|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-F-032 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | DAISAM | Zanzibar, 3-5 September 2019 |
| **DOCUMENT** |
| **Source:** | WG-DAISAM Chair |
| **Title:** | Quality Criteria: Regulatory Synthesis and TG-Questionnaire |
| **Purpose:** | Discussion |
| **Contact:** | Pat BairdPhilips, USA | Tel: +1 262 239-9321Email: pat.baird@philips.com  |
| **Contact:** | Luis OalaFraunhofer HHI, Germany | Tel: +49 30 31002-104 Email: luis.oala@hhi.fraunhofer.de  |
| **Contact:** | Pradeep BalachandranIndia | Email: pbn.tvm@gmail.com  |

|  |  |
| --- | --- |
| **Abstract:** | In this document we outline the preliminary results and current working packages of the working group on Data and AI solution assessment methods (WG-DAISAM). They comprise (1) the compilation of quality dimensions based on the review of regulatory documents, (2) an analysis of the TG types along with a questionnaire template to fill gaps in the TG TDDs relevant to our WG, and finally (3) a collection of questions we encountered in our work with respect to the FG scope that require negotiation with the overall FG. |

[1 Regulation-driven quality principles 2](#_Toc17823017)

[1.1 Introduction / Purpose 2](#_Toc17823018)

[1.2 Description 2](#_Toc17823019)

[1.3 IMDRF Classification 2](#_Toc17823020)

[1.4 Data Quality 2](#_Toc17823021)

[1.5 Risk 3](#_Toc17823022)

[1.6 Security/Privacy 3](#_Toc17823023)

[1.7 Interfaces, Dependencies 3](#_Toc17823024)

[1.8 Verification, Validation, Testing 3](#_Toc17823025)

[1.9 Change Management 4](#_Toc17823026)

[2 TG-Questionnaire 5](#_Toc17823027)

[2.1 Questionnaire 5](#_Toc17823028)

[2.3 Example: TG-Psy 5](#_Toc17823029)

[2.2 Glossary 10](#_Toc17823030)

[3 Appendix 16](#_Toc17823031)

[3.1 Annex 1 – IMDRF Classification Scheme 16](#_Toc17823032)

[3.2 Open items to resolve with FG 19](#_Toc17823033)

# 1 Regulation-driven quality principles

## 1.1 Introduction / Purpose

As part of our effort to come up with meaningful quality dimensions we continuously review regulatory and guidance documents on this topic. This first iteration included documents from the Xavier Health AI Summit, Radiology Ethics, FDA Genome IVD, MS-Future, the IMDRF and a very recent Chinese guidance document on the design of aided decision-making medical device software using deep learning techniques. The review resulted in an assessment checklist containing more than 150 entries. We grouped, sorted out and synthesized these entries into the question groups that can be found below.

The purpose of this document is to provide a quality assessment framework for the evaluation of data and AI solutions.

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

## 1.2 Description

This section provides an overview of the product being evaluated.

1. What is the intended use of this product ?
2. Does it diagnose, treat, mitigate, prevent, or monitor a disease, disorder, or injury? Does it support operations or assist in managing public health issues? *This could be re-worked into a series of check-boxes if that would be easier – one box for diagnose, one box for treat, … etc.*
3. Is it a locked algorithm or an adaptive algorithm?
	1. If adaptive, what is the frequency or the trigger for an update (e.g. does it update after every use, or does it update every 6 months) ?
	2. If adaptive, are there boundaries around the amount of allowed autonomous learning?
4. Describe the principles of operation of the product?
5. How is the design appropriate for its use (e.g. why did you choose neural network?)
6. What contra-indications, warnings, and precautions have been identified?
7. How is this product helping healthcare? What critical aspects support that intended use. E.g. "Triage software, to be successful, needs to have a low rate of false-negative results."

## 1.3 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.

Annex 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?
2. What is the worst-case state of healthcare situation or condition?
3. What class (I – VI) does this product belong?

## 1.4 Data Quality

This section is used to determine the quality of the training data.

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
	5. Validity – are the values correct? Are they up to date? Describe the protocol 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. What demographic does this serve? Does the patient/study population match the intended use?
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?
	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 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 ? (e.g. different model cell phones have different resolution cameras)
5. What was the method of ground truth determination?

## 1.5 Risk

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

1. Has a risk management process been used to identify and mitigate potential risks?
2. Has a benefit-risk analysis been completed?
3. What controls are in place to ensure the patient is properly matched with the training data (e.g. a 12 year old patient probably doesn’t have breast cancer..) ?
4. What unintended consequences have been identified and how will the system deal with them?
5. 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?
6. 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.

## 1.6 Security/Privacy

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)?
2. If yes, what controls have been put in place to ensure
	1. Data confidentiality ?
	2. Data integrity ?
	3. Product availability ?

## 1.7 Interfaces, Dependencies

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

1. What other products or systems does this interface with ?
2. Does this product depend on those interfaces for proper functioning?

## 1.8 Verification, Validation, Testing

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

1. How will this product be verified?
	1. Describe the software test procedures
	2. How do you know that there are no significant software bugs?
2. How well did the model perform?
3. What performance limitations were discovered?
4. What results were 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. “postmarket surveillance”)

## 1.9 Change Management

This section is used to determine how changes are handled.

* 1. Describe the controls in place for traditional software changes – e.g. feature updates, bug fixes, version control, etc.
	2. Describe the controls in place for learning updates (e.g. unlock the system, add more training data, lock the system.)
	3. Are there features in place to help debug issues that may be found?
	4. Healthcare delivery changes over time – hospital populations change, clinical protocols change, etc., what monitoring is required for these changes?

# 2 TG-Questionnaire

The initial work of WG-DAISAM comprised a review of all TG documents to obtain a better overview on the types of tasks, data and models represented in the FG use cases. The motivation behind this approach is our strategy to converge to a catalogue of quality criteria by considering both: general criteria assuming black-box models as well as concrete criteria based on the actual learning systems employed in the TGs.

Since we noticed that many TDDs are far from complete we selected the TDD categories that we need for our work and made them into a questionnaire. The template as well as a preliminary glossary and dry run on the TG-Psy, for illustration purposes, can be found below. We plan to send this questionnaire out to the TGs after meeting F. The results may also be used by the TG to complete their TDDs.

## 2.1 Questionnaire

|  |
| --- |
| **Template** |
|  |  | **Specification** | **Specification reference***In what document and what section can the specification be found?* | **Comment** |
| **Model - Basics** | Underlying task |  |  |  |
| Model type |  |  |  |
| Input data |  |  |  |
| Output data |  |  |  |
| Target data |  |  |  |
| Optimization objective(s) |  |  |  |
| Evaluation metric(s) |  |  |  |
|  |
| **Model – Safety Tools** | Safety tool(s) training |  |  |  |
| Safety tool(s) deployment |  |  |  |
|  |
| **Test data** | Quality tests for test data |  |  |  |

## 2.3 Example: TG-Psy

|  |
| --- |
| **TG-PSYCHIATRY** |
|  |  | **Specification** | **Specification reference***In what document and what section can the specification be found?* | **Comment** |
| **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 |
| 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:** 1. 1600 samples (initial HBN release )
2. 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
 | 1. Pediatric Clinical EEG database -Healthy Brain Network –Biobank

( 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). |  |  |
| Optimization objective(s) |  | - | - |
| 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.
 | - | - |
|  |
| **Model – Safety Tools** | Safety tool(s)training |  | - | - |
| Safety tool(s) deployment |  | - | - |
|  |
| **Test data** | Quality tests for test data | Data Quality Assurance protocols include:* 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
* All the tests were conducted by licensed clinicians
* 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 |

## 2.2 Glossary

|  |  |  |  |
| --- | --- | --- | --- |
| **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 ML 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 and Output 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
 |
| 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
 |
| 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
 |
| 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
 |
| 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
 |

# 3 Appendix

## 3.1 Annex 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:

Table x – SAMD Categories

| State of healthcare situation or condition | Significance of information provided by SaMD to healthcare decision |
| --- | --- |
| Treat or diagnose | Drive clinical management | Inform clinical management |
| Critical | IV | III | II |
| Serious | III | II | I |
| Non-Serious | II | I | I |

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., paediatrics, 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 non-invasive 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 analysing 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).

## 3.2 Open items to resolve with FG

The following questions were identified as needing further discussion; they were raised with the steering committee but never resolved.

Some candidate questions relate to content that would probably be required for regulatory approval (varies by country) but we are not sure if they are applicable to the TDDs. The table below shows those questions; we would appreciate feedback if these questions are in or out of scope.

| ID | Category | Question | Keep or Delete? |
| --- | --- | --- | --- |
| 1 | 1 - Description | What is the scope of the AI app? Global/Corporate? Local/site? Individual? Group? |  |
| 2 | 1 - Description | What are computer system hardware specifications? Processor (manufacturer, speed, and features), RAM (memory size), hard disk size, other storage, communication, display interface, sensors, energy sources, safety features, etc. |  |
| 3 | 1 - Description | What are computer system software specifications? Include name, specific version level, and complete list of patches as warranted. |  |
| 4 | 1 - Description | Provide an architecture design chart depending on Level of Concern |  |
| 5 | 1 - Description | Is there a mobile app associated? |  |
| 6 | 1 - Description | Does the AI use wireless communication of any form? |  |
| 7 | 1 - Description | Provide a diagram of the layout and interconnections. Any interfaces are to be labelled. Interconnections which transmit compressed data should be labelled along with compression ratios. |  |
| 8 | 3 - Risk | Where application, FDA also intends to include such information in the Summary of Safety and Effectiveness Data (SSED) and to flag post-market studies that are a condition of approval for the device on our website. *Pat's thought - for a Continuous Learning System, could be consider having a "Continuous SSED" ? Do we create a report that monitors performance over time?* |  |
| 9 | 3 - Risk | Some AI applications may have a difficult time demonstrating benefit-risk; I suggest that the bulleted list on page 11-12 could be used as further guidance (e.g. "... the extent of the benefits of the device, including type, magnitude, probability, duration, and frequency of those benefits". ) However, I don't want to take up an entire page of this spreadsheet to document everything there, so this is just a placeholder than when it is time to explore benefit-risk in depth, we should look at this document.  |  |
| 10a | 4 - Security/Privacy | Since security/privacy rules vary by country, do we want to ask security/privacy questions? |  |
| 10b | 4 - Security/Privacy | Consider Adversarial Poisoning - bad guy contaminating the data |  |
| 11 | 6 - Verification / Validation / Testing | Provide traceability to link together design, implementation, testing, and risk management |  |
| 12 | 6 - Verification / Validation / Testing | For predictive models, developers must evaluate how far in advance the algorithm identifies positive cases. |  |
| 13 | 6 - Verification / Validation / Testing | Autonomy: Proposers should describe: simulation environment (validation??), types of missions (intended use??), decision policies and mission tasks to be learned, types of explanations to be provided, and user decisions to be supported. |  |
| 14 | 6 - Verification / Validation / Testing | Concept of Learning Rates - take a look at accuracy of model (vs test set) as you are training it -- look at how correct the AI is (low rate = bad) and look at graph to see if system is still learning or if it's pretty much peaked. |  |
| 15 | 6 - Verification / Validation / Testing | There is a suggestion that could be made more generic -- in addition to creating a confusion matrix for the overall test set, also create matrix for subsets (e.g. compare confusion for population as a whole vs. confusion for male populations and female populations) - this technique might be a way to find some hidden bias or errors in the data. |  |
| 16 | 7 - Change Management | What are the reported problems (bug list) and how are they relevant to the application? |  |
| 17 | 7 - Change Management | How is the AI retired? |  |
| 18 | 7 - Change Management | SOPs should define how variant information is aggregated, curated, and evaluated. These SOPs should be documented and versioned…. To maintain quality variant assertions… operations and SOPs should be reviewed at least annually. |  |
| 19 | 7 - Change Management | For CLS, should we tell users that there is an update? Do users have the right to reject or rollback the version in CLS systems? |  |
| 20 | 7 - Change Management | Need to have traceability for what data was used when - this is needed for root cause analysis and support of corrective actions when performance fails -- need to know what training set version was in use. |  |
| 21 | 7 - Change Management | Will we need to do anything different in our post-market plans? E.g. human judgment will be key to identifying potential blind spots and biases - is current PM processes sufficient for this |  |
| 22 | 7 - Change Management | Perhaps for high autonomy systems (or high-risk systems), we require this sort of anomaly detector / active watchdog |  |
| 23 | 8 - Explainability/Trust | How is traceability to the final answer maintained? |  |
| 24 | 8 - Explainability/Trust | Says that GDPR states that individual has the right to an explanation of how the automated decision was arrived at, and the consequence of that decision. This has been interpreted to mean that AI decisions should be able to be rationalized in human-understandable terms.  |  |
| 25 | 8 - Explainability/Trust | Patients should be made aware of: 1. the ways in which humans oversee the decisions made by ai, 2. the controls in place to assess, validate, and monitor the AI tools |  |
| 26 | 98 - Other | Has appropriate labelling been drafted? |  |

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