|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-D-021 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | N/A | Shanghai, 2-5 April 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | Chongqing University | | |
| **Title:** | | Proposal to add new contents into [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf): FG-AI4H Data handling and data acceptance | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Daidi Zhong Chongqing University P.R.China | | Tel: +86-13696454858 E-mail: [daidi.zhong@hotmail.com](mailto:daidi.zhong@hotmail.com) |

|  |  |
| --- | --- |
| **Abstract:** | This document suggests some modifications and additions to the FG-AI4H Data handling and data acceptance policy in document [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf). The purpose is to further help the data users and regulators to better understand the nature, origin and quality of the data leveraged by the very AI product. |

# Conformance to health data standards

The Clause 3.1 of original [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf) document contains the following requirement -- “Output variables are characterised, if they exist”. This clause aims to provide a **mandatory** requirement on the representation of the dataset, but it doesn’t make the reference to existing health data standards sufficiently explicit.

Meanwhile, the 6th paragraph of Clause 3.3 contains the sentence “Do the data conform to any standards for health data?”. This is an **optional** requirement to the dataset suppliers.

The above two occurrences seem to be conflict to each other.

The original wording in Clause 3.1 mentioned ICD, which is a well-known example of various health data standards that had been created by the clinician community or medical device industry. Similar standards include, but are not limited to, SNOMED, LOINC, 11073-10101, etc. These health data standards represent the well-established domain knowledge shared by the professionals; and they also are the common languages used ultimately by our stakeholders. Behind these standards, much efforts had been put together in order to clearly define or distinguish various semantics used in the field.

Therefore, it would be wise if FG-AI4H could directly leverage such standardized data infrastructure and build our standards on top them. Based on this motivation, we hereby **propose to refine the wording in order to make the reference towards health data standard more explicit**. And such requirement shall be placed in either Mandatory (Clause 3,1) or Conditional (Clause 3.2), but not Recommended (Clause 3.3).

In addition, the same logic applies to input data as well.

A sample wording, subjected to group discussion, is provided as below: “Input and output variables are characterized (possibly with codes, classifications, triage tags, pixel or voxel labels, annotations, etc.). Whenever applicable, such characterization shall be conformed to existing health data standards, with the understanding that local or regional extension, restriction, profiling or adaption may be applied.”

# Handling missing or incomplete data

In the Clause 3.1 of original [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf) document, after the sentence “Any data (pre-)processing methods are explained”, a possible sub-scenario was listed. This is a **Mandatory** requirement.

Meanwhile, the 5th paragraph of Clause 3.3 contains the following sentence -- “how missing or incomplete data have been treated if the case?”. This is a **Recommended** requirement.

However, the 2nd sub-scenario mentioned above occur in the field frequently, and it is equally important as the sub-scenario listed in the 1st bullet.

For example, in certain remote monitoring scenario, a series of data without synchronized timeline basis might still be able to provide some useful hints for diagnosis/prevention purpose. In some other scenarios, the monitoring devices or data collectors are not able to fully interpret or label the data they obtained, but they did annotate such exceptions with proper tags. Therefore, data that is useless for user A might still be useful or partially useful to user B. Such details can be described when applicable.

Therefore, **we propose to move the 2nd sub-scenario from Clause 3.3 into Clause 3.1, in order to make it mandatory.** In addition, I propose to change the wording to -- “How the missing, doubtable, incomplete or NULL but partially-useful data are treated?”

# Data creation and validation

The Clause 3.1 of original [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf) document contains following content – “Who has created the labels/ground truths? Who has assessed the data, e.g. with respect to quality?”. This is a **Mandatory** requirement.

The 1st paragraph of Clause 3.3 contains the following sentence – “Data and annotations/labels/ output variables have been validated by an independent domain expert/specialist in terms of quality and suitability”. This is a **Recommended** requirement.

The above two occurrences seem to be conflict to each other. The concept of the 2nd requirement has already been included by the 1st requirement.

Therefore, **we propose to move the 2nd requirement from Clause 3.3 into Clause 3.1, in order to make it mandatory.**

# Checklist as Appendix

The original [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf) document contains multiple mandatory and optional requirements on how to describe a dataset when submitting it to ITU. However, the current layout of [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf) is not sufficiently intuitive for data suppliers to act on.

To make it easier to use and implement, **we propose to define a checklist based on current contents, with corresponding identifiers** (Mandatory, Optional, Conditional, etc.). Such checklist can be added as an Appendix of [FGAI4H-C-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-C-103.pdf). This may also help us to identify potential issues like the above three items.

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