|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FGAI4H-K-016-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plen | E-meeting, 27-29 January 2021 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Neuro Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Neuro) [same as Meeting J] | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Marc Lecoultre MLLab.ai Switzerland | | Tel: +41 79 321 09 29 Fax: +41 22 364 30 69 Email: [ml@mllab.ai](mailto:ml@mllab.ai) |
| **Contact:** | | Kherif Ferah, vice-director  LREN, CHUV Switzerland | | Tel: +41 79 556 11 06 Email: [Ferath.kherif@chuv.ch](mailto:Ferath.kherif@chuv.ch) |

|  |  |
| --- | --- |
| **Abstract:** | This document is the Topic Description Document (TDD) containing the standardized benchmarking approach for the use of AI for Neuro-Cognitive diseases. It follows the structure defined in FGAI4H-C-105 relevant for setting up this benchmarking.  This version of the TDD is the same as seen in Meeting J (FGAI4H-J-016-A01), reproduced for easier reference as a Meeting K document. |

**Table of Contents**

[1 Introduction 3](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725459)

[1.1 Topic Description 3](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725460)

[1.1.1 Relevance 3](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725461)

[1.1.2 Current approaches and gold standards for detection of AD 3](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725462)

[1.1.3 Impact of benchmarking AI Solutions 3](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725463)

[1.2 Ethical considerations 4](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725464)

[1.3 Existing AI solutions (includes datasets, systems and benchmarks) 4](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725465)

[2 AI4H Topic Group 5](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725466)

[2.1 General mandate of the Topic Group 5](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725467)

[2.2 Topic description document 6](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725468)

[2.3 Subtopics 6](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725469)

[2.4 Topic group participation 6](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725470)

[2.5 Status of this Topic Group 6](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725471)

[2.6 Next meetings 7](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725472)

[3 Method 7](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725473)

[3.1 Overview of the benchmarking 7](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725474)

[3.2 AI Input Data Structure 7](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725475)

[3.3 AI Output Data Structure 8](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725476)

[3.3.1 Test Data Labels 8](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725477)

[3.4 Scores and metrics 8](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725478)

[3.5 Undisclosed test data set collection 9](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725479)

[3.6 Benchmarking methodology and architecture 9](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725480)

[4 Reporting methodology 10](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725481)

[5 Results 10](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725482)

[5.1 Data Quality and curation 10](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725483)

[5.2 Data catalogue 14](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725484)

[5.3 Explainability of Deep Learning Models Trained on MRI Scans 16](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725485)

[5.4 Robustness and Fairness Measure 19](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725486)

[6 Discussion 19](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725487)

[7 Declaration of conflict of interest 19](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725488)

[Appendix A: Glossary 20](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725489)

[Appendix B: Data example 21](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725490)

[References 25](file:///C:\Users\banno\AppData\Local\Temp\Temp2_Fillers.zip\Fillers\FGAI4H-J-016-A01.docx#_Toc52725491)

# Introduction

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a standardized benchmarking approach for AI-based diagnostic applications for neuro-cognitive disorders.

## Topic Description

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.

### Relevance

1. 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.

### Current approaches and gold standards for detection of AD

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.

### Impact of benchmarking AI Solutions

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.

## Ethical considerations

…

## Existing AI solutions (includes datasets, systems and benchmarks)

We have a proven track record in applying supervised classification methods for prediction of clinical outcome and explaining the variance of the data. We previously applied support-vector machine (SVM) classification methods to anatomical data for diagnosis of different dementia subtypes. However, multivariate pattern recognition methods have been applied primarily to uni-modal data, motivating a novel methodological approach to accommodate multi-modal data. Recently, we used this methodology to build predictive models for healthy ageing and showed that the mean prediction error was significantly lower when combining all measurements. 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 |
| [Anitha‌2017] | 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 |
| [Mitchell‌2018] | 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 |
| [Tolonen‌2018] | 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 |
| [Vashistha‌2019] | AI-based clinical decision systems (CDSs) along with POC diagnosis | Neurodegenerative disorders such as Parkinson’s disease, amyo-trophic lateral sclerosis (ALS), Alzheimer’s disease, epilepsy | * Machine learning and wearables based Therapeutics * A combinatorial intelligent system for the prediction of PD development by ML | Markov decision processes (MDP) and dynamic decision net-works | Neurodegenerative disorders Specialist |

# AI4H Topic Group

Over the past decade, considerable resources have been allocated to exploring the use of AI for health, which has revealed an immense potential. Yet, due to the complexity of AI models, it is difficult to understand their strengths, weaknesses, and limitations. If the technology is poorly designed or the underlying training data are biased or incomplete, errors or problematic results can occur. AI technology can only be used with complete confidence if it has been quality controlled through a rigorous evaluation in a standardized way. Towards developing this standard assessment framework of AI for health, the ITU has established FG-AI4H in partnership with the WHO.

Thus far, FG-AI4H has established thirteen topic groups. These are concerned with: AI and cardiovascular disease risk prediction, child growth monitoring, dermatology, falls among the elderly, histopathology, neuro-cognitive diseases, ophthalmology (retinal imaging diagnostics), psychiatry, radiotherapy, snakebite and snake identification, symptom checkers, tuberculosis, and volumetric chest computed tomography.

As the work by the Focus Group continues, new Topic Groups will be created. To organize the Topic Groups, for each topic the Focus Group chose a topic driver. The exact responsibilities of the topic driver are still to be defined and are likely to change over time. The preliminary and yet-to-confirm list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document.
* Reviewing the input documents for the topic and moderating the integration in a dedicated session at each Focus Group meeting.
* Organizing regular phone calls to coordinate work on the topic description document between meetings.

## General mandate of the Topic Group

The Topic Group is a concept specific to the AI4H-FG. The preliminary responsibilities of the Topic Groups are:

1. Provide a forum for open communication among various stakeholders
2. Agree upon the benchmarking tasks of this topic and scoring metrics
3. Facilitate the collection of high quality labeled test data from different sources
4. Clarify the input and output format of the test data
5. Define and set-up the technical benchmarking infrastructure
6. Coordinate the benchmarking process in collaboration with the Focus Group management and working groups

## Topic description document

The primary output of each Topic Group is the topic description document (TDD) specifying all relevant aspects of the benchmarking for the individual topics. **This document is the TDD for the Topic Group on “AI against neuro-cognitive diseases” (TG-Cogni)** The document will be developed cooperatively over several FG-AI4H meetings starting from meeting D in Shanghai. Suggested changes to the document will be submitted as input documents for each meeting. The relevant changes will then be discussed and integrated into an official output document until the TDD ready for the first official benchmarking.

More details about the activities of the topic group can be found in the documents: FGAI4H-C-020-R1: Status report for Alzheimer’s disease use case FGAI4H-B-013-R1: Proposal: Using machine learning and AI for validation of Alzheimer’s disease biomarkers for use in the clinical practice

## Subtopics

Topic groups summarize similar AI benchmarking use cases to limit the number of use case specific meetings at the Focus Group meetings and to share similar parts of the benchmarking. However, in some cases, it is expected that inside a Topic Group different subtopic Groups can be established to pursue different topic-specific specializations.

## Topic group participation

The participation in both the focus and Topic Group is generally open and free of charge. Anyone who is from a member country of the ITU may participate. On the 14. of March 2019 the ITU published an official “call for participation” document outlining the process for joining the Focus Group and the Topic Group. For this topic, the corresponding call can be found [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/cogni.aspx).

## 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](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Cogni.pdf).
* 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.

## Next meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* F: Zanzibar, Tanzania; 2-5 September 2019
* G: New Delhi, India; November 2019
* H: Brasilia, Brazil; January 2020

An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

# Method

## Overview of the benchmarking

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.

## AI Input Data Structure

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

## AI Output Data Structure

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:
* Alzheimer's Disease
* Mild cognitive impairment (MCI)
* Cognitively normal (CN)
* Other Mixed Dementia (MD)

### Test Data Labels

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

## Scores and metrics

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)

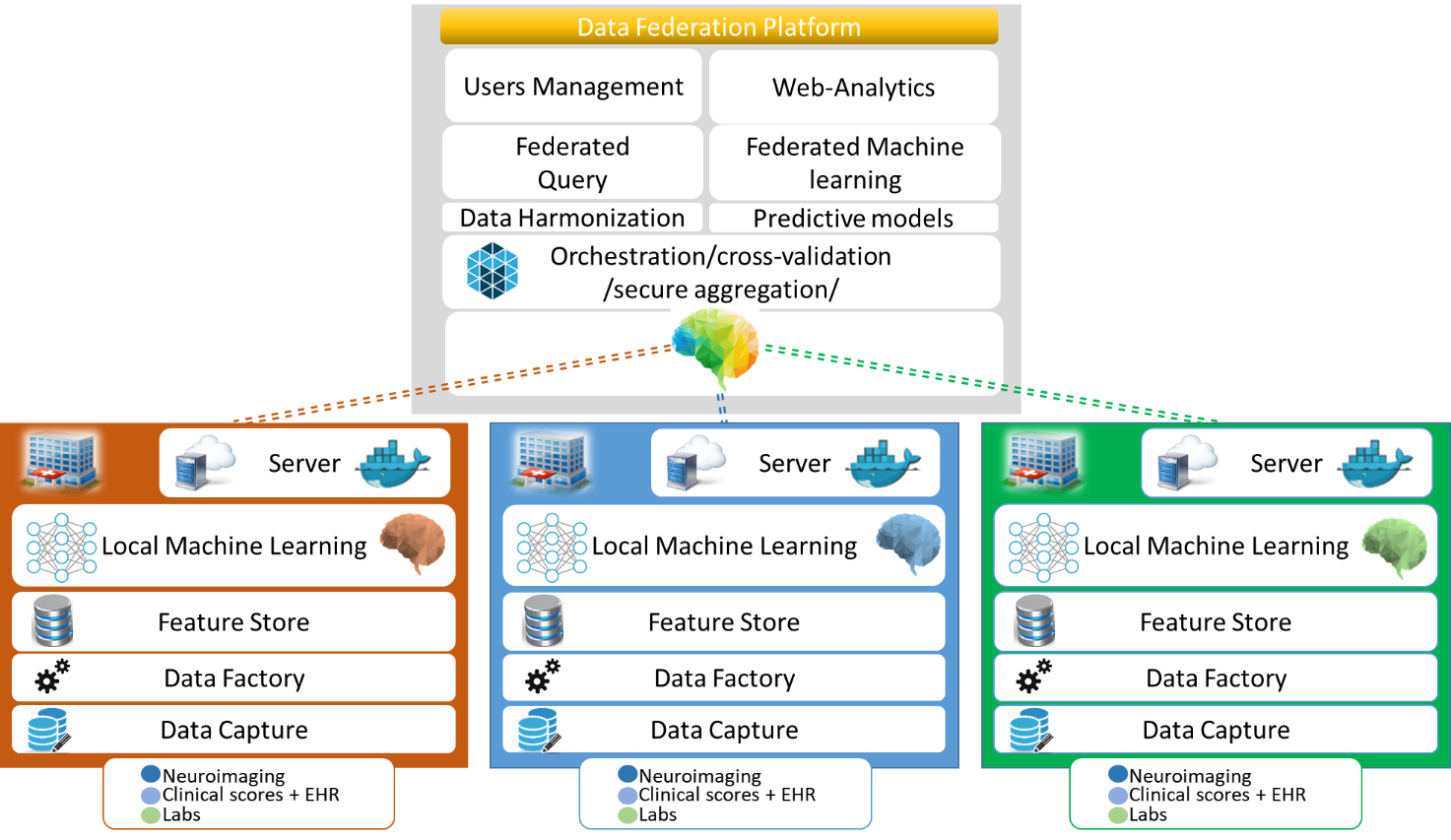
Clinical precision the probability that the disease is present when the test is positive = sensitivity x prevalence / (sensitivity x prevalence + (1-specificity) x (1-sensitivity)

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 …).

## Undisclosed test data set collection

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 6000 patients on dementia (one of the largest patients’ cohort) different stages of the disease (subjective complains, mild impairments or demented) raw data acquisition / acceptance

## Benchmarking methodology and architecture



(TBC)

* technical architecture
* hosting (IIC, etc.)
* possibility of an online benchmarking on a public test dataset
* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations

# Reporting methodology

# Results

## Data Quality and curation

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”*

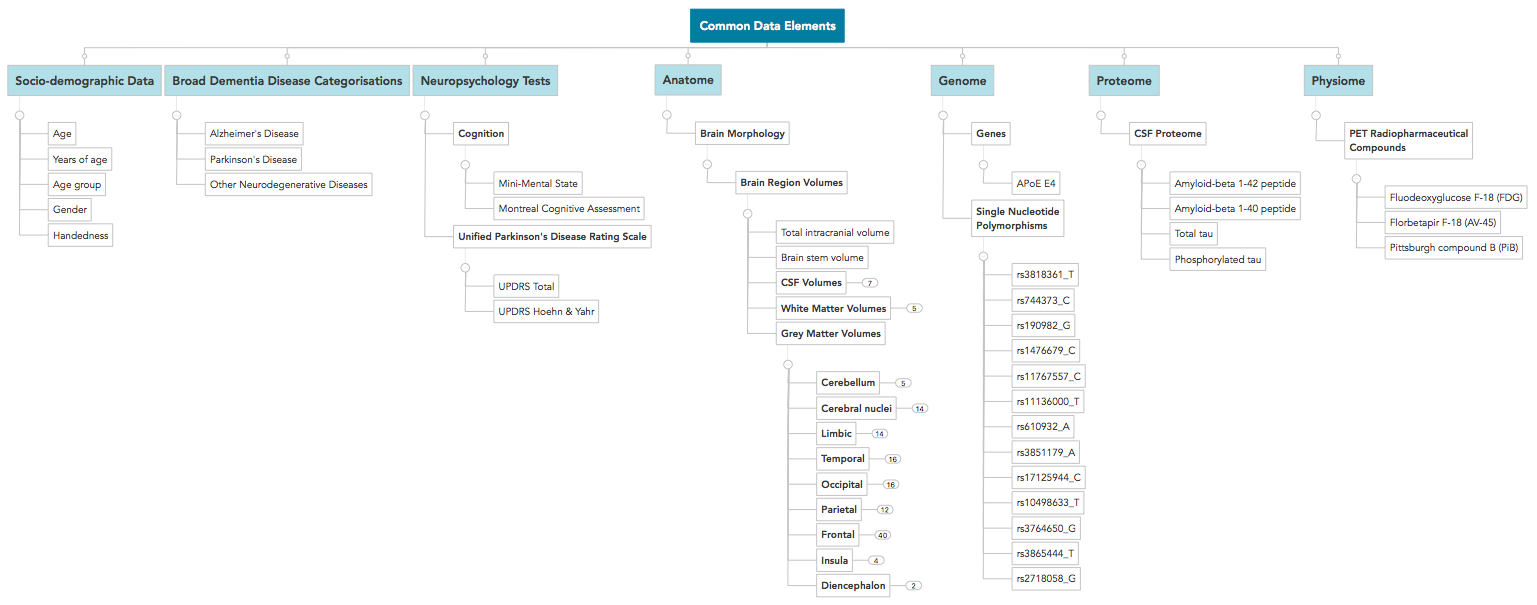
**DACQORD Indicators.**

|  |  |  |
| --- | --- | --- |
| **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. |

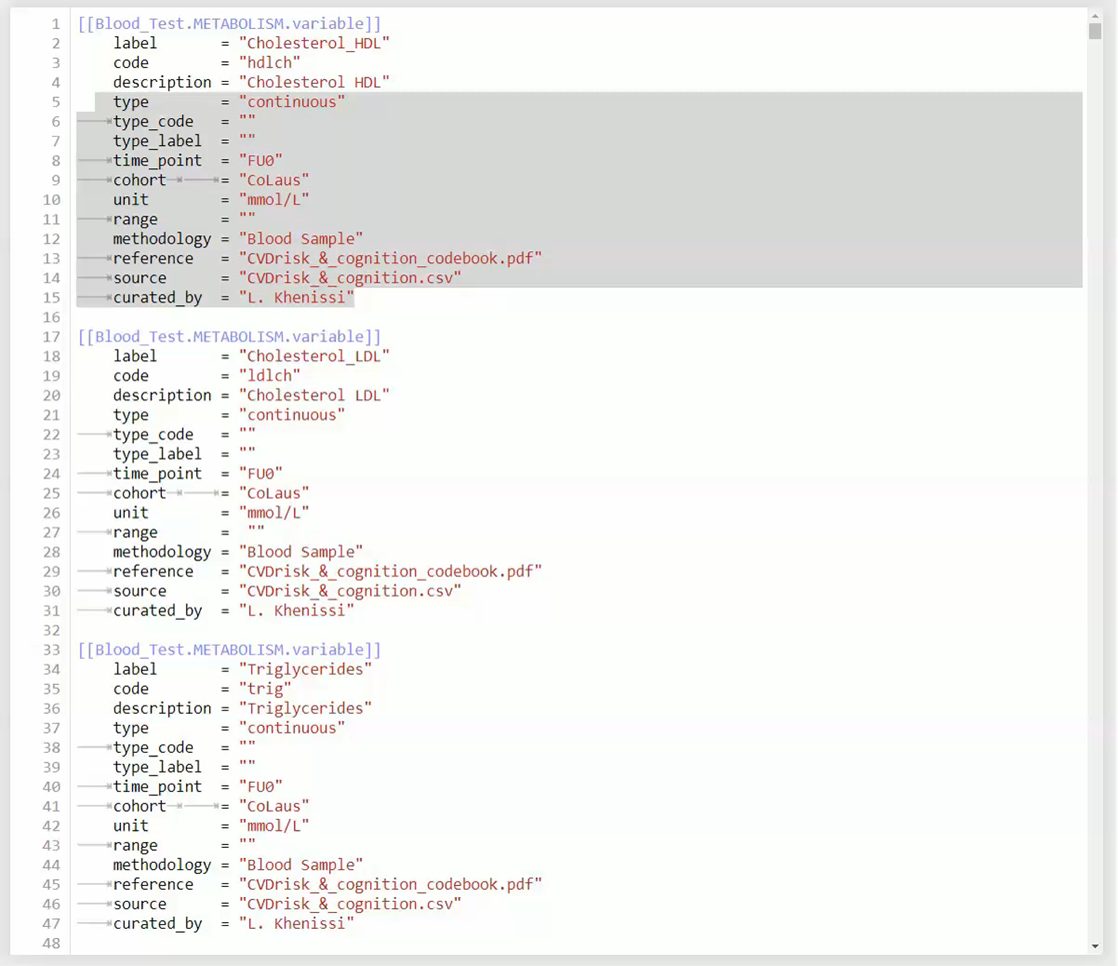
|  |  |  |
| --- | --- | --- |
| Design-time | Correctness | 16. Data collection includes fields for documenting that participants meet inclusion/ exclusion criteria. |
| Representation | 17. The data entry tool does not perform rounding or truncation of entries that might result in precision-loss. |
| 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. |

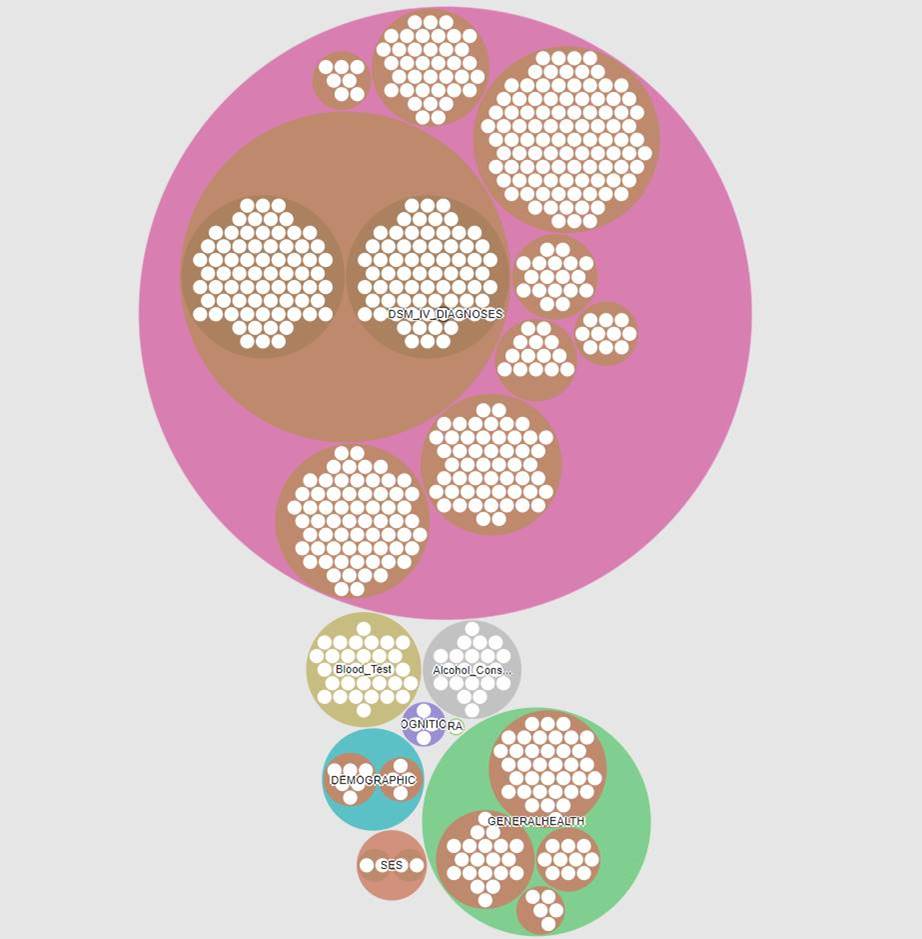
## Data catalogue

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).

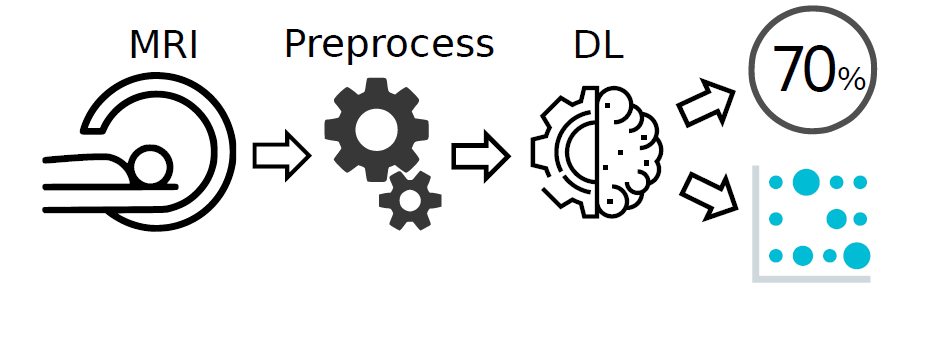


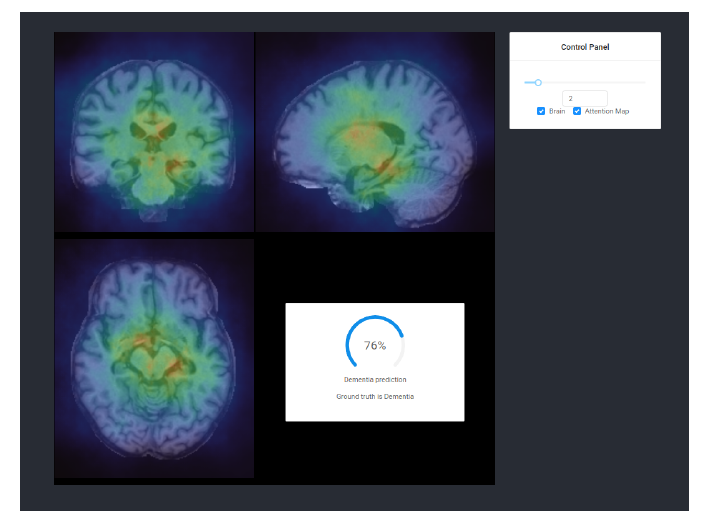


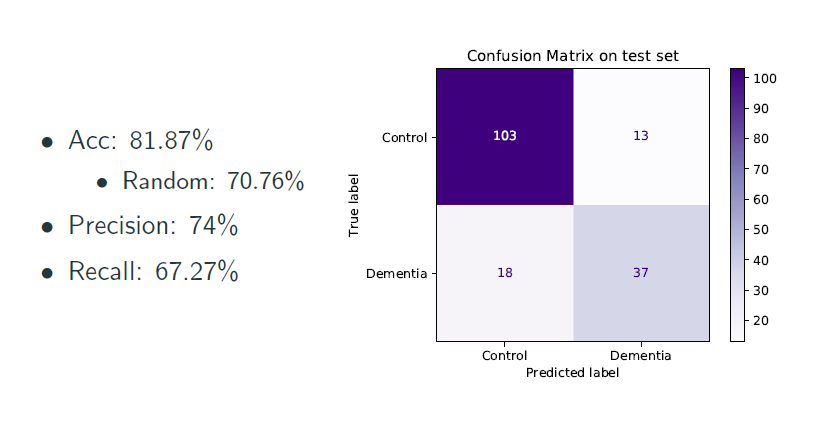
## Explainability of Deep Learning Models Trained on MRI Scans

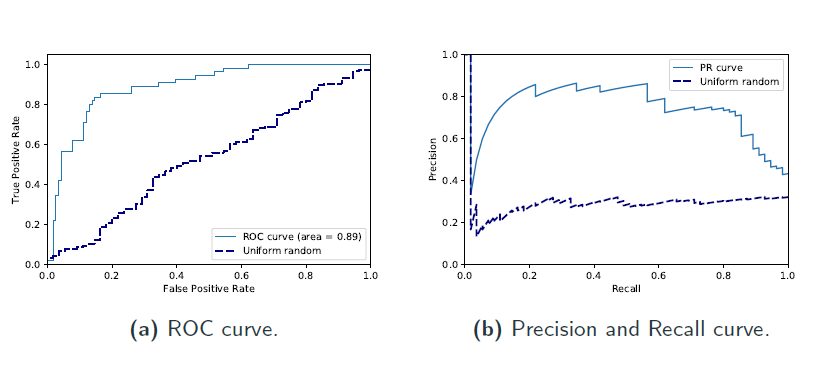
Problem Statement

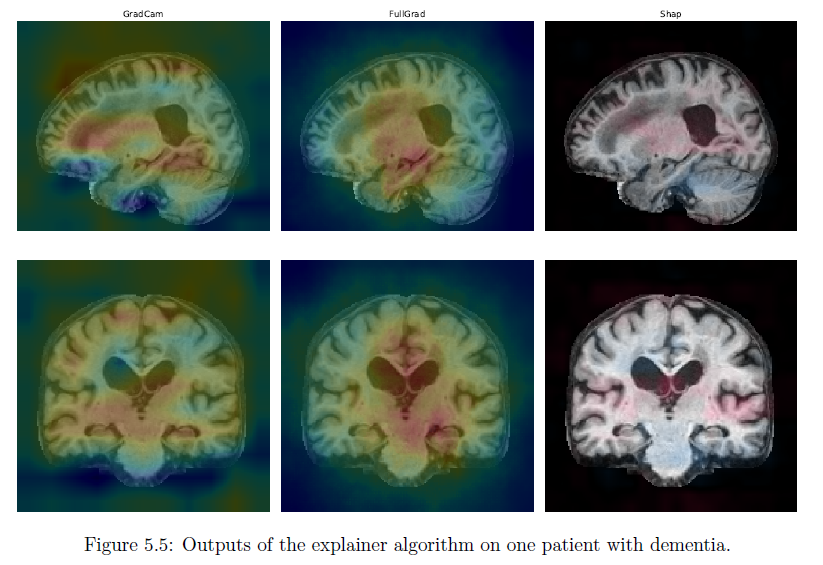
The aim is to provide a machine learning model to automatically detect dementia. The outcome model with the requirement of having reasonable performances in terms of the different losses and metrics defined and must be able to explain its predictions. In our approach, we chose to work with a three-dimensional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











## Robustness and Fairness Measure

To be included

# Discussion

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.

# Declaration of conflict of interest

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

# Appendix A: Glossary

This section lists all the relevant abbreviations and acronyms used in the document. If there is an external source

* **AI** - [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) – an umbrella term that refers to one or more of the various fields of computer science including machine learning, neural networks and deep learning.
* **AI4H** - AI for health - An [ITU-T SG16 Focus Group](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx) founded in cooperation with the WHO in July 2018.
* **API** - [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) - the software interface systems communicate through.
* **FG** - [Focus Group](https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx) - An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas.
* **IIC** - International Computing Centre - The United Nations data center that will host the benchmarking infrastructure.
* **ITU** - [International Telecommunication Union](https://www.itu.int) - The United Nations specialized agency for information and communication technologies – ICTs.
* **LMIC** - Low and Middle Income Countries
* **NGO** - [Non Governmental Organization](https://en.wikipedia.org/wiki/Non-governmental_organization) - NGOs are usually non-profit and sometimes international organizations independent of governments and international governmental organizations that are active in humanitarian, educational, health care, public policy, social, human rights, environmental, and other areas to affect changes according to their objectives. (from Wikipedia.en)
* **SDG** - [Sustainable Development Goals](https://www.un.org/sustainabledevelopment/) - The United Nations Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. Currently there are 17 goals defined. SDG 3 is to “Ensure healthy lives and promote well-being for all at all ages” and is therefore the goal that will benefit from the AI4H Focus Groups work the most.
* **TBC** - A topic group item to be completed.
* **TBD** - A topic group item to be discussed / determined
* **TDD** - Topic Description Document - Document specifying the standardized benchmarking for a topic FG AI4H Topic Group works on. This document is the TDD for the Topic Group “AI for Ophthalmology (retinal imaging diagnostics)”.
* **TG** - Topic Group - Structures inside AI4H FG summarizing similar use cases and working on a TDD specifying the setup of a standardized benchmarking for the corresponding topic. The Topic Groups have been first introduced by the FG at the Meeting C, January 2019 in Lausanne. See protocol FG-AI4H-C-10x for details.
* **WHO** - [World Health Organization](https://www.who.int) - The United Nations specialized agency for international public health.

# Appendix B: Data example

| 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 |
| 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 |
| 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 |
| 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 |

# References

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_