INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2017-2020

FG-AI4H-B-105

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plenary	New York, 15-16 November 2018

DOCUMENT

Source:	FG-AI4H
Title:	Draft FG-AI4H data handling policy
Purpose:	Discussion

Contact:	Marc Lecoultre - Editor MLLab.ai Switzerland	Tel: +41 79 321 09 29 Fax: +41 22 364 30 69 Email: ml@mllab.ai

Abstract: This document contains a preliminary outline of the data handling policy for the FG-AI4H work, as approved at the second FG-AI4H meeting (New York, 15-16 November 2018).

Health data is one of the most valuable and sensitive types of data. Handling this kind of data is often associated with a strict and factual framework defined by data protection laws. It is important to set a strict data policy which will ensure confidence in FG-AI4H not only among contributors, but across all stakeholders. There are two major issues that the data handling policy should address:

a) compliance with regulations dealing with the use of personal health data; and

b) non-disclosure of the *Test data* held by FG-AI4H for the purpose of model evaluation.

1. Definitions

In this document, we refer to different types of datasets. To clarify, we suggest the following definitions:

Received data: Any dataset sent by a trusted source (tbd) and received by FG-AI4H;

Example data: Subset of the *Received data* that is made public by FG-AI4H to help submitting organizations to train their models;

Test data: Two possibilities: 1) Corresponds to the remaining *Received data* after removing *Example data.* In a simpler way, it is *Received data - Example data*. 2) This set could be submitted by a different trusted source. This set is kept strictly private to evaluate submitted models.

2. Introduction

Understanding the importance of data to our initiative and how that information is handled reflects our commitment as a secure organization. The purpose of a data handling policy is to ensure that all sensitively data is confidentially controlled whether being transmitted within the organization or to a trusted third party.

When handling data, all users should be in accordance with and be responsible for adherence to the Data Handling Policy. Periodic auditing of adherence to this policy shall be the responsibility of one Information Governance Team.

Data should be handled in the context of a multi-tiered security system that safeguards patient data according to government statute and regulations. Data should be hosted in secured data centres.

The system shall comply with all applicable regulations over the targeted countries (EU regulations, GDPR, US HIPAA, individual countries healthcare privacy regulations, etc.). Regulations include information security, privacy and quality laws, guidelines and standards. We should design a regulatory compliance framework to ensure conformance with these regulations.

3. Legal context

Where national data protection laws may differ significantly, it is important to cover the most restrictive matters to allow the greatest number of entities to share their datasets. This includes data security, anonymization, access control and many other discussed in this document.

4. Data security

The infrastructure for data storage and processing should be based on state of the art security policies, practices and located in a secure location. Information should be securely received, stored and transferred. Encrypted transmission of datasets and encryption at rest (data stored encrypted) are among many other requirements. Only well-established and approved by FG-AI4H transfer methods should be used (tbd).

Where possible, data transfers should be carried out, using existing, protected and trusted networks (internal to FG-AI4H or over VPN: Virtual Private Network with dedicated IPSec & SSL encrypted channels). However, there may be occasions where data will need to be transferred via other networks such as Internet or any other open networks. On these occasions, the data files should be protected by encryption to prevent usage by unauthorised parties.

In case of a physical data transfer, e.g. USB or hard disks, all data should be securely stored in an encrypted format using an approved by FG-AI4H method. Transfers of data in hard copy format should be protected, using such methods as approved secure couriers.

5. Data Integrity

Data integrity should be enforced when the data travels from one component to other using checksum mechanisms that guarantee that the data has not been corrupted or modified. Any data files transferred or generated should be digitally signed and the data integrity of the payload should be validated at the edge of the network prior to storing the data in the database. This would ensure validation of data integrity of all raw and interpreted patient data.

Any corrupt data (inaccurate or incomplete) should either be rejected by the system or removed from it.

The security & privacy architecture should be designed to ensure a high level of data integrity and privacy for Protected Health Information in compliance with GDPR, US HIPAA, or any other participating country healthcare privacy, security, and quality regulations. This may be dependent on where the data was transferred from, where the data will be processed and by which entity.

6. Access control

Authorised stakeholders need to access the data for their own defined purpose and infrastructure administrators for maintenance. The receiving parties such as the Working Groups should evaluate and work on the datasets. The organizations that are willing to submit their algorithms need to access the *Example data* to develop their models. To guarantee absolute fairness among submitting organizations and ensure the credibility of the Focus Group, the *Test data* (undisclosed) should remain undisclosed.

Clear access control should be defined and a database with detailed access rights policies should be implemented.

The system should authenticate users before any access to the system and its resources. The system should support standard authentication technique that can verify the identity claimed by the user (Claims based, Federated authentication…)

Everyone willing to submit an algorithm should have access to the *Example data*. The only restriction might be for the party submitting the *Test data*.

7. Auditing / Logging

All transactions should be authenticated, authorized, monitored, and logged and audited regularly to detect unauthorized events. The system should detect events that can affect the confidentiality of personal health data or content of the *Test data*. The system should also record a trail of all processing of personal health information or *Test data*, such as viewing, creation, modification, validation, printing, copying, import, export, transmission, reception.

Unauthorized access attempts should be denied, and all requests should be logged and retained for audit purposes. Audit logs should be stored in encrypted form and decrypted only by recorded authorized requests and analysed as potential breaches.

8. Data processing

During evaluation phase, *Test data* needs to be decrypted. We should ensure non-disclosure of the data during this critical phase. Homomorphic encryption, a technology for performing operations on encrypted data, could be a solution. This technology still needs development to be applicable to all problems.

9. Data lifecycle

Managing data requires effective data governance. Data governance refers to the overall management and caretaking of data, from creation to deletion, covering usability integrity and security. The data governance process should be defined to determine what data is retained or deleted. Data should be kept, so in the case of the creation of a new benchmark, models could be retested.

Once the data is received, it should be stored in a temporary location until data quality validation (verification or detection of any data abnormalities) is completed before transfer to the production environment.

When required, data should be securely erased in accordance with a data destruction policy.

10. Data ownership

The use and ownership of *Received data* should be clearly defined in a licence agreement between the party providing the data (the owner of the data) and the FG-AI4H.

11. Backup and archiving

Backed-up or archived data should have at least the same level of protection as those in use, it should be encrypted. Backup should be in separate secure location.

12. Compliance with international standards

Yearly audits should be conducted by internationally accredited auditors to confirm ICT/WHO observe obligatory security, data protection, continuity and compliance guidelines and procedures. This could comply with international standards such as ISO 27001.

The security architecture for the Data Repositories should comply to security policies and privacy policies. The security solutions should be in alignment with ISO 7498-2 Security Model best practice recommendations on information security management.

13. Risk assessment

There should be periodic assessment of the potential risks and vulnerabilities to the confidentiality, integrity, and availability of electronic protected information held in the repository.

We should conduct a proactive periodic risk analysis of the audit logs and should take corrective action when unacceptable risks are identified. Proactive security measures sufficient to reduce risks and vulnerabilities to the level required by the data’s high sensitivity shall be maintained throughout the programme lifecycle.