|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-G-037 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | WG-DAISAM | New Delhi, 13-15 November 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | WG DAISAM chairs | | |
| **Title:** | | Regulatory review: six quality criteria for data and AI solution assessment | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Pat Baird Philips USA | | Tel: +1 262 239-9321 Email: [pat.baird@philips.com](mailto:pat.baird@philips.com) |
| **Contact:** | | Luis Oala Fraunhofer HHI Germany | | Tel: +49 30 31002-104  Fax: +49 30 31002-190 Email: [luis.oala@hhi.fraunhofer.de](mailto:luis.oala@hhi.fraunhofer.de) |
| **Contact:** | | Pradeep Balachandran Technical Consultant (eHealth) India | | Email: [pbn.tvm@gmail.com](mailto:pbn.tvm@gmail.com) |

|  |  |
| --- | --- |
| **Abstract:** | In this document we outline the preliminary results in the working package “Regulatory Review” of WG-DAISAM. This includes a tentative list of six quality criteria for data and AI solution assessment: (1) description, (2) risk, (3) security/privacy, (4) interfaces/dependencies, (5) verification/validation/testing, and (6) change management. These quality criteria were sourced from a selection a of regulatory and guidance documents. Each quality criterion is accompanied by a selection of questions that were collected from the regulatory and guidance documents as well. |

*Note this is a draft document intended for review and feedback.*

# Abbreviations

|  |  |
| --- | --- |
| AI | Artificial intelligence |
| AI4H | Artificial intelligence for health |
| API | Application Programming Interface |
| CIA | Confidentiality, Integrity, and Availability |
| CMDE | Center for Medical Device Evaluation |
| DAISAM | Data and AI Solution Assessment Methods |
| FDA | Food and Drug Administration |
| FG | Focus group |
| HIPAA | [Health Insurance Portability and Accountability Act](https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act) |
| IMDRF | International Medical Device Regulators Forum |
| ITU | International Telecommunication Union |
| IVD | In Vitro Diagnostics |
| PII | Personally Identifiable Information |
| RAID | Redundant Array of Inexpensive Disks |
| SaMD | Software as a medical device |
| SiMD | Software in a medical device |
| WG | Working group |
| WHO | World Health Organization |

**Table of Contents**

[Abbreviations 2](#_Toc24148983)

[1 Introduction 4](#_Toc24148984)

[2 Purpose of This Document 4](#_Toc24148985)

[3 Scope and Method Behind the Outputs of This Document 4](#_Toc24148986)

[4 Outputs 5](#_Toc24148987)

[4.1 Application Description 5](#_Toc24148988)

[4.2 IMDRF Classification 6](#_Toc24148989)

[4.3 Data Quality Assessment 7](#_Toc24148990)

[4.3.1 Data Set Description 7](#_Toc24148991)

[4.3.2 Risk Management 8](#_Toc24148992)

[4.3.3 Security/Privacy 8](#_Toc24148993)

[4.3.4 Dependencies/Interfaces 8](#_Toc24148994)

[4.3.5 Verification/Validation/Testing 8](#_Toc24148995)

[4.3.6 Change Management 9](#_Toc24148996)

[4.4 AI Solution Quality Assessment 9](#_Toc24148997)

[4.4.1 Algorithm Description 9](#_Toc24148998)

[4.4.2 Risk Management 10](#_Toc24148999)

[4.4.3 Security/Privacy 10](#_Toc24149000)

[4.4.4 Dependencies/Interfaces 11](#_Toc24149001)

[4.4.5 Verification/Validation/Testing 11](#_Toc24149002)

[4.4.6 Change Management 12](#_Toc24149003)

[5 Outlook 12](#_Toc24149004)

[6 Appendix 13](#_Toc24149005)

[6.1 Appendix 1 – IMDRF Classification Scheme 13](#_Toc24149006)

[6.2 Appendix 2 – Quality Assessment Questionnaire & Glossary 16](#_Toc24149007)

[References 26](#_Toc24149008)

[Acknowledgments 26](#_Toc24149009)

# Introduction

Health technologies require careful evaluation under consideration of both technical and health-related aspects, prior to wider usage. Several factors complicate this evaluation and, thus, the deployment of artificial intelligence (AI) solutions in the health context. Therefore, the International Telecommunication Union (ITU) and the World Health Organization (WHO) have initiated an international standard-seeking effort to address these challenges by creating a joint focus group on “Artificial Intelligence for Health” (FG-AI4H). The ITU and the WHO are two specialized agencies of the United Nations authorized for creating global standards in the fields of information technology and health, respectively. The mandate of the ITU/WHO focus group is to undertake crucial, exploratory steps towards evaluation standards that are applicable on a global scale. FG-AI4H has begun working towards establishing a rigorous evaluation process for AI4H solutions, under the supervision of ITU and WHO, with a global community of experts from health, machine learning, AI, both from academia and industry, and regulation.

Cross-sectional problems are tackled by dedicated working groups (WG) of FG-AI4H. The working group data and AI solution assessment methods (DAISAM) is tasked to develop quality criteria together with the individual topic groups as well as independent experts[[1]](#footnote-1). This document presents a snapshot of the current state of one of the activities undertaken by WG-DAISAM: the review of regulatory and guidance documents by national and international oversight organizations.

# Purpose of This Document

As part of the WG-DAISAM effort to identify quality dimensions for the assessment of data and AI solutions, we continuously review regulatory and guidance documents. The motivation behind this review process is that existing regulations for related technologies can provide blueprints for our work and that the inclusion of guidance documents ensures that we consider the concerns of important stakeholders, like national and international medical device oversight organizations, from the start.

In the following, we briefly outline the method and scope for this first review iteration, present the outputs, and finally provide an illustration and additional material in the appendix.

# Scope and Method Behind the Outputs of This Document

A number of documents have already been published that provide a series of good practices that are directly or indirectly related to AI. A review was performed of these existing documents, and where relevant, they were used as a starting point for the DAISAM approach:

1. Xavier Health held a conference in 2017 regarding healthcare AI; conference attendees developed a white paper, published in 2018, “Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare” [1] which contains suggestions for good machine learning practices.
2. Radiologists have been early adopters for using AI in healthcare, and have published and international paper regarding their experiences and potential ethical issues in “Ethics of AI in Radiology: European and North American Multi Society Statement” [2].
3. Although the US FDA paper “Use of Public Human Genetic Variant Databases to Support Clinical Validity for Genetic and Genomic-Based In Vitro Diagnostics” [3] is not specific to AI, it has several recommendations regarding data collection, testing, curation, etc.
4. Similarly, Microsoft has published a paper of “The Future Computed” [4] which contains additional considerations to assure quality data.

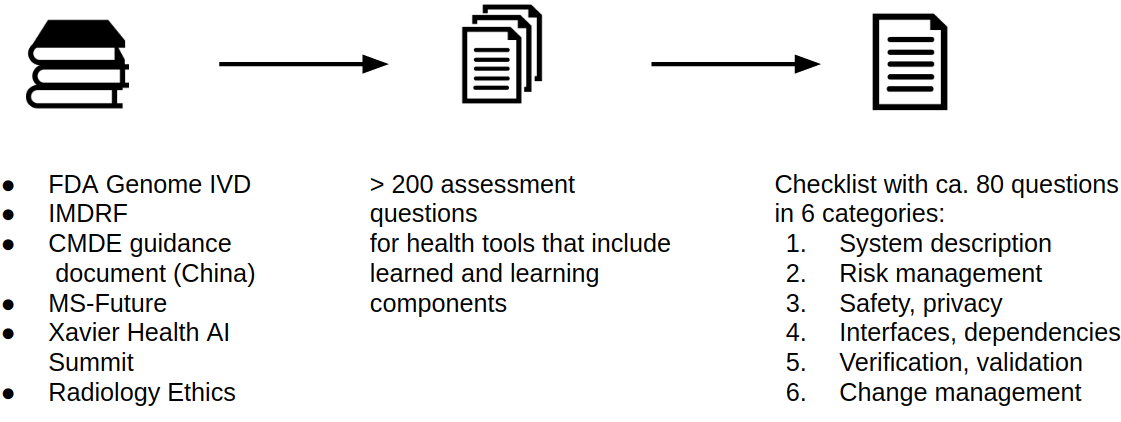


Illustration 1: Flow chart of the first document review iteration

1. The Chinese Center for Medical Device Evaluation (CMDE) published a guidance document regarding some regulatory expectations for AI in healthcare - “Review Highlights of Aided Decision-making Medical Device Software Using Deep Learning Technique (Exposure Draft)” [5].
2. The International Medical Device Regulators Forum (IMDRF) had published a guidance document [6] regarding classification of risk for Software as a Medical Device (SaMD) applications in 2014. This classification scheme was then extended by DAISAM to include levels of AI autonomy.

Thus documents of US ([1], [3], [4]), Chinese ([5]), and international ([2], [6]) provenance were considered. We note that the geographical and organizational coverage is not complete. This document serves as a snapshot of the current status of our work and is not published as a final report. As work progresses, the coverage, both geographical and organizational, will be expanded.

As displayed in *Illustration 1*, we extracted questions and criteria for quality assessment from the above documents. These entries were collected in a grand assessment checklist which totaled more than 150 entries which was subsequently screened for duplicates and flagged for out-of-scope criteria. These decisions were discussed in the FG-AI4H management call as well as the public meeting F of FG-AI4H. As a final step the remaining entries on the grand assessment checklist were grouped into several topic areas which are explained below.

# Outputs

The first review iteration resulted in a grand assessment checklist containing more than 150 entries. We first identified questions that enable the tester to better understand the application context. These questions are manifested in the overall application description (section 4.1) and the IMDRF classification (section 4.2; more detail on the IMDRF classification scheme can be found in section 6.1 of the appendix). The remaining entries on the grand assessment checklist fed into each area of quality assessment, that is data (section 4.3) and AI solution (section 4.4). We grouped, cleared and synthesized these entries into six clusters namely (1) description, (2) risk, (3) security/privacy, (4) interfaces/dependencies, (5) verification/validation/testing, and (6) change management, which are detailed below.

## Application Description

This section provides an overview of the application being evaluated.

1. What is the intended use of this application?
   1. ‘Medical grade’ application/service
   2. ‘Non- medical grade’ application/service
2. Does it diagnose, treat, mitigate, prevent, or monitor a disease, disorder, or injury?
3. Does it support operations, such as
   1. Clinical process optimization,
   2. Data pre- or post-processing,
   3. Automated transcription, annotation / labeling
4. Does it assist in managing public health issues?
5. Does the application use a locked AI algorithm (static model) or an adaptive AI algorithm (dynamic or adaptive model)?
   1. If adaptive, what is the frequency or the trigger for a model update, e.g.
      1. Does it update after every use, or does it update every 6 months, does it update with a new set of training data?
   2. If adaptive, are there constraints on the amount of allowed autonomous model learning?
6. Describe the principles of operation of the application.
7. How is the underlying AI model design appropriate for its use? (e.g. Why did you choose neural network?)
8. What control measures and safeguards have been identified for detecting the operating environment risks of the product, such as
   1. Contra-indications,
   2. Warnings,
   3. Precaution statements,
   4. Labels,
   5. Displays
9. How is this application helping healthcare? What critical aspects support the intended use, such as
   1. Low rate of false-negative results,
   2. Process efficiency improvement and cost reduction,
   3. Improved patient satisfaction
10. What is the application deployment environment? For example, is the AI application going to be deployed as
    1. Software as a medical device (SaMD),
    2. Software in a medical device (SiMD),
    3. Mobile application,
    4. Desktop application,
    5. Web/Cloud Application

## IMDRF Classification

In 2014, the International Medical Device Regulators Forum had published a risk classification scheme for software; products with a higher risk classification require more diligence than those with lower risk.

*Appendix 1* contains a description of a modified version of this classification scheme. Please refer to that annex when answering the following questions.

1. What is the significance of the information provided by the software to healthcare?
   1. Inform
   2. Drive
   3. Treat or diagnose with Approval
   4. Treat or diagnose with Override
   5. Treat or diagnose with no intervention possible
2. What is the worst-case state of healthcare situation or condition?
   1. Critical
   2. Serious
   3. Non-serious
3. What class (I – VI) does this product/application belong to?

## Data Quality Assessment

This subsection lists concrete questions to determine the quality of training and test data in the six categories (1) data set description, (2) risk, (3) security/privacy, (4) interfaces/dependencies, (5) verification/validation/testing, and (6) change management.

### Data Set Description

1. Describe data sets and training data, including:
   1. Volume - how much data is there?
   2. Velocity - how quickly is the data being created?
   3. Variety – how many data sources are there?
   4. Veracity – why do you think you can trust the data /data source?
   5. Validity – are the values correct? Are they up to date? Describe the protocol (explaining ‘When’, ‘How’ & ‘by Whom’ information ) used to gather and validate the data.
   6. Viability – how is the data relevant to the use case?
   7. Volatility – how often does the data change?
2. Demographics
   1. What demographic does this serve?
   2. Does the patient/study population match the intended use?
   3. Does it use any ‘re-calibration’ or ‘compensation’ technique(s) to adapt input data to that of a target demographic setting?
3. Bias
   1. What kinds of bias might exist in your data?
   2. What have you done to evaluate your data for bias, and how does it affect your model?
   3. What are possible risks due to bias and what is your mitigation plan?
   4. What residual bias might remain and how should users take this into account?
4. Quality
   1. What factors affect data quality and what has been put in place to control that quality?
   2. Has the data been evaluated for data completeness? (in terms of ranges, variations, outliers, missing values, etc.)? Has the data been reviewed to remove duplicate entries?
   3. How will different hardware platforms affect the quality of the training and use data, and how has this been accounted for? (For example, different model cell phones have different resolution cameras) What kind of normalization technique is used in case of data capture from different hardware specifications?
   4. What is the method of ground truth determination? What type of data classification & data labeling processes were followed? (e.g. (i) manual, (ii) semi-automated or (iii) fully-automated), Define the existing ‘gold/reference/benchmark standard’ being followed?

*Note: A glossary for filling out the data description as well as an example can be found in Appendix 2 (section 6.2).*

### Risk Management

This list of questions concerns matters of risk management for the quality assessment of data sets.

1. What kind of risk management process is used to identify and mitigate potential data-related risks? (data risk management may take into account (a) ‘risk severity classification ‘ (e.g. minor, moderate, major), (b) risk assessment criteria. (For example, how to assess the data related processes like data collection, data curation, data modeling, etc follow standardized and regulated procedures to safeguard data integrity and data usability?)
2. What controls are in place to ensure the patient is properly matched with the training data? (For example, by doing a comparative analysis of training data with ‘target class or population’ specific data of types- demographic, epidemiological, clinical)

### Security/Privacy

This section is used to determine what, if any, data security and privacy concerns are relevant concerning the assessment of data quality.

1. Does it manage patient-personally identifiable information (PII)?
2. If yes, what controls have been put in place to ensure
   1. Data Confidentiality? (Does it follow any data anonymization/pseudo-anonymization or data de-identification process?, Does it comply with applicable regulations such as GDPR, US HIPAA, or other country-specific healthcare privacy regulations?)
   2. Data Integrity? Does it follow any data integrity mechanism (for example RAID, mirroring, checksum, digital signature, block chain)?
   3. Data Security: Is the data hosted at secured data centre(s)? Is the data transfer protected by any encryption mechanism? Does it have role based access control? Does it have data backup and data recovery plan?
   4. Data Availability? (Is there any legal obstacle in sharing the data with the FG-AI4H in terms of data protection, privacy laws, copyright?)

### Dependencies/Interfaces

Some datasets work as standalone items, others connect to external systems or interface with other databases. These interfaces can affect data interoperability and data exchange/porting.

1. Does the system use any data interface (for example APIs/web services, mapping tools) in order to connect with other databases/systems (for example third party validation/test databases, open databases, other domain specific databases) for data exchange/porting/migration?
2. Does the data interfaces used ensure data porting compatibility in terms of data formats, data structure, data encoding/decoding, data value ranges, data resolution?

### Verification/Validation/Testing

This section is used to document the verification and validation processes for data quality assessment as part of FG-AI4H Data Acceptance Policy.

1. How will the data quality be verified?
   1. Describe the software test plan & procedure for data quality assessment
      1. Test plan and procedure to cover the following aspects:
         1. Adherence to the FG-AI4H Data Acceptance Policy
         2. Compliance with respect to data integrity, data completeness, and data bias.
         3. Data access control with respect to authentication, authorization, monitoring, logging and auditing
         4. Data security/privacy compliance
         5. Ethical compliance
   2. How do you know that there are no significant software bugs?
2. What limitations are discovered from data quality test report?
3. What results from data quality test report are unexpected?

### Change Management

This section is used to determine how changes with respect to data quality are handled.

1. Describe the controls in place for traditional software tools and protocol changes with respect to data sensing, data representation and storage, data processing, data encoding and transmission – for example feature updates, bug fixes, version control, data migration to and validation of target databases.
2. Describe the techniques and tools to be used for configuration of new training and test data sets, aggregation of historical training data. Are there mechanisms in place to help debug data quality related issues that may be found?
3. Healthcare delivery changes over time – hospital populations change, clinical protocols change. What monitoring is required to make suitable adaptations with respect to data quality assessment procedures in anticipation of these changes?

## AI Solution Quality Assessment

This subsection lists concrete questions to determine the quality of AI solution in the six categories (1) algorithm description, (2) risk, (3) security/privacy, (4) interfaces/dependencies, (5) verification/validation/testing, and (6) change management.

### Algorithm Description

Describe the AI algorithm (or algorithms) that were used in this application.

1. Basics
   1. Describe the underlying task the AI algorithm is solving (for example classification, regression).
   2. What is the underlying AI algorithm used in the AI application?
   3. How does the input data look like?
   4. How does the output data look like?
   5. How does the target data look like?
   6. State the optimization objective(s) that were used during training.
   7. State the evaluation metric(s) that were used during training.
2. Safety by design
   1. List the safety tools that are used during training of the AI solution.
   2. List the safety tools that are to be used during deployment of the AI solution.

*Note: A glossary for filling out the algorithm description as well as an example can be found in Appendix 2 (section 6.2).*

### Risk Management

This section is used to discuss risk management activities and results.

1. What kind of risk management process is used to identify and mitigate potential risks?
2. What is the outcome of the benefit-risk analysis conducted? (Proposed outcomes may be compared with benchmarking parameters of benefits and risks. Benefit benchmarks may include improved patient outcomes, patient quality of life, practicality in use, and reduced medical system costs)
3. Does the application estimate the prediction error in terms of bias and variance errors?
4. How does the application ensure that the AI algorithm is not (a) under-fitting or (b) over-fitting the training data?
5. Does the application exhibit consistency with respect to AI algorithm behavior, metrics, accuracy and reproducibility of results?
6. How does the application ensure that domain relevant or representative dataset is used for testing the AI algorithm performance?
7. What all approval procedures are needed for acceptance of the application for clinical practice?
8. What unintended consequences have been identified and how will the system deal with them? How does the system deal with unknown variables and irreducible errors?
9. Risk should take into account the “reversibility” of a decision – if the product is wrong, can the patient recover or is there a permanent effect on the patient? How many clinician or specialist control/intervention levels or points are defined in the actual deployment model of the application?
10. An acceptable level of specificity may change due to the clinical consequences of misdiagnoses - might be okay to accept false-positives because a false-negative could have a higher risk. Therefore, there should be a discussion about why the confusion matrix is acceptable for the particular clinical use case.
11. Does the application aid the user in ‘how to properly interpret the result’?

### Security/Privacy

When discussing cybersecurity, the concepts of “Confidentiality, Integrity, and Availability”, also known as “CIA” often come up. A successful cybersecurity breach may impact one or more elements of CIA, and therefore measures should be taken to defend against potential CIA-related failures. Confidentiality is related to preserving restrictions on information access and disclosure; a hacker may try to gain access to confidential patient data. Integrity is guarding against improper information modification or destruction; this is of concern because corrupted data may lead to incorrect learning by an AI application. Availability is related to timely and reliable access to information; a hacker may attach a server or network that the AI application is using, and therefore the application is not able to perform its intended function – for some applications, un-availability could have a significant impact to patient safety.

This section is used to determine what, if any, security and privacy concerns are relevant to this project.

1. Does it manage patient personally identifiable information (PII)?
   1. If yes, what controls have been put in place to ensure
      1. Data confidentiality?
      2. Data integrity?
      3. Product availability?
2. Does the application employ any user safety risk control measures to identify and avert potential human safety risks caused by accidental or malicious misuse (cyber attacks) of the technology involved?
3. Does the application employ any application safety risk control measures to identify tampering, manipulation, theft, code corruption or any other unauthorized practices of the output of the AI algorithm?
4. Does the application employ any resilience measures to reduce the risk (negative impact) in the occurrence likelihood of 'false-positives' and human errors’?
5. Does the application follow any control baselines for system-level risk assessment?

### Dependencies/Interfaces

Some applications work as standalone items, others connect to external systems or interface with other products. These interfaces can affect performance.

1. What other application or systems does your application interface with?
2. Does this application depend on those interfaces for proper functioning?
3. Does the application has any (a) sensing and/or (b) actuation (c) control interfaces with other systems/subsystems?
4. Does the application make any assumptions on the selection of the AI algorithm based on the input dataset? Does it issue any transparent disclaimers on potential vulnerabilities, risks or biases of the particular algorithm selected?

### Verification/Validation/Testing

This section is used to document the verification and performance of the application.

1. How will this application be verified?
   1. Describe the application software test procedures
   2. How do you know that there are no significant software bugs?
2. How well does the AI algorithm perform?
3. What performance limitations are discovered?
4. What results are unexpected?
5. Would you be willing to share your metrics & be willing to have those numbers validated? Can you direct us to publications/whatever external peer review version of their results?
6. Is there a system in place to monitor user feedback? (e.g. “post-market surveillance”)
7. Describe the test plan and procedure for application quality assessment
   1. Test plan and procedure to cover the following aspects:
      1. Verification of standardized metrology(proper metrics) selection for AI algorithm in terms of performance, robustness, generalizability, explainability, interpretability and transparency
      2. Performance evaluation of the application under conditions that differ from the training set.
      3. Robustness assessment of application in the context of its migration from staging servers to production servers
      4. Verification of interface requirements with other systems
      5. Verification of AI application auditing procedures, risk mitigation and management procedures
      6. Verification of human factor/ usability features
      7. Verification of training requirements and target user skill sets

### Change Management

This section is used to determine how changes are handled.

1. Describe the controls in place for traditional application software changes – for example feature updates, bug fixes, version control.
2. Describe the controls in place for AI algorithm learning updates (for example unlock the system, add more training data, lock the system, hyper-parameter tuning).
3. Are there mechanisms in place to help debug application quality issues that may be found?
4. Healthcare delivery changes over time – hospital populations change, clinical protocols change, etc., what monitoring is required to make suitable adaptations with respect to application quality assessment procedures in anticipation of these changes?
5. Is the AI algorithm designed as static model or dynamic/incremental learning model? In case of dynamic learning model, what is the expected frequency of model update in terms of model re-training on new training dataset, model hyper parameter tuning , adaptation of generic model to target data model,?
6. How does the application deal with the performance evaluation of algorithms under a wide range of prospective clinical situations e.g. change in clinical specifications with respect to sensitivity and specificity, change in training data set?

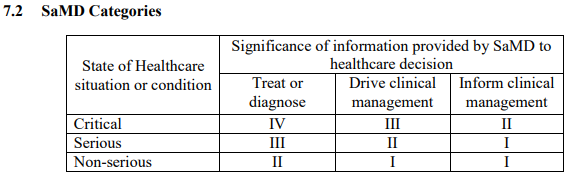
# Outlook

This document provides an overview on the status-quo of WG-DAISAM’s work in monitoring the concerns of national and international oversight organizations with respect to AI4H. We put forward a tentative group of quality dimensions accompanied by concrete assessment questions. As next steps the organizational and geographic coverage needs to be increased.

# Appendix

## Appendix 1 – IMDRF Classification Scheme

In 2014, the IMDRF published “Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations”, which assigned different risk levels to a product based on the state of the patient and the significance of the contribution that the software was making. The original risk table from that document is copied below:



However, being published in 2014, there was not much discussion that accounted for various levels of autonomy. For the purposes of this evaluation, additional levels have been added to the “Treat or diagnose” category, resulting in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Significance of information provided by software to healthcare decision | | | | |
| State of Healthcare situation or condition | Treat or diagnose w/no intervention possible | Treat or diagnose w/Override | Treat or diagnose w/Approval | Drive Clinical Management | Inform Clinical Management |
| Critical | VI | V | IV | III | II |
| Serious | V | IV | III | II | I |
| Non-serious | IV | III | II | I | I |

Three different levels of autonomy are proposed:

1. Approval: the software may make suggestions to the user, but either it cannot take action on its own, or it requires operator approval before taking action.
2. Override: the software can take action without approval, but the operator has the ability to over-ride (cancel) the software if need be. For example, a human driver in a self-driving car can take control.
3. No Intervention: the operator is not involved in the treatment and has no ability to override the software.

The IMDRF document had the following definitions:

**Critical situation or condition**

Situations or conditions where accurate and/or timely diagnosis or treatment action is vital to

avoid death, long-term disability or other serious deterioration of health of an individual patient

or to mitigating impact to public health. SaMD is considered to be used in a critical situation or

condition where:

• The type of disease or condition is:

o Life-threatening state of health, including incurable states,

o Requires major therapeutic interventions,

o Sometimes time critical, depending on the progression of the disease or condition that could affect the user’s ability to reflect on the output information.

• Intended target population is fragile with respect to the disease or condition (e.g., pediatrics, high risk population, etc.)

• Intended for specialized trained users.

**Serious situation or condition**

Situations or conditions where accurate diagnosis or treatment is of vital importance to avoid unnecessary interventions (e.g., biopsy) or timely interventions are important to mitigate long term irreversible consequences on an individual patient’s health condition or public health. SaMD is considered to be used in a serious situation or condition when:

• The type of disease or condition is:

o Moderate in progression, often curable,

o Does not require major therapeutic interventions,

o Intervention is normally not expected to be time critical in order to avoid death, long term disability or other serious deterioration of health, whereby providing the user an ability to detect erroneous recommendations.

• Intended target population is NOT fragile with respect to the disease or condition.

• Intended for either specialized trained users or lay users.

Note: SaMD intended to be used by lay users in a "serious situation or condition" as described here, without the support from specialized professionals, should be considered as SaMD used in a "critical situation or condition".

**Non-Serious situation or condition**

Situations or conditions where an accurate diagnosis and treatment is important but not critical for interventions to mitigate long term irreversible consequences on an individual patient's health condition or public health. SaMD is considered to be used in a non-serious situation or condition when:

• The type of disease or condition is:

o Slow with predictable progression of disease state (may include minor chronic illnesses or states),

o May not be curable; can be managed effectively,

o Requires only minor therapeutic interventions, and

o Interventions are normally noninvasive in nature, providing the user the ability to detect erroneous recommendations.

• Intended target population is individuals who may not always be patients.

• Intended for use by either specialized trained users or lay users.

**Inform clinical management**

Informing clinical management infers that the information provided by the SaMD will not trigger an immediate or near term action:

• To inform of options for treating, diagnosing, preventing, or mitigating a disease or condition.

• To provide clinical information by aggregating relevant information (e.g., disease, condition, drugs, medical devices, population, etc.)

**Drive clinical management**

Driving clinical management infers that the information provided by the SaMD will be used to aid in treatment, aid in diagnoses, to triage or identify early signs of a disease or condition will be used to guide next diagnostics or next treatment interventions:

• To aid in treatment by providing enhanced support to safe and effective use of medicinal products or a medical device.

• To aid in diagnosis by analyzing relevant information to help predict risk of a disease or condition or as an aid to making a definitive diagnosis.

• To triage or identify early signs of a disease or conditions.

**Treat or to diagnose**

Treating and diagnosing infers that the information provided by the SaMD will be used to take an immediate or near term action:

• To treat/prevent or mitigate by connecting to other medical devices, medicinal products, general purpose actuators or other means of providing therapy to a human body

• To diagnose/screen/detect a disease or condition (i.e., using sensors, data, or other information from other hardware or software devices, pertaining to a disease or condition).

## Appendix 2 – Quality Assessment Questionnaire & Glossary

This appendix consists of two parts – first, the quality assessment questionnaire and, second, the quality assessment questionnaire glossary

The quality assessment questionnaire serves as a preliminary checklist for data acceptance and is prepared in compliance with the data and AI solution quality assessment framework, governed by FG-AI4H Data Acceptance Policy. This questionnaire is intended to guide the various FG-AI4H Topic Groups in following a uniform procedure for preparing the data and AI solution specifications and submitting them in a common reporting format. Future revisions of this questionnaire template may involve extension and customization of quality assessment criteria by the FG-AI4H Topic Groups to suit their respective use cases. A sample data sheet for the Topic Group- Psychiatry is shown below.

The quality assessment questionnaire glossary contains definitions for technical terms specific to data and AI solution quality criteria. This is provided to guide the FG-AI4H Topic Groups in interpreting the quality assessment questionnaire in a clear and concise manner and in mapping the respective data and AI solution specifications.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TG-PSYCHIATRY** | | | | |
|  |  | **Specification** | **Specification reference**  *In what document and what section can the specification be found?* | **Comment** |
| **AI Model Basics** | Underlying task | Multi task Classification / Regression problem | FGAI4H-E-015, Section 1.1 | Workflow / Use Case- Automated assessment of Psychiatric Multimorbidity with the help of neuro-physiological measures. behavioral behavioral/cognitive, demographic measures |
| AI Model type | State-of-the-Art Multi-Class /Multi-Label Classifier Algorithms and /or Deep Learning approaches |  |  |
| Input data | **Data Source :**   1. Neuro-physiological Data –Resting EEG (raw, pre- processed & features) 2. Demographics Data (age, gender, etc) 3. Phenotypical / Behavioral data   Data (e.g., responses and outcomes of an intelligence scale) / Questionnaires  **Data Type:**  (a) Real Value  (b) Ordinal Value  (c) Categorical Value  **Data Format:**   1. Behavioral data: ‘.csv ‘ 2. Raw and preprocessed EEG data: ‘.mat’ and ‘.csv’   **Data Size:**  1600 samples (initial HBN release )  Future HBN releases: to be procured @ ~500 new samples per year  **Data Sharing Ethical Norms:**   1. Written informed consent obtained from subjects 2. Data shared in pseudonymized format | Pediatric Clinical EEG database -Healthy Brain Network –Bioban  ( c) Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., ... & Wigal, T. (2012). Categorical and Dimensional Definitions and Evaluations of Symptoms of ADHD: History of the SNAP and the SWAN Rating Scales. *The International journal of educational and psychological assessment*, *10*(1), 51 |  |
| Output data | **Data Format- ‘.**csv ‘  **Data Type:**  (a)Predicted Data Matrix-  N x D binary valued(1 / 0) matrix, where N is No. of test subjects/samples & D is No. of Psychiatric Disorders(as per DSM-V Diagnostic Categories)  (b)Continuous labels for all disorder types will be normalized to a common dynamic range (e.g. between 0 and 10). | - | The binary variable, coded as 1/0, indicates the presence or absence respectively of the d-th disorder in the n-th test subject as predicted by the model |
| Target data | **Data Format- ‘.**csv ‘  **Data Type:**  (a) Ground Truth Data Matrix-  N x D binary valued(1 / 0) matrix, where N is No. of test subjects/samples & D is No. of Psychiatric Disorders(as per DSM-V Diagnostic Categories)  (b)Continuous labels for all disorder types will be normalized to a common dynamic range (e.g. between 0 and 10). |  |  |
| AI Model Optimization objective(s) |  | - | - |
| AI Model Evaluation metric(s) | **I. Multi-task Classification Metrics**  (a)Primary Metric: Multi-task Accuracy ACC (*for multi-task classification problem*) defined as  where ‘Ytrue’ is the matrix of labels, where ‘Yn,dtrue’ is a binary variable (coded as 0/1) indicating the presence of the d-th disorder in the n-th test subject, and ‘Ypred’ is the corresponding binary matrix of disorder occurrences predicted by the model  (b)Secondary Metrics:  (b.1) Multi-task sensitivity and specificity  (b.2) Accuracy, Sensitivity and Specificity separately for each disorder type  **II.Multi-task Regression Metrics**  Primary Metric: Average of Squared Prediction Errors across samples and disorders  Secondary Metric: Single-Disorder Mean-Squared Errors in the continuous label case. | - | - |
|  | | | | |
| **AI Model – Safety Tools** | Safety tool(s)training |  | - | - |
| Safety tool(s) deployment |  | - | - |
|  | | | | |
| **Test data** | Quality tests for test data | Data Quality Assurance protocols include:   1. Reduced Training Label Errors achieved by consensus clinical diagnosis assessed for each child based on the decision of a clinical team with support of interviews and materials conducted as basis for the DSM-5 consensus diagnosis 2. All the tests were conducted by licensed clinicians 3. All test scores from clinical interviews are double entered into the database by two (different) trained research assistants | - | (b)The clinical staff consists of a combination of psychologists and social workers, with psychopharmacological consultation support provided by psychiatrists |

**Quality Assessment Questionnaire – Glossary**

|  |  |  |  |
| --- | --- | --- | --- |
| **AI Model Development Workflow** | **Assessment Criteria** | **Description** | **Examples** |
| Problem Definition | Underlying Task | Underlying Task refers to the broad taxonomy followed in organizing Machine Learning (ML) Tasks based on how the solution will be applied to solve or address the specific business problem of the respective practice domain use cases. Please refer to sections- Level-1A and Level-1B of FGAI4H-C-104 for domain use-case thematic classifications) | * Classification * Regression/Prediction * Clustering * Association rule learning * Decision Support / Virtual Assistance / Recommendation systems * Matching * Labeling * Detection * Segmentation * Sequential data models * Anomaly detection and Fraud Prevention * Compliance Monitoring / Quality Assurance * Process optimization / Automation * Other |
| Data Preparation | Input Data Sources, Types & Formats | Input Data refers to the subset of the dataset that is used to train the AI model  Data Type refers to the type of the different data attributes involved  Data Format refers to the standard representation formats of the different data attributes involved | Input Data Sources include:   * Electronic Health Records(Anonymised) * Medical Images * Vital signs signals * Lab test results * Photographs * Non medical data-Socioeconomic, Environmental, etc) * Questionnaire responses * Free Text (Discharge / Summary, Medical History / Notes, etc.) * Other   Input Data Types include:   * Real valued * Integer-valued * Categorical value * Ordinal value * Strings * Dates * Times * Complex data type * Other   Standard Input Data Formats include:   * DICOM PS3.0 (latest versions)- for Diagnostic Image ( X-Ray, CT,MRI, PET, other pathological slides, etc) * JPEG / PNJ – for Static Image * MP3 / OGG – for Audio: * MP4 / MOV- for Video * SNOMED – for clinical observations/terminology * LOINC- for laboratory observations * WHO ICD-10 for disease classifications * RxNORM for Medication Code * Other |
| Data Preparation | Output Data Types | Output Data refers to type of data generated by the AI Model, when a particular ML algorithm is applied on the Input Data | * Binary/Class output (0 or 1) as in case of classification problems * Probability output(0-1) as in case of classification problems * Continuous valued output as in case of regression problems |
| Data Preparation | Target Data Types | Target Data refers to the output data in the training dataset that is defined as the reference (ground truth) for AI Model validation/testing | * Binary/Class output (0 or 1) as in case of classification problems * Probability output(0-1) as in case of classification problems * Continuous valued output as in case of regression problems |
| AI Model Selection | Model Type | Model Type refers to the specific machine learning algorithm and its configuration that is applied on the training dataset in order to learn the Model | Broad Classification of ML Algorithms include :   * Supervised Learning based algos * Linear Regression * Logistic Regression * k-nearest neighbors * Decision Trees * Random Forest * Gradient Boosting Machines * XGBoost * Support Vector Machines (SVM) * Neural Network * other * Unsupervised Learning based algos * k means clustering * Hierarchical clustering * Neural Network * other * Reinforcement Learning based algos * Association rule learning based algos * Apriori algorithm * Eclat algorithm * Deep learning based algos * Convolutional Neural Network (CNN) * Recurrent Neural Networks (RNNs) * Long Short-Term Memory Networks (LSTMs) * Stacked Auto-Encoders * Deep Boltzmann Machine (DBM) * Deep Belief Networks (DBN) * other |
| AI Model Evaluation | Evaluation Metrics | Metrics used to quantify the errors and to evaluate the performance quality of the trained model on the test dataset  Selection of metrics depend on the type of the problem & the type of the model under consideration | * Model Accuracy (%) * Model Accuracy -Mean & Standard Deviation * Model Accuracy –Box Plot Summarization * Root Mean Squared Error(RMSE) * Sensitivity (True Positive Rate) * Specificity (True Negative Rate) * F1-Score (class wise performance determination) * Confusion matrix * K-fold Cross-validation * Gain and Lift Charts * Kolmogorov Smirnov Chart * Gini Coefficient * Log [Loss](https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss) * [Area under the ROC curve (AUC)](https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc) * Concordant – Discordant Ratio * Other user defined performance measures * Other |
| AI Model Optimization | Optimization Objective(s) | This deals with the iterative process (feedback principle) of reconfiguring or tweaking the Model Parameters to their optimal values in order to achieve the desired level of accuracy or performance score in comparison with the baseline definition  Model performance can be systematically tracked by maintaining progressive versions of Code, Model, and Data | Optimization techniques include:   * Adding or deleting Features /Attributes of the input data * Aggregating or Decomposing Features /Attributes of the input data * Tuning Model Hyper-parameters * Normalization & Standardization of input data * Changing the learning rate of the algorithm * Examining the Statistical Significance of results * Recruiting Ensemble Methods for combining / augmenting the prediction scores of multiple models * Monitoring and tracking API response times and Computational Memory requirements of the serving infrastructure * etc |
| Safety Standards Compliance | Safety tool(s)training | This deals with the user training/orientation given on how to identify potential human safety risks occurring due to accidental or malicious misuse of the technology involved in AI Model deployment | Safety Risk Mitigation and Management Plan & Procedure |
| Safety tool(s) deployment | This deals with the incorporation of necessary preventative system measures/tools as per the defined Risk Mitigation Plan to ensure that no damage or harm is caused to human safety out of potential physical or cyber attacks on the AI Model being applied | * Adopting governance procedures to assert alternative system fault tolerance plans * Adopting security mechanisms like   + Authentication   + Role based Access Control   + Encryption   + Transport Level Security   + Informed Consent   + Anonymisation   + etc * Maintaining Data Audit Logs for secure content verification, based on   + Blockchain Technology   + Merkle Trees   + etc * Implementing Security Standards based on Digital Certificate, SSL, SHA-256, etc |
| AI Model Testing | Test Data Quality Tests | Test Data refers to the subset of the dataset and not part of the training dataset that is used to evaluate the ML Model accuracy after its primary vetting by the validation dataset  Quality tests are performed to minimize the noise and variance of the test data in order to maximize the performance accuracy of ML algorithm applied on it | Standard Test Options include:   * Training and testing on the same dataset * Split tests * Multiple split tests * Cross validation * Multiple cross validation * Statistical significance |

# References

[1] Xavier Health Organization. (2018). Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare*.* Retrieved October  
8, 2019, from [*https://www.advamed.org/sites/default/files/resource/perspectives\_and\_good\_practices\_for\_ai\_and\_continuous\_learning*](https://www.advamed.org/sites/default/files/resource/perspectives_and_good_practices_for_ai_and_continuous_learning)[*\_  
systems\_in*](https://www.advamed.org/sites/default/files/resource/perspectives_and_good_practices_for_ai_and_continuous_learning_systems_in)[*\_healthcare.pdf*](https://www.advamed.org/sites/default/files/resource/perspectives_and_good_practices_for_ai_and_continuous_learning_systems_in_healthcare.pdf)

[2] Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranshaert, E., Jaremko, J. L., ... & van den Hoven van Genderen, R. (2019). Ethics of artificial   
intelligence in radiology: summary of the joint European and North American multisociety statement. *Radiology*, 191586. doi: <https://doi.org/10.1016/j.jacr.2019.07.028>

[3] U.S. Food & Drug Administration. (2018). Use of Public Human Genetic Variant Databases to Support Clinical Validity for Genetic and Genomic-  
Based In Vitro Diagnostics - Guidance for Stakeholders and Food and Drug Administration Staff. Retrieved October 8, 2019, from <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-public-human-genetic-variant-databases-support-clinical-validity-genetic-and-genomic-based-vitro>

[4] Microsoft. (2018). The Future Computed – Artificial Intelligence and its role in society. Retrieved October 8, 2019, from   
<https://news.microsoft.com/cloudforgood/_media/downloads/the-future-computed-english.pdf>

[5] Center for Medical Device Evaluation. (2019). Review Highlights of Aided Decision-making Medical Device Software Using Deep Learning   
Technique (Exposure Draft). Retrieved October 8, 2019, from <https://www.cmde.org.cn/CL0004/18639.html>

[6] International Medical Device Regulators Forum. (2014). Software as a Medical Device: Possible Framework for Risk Categorization and   
Corresponding Considerations. Retrieved October 8, 2019, from <http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf>

# Acknowledgments

* Icons in *Illustration 1* are by [Dave Gandy](https://www.flaticon.com/authors/dave-gandy) and [freepik](https://www.flaticon.com/authors/freepik) from flaticon.com

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

1. Detailed information on the organizational structure and other work of FG-AI4H can be found in the online collaboration environment at *https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx* [↑](#footnote-ref-1)