE TDD WORKING VERSION TO THE SHARE POINT]

Select the following link:

- [INSERT THE **LINK** TO THE **TDD WORKING VERSION HERE**]

2.2 Status of this topic group

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

The following is an update of activities since meeting D:

- ✓ The updated Call for Topic Group participation for TG-Cogni was published on the ITU website and can be downloaded here.
- ✓ We had several email exchanges with the topic group members to request inputs and updates to the TDD.

- ✓ We reached out to our networks via email and social media (LinkedIn, Twitter), sharing the call for topic group participation and to spread the word.
- ✓ We have had preliminary interest from several groups and individuals interested in contributing to the topic group and are following up with them individually.

The following is an update of activities since meeting E:

- ✓ We received a new submission regarding Standardization of MRI Brain Imaging for Parkinson Disease by Biran Haacke, Prof. Mark Haacke, Mark Messow from The MRI Institute for BMR in Canada.
- ✓ We added 300 patients' datasets to the Alzheimer's data that will be available for AI solutions. We included new quantitative and semi-quantitative methods for assessing image quality.
- ✓ We held several discussions with clinical research groups and hospitals that will be interested to join the Neuro-cognitive disease. The discussion is ongoing and still, at a preliminary stage, we think that we will be able to integrate new groups from Italy and Bulgaria.
- ✓ We are onboarding Prof. Alexander Tsiskaridze (neurologist) from Ivane Javakhishvili Tbilisi State University | TSU · Faculty of Medicine in Georgia. He might be providing data, new topics and AI solutions.
- ✓ We had two meetings with the Norwegian Ministry of Health and Care Services to include stakeholders from northern Europe in the FG.
- ✓ We had a discussion with EU official on the topic of defining cloud/compute infrastructure needs for health research. A meeting/workshop is planned for October, final date TBD. Ferath Kherif will be presenting the neurocognitive disease group.

2.2.1 Status update for meeting [MEETING LETTER]

Topic Driver: Please insert a one-page summary of the work since the last focus group meeting. This can include:

- Work on this document
- Work on the benchmarking software
- Progress with data acquisition, annotation, etc.
- Overview of the online meetings including links to meeting minutes
- Relevant insights from interactions with other working groups or topic groups
- Partners joining the topic group
- List of current partners
- Relevant next steps

2.2.2 Status update for meeting [MEETING LETTER]

[…]

2.3 Topic group participation

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

- [INSERT THE LINK TO YOUR 'CALL FOR TG PARTICIPATION' (CfTGP)]

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

- [INSERT THE LINK TO YOUR TOPIC GROUP SUBPAGE]

Topic Driver: Please set up a regular (e.g., bi-weekly) online meeting with rotating and considerate time windows (to account for participants in different time zones) and inform the ITU secretariat to schedule the meeting in the FG-AI4H calendar.

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—like 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

In addition to the general FG-AI4H mailing list, each topic group can create an *individual mailing list:*

Topic Driver: Please contact the ITU secretariat tsbfgai4h@itu.int to create a mailing list for your TG.

Delete this passage if you are starting without a specific mailing list for your TG.

- [INSERT YOUR TOPIC GROUP MAILING LIST HERE]

Regular FG-AI4H workshops and meetings proceed 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 section contains a detailed description and background information of the specific health topic for the benchmarking of AI in Neuro-Cognitive 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. The TG-NEURO currently has no subtopics. Future subtopics for [SUBTOPIC NAME] might be introduced.

This topic group is dedicated to AI against neuro-cognitive diseases. We provide an empirical basis for testing the clinical validity of machine learning-based diagnostics for neurodegerative disease (Alzheimer's disease or Parkinson 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 Dementia

3.1.1 Definition of the AI task

This section 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 chapter 4). This section corresponds to DEL03 *"AI requirements 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 become demented is growing proportionally. Current estimates count world-wide 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 Organisation 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 diagnosis to provide an output that is readily usable and interpretable by clinicians. Beyond this immediate impact on clinical decision making and patients care, our 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 section provides a description of the established gold standard of the addressed health topic.

Compared to visual assessment, automated diagnostic methods based on brain imaging are more reproducible and have demonstrated a high accuracy in separating AD from healthy aging, but also 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.

Our 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 section 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 Big Data and validate their accuracy as biomarkers. It will not only advance our 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 section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain 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 chapter 6.

We have successfully used supervised classification methods for predicting clinical outcome and analyzing variance in the data. 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 uni-modal data, motivating the development of a methodological approach to accommodate multiple-modal data. Recently, we have applied this methodology in order to develop predictive models for healthy aging and found that the mean prediction error was significantly reduced when all measurements were combined. The table below provides summaries of other AI solutions.

Reference	Supporting System	Domain	Features	Methodology	Target Users
[Bruun2019]	Clinical Decision Support System, PredictND tool	Dementia: Vascular, Frontotemporal, Alzheimer, Subjective cognitive decline.	– Clinical test – MRI visual – Data Analytics	Objective comparison of data	Clinicians, neurologist
[Anitha 2017]	CDS-CPL: Clinical Decision Support and Care Planning Tool	Alzheimer's Disease and Related Dementia: ADRD	– Online questionnaire – Evidence-based recommendations – physical exam techniques – referrals medications	differential diagnosis, individualized care plans	Caregivers, NPs, and PAs
[Mitchell 2018]	An advance care planning Video Decision Support tool	Promote goal-directed care for advanced dementia patient	– Medical Records – Bedford Alzheimer Nursing Severity-Subscale	Providing care after viewing the video	Nursing Home Residents
[Tolonen 2018]	Clinical Decision Support System, PredictND tool	Designed for differential diagnosis of different types of dementia	– multiple diagnostic tests such as neuropsychological tests, MRI and cerebrospinal fluid samples	multiclass Disease State Index classifier, visualization of its decision making	Support Physician
[Vashistha 2019]	AI-based clinical decision systems (CDSs) along with POC diagnosis	Neurodegenerative disorders such as Parkinson's disease,	– Machine learning and wearables based Therapeutics	Markov decision processes (MDP) and dynamic	Neurodegenerative disorders Specialist

[redacted]		amyo-trophic lateral sclerosis (ALS), Alzheimer's disease, epilepsy	– A combinatorial intelligent system for the prediction of PD development by ML	decision net-works	

3.2 Subtopic [B]

Topic driver: If you have subtopics in your topic group, describe how the existing AI solutions in the second subtopic [B] deviate from the description in the previous section. Please use the same subsection structure as above for the first subtopic [A]. If there are no subtopics in your topic group, you can remove the "Subtopic" outline level, but - of course - you need to keep the subsections! In this case, please adapt the lower outline levels accordingly (section numbering).

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 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section 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?
- 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)?
- 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)?
- What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
- 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)?
- How is 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)?
- What are your experiences and learnings from addressing ethics in your TG?

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and Neuro-Cognitive 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 learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

5.1 Subtopic Dementia

Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the first subtopic [A] in this section. If there are no sub-topics, you can remove the "Subtopic" outline level, but - of course - you need to keep the subsections below!

5.1.1 Publications on benchmarking systems

While a representative comparable benchmarking for Neuro-Cognitive disorders does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *"AI for health evaluation considerations,"* DEL07_1 *"AI4H evaluation process description,"* DEL07_2 *"AI technical test specification,"* DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL07_4 *"Clinical Evaluation of AI for health"*.

- What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
- State 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 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 Neuro-Cognitive disorders implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

- What are the most relevant learnings from the benchmarking by AI developers in this field (e.g., ask the members of your topic group what they want to share on their benchmarking experiences)?
- Which scores and metrics have been used?
- How did they approach the acquisition of test data?

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 section 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 DEL07_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).

- Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
- Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
- Are there important features in this topic group that require special attention?
- Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
- What are the relative advantages and disadvantages of these diverse solutions?

5.2 Subtopic [B]

Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the second subtopic [B] in this section using the same subsection structure as above. (If there are no sub-topics, you can remove the "Subtopic" outline level.)

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the Neuro-Cognitive disorders AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL05 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL05_1 *"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL05_2 *"Data acquisition"*, DEL05_3 *"Data annotation specification"*, DEL05_4 *"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), DEL05_5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL05_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), DEL07 *"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL07_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), DEL07_2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL07_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), DEL07_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), DEL07_5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL09 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL09_1 *"Mobile based AI applications,"* and DEL09_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 [A]

Topic driver: Please refer to the above comments concerning subtopics.

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

- Which benchmarking iterations have been implemented thus far?
- What important new features are introduced with each iteration?
- What are the next planned iterations and which features are they going to add?

6.1.1 Benchmarking version [Y]

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's 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.

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [Y].

- What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
- What features have been added to the benchmarking in this iteration?

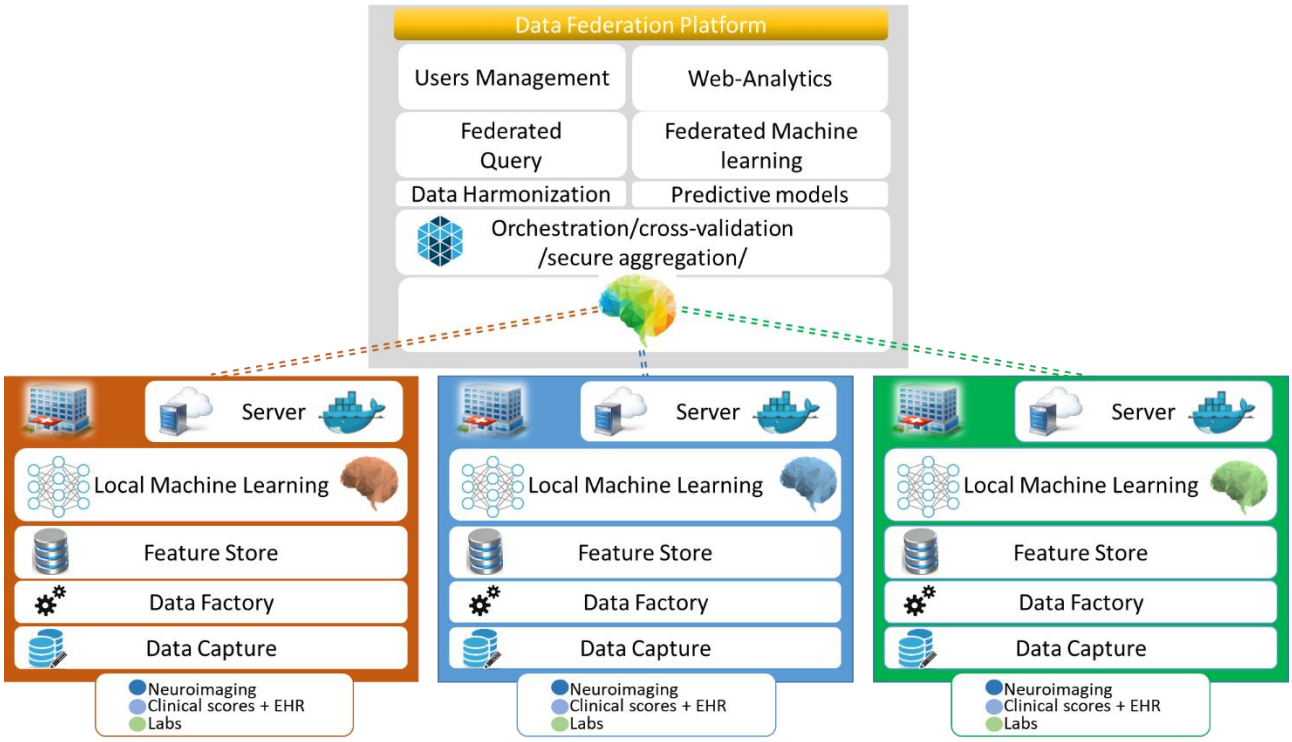
6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. 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 section 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.

- How does the architecture look?
- What are the most relevant components and what are they doing?
- How do the components interact on a high level?
- What underlying technologies and frameworks have been used?
- How does the hosted AI model get the required environment to execute correctly? What is the technology used (e.g., Docker/Kubernetes)?



6.1.1.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

- How do benchmarking data access the system?
- Where and how (data format) are the data, the responses, and reports of the system stored?
- How are the inputs and the expected outputs separated?
- How are the data sent to the AI systems?
- Are the data entries versioned?
- How does the lifecycle for the data look?

6.1.1.2.3 Safe and secure system operation and hosting

From a technical point of view, the benchmarking process is not particularly complex. It is more about agreeing on something in the topic group with potentially many competitors and implementing the benchmarking in a way that cannot be compromised. This section describes how the benchmarking system, the benchmarking data, the results, and the reports are protected against manipulation, data leakage, or data loss. Topic groups that use ready-made software might be able to refer to the corresponding materials of the manufacturers of the benchmarking system.

This section 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 (vs. self-generated data), it is possible to refer to the manufacturer's prescriptions.

- Based on the architecture, where is the benchmarking vulnerable to risk and how have these risks been mitigated (e.g., did you use a threat modelling approach)? A discussion could include:
  - Could someone access the benchmarking data before the actual benchmarking process to gain an advantage?

    - What safety control measures were taken to manage risks to the operating environment?
    - Could someone have changed the AI results stored in the database (your own and/or that of competitors)?
    - Could someone attack the connection between the benchmarking and the AI (e.g., to make the benchmarking result look worse)?
    - How is the hosting system itself protected against attacks?
- 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 section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

- 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 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 section describes the input data provided to the AI solutions as part of the benchmarking of [YOUR TOPIC]. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

The following input data structure is being proposed for all eye conditions - DR, AMD, GC.

Whole Brain images from MRI, PET or CT scans.

- Image File Format: DICOM or NIFTI format
- Image File Names: Images names will be anonymised to exclude any patient identifying information.
- Image Resolution: the 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 neuro-anatomical data extraction using voxel-based statistical parametric mapping of brain image data sequences:

– Each T1-weighted image is normalised 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 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 the Creative Commons Attribution-Non-Commercial (CC BY-NC. 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 txt delimited format :

– Count Vascular lesion

– History

– Genetic

– Memory Score

– Executive functioning scores

– Co-morbidity symptoms
– Verbal fluency
– Delayed memory scores
– Motor scores
– Psychiatric questionnaires
– Alcohol Use
– Temperature

6.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. 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 an disease severity:
  o Alzheimer's Disease
  o Mild cognitive impairment (MCI)
  o Cognitively normal (CN)
  o Other Mixed Dementia (MD)

6.1.1.5 Test data label/annotation structure

Topic driver: Please describe how the expected AI outputs are encoded in the benchmarking test data. Please note that it is essential that the AIs never access the expected outputs to prevent cheating. The topic group should carefully discuss whether more detailed labelling is needed. Depending on the topic, it might make sense to separate between the best possible output of the AI given the input data and the correct disease (that might be known but cannot be derived from the input data alone). Sometimes it is also helpful to encode acceptable other results or results that can be clearly ruled out given the evidence. This provides a much more detailed benchmarking with more fine-grained metrics and expressive reports than the often too simplistic leader boards of many AI competitions.

While the AI systems can only receive the input data described in the previous sections, 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 section.

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

*Topic drivers: This section describes the scores and metrics that are used for benchmarking. It includes details about the testing of the AI model and its effectiveness, performance, transparency, etc. Please note that this is only the description of the scores and metrics actually used in **this** benchmarking iteration. A general description of the state of the art of scores and metrics and how they have been used in previous work is provided in section 3.*

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

All metrics will be computed based on the performance of the algorithm on the undisclosed test data-set. Thus, assessment of clinical validity involves measurement of the following metrics derived from the confusion matrix:

– Test accuracy: F1 score

– Clinical sensitivity: ability to identify those who have or will get the disease = TP/(TP+FN)

– Clinical specificity ability to identify those who do not have or will not get the disease =TN/(FP+FN)

Clinical precision the probability that the disease is present when the test is positive = sensitivity x prevalence / (sensitivity x prevalence + (1-specificity) x (1-sensitivity)

In addition, we propose 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 …).

- Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
- What general criteria have been applied for selecting scores and metrics?
- What scores and metrics have been chosen/defined for robustness?
- What scores and metrics have been chosen/defined for medical performance?
- What scores and metrics have been chosen/defined for non-medical performance?
    - Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
- What scores and metrics have been chosen/defined for model explainability?
- Describe for each aspect
    - The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
    - Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a

real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)

- What are the origins of these scores and metrics?
- Why were they chosen?
- What are the known advantages and disadvantages?
- How easily can the results be compared between or among AI solutions?
- Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?

- How does this consider the general guidance of WG-DAISAM in DEL07_3 "Data and artificial intelligence assessment methods (DAISAM)"?
- Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

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.

Applied the DACQORD framework for the design, documentation and reporting of data curation methods.

"The Data Acquisition, Quality and Curation for Observational Research Designs (DAQCORD) Guidelines were developed for investigators conducting large observational research studies to aid the design, documentation and reporting of practices for assuring data quality within their studies. This information is intended to provide guidance and a transparent reporting framework for improving data quality and data sharing"

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 normalised: 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 / 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.
	Representation	19. Internationalisation 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.

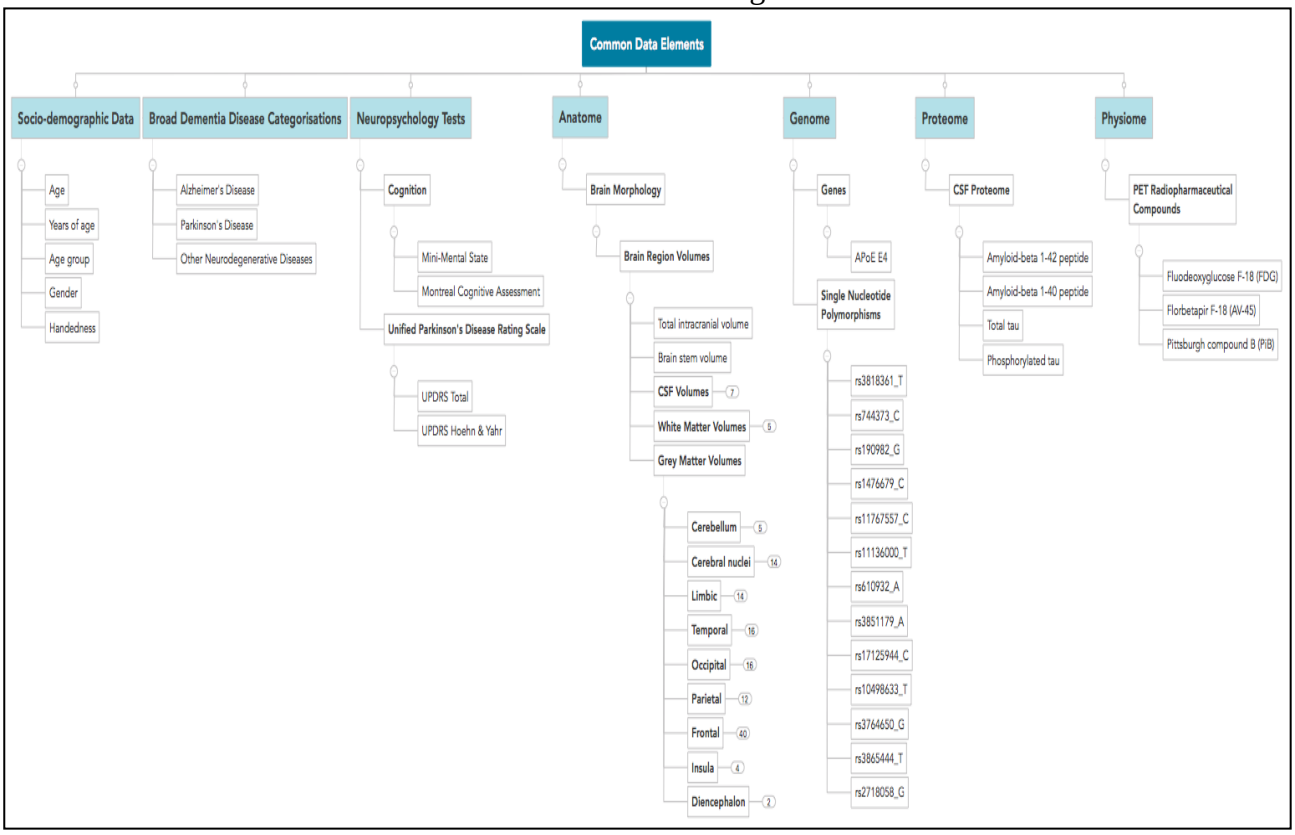
	Correctness	21. All personnel responsible for entering data receive training and testing on how to complete the CRF.
	Correctness	22. The CRF / eCRF are easy to use and include a detailed description of the data collection guidelines and how to complete each field in the form. They are pilot tested in 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 standardised (e.g. magnetic resonance imaging).
	Correctness	30. Biospecimen preparation techniques are standardised.
	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.
	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.

- How does the overall dataset acquisition and annotation process look?
- How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
- Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
- How was the dataset documented and which metadata were collected?

  - Where were the data acquired?
  - Were they collected in an ethical-conform way?
  - Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
  - Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
  - What kind of data anonymization or deidentification has been applied?
  - Are the data self-contained (i.e., independent from externally linked datasets)?
  - How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
  - What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
- Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
- Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
- How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
- Specific data governance derived by the general data governance document (currently F-103 and the deliverables beginning with DEL05)
- How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
- Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
- How was the annotation organized?
  - How many annotators/peer reviewers were engaged?
  - Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
  - How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
  - If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
  - How was the annotation of each case documented?
  - Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
- Were data/label update/amendment policies and/or criteria in place?

- How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
- How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
- Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
- What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)
- The available data (Appendix A) are described sung the concept of Common Data Element, that we enriched with new hierarchical definition for biological data.



- Data catalogue format is a TOML file. Clinicians (Neurologist, neuropsychologists, …) complemented the Variable descriptions with attributes according to FDA standards for clinical trial (see example below).

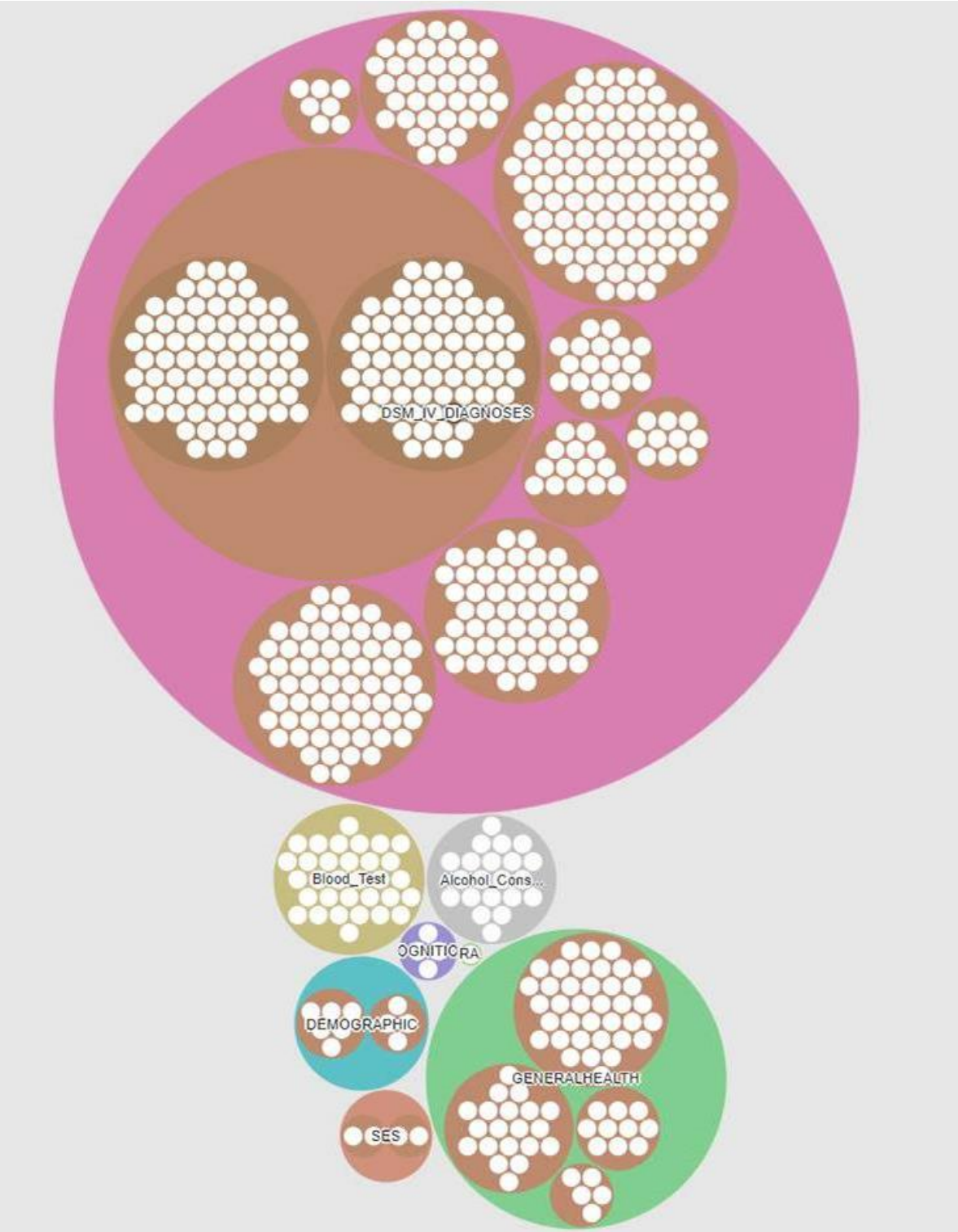
```
[[Blood_Test.METABOLISM.variable]]
    label       = "Cholesterol_HDL"
    code        = "hdlch"
    description = "Cholesterol HDL"
    type        = "continuous"
    type_code   = ""
    type_label  = ""
    time_point  = "FU0"
    cohort      = "CoLaus"
    unit        = "mmol/L"
    range       = ""
    methodology = "Blood Sample"
    reference   = "CVDrisk_&_cognition_codebook.pdf"
    source      = "CVDrisk_&_cognition.csv"
    curated_by  = "L. Khenissi"

[[Blood_Test.METABOLISM.variable]]
    label       = "Cholesterol_LDL"
    code        = "ldlch"
    description = "Cholesterol LDL"
    type        = "continuous"
    type_code   = ""
    type_label  = ""
    time_point  = "FU0"
    cohort      = "CoLaus"
    unit        = "mmol/L"
    range       =  ""
    methodology = "Blood Sample"
    reference   = "CVDrisk_&_cognition_codebook.pdf"
    source      = "CVDrisk_&_cognition.csv"
    curated_by  = "L. Khenissi"

[[Blood_Test.METABOLISM.variable]]
    label       = "Triglycerides"
    code        = "trig"
    description = "Triglycerides"
    type        = "continuous"
    type_code   = ""
    type_label  = ""
    time_point  = "FU0"
    cohort      = "CoLaus"
    unit        = "mmol/L"
    range       = ""
    methodology = "Blood Sample"
    reference   = "CVDrisk_&_cognition_codebook.pdf"
    source      = "CVDrisk_&_cognition.csv"
    curated_by  = "L. Khenissi"

```

-
-



6.1.1.8 Data sharing policies

This section 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 DEL05_5 on *data handling* and DEL05_6 on *data sharing practices*).

- Which legal framework was used for data sharing?
- Was a data sharing contract signed and what was the content? Did it contain:

    - Purpose and intended use of data
    - Period of agreement
    - Description of data
    - Metadata registry
    - Data harmonization
    - Data update procedure
    - Data sharing scenarios
        - Data can be shared in public repositories
        - Data are 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
- 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.

- 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 section.
- 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

After the benchmarking, the next step is to describe how the results are compiled into reports that allow stakeholders to make decisions (e.g., which AI systems can be used to solve a pre-diagnosis task in an offline –field –clinic scenario in central America). For some topic groups, the report might be as simple as a classical AI competition leader board using the most relevant performance indicator. For other tasks, it could be an interactive user interface that allows stakeholders to compare the performance of the different AI systems in a designated context with existing non-AI options. For the latter, statistical issues must be carefully considered (e.g., the multiple comparisons problem). Sometimes, a hybrid of prepared reports on common aspects are generated in addition to interactive options. There is also the question of how and where the results are published and to what degree benchmarking participants can opt in or opt out of the publication of their performance.

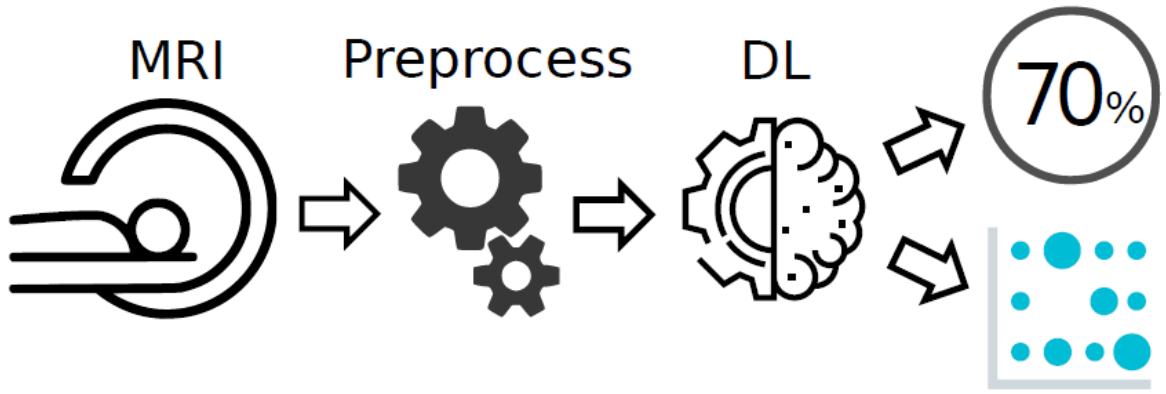
This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

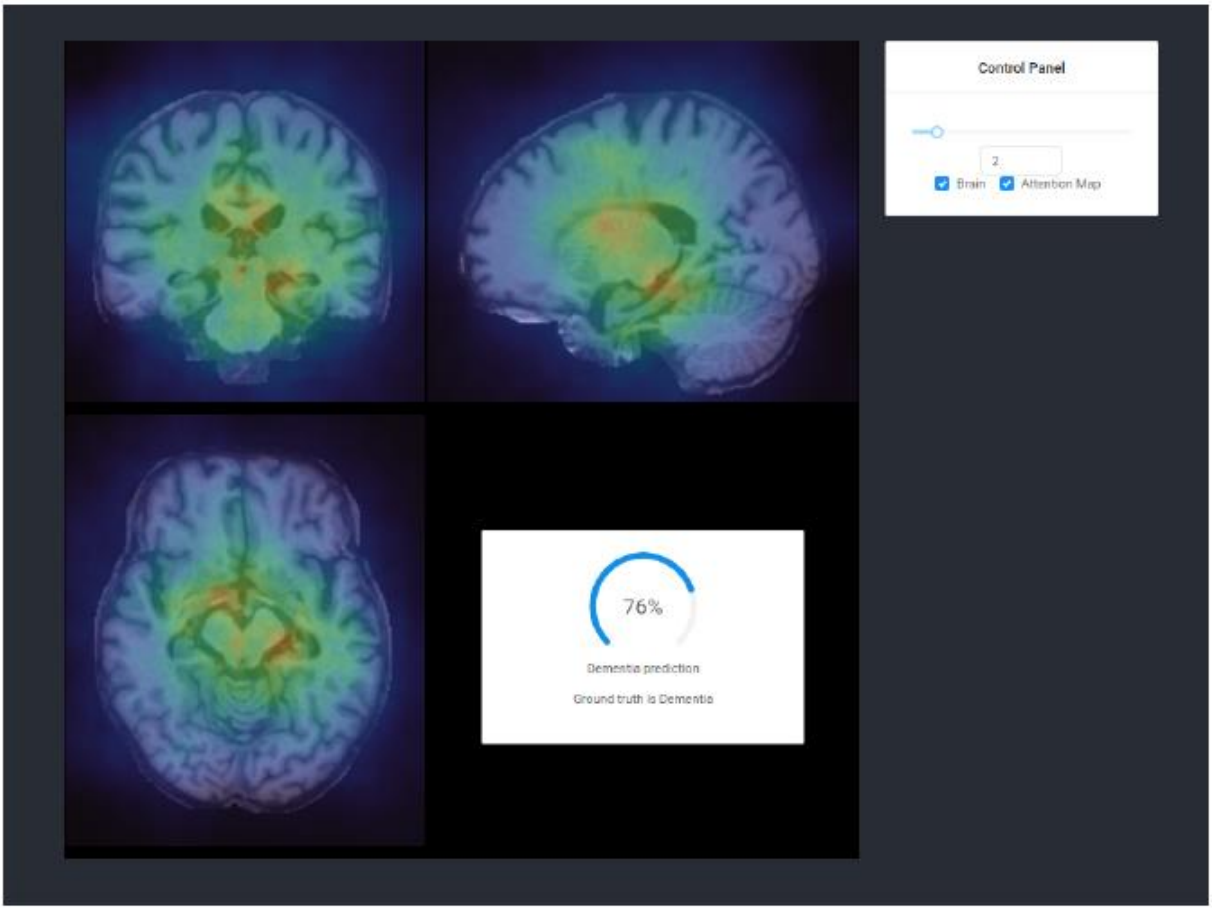
- What is the general approach for reporting results (e.g., leader board vs. drill down)?
- How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
- How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
- 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

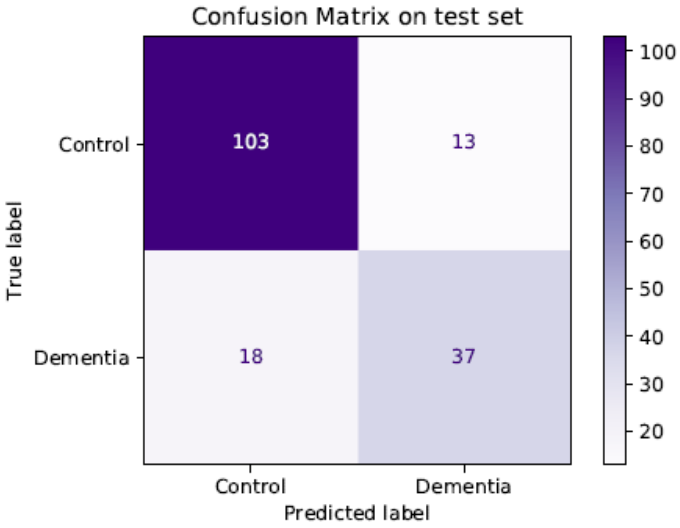
This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

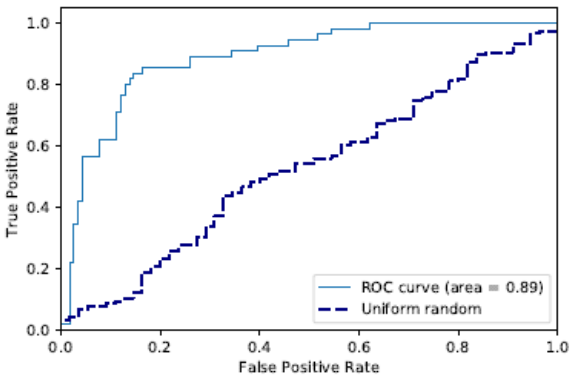
The aim is to provide a machine learning model to automatically detect dementia. The outcome model with the requirement of having reasonable performances in terms of the different losses and metrics defined and must be able to explain its predictions. In our approach, we chose to work with a three-dimensional scan of the brain as input. Namely the raw T1-weighted Magnetic Resonance Images (MRI) of the patient brain.


MRI
Preprocess
DL
70%

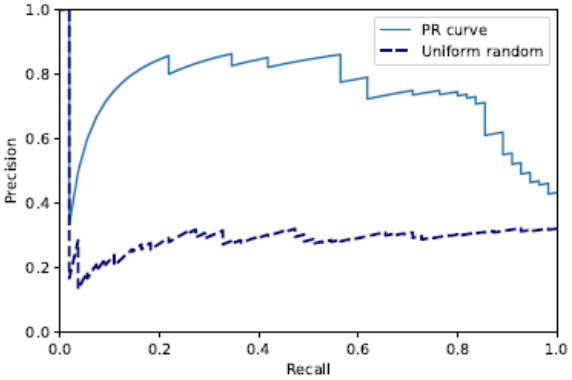

Control Panel
Brain
Attention Map
76%
Dementia prediction
Ground truth is Dementia

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





(a) ROC curve.



(b) Precision and Recall curve.

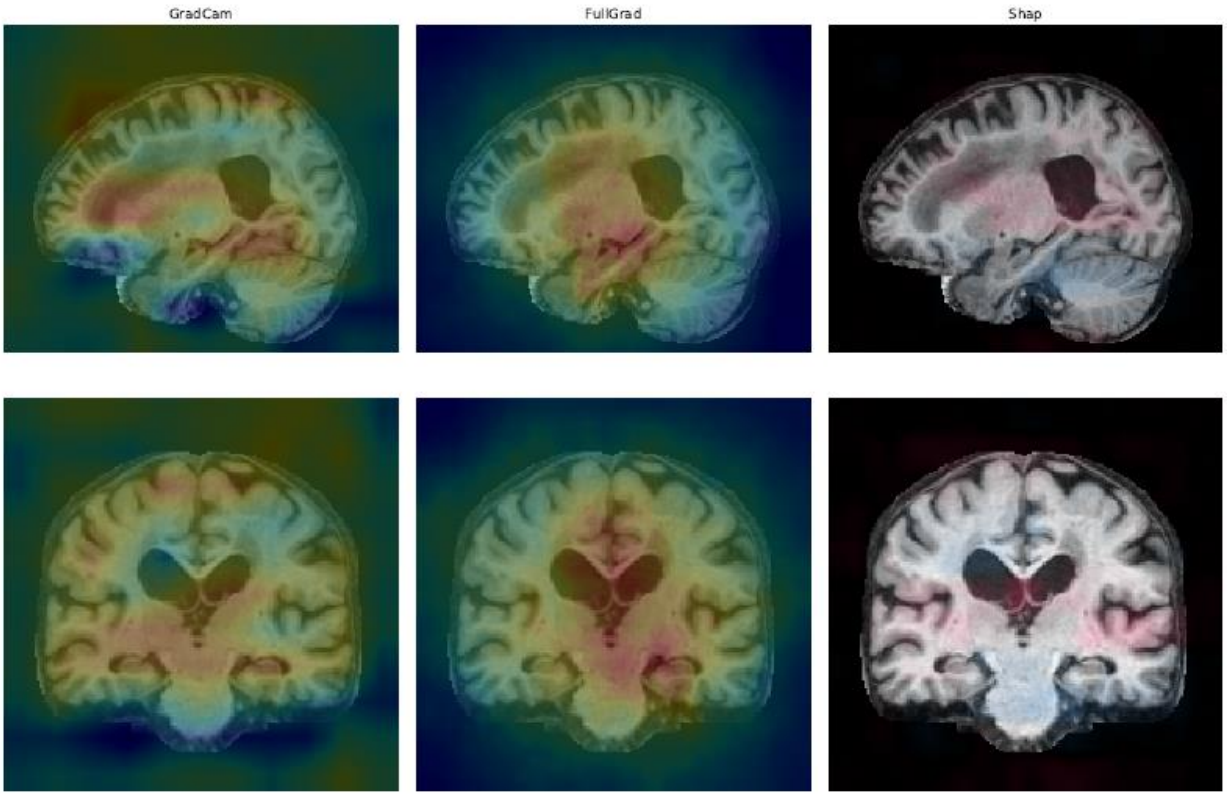


Figure 5.5: Outputs of the explainer algorithm on one patient with dementia.

- When was the benchmarking executed?
- Who participated in the benchmarking?
- What overall performance of the AI systems concerning medical accuracy, robustness, and technical performance (minimum, maximum, average etc.) has been achieved?
- What are the results of this benchmarking iteration for the participants (who opted in to share their results)?

6.1.1.12 Discussion of the benchmarking

This section 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).

- What was the general outcome of this benchmarking iteration?
- How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
- Are there real benchmarking results and interesting insights from this data?
    - How was the performance of the AI system compared to the baseline?
    - How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
    - How did the results change in comparison to the last benchmarking iteration?
- Are there any technical lessons?

    - Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
    - How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
- Are there any lessons concerning data acquisition?
    - Was it possible to collect enough data?
    - Were the data as representative as needed and expected?
    - How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
    - Was it possible to find annotators?
    - Was there any relevant feedback from the annotators?
    - How long did it take to create the dataset?
- Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
    - Are metrics missing?
    - Do the stakeholders need different reports or additional metadata (e.g., do they need the "offline capability" included in the AI metadata so that they can have a report on the best offline system for a certain task)?
- Are there insights on the benchmarking process?
    - How was the interest in participation?
    - Are there reasons that someone could not join the benchmarking?
    - What was the feedback of participants on the benchmarking processes?
    - How did the participants learn about the benchmarking?

6.1.1.13 Retirement

Topic driver: describe what happens to the benchmarking data and the submitted AI models after the benchmarking.

This section 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 section DEL04 "*AI software lifecycle specification"* (identification of standards and best practices that are relevant for the AI for health software life cycle).

- 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?

6.1.2 Benchmarking version [X]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [X].

Topic driver: Provide details of previous benchmarking versions here using the same subsection structure as above.

6.2 Subtopic [B]

Topic driver: If there are subtopics in your topic group, please provide the details about the benchmarking of the second subtopic [B] here using the same subsection structure as above (please refer to earlier comments – in red fonts - concerning subtopics).

7 Overall discussion of the benchmarking

This section 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 6.1.1.12).
We built a complete pipeline composed of preprocessing, training, evaluation and explanation to detect dementia from raw MRI scans. The models obtained by training on the OASIS dataset did not attain state-of-the-art performances but have the advantage of providing not only a diagnostic but an explanation about which region of the MRI made the model do such a prediction.

- What is the overall outcome of the benchmarking thus far?
- Have there been important lessons?
- Are there any field implementation success stories?
- Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
- 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?
- What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

8 Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on **Regulatory considerations on AI for health (WG-RC)** and their various deliverables. It is **NOT requested to re-produce regulatory frameworks**, but to show the regulatory frameworks that have to be applied in the context of your AIs and their benchmarking (**2 pages max**).*

For AI-based technologies in healthcare, 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 DEL02 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL02_1 *"Mapping of IMDRF essential principles to AI for health software",* and DEL02_2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (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 sections 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 section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for [YOUR TOPIC].

- 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 section 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).

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

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 section outlines the regulatory requirements for software used for benchmarking in this topic group.

- 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

Topic Driver: Please select the points relevant for your type of AI and the corresponding benchmarking systems. If your AIs and your benchmarking are not a medical device, this might be quite short.

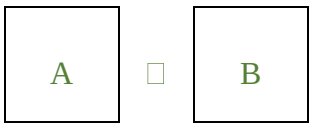
Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL02 *"AI4H regulatory considerations."*

- Documentation & Transparency
    - How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
- Risk management & 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., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
    - How are the corresponding processes document?
- Intended Use & 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 & 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 & Collaboration
    - How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

9 References

Topic driver: Add the bibliography here.

Topic driver: If you include figures in this document, please use the following MS Word format/style (otherwise the figure won't be included in the table of figures).



Captions for figures use WinWord style "Figure_No & title"

Figure 1: Example of a figure

Annex A:
Glossary

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

Acronym/Term	Expansion	Comment
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 Topic Group [YOUR TOPIC GROUP]
TG	Topic Group	
WG	Working Group	
FGAI4H	Focus Group on AI for Health	
AI	Artificial intelligence	
ITU	International Telecommunication Union	
WHO	World Health Organization	
DEL	Deliverable	
CfTGP	Call for topic group participation	
AI4H	Artificial intelligence for health	
IMDRF	International Medical Device Regulators Forum	
MDR	Medical Device Regulation	
ISO	International Standardization Organization	
GDPR	General Data Protection Regulation	
FDA	Food and Drug administration	
SaMD	Software as a medical device	
AI-MD	AI based medical device	
LMIC	Low-and middle-income countries	
GDP	Gross domestic product	
API	Application programming interface	
IP	Intellectual property	
PII	Personal identifiable information	
[…]		

Annex B:
Declaration of conflict of interests

In accordance with the ITU transparency rules, this section lists the conflict-of-interest declarations for everyone who contributed to this document. Please see the guidelines in FGAI4H-F-105 "ToRs for the WG-Experts and call for experts" and the respective forms (Application form & Conflict of interest form).

Company/Institution/Individual XYZ

A short explanation of the company's area of activity and how the work on this document might benefit the company and/or harm competitors. A list of all people who contributed to this document on behalf of this company and any personal interest in this company (e.g., shares).

Diagnostic	Dementia stage (HC; MCI, AD)	categorical
Demography	Age	continuous
	Gender	categorical
	Education level	categorical
	Education years	continuous
CSF-Biomarkers	Ab1_40	continuous
	Ab1_42	continuous
	Tau	continuous
genetic	Apoe4	categorical
Neuropsychology Score	ADAS	continuous
	MMSE	continuous
	MOCA	continuous
Brain Features (Volumes)	Left Accumbens Area	continuous
	Left Anterior Cingulate Gyrus	continuous
	Left Anterior Insula	continuous
	Left Amygdala	continuous
	Left Angular Gyrus	continuous
	Left anterior Orbital Gyrus	continuous
	Left Basal Forebrain	continuous
	Left Calcarine cortex	continuous
	Left caudate	continuous
	Left Cerebellum Exterior	continuous
	Left cerebellum White Matter	continuous
	Left cerebral White Matter	continuous
	Left co Central Operculum	continuous
	Left cun Cuneus	continuous
	Left Ententorhinal Area	continuous
	Left fo Frontal Operculum	continuous

Diagnostic	Dementia stage (HC; MCI, AD)	categorical
	Left frp Frontal Pole	continuous
	Left fug Fusiform Gyrus	continuous
	Left gre Gyrus Rectus	continuous
	Left hippocampus	continuous
	Left inflatvent	continuous
	Left iog Inferior Occipital Gyrus	continuous
	Left itg Inferior Temporal Gyrus	continuous
	Left Lateralventricle	continuous
	Left liglingual Gyrus	continuous
	Left lorg Lateral Orbital Gyrus	continuous
	Left mcgg Middlecingulate Gyrus	continuous
	Right mfc Medial Frontalcortex	continuous
	Left mfc Medial Frontalcortex	continuous
	Left mfg Middle Frontal Gyrus	continuous
	Left mog Middle Occipital Gyrus	continuous
	Left morg Medial Orbital Gyrus	continuous
	Left mpog Post-Central Gyrus Medial Segment	continuous
	Left mprg PreCentral Gyrus Medial Segment	continuous
	Left msfg Superior Frontal Gyrus Medial Segment	continuous
	Left mtg Middle Temporal Gyrus	continuous
	Left ocp Occipital Pole	continuous
	Left ofug Occipital Fusiform Gyrus	continuous
	Left opifgopercularpartofthe Inferior Frontal Gyrus	continuous
	Left orifg Orbitalpartofthe Inferior Frontal Gyrus	continuous
	Left pallidum	continuous
	Left pcggposteriorcingulate Gyrus	continuous
	Left pcuprecuneus	continuous
	Left phgparahippocampal Gyrus	continuous
	Left pinsposteriorinsula	continuous
	Left pog Post-Central Gyrus	continuous
	Left poparietal Operculum	continuous
	Left porgposterior Orbital Gyrus	continuous
	Left ppplanumpolare	continuous
	Left prg PreCentral Gyrus	continuous
	Left pt Planum Temporale	continuous
	Left Putamen	continuous
	Left sca subcallosal Area	continuous
	Left sfg Superior Frontal Gyrus	continuous

Diagnostic	Dementia stage (HC; MCI, AD)	categorical
	Left sm csupplementarymotorcortex	continuous
	Left smg supramarginal Gyrus	continuous
	Left sog Superior Occipital Gyrus	continuous
	Left spl Superior Parietallobule	continuous
	Left stg Superior Temporal Gyrus	continuous
	Left thalamus Proper	continuous
	Left tmp Temporal Pole	continuous
	Left trifg Triangular part of the Inferior Frontal Gyrus	continuous
	Left ttg Transverse Temporal Gyrus	continuous
	Left ventraldc	continuous
	Lipidemia comorbidity	continuous
	minimentalstate	continuous
	Right accumbens Area	continuous
	Right acgganteriorcingulate Gyrus	continuous
	Right ainsanteriorinsula	continuous
	Right amygdala	continuous
	Right angangular Gyrus	continuous
	Right aorganterior Orbital Gyrus	continuous
	Right basalforebrain	continuous
	Right calccalcarinecortex	continuous
	Right caudate	continuous
	Right cerebellum Exterior	continuous
	Right cerebellum White Matter	continuous
	Right cerebral White Matter	continuous
	Right co central Operculum	continuous
	Right cuncuneus	continuous
	Right ententorhinal Area	continuous
	Right fo Frontal Operculum	continuous
	Right frp Frontal Pole	continuous
	Right fug Fusiform Gyrus	continuous
	Right gre Gyrus Rectus	continuous
	Right hippocampus	continuous
	Right inflatvent	continuous
	Right iog Inferior Occipital Gyrus	continuous
	Right itg Inferior Temporal Gyrus	continuous
	Right Lateral ventricle	continuous
	Right lig lingual Gyrus	continuous
	Right lorg Lateral Orbital Gyrus	continuous

Diagnostic	Dementia stage (HC; MCI, AD)	categorical
	Right mcgg Middlecingulate Gyrus	continuous
	Right mfc Medial Frontalcortex	continuous
	Right mfg Middle Frontal Gyrus	continuous
	Right mog Middle Occipital Gyrus	continuous
	Right morg Medial Orbital Gyrus	continuous
	Right mpog Post-Central Gyrus Medial Segment	continuous
	Right mprg PreCentral Gyrus Medial Segment	continuous
	Right msfg Superior Frontal Gyrus Medial Segment	continuous
	Right mtg Middle Temporal Gyrus	continuous
	Right ocp Occipital Pole	continuous
	Right ofug Occipital Fusiform Gyrus	continuous
	Right opifgopercularpartofthe Inferior Frontal Gyrus	continuous
	Right orifg Orbitalpartofthe Inferior Frontal Gyrus	continuous
	Right pallidum	continuous
	Right pcgg Posteriorcingulate Gyrus	continuous
	Right pcu pPrecuneus	continuous
	Right phg parahippocampal Gyrus	continuous
	Right pinsposteriorinsula	continuous
	Right pog Post-Central Gyrus	continuous
	Right po Parietal Operculum	continuous
	Right porg Posterior Orbital Gyrus	continuous
	Right ppplanumpolare	continuous
	Right prg PreCentral Gyrus	continuous
	Right ptplanum Temporale	continuous
	Right putamen	continuous
	Right scasubcallosal Area	continuous
	Right sfg Superior Frontal Gyrus	continuous
	Right smc Supplementary motorcortex	continuous
	Right smg Supramarginal Gyrus	continuous
	Right sog Superior Occipital Gyrus	continuous
	Right spl Superior Parietallobule	continuous
	Right stg Superior Temporal Gyrus	continuous
	Right thalamus proper	continuous
	Right tmp Temporal Pole	continuous
	Right trifgtriangularpartofthe Inferior Frontal Gyrus	continuous
	Right ttgtransverse Temporal Gyrus	continuous