International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

16 March 2023

PRE-PUBLISHED VERSION

DEL5.1

Data requirements

International
Telecommunication
Union

Summary

Deliverable 5.1 lists *acceptance criteria for data* submitted to the FG-AI4H benchmarking platform and states the governing principles and rules. These principles are crucial because the core of the benchmarking framework for AI for health methods will be an *undisclosed test data* set – per use case of each topic area – that will not be made accessible to the AI developers.

Keywords

Artificial intelligence; health; data requirements; benchmarking platform; test data; governance

Change Log

This document contains Version 1 of the Deliverable DEL5.1 on "*Data requirements*" approved on 16 March 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Marc Lecoultre
ML|LAB.AI
Switzerland
Email: ml@mllab.ai

Contributor:

Markus Wenzel
Fraunhofer HHI
Germany
Email: markus.wenzel@hhi.fraunhofer.de

CONTENTS

Page

1 Scope ... 3

2 References ... 4

3 Terms and definitions ... 4

4 Abbreviations ... 4

5 Conventions ... 4

6 Criteria for data acceptance ... 4

6.1 Mandatory ... 4

6.2 Conditional ... 5

6.3 Recommended ... 5

ITU-T FG-AI4H Deliverable DEL5.1

Data requirements

Summary

Deliverable 5.1 lists *acceptance criteria for data* submitted to the FG-AI4H benchmarking platform and states the governing principles and rules. These principles are crucial because the core of the benchmarking framework for AI for health methods will be an *undisclosed test data* set – per use case of each topic area – that will not be made accessible to the AI developers.

1 Scope

Artificial Intelligence (AI) can help achieving the important objective of ensuring health for everyone in many ways, worldwide, often at reduced costs and enhanced speed. In the case of modern AI, it is important to notice that practitioners, patients and medical device regulators are confronted with a new kind of machine. While mechanical devices, electronics and software tools from the past have been typically designed from fully understood first principles, it is difficult to anticipate the behaviour of modern AI algorithms, because of the enormous complexity of the algorithms, and because the performance depends not only on the learning algorithm, but also on the underlying training data. These properties let the users raise doubts about whether they can trust AI models, when they face critical decisions in the health domain. Crucially, these reasonable doubts cannot be resolved at present, because there are no established ways to assess the quality of AI models for health.

The Focus Group on "Artificial Intelligence for Health" (FG-AI4H) will meet this need by demonstrating how the performance of AI solutions for health can be evaluated in a systematic fashion. For this purpose, a benchmarking framework will be developed in a best practice type of approach for representative use cases. Having successfully demonstrated the benefits of benchmarking for selected representative use cases, will allow for expanding the approach to a wider range of use cases. Exemplary use cases may include AI-based diagnostics, treatment decision making, triage, patient self-management, risk assessment, image segmentation or annotation, early detection, among others. Obviously not all possible use cases can be addressed considering the limited timespan and resources of the Focus Group.

The core of the benchmarking framework consists of *undisclosed test data* sets - per use case of each topic area to be defined – that will not be made accessible to the AI developers. In addition, (relatively small or large sets of) public data may be made available by FG-AI4H. We would like to note that data publication is not essential for the core idea of the benchmarking framework, but merely an optional extra, and that related problems have already been addressed by others before. Data sets are not limited to any modality such as images, time series, laboratory tests, "omics", text, or electronic health records, but a wide variety is welcome. Details of the envisioned benchmarking procedure are presented in the White Paper of FG-AI4H.

The document specifies the criteria for data acceptance. Decisions whether to accept or reject submitted data will be taken according to these criteria.

For sensible benchmarking, the topic drivers will address the following three dilemmas: (1) Benchmarking is not valid if AI-techniques developed by data donors are tested on their own donated data, because they know the data and associated output variables/labels. (2) Excluding data donors from benchmarking will considerably reduce the willingness to donate data, which are essential for a reasonable evaluation. (3) Having a data pool from several sources and testing each AI-technique only on data from other sources (i.e. testing AI-technique developed by x only on data donated by y and z) may tempt data donors that also develop AI-technology to contribute as "difficult" data (low quality data, wrong annotations, etc.) as possible to the data pool, in a competitive setting.

2 References

[WMA Helsinki]	*World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects*. JAMA. 2013; 310(20):2191-2194. https://doi.org/10.1001/jama.2013.281053

3 Terms and definitions

In this document, we refer to different types of datasets, as follows:

3.1 Received data: Any dataset submitted from a *trusted source* to be used in the benchmarking platform;

3.2 Public data: Subset of the *received data* that is made public by FG-AI4H to help AI developers to understand the structure of the undisclosed test data, or to train AI technology if enough data are provided;

3.3 Undisclosed test data: Corresponds to the remaining *received data* after removing *public data*. This set is kept strictly private to evaluate submitted AI technology.

3.4 Trusted source: A party that is considered reliable and dependable for providing data to be used in the benchmarking platform. Data from such a source is typically regarded as credible, accurate, and can be used with confidence for benchmarking purposes.

4 Abbreviations

EKG	Electrocardiogram
API	Application Programming Interface

5 Conventions

The following conventions apply in this document:

- "Shall": states a **mandatory** requirement of this policy.
- "Should": states a **recommended** requirement of this policy.
- "May": states an **optional** requirement.

6 Criteria for data acceptance

6.1 Mandatory

The data set and the targeted use case are described clearly and concisely. The use case is relevant and of interest for FG-AI4H (selected after prioritization among the various possible use cases). The data acquisition procedure is described in detail, such that experts from independent trusted institutions can acquire more *undisclosed test data* according to this protocol.

The data type (e.g. images, time series, laboratory tests, "omics", text, electronic health records, etc. or combinations thereof), size (e.g. file size, number of samples), structure (e.g. database type, file format and content etc.), and properties (depending on the data type) are indicated. Any data (pre-)processing methods are explained: It is explained how missing, uncertain, or incomplete data have been treated if they occur. (E.g.: are there any gaps or redundancies, if the submitter provides time series patient or clinical data - in the sense of data with a unique identifier collected over certain time intervals, not continuous EKGs data type? Are imputations or projections of the data available?)

Have the raw data been preserved or have the submitters applied any cleaning mechanism or transformation on the collected data? The data provenance/source is named: Who has collected and/or aggregated the data and where? Who has created the labels/ground truths? Who has assessed the data, e.g. with respect to quality? What were the objectives of the data acquisition? What is the current ownership of the data? Data and annotations/labels/output variables have been validated by an independent domain expert/specialist in terms of quality and suitability, especially in the case of automated data annotation procedures.

The data follow the applicable laws and regulations for data acquisition, processing and sharing, such as privacy laws, copyright laws etc. Contact details and relevant information about the submitter are given. Any potential conflicts of interest are clearly indicated.

The *undisclosed test data* are crucial for the benchmarking procedure. Therefore, the safe storage has to be assured (cf. section 4 on data handling). The measures that guarantee secrecy are described and it is specified who has had access to the *undisclosed test data* in the past, at present, in the future (e.g. published or plan to share with other researchers). How and where are the *undisclosed test data* currently hosted/stored? Consent is given to keep the *undisclosed test data* undisclosed. Clearance is demonstrated for the use in benchmarking (under compliance with the relevant laws, e.g. copyright, privacy). The *undisclosed test data* are suited for benchmarking (to be defined by respective working groups and topic drivers).

6.2 Conditional

If the data originate from humans or are related to humans, one has to adhere the principles of the Declaration of Helsinki [WMA Helsinki]. Approval was obtained from the local ethics committee where the data were generated (if applicable). Informed written consent to data acquisition, processing and sharing was obtained from the respective person. The anonymization/ pseudonymization and privacy procedure is detailed, and follows the best practices from hospitals or other institutions.

If a subset of the submitted data will be published (public data). Clearance is demonstrated for data publication under compliance with the relevant laws.

Input and output variables are characterized (with codes, classifications, triage tags, pixel or voxel labels, annotations, where they exist in the dataset). 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.

6.3 Recommended

Submitters transparently describe potential biases. (Arguably, bias cannot be avoided in typical cases of data acquisition and can be expected even in an expert setting. In hospitals, diagnoses and treatment decision are made by experts but might be biased towards reimbursement from health insurances.)

The data qualifiers are described (degree of measurement precision, definition of the quality standards). The data are of sufficient size to create a statistically valid output report. (Otherwise, further data donors need to be found and added.) Further criteria of the data (heterogeneity, real-world relevance etc.) might be considered depending on the use case (to be defined by respective working groups and topic drivers).

It is described how the data can be loaded. (Special software required? Data loader/importer functions available for common programming languages? In case of API access to the data, are there any limitations of the APIs in terms of response time or size of data packages the API endpoints are expected to return? How is the release of a new API version handled? If the APIs are used directly in the tests, will a new release maintain backwards compatibility? This information ensures that the tests will not break.)

Can the submitter help to record more data, in principle? Are the data comparable to other similar data sets? What are the submitter’s data handling procedures and data governance processes? How does the submitter handle data versioning? What is the frequency of data updates, if applicable to the particular use case?