|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-J-004 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | Virtual, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | Editors |
| **Title:** | TDD template: Proposed draft update |
| **Purpose:** | Discussion |
| **Contact:** | Eva WeickenFraunhofer HHIGermany | E-Mail: eva.weicken@hhi.fraunhofer.de |
| **Contact:** | Henry Hoffmann Ada Health GmbHGermany | E-Mail: henry.hoffmann@ada.com |
| **Contact:**  | Markus WenzelFraunhofer HHIGermany | E-Mail: markus.wenzel@hhi.fraunhofer.de |
| **Contact:** | FG-AI4H Secretariat | E-Mail: tsbfgai4h@itu.int |

|  |  |
| --- | --- |
| **Abstract:** | This template defines a common structure for all topic description documents (TDD) of the various topic groups of the ITU/WHO Focus Group on AI for Health (FG-AI4H). Each TDD specifies a standardized benchmarking of AI-based solutions for a specific topic that covers all scientific, technical, and other relevant aspects. |
| **Font colour legend:** | *Red: Instructions (read and delete)*Black: Pre-filled text (keep or adapt)Green: Placeholder (insert the requested information; finally: change the font colour to black; you can use the shortcut *CTRL+H*, to replace [YOUR TOPIC] with, e.g., “symptom assessment” everywhere in the document. The same applies to other placeholders such as [TOPIC GROUP NAME] that you can replace with, e.g., “symptom,” or [YOUR DEL ID] with, e.g., “10.3”) Grey: Suggestions for the content (please add the requested content, keep the grey suggestions for later collaborators and reviewers) |
| **Change Notes:** | This TDD template is adapted from the first version found in document [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B50606D7D-9BF3-4019-8B64-23E4D5BABBE6%7D&file=FGAI4H-C-105.docx&action=default) (Lausanne, January 2019). Since then, it has been updated according to the feedback of FG-AI4H topic drivers, working groups, and deliverable editors; and according to changes requested during meeting “H” in Brasilia (cf. [FGAI4H-H-101-R01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8B9DBABD-D045-42D6-981E-E1173BE555B5%7D&file=FGAI4H-H-101-R01.docx&action=default)). It is expected that this new version includes the requirements of all working groups and their ‘horizontal’ deliverables (No. 1-9), is applicable for all topic groups, and facilitates a review process of the TDDs. Updates include: subtopics in the TDD structure (as proposed in [FGAI4H-D-022](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8980136E-DAED-424A-845F-6E3A74A2F3A0%7D&file=FGAI4H-D-022.docx&action=default)), change notes, mentioning of contributors, restructuring and amendment of several sections, lists of figures and tables, a glossary, a regulatory section, and improved guidance. |

|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-X-# |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | [Location, Day Month Year] |
| **DOCUMENT** |
| **Source:** | TG-[TOPIC GROUP NAME] Topic Driver |
| **Title:** | TDD update: TG-[TOPIC GROUP NAME] |
| **Purpose:** | Discussion |
| **Contact:** | [NAME TOPIC DRIVER]TG-[TOPIC GROUP NAME]Some Institute, Country | E-Mail: [your.mail@somemail.com] |
| **Contact:** | Co-drivers, if any. If not, remove this table row |  |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based [YOUR TOPIC]. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking (and follows the template structure defined in [FGAI4H-I-004](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6701AD5B-8504-4C34-B286-D0388FDAB7B9%7D&file=FGAI4H-I-004.docx&action=default)). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable No. [YOUR DEL ID]. This draft will be a continuous input- and output document. |
| **Change notes:** | *Topic Driver: Please list the changes of the current TDD version in comparison to earlier versions*. **Version [x.x] (submitted as FGAI4H-X-# to meeting X in location Y)*** Added xyz
* Updated xyz
* …

**Version [y.y] (submitted as FGAI4H-X-# to meeting X in location Y)*** Added xyz
* Updated xyz

.. |

Contributors

|  |  |
| --- | --- |
| Name Some Company/InstituteCountry | Tel: +4x xx xxxxxxxxxEmail: some.name@somecompany.com |
| Name Some Company/InstituteCountry | Tel: +4x xx xxxxxxxxxEmail: some.name@somecompany.com |
|  |  |

**CONTENTS**

|  Page |
| --- |
| [1 Introduction 6](#_Toc50479660)[1.1 FG-AI4H topic group [TOPIC GROUP NAME] 6](#_Toc50479661)[1.1.1 Documentation 6](#_Toc50479662)[1.1.2 Status of this topic group 7](#_Toc50479663)[1.1.3 Topic group participation 7](#_Toc50479664)[1.1.4 Topic description 8](#_Toc50479665)[1.1.5 Subtopics 8](#_Toc50479666)[1.1.6 Definition of the AI task 8](#_Toc50479667)[1.1.7 Current gold standard 9](#_Toc50479668)[1.1.8 Relevance and impact of an AI solution 9](#_Toc50479669)[1.2 Ethical considerations 9](#_Toc50479670)[1.3 Regulatory considerations 10](#_Toc50479671)[1.3.1 Existing applicable regulatory frameworks 10](#_Toc50479672)[1.3.2 Regulatory features to be reported by benchmarking participants 10](#_Toc50479673)[1.3.3 Regulatory requirements for the benchmarking systems 11](#_Toc50479674)[1.3.4 Regulatory approach for the topic group 11](#_Toc50479675)[2 Existing AI solutions 12](#_Toc50479676)[3 Existing work on benchmarking 12](#_Toc50479677)[3.1 Publications on benchmarking systems for [YOUR TOPIC] 12](#_Toc50479678)[3.2 Benchmarking by AI developers 13](#_Toc50479679)[3.3 Relevant existing benchmarking frameworks 13](#_Toc50479680)[4 Benchmarking 13](#_Toc50479681)[4.1 Benchmarking iterations 14](#_Toc50479682)[4.2 Benchmarking version [X] 14](#_Toc50479683)[4.2.1 Overview 14](#_Toc50479684)[4.2.2 Benchmarking methodology and architecture 14](#_Toc50479685)[4.2.3 AI input data structure for the benchmarking 17](#_Toc50479686)[4.2.4 AI output data structure 17](#_Toc50479687)[4.2.5 Test data label/annotation structure 17](#_Toc50479688)[4.2.6 Scores and metrics 18](#_Toc50479689)[4.2.7 Test dataset acquisition 19](#_Toc50479690)[4.2.8 Baseline acquisition 21](#_Toc50479691)[4.2.9 Reporting methodology 22](#_Toc50479692)[4.2.10 Result 22](#_Toc50479693)[4.2.11 Discussion of the benchmarking 23](#_Toc50479694)[4.2.12 Retirement 24](#_Toc50479695)[5 Discussion 24](#_Toc50479696)[Annex A: Glossary 25](#_Toc50479697)[Annex B: Declaration of conflict of interests 26](#_Toc50479698) |

**List of Tables**

|  Page |
| --- |
| [Table 1: Topic group output documents 6](#_Toc50469729) |

**List of Figures**

|  Page |
| --- |
| [Figure 1: Example of a figure 25](#_Toc50469730) |

FG-AI4H Topic Description Document

Topic group-[TG-GROUP NAME]

# Introduction

*Topic Driver: Add a short (half page) introduction to the topic. The introduction should provide a general overview of the addressed health topic and basic information about the AI task, including the input data and the output of the AI. The objective and expected impact of the benchmarking should also be described. More detailed information about the topic will appear in section 1.3.*

This topic description document specifies the standardized benchmarking for [YOUR TOPIC] systems. It serves as deliverable No. [YOUR DEL ID] of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

## FG-AI4H topic group [TOPIC GROUP NAME]

The introduction highlights the potential of a standardized benchmarking of AI systems for [YOUR TOPIC] to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-[TOPIC GROUP NAME] at the meeting [MEETING NO.] in [LOCATION AND DATE OF YOUR FOUNDING MEETING].

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD.During FG-AI4H meeting [LOCATION AND DATE OF YOUR FOUNDING MEETING], [TOPIC DRIVER] from [TOPIC DRIVER INSTITUTION] was nominated as topic driver for the TG-[TOPIC GROUP NAME]. *[Topic Driver: If there are modifications of the topic group driver, please indicate any changes here.]*

### Documentation

This document is the TDD for the TG-[YOUR TOPIC]. It will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable “DEL 10.[YOUR DEL ID] [YOUR TOPIC] (TG-[TOPIC GROUP NAME]).” The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic group output documents

| Number | Title |
| --- | --- |
| FGAI4H-x-0y-A01 | Latest update of the Topic Description Document of the TG-[TOPIC GROUP NAME]  |
| FGAI4H-x-0y-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| FGAI4H-x-0y-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-[TOPIC GROUP NAME] |

The working version of this document can be found in the official topic group SharePoint directory.

* [INSERT THE **LINK** TO YOUR **TOPIC GROUP SHAREPOINT FOLDER HERE** AND UPLOAD THE TDD WORKING VERSION TO THE SHARE POINT]

Select the following link:

* [INSERT THE **LINK** TO THE **TDD WORKING VERSION HERE**]

*Topic Driver: As the structure of the TDD document is the same for all topic groups, you only need to fill in the green placeholders []. Please update this section in case you change the document structure.*

The TDD document is structured into the following sections:

* **Section 1** introduces the health topic and outlines its relevance and the potential impact that the benchmarking will have on the health system, patient outcome, etc.
* **Section 2** provides an overview of the existing AI solutions for [YOUR TOPIC].
* **Section 3** describes the existing approaches for assessing the quality of [YOUR TOPIC] systems and details that are likely relevant for setting up a new standardized benchmarking.
* **Section 4** specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking.
* **Section 5** summarizes the results of the topic group’s benchmarking initiative and benchmarking runs.

### Status of this topic group

The following subsections describe the update of the collaboration within the TG-[TOPIC GROUP NAME] for the official focus group meetings.

#### Status update for the meeting [MEETING LETTER] in ([MEETING LOCATION]) submission

*Topic Driver: Please insert a one-page summary of the work since the last focus group meeting. This can include:*

* Work on this document
* Work on the benchmarking software
* Progress with data acquisition, annotation, etc.
* Overview of the online meetings including links to meeting minutes
* Relevant insights from interactions with other working groups or topic groups
* Partners joining the topic group
* List of current partners
* Relevant next steps

#### Status update for the meeting [MEETING LETTER] in ([MEETING LOCATION]) submission

[…]

### Topic group participation

The participation in both the Focus Group on AI for Health and in a TG is generally open to anyone (with an ITU account). For this TG, the corresponding ‘Call for TG participation’ (CfTGP) can be found here:

* [INSERT THE LINK TO YOUR ‘CALL FOR TG PARTICIPATION’ (CfTGP)]

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

* [INSERT THE LINK TO YOUR TOPIC GROUP SUBPAGE]

*Topic Driver: Please set up a regular (e.g., bi-weekly) online meeting with rotating and considerate time windows (to account for participants in different time zones) and inform the ITU secretariat to schedule the meeting in the FG-AI4H calendar.*

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG ‘zoom’ link:

* <https://itu.zoom.us/my/fgai4h>

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the ‘Call for Topic Group participation’ and this link:

* <https://itu.int/go/fgai4h/join>

In addition to the general FG-AI4H mailing list, each topic group can create an *individual mailing list:*

*Topic Driver: Please contact the ITU secretariat* *tsbfgai4h@itu.int* *to create a mailing list for your TG.*

*Delete this passage if you are starting without a specific mailing list for your TG.*

* [INSERT YOUR TOPIC GROUP MAILING LIST HERE]

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

* <https://itu.int/en/ITU-T/focusgroups/ai4h/Pages/all.aspx>

### Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in [YOUR TOPIC] and how this can help to solve a relevant ‘real-world’ problem.

### Subtopics

*Topic Driver: Topic groups typically begin without subtopics. Please write a few lines indicating future subtopics that might become relevant. Once you have defined subtopics, their focus/mandate should be explained in this section.*

Topic groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-[TOPIC GROUP NAME] currently has no subtopics. Future subtopics for [SUBTOPIC NAME] might be introduced.

### Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to [DEL03](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) *“AI requirements specifications*,” which describes the functional, behavioural, and operational aspects of an AI system.

* What is the AI doing?
* What kind of AI task is implemented (e.g., classification, prediction, clustering, or segmentation task)?
* Which input data are fed into the AI model?
* Which output is generated?

### Current gold standard

This section provides a description of the established gold standard of the addressed health topic.

* How is the task currently solved without AI?
* Do any issues occur with the current gold standard? Does it have limitations?
* Are there any numbers describing the performance of the current state of the art?

### Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

* Why is solving the addressed task with AI relevant?
* Which impact of deploying such systems is expected (e.g., impact on the health system, overall health system cost, life expectancy, or gross domestic product)?
* Why is benchmarking for this topic important (e.g., does it provide stakeholders with numbers for decision-making; does it simplify regulation, build trust, or facilitate adoption)?

## Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default) “*AI4H ethics considerations,”* which was developed by the working group on “Ethical considerations on AI4H” (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-[TOPIC GROUP NAME].

* What are the ethical implications of applying the AI model in real-world scenarios?
* What are the ethical implications of introducing benchmarking (having the benchmarking in place itself has some ethical risks; e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk)?
* What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
* What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
* How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
* How do we ensure that benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can we guarantee the same performance in high, low-, and middle-income countries; are differences in race, sex, and minority ethnic populations captured; do we consider biases when implementing the same AI application in a different context; is there a review and clearance of ‘inclusion and exclusion criteria’ for test data)?
* What are your experiences and learnings from addressing ethics in your TG?

## Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on* [***Regulatory considerations on AI for health (WG-RC)***](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) *and their various deliverables. It is* ***NOT requested to re-produce regulatory frameworks****, but to show the regulatory frameworks that have to be applied in the context of your AIs and their benchmarking (****2 pages max****).*

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“*[*Regulatory considerations on AI for health”* *(WG-RC)*](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-RC.aspx) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), [DEL02\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6AF7C004-8BCE-4151-9F44-45F041A1EB1D%7D&file=DEL02_1.docx&action=default) *“Mapping of IMDRF essential principles to AI for health software”,* and[DEL02\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1ED0D4D1-876C-4A0F-AEF7-06D3F445F5E6%7D&file=DEL02_2.docx&action=default) *“Guidelines for AI based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) identifies standards and best practices that are relevant for the “*AI software lifecycle specification*.*”* The following sections discuss how the different regulatory aspects relate to the TG-[TOPIC GROUP NAME].

### Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for [YOUR TOPIC].

* What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR, and ISO; maybe the systems in this topic group always require at least “MDR class 2b” or maybe they are not considered a medical device)?
* Are there any aspects to this AI system that require additional specific regulatory considerations?

### Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

* Which certifications and regulatory framework components of the previous section should be part of the metadata (e.g., as a table with structured selection of the points described in the previous section)?

### Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

* Which regulatory frameworks apply to the benchmarking system itself?
* Are viable solutions with the necessary certifications already available?
* Could the TG implement such a solution?

### Regulatory approach for the topic group

*Topic Driver: Please select the points relevant for your type of AI and the corresponding benchmarking systems. If your AIs and your benchmarking are not a medical device, this might be quite short.*

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the [DEL02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF2F46A99-7457-4BC8-81A3-0E1E63D6072A%7D&file=DEL02.docx&action=default) *“AI4H regulatory considerations.”*

* Documentation & Transparency
	+ How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
* Risk management & Lifecycle approach
	+ How will the risk management be implemented?
	+ How is a life cycle approach throughout development and deployment of the benchmarking system structured?
* Data quality
	+ How is the test data quality ensured (e.g., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
	+ How are the corresponding processes document?
* Intended Use & Analytical and Clinical Validation
	+ How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
* Data Protection & Information Privacy
	+ How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis, and destruction)? This especially relevant if real patient data is used for the benchmarking.
	+ Note: it is acceptable to keep the details in chapter 4
* Engagement & Collaboration
	+ How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

# Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered by chapter 3.

* Description of the general status and the maturity of AI systems for the health topic of your TG (e.g., exclusively prototypes, applications, and validated medical devices)
* Which are the currently known AI systems and their inputs, outputs, key features, target user groups, and intended use (if not discussed before)? This can also be provided as a table.
* What are the common features found in most AI solutions that might be benchmarked?
* What are the relevant metadata dimensions characterizing the AI systems in this field and with relevance for reporting (e.g., systems supporting offline functions, availability in certain languages, and the capability to process data in a specific format)?
* Description of existing AI systems and their scope, robustness, and other dimensions.

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and [YOUR TOPIC] for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

## Publications on benchmarking systems for [YOUR TOPIC]

While a representative comparable benchmarking for [YOUR TOPIC] does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *“AI for health evaluation considerations,”* [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description,”* [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification*,*”* [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM),”* and [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default) *“Clinical Evaluation of AI for health”*.

* What are the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
* State what are the most relevant approaches used in literature?
* Which scores and metrics have been used?
* How were test data collected?
* How did the AI system perform and how did it compare the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
* How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?
* Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?

## Benchmarking by AI developers

All developers of AI solutions for [YOUR TOPIC] implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

* What are the most relevant learnings from the benchmarking by AI developers in this field (e.g., ask the members of your topic group what they want to share on their benchmarking experiences)?
* Which scores and metrics have been used?
* How did they approach the acquisition of test data?

## Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

* Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
* Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
* Are there important features in this topic group that require special attention?
* Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
* What are the relative advantages and disadvantages of these diverse solutions?

# Benchmarking

This section describes all technical and operational details regarding the benchmarking process for the [TOPIC NAME] AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default) *“Data specification”* (introduction to deliverables 5.1-5.6), [DEL05\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B19830259-F63B-42D4-A408-48C854D6C124%7D&file=DEL05_1.docx&action=default)*“Data requirements”* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL05\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B25141F77-E59A-45F1-B081-185C2194FE67%7D&file=DEL05_2.docx&action=default) *“Data acquisition”* (which presents a framework for public healthcare data acquisition and a management model based on a standard protocol for easy adoption by any country or international health organization), [DEL05\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B05D8938E-BC2A-4A62-BCB0-1FD46AA72235%7D&file=DEL05_3.docx&action=default) *“Data annotation specification”* (which provides a general guideline of data annotation specification as well as recommended metadata including a questionnaire to seek input and collaboration with topic groups regarding data annotation), [DEL05\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF267A95C-4C5B-4D63-A135-58AF487C3AD3%7D&file=DEL05_4.docx&action=default) *“Training and test data specification”* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) *“Data handling”* (which outlines how data will be handled once they are accepted), [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) *“Data sharing practices”* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), [DEL06](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF5967277-90C8-4252-A0B9-43A5692F35E2%7D&file=DEL06.docx&action=default) *“AI training best practices specification”* (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL07](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B47E77197-F87B-49F4-80B3-2DD949A5F185%7D&file=DEL07.docx&action=default)*“AI for health evaluation considerations”* (which discusses aspects in the validation and evaluation of AI for health that are novel or otherwise essential, collects best practices for the assessment of AI for health from selected sources, and considers requirements for a benchmarking platform), [DEL07\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B565EEC0A-D755-41C8-AC68-37B4C38C953F%7D&file=DEL07_1.docx&action=default) *“AI4H evaluation process description”* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), [DEL07\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B58679341-C738-40F0-A822-3AC2B24DD09F%7D&file=DEL07_2.docx&action=default) *“AI technical test specification”* (which specifies how an AI can and should be tested *in silico*), [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) *“Data and artificial intelligence assessment methods (DAISAM)”* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), [DEL07\_4](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB846B260-373A-41FC-A892-EE5BBCFE3CF8%7D&file=DEL07_4.docx&action=default)*“Clinical Evaluation of AI for health”* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), [DEL07\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8BFCFF21-3908-4BAD-AB9C-9814EB3F9B36%7D&file=DEL07_5.docx&action=default) *“FG-AI4H assessment platform”* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL09](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3E940987-8D75-44B8-85E4-F0E475964F15%7D&file=DEL09.docx&action=default) *“AI for health applications and platforms”* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL09\_1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B1A2EC8D5-53CA-4C8C-9B09-B61CA6F428C5%7D&file=DEL09_1.docx&action=default) *“Mobile based AI applications,”* and [DEL09\_2](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3B5A31DE-D3B1-4EC1-A261-2C2E19F73810%7D&file=DEL09_2.docx&action=default) *“Cloud-based AI applications”* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

## Benchmarking iterations

The benchmarking of [YOUR TOPIC] is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to section 4.2, where the actual benchmarking methodology for each version will be described.

* Which benchmarking iterations have been implemented thus far?
* What important new features are introduced with each iteration?
* What are the next planned iterations and which features are they going to add?

## Benchmarking version [X]

This section includes all technological and operational details of the benchmarking process for the TDD version [X].

### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [X].

* What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
* What features have been added to the benchmarking in this iteration?

### Benchmarking methodology and architecture

This section provides details about the methods and architecture of the benchmarking version [X]. It contains detailed information about the data and software for the benchmarking process (e.g., test scenarios, data sources, and legalities). The information should also be provided for each subtopic as needed.

#### Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

* How does the architecture look?
* What are the most relevant components and what are they doing?
* How do the components interact on a high level?
* What underlying technologies and frameworks have been used?

#### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

* How do benchmarking data access the system?
* Where and how (data format) are the data, the responses, and reports of the system stored?
* How are the inputs and the expected outputs separated?
* How are the data sent to the AI systems?
* Are the data entries versioned?
* How does the lifecycle for the data look?

#### Safe and secure system operation and hosting

*From a technical point of view, the benchmarking process is not particularly complex. It is more about agreeing on something in the topic group with potentially many competitors and implementing the benchmarking in a way that cannot be compromised. This section describes how the benchmarking system, the benchmarking data, the results, and the reports are protected against manipulation, data leakage, or data loss. Topic groups that use ready-made software might be able to refer to the corresponding materials of the manufacturers of the benchmarking system.*

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer’s prescriptions.

* Based on the architecture, where is the benchmarking vulnerable to risk and how have these risks been mitigated (e.g., did you use a threat modelling approach)? A discussion could include:
* Could someone access the benchmarking data before the actual benchmarking process to gain an advantage?
* What safety control measures were taken to manage risks to the operating environment?
* Could someone have changed the AI results stored in the database (your own and/or that of competitors)?
* Could someone attack the connection between the benchmarking and the AI (e.g., to make the benchmarking result look worse)?
* How is the hosting system itself protected against attacks?
* How are the data protected against data loss (e.g., what is the backup strategy)?
* How is it ensured that the correct version of the benchmarking software and the AIs are tested?
* How are automatic updates conducted (e.g., of the operating system)?
* How and where is the benchmarking hosted and who has access to the system and the data (e.g., virtual machines, storage, and computing resources, configurational settings)?
* How is the system’s stability monitored during benchmarking and how are attacks or issues detected?
* How are issues (e.g., with a certain AI) documented or logged?
* In case of offline benchmarking, how are the submitted AIs protected against leakage of intellectual property?

#### Benchmarking process

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

* How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
* How do possible participants learn about an upcoming benchmarking?
* How can one apply for participation?
* What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
* Are there any contracts or legal documents to be signed?
* Are there inclusion or exclusion criteria to be considered?
* How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
* How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
* Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
* If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
* How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
* In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how their benchmarking methods and architecture details deviate from the general description in the previous sections.* ***Otherwise, remove this section****.*

### AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of [YOUR TOPIC]. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

* What are the general data types that are fed in the AI model?
* How exactly are they encoded? For instance, discuss:
	+ The exact data format with all fields and metadata (including examples or links to examples)
	+ Ontologies and terminologies
	+ Resolution and data value ranges (e.g., sizes, resolutions, and compressions)
	+ Data size and data dimensionality

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how the inputs given to the AIs differ from the above (as needed). If it does not differ, please make this explicitly clear.*

### AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

* What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
	+ How exactly are they encoded? Discuss points like:
	+ The exact data format with all fields and metadata (including examples or links to examples)
	+ Ontologies and terminologies
* What types of errors should the AI generate if something is defective?

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how the output of the AI differs from the above (as needed). If it does not differ, please make this explicitly clear.*

### Test data label/annotation structure

*Topic driver: Please describe how the expected AI outputs are encoded in the benchmarking test data. Please note that it is essential that the AIs never access the expected outputs to prevent cheating. The topic group should carefully discuss whether more detailed labelling is needed. Depending on the topic, it might make sense to separate between the best possible output of the AI given the input data and the correct disease (that might be known but cannot be derived from the input data alone). Sometimes it is also helpful to encode acceptable other results or results that can be clearly ruled out given the evidence. This provides a much more detailed benchmarking with more fine-grained metrics and expressive reports than the often too simplistic leader boards of many AI competitions.*

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

* What are the general label types (e.g., expected results, acceptable results, correct results, and impossible results)?
* How exactly are they encoded? Discuss points like:
	+ The exact data format with all fields and metadata (including examples or links to examples)
	+ Ontologies and terminologies
* How are additional metadata about labelling encoded (e.g., author, data, pre-reviewing details, dates, and tools)?
* How and where are the labels embedded in the input data set (including an example; e.g., are there separate files or is it an embedded section in the input data that is removed before sending to the AI)?

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how the labels differ from above (as needed). If they do not differ, please make this explicitly clear.*

### Scores and metrics

*Topic drivers: This section describes the scores and metrics that are used for benchmarking. It includes details about the testing of the AI model and its effectiveness, performance, transparency, etc. Please note that this is only the description of the scores and metrics actually used in* ***this*** *benchmarking iteration. A general description of the state of the art of scores and metrics and how they have been used in previous work is provided in section 3.*

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

* Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
* What general criteria have been applied for selecting scores and metrics?
* What scores and metrics have been chosen/defined for robustness?
* What scores and metrics have been chosen/defined for medical performance?
* What scores and metrics have been chosen/defined for non-medical performance?
	+ Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
* What scores and metrics have been chosen/defined for model explainability?
* Describe for each aspect
	+ The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
	+ Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)
	+ What are the origins of these scores and metrics?
	+ Why were they chosen?
	+ What are the known advantages and disadvantages?
	+ How easily can the results be compared between or among AI solutions?
	+ Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?
* How does this consider the general guidance of WG-DAISAM in [DEL07\_3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA3088882-F82B-493B-B1C5-49CFF0EEEFA8%7D&file=DEL07_3.docx&action=default) “Data and artificial intelligence assessment methods (DAISAM)”?
* Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

#### Sub-topic [X] details

*Topic driver: If you have subtopics, describe how the scores and metrics for this subtopic differ from above (as needed).*

### Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure (also for each subtopic) including quality control of the dataset, control mechanisms, data sources, and storage.

* How does the overall dataset acquisition and annotation process look?
* How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
* Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
* How was the dataset documented and which metadata were collected?
	+ Where were the data acquired?
	+ Were they collected in an ethical-conform way?
	+ Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
	+ Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
	+ What kind of data anonymization or deidentification has been applied?
	+ Are the data self-contained (i.e., independent from externally linked datasets)?
	+ How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
	+ What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
* Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
* Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
* How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
* Specific data governance derived by the general data governance document (currently [F-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-F-103-DataPolicy.pdf) and the deliverables beginning with [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default))
* How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
* Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
* How was the annotation organized?
	+ How many annotators/peer reviewers were engaged?
	+ Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
	+ How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
	+ If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
	+ How was the annotation of each case documented?
	+ Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
* Were data/label update/amendment policies and/or criteria in place?
* How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
* How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?

#### Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also [DEL05\_5](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B71FE8B9D-ACB3-48CE-AA3F-136409B550A4%7D&file=DEL05_5.docx&action=default) on *data handling* and [DEL05\_6](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5C95327E-96A5-4175-999E-3EDB3ED147C3%7D&file=DEL05_6.docx&action=default) on *data sharing practices*).

* Which legal framework was used for data sharing?
* Was a data sharing contract signed and what was the content? Did it contain:
	+ Purpose and intended use of data
	+ Period of agreement
	+ Description of data
	+ Metadata registry
	+ Data harmonization
	+ Data update procedure
	+ Data sharing scenarios
		- Data can be shared in public repositories
		- Data are stored in local private databases (e.g., hospitals)
	+ Rules and regulation for patients’ consent
	+ Data anonymization and de-identification procedure
	+ Roles and responsibilities
		- Data provider
		- Data protection officer
		- Data controllers
		- Data processors
		- Data receivers
* Which legal framework was used for sharing the AI?
* Was a contract signed and what was the content?

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how the test set acquisition/annotation differs from the above (as needed).*

### Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

* Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
* Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
* Was a baseline assessed as part of the benchmarking?
* How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
* What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

### Reporting methodology

*After the benchmarking, the next step is to describe how the results are compiled into reports that allow stakeholders to make decisions (e.g., which AI systems can be used to solve a pre-diagnosis task in an offline –field –clinic scenario in central America). For some topic groups, the report might be as simple as a classical AI competition leader board using the most relevant performance indicator. For other tasks, it could be an interactive user interface that allows stakeholders to compare the performance of the different AI systems in a designated context with existing non-AI options. For the latter, statistical issues must be carefully considered (e.g., the multiple comparisons problem). Sometimes, a hybrid of prepared reports on common aspects are generated in addition to interactive options. There is also the question of how and where the results are published and to what degree benchmarking participants can opt in or opt out of the publication of their performance.*

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

* What is the general approach for reporting results (e.g., leader board vs. drill down)?
* How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
* How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
* What additional metadata describing the AI models have been selected for reporting?
* How is the relationship between AI results, baselines, previous benchmarking iterations, and/or other benchmarking iterations communicated?
* What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
* What is the publication strategy for the results (e.g., website, paper, and conferences)?
* Is there an online version of the results?
* Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
* Are there any known limitations to the value, expressiveness, or interpretability of the reports?

#### Subtopic [X] details

*Topic driver: If you have subtopics, describe how the reporting approach is different for them (as needed).*

### Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

* When was the benchmarking executed?
* Who participated in the benchmarking?
* What overall performance of the AI systems concerning medical accuracy, robustness, and technical performance (minimum, maximum, average etc.) has been achieved?
* What are the results of this benchmarking iteration for the participants (who opted in to share their results)?

### Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

* What was the general outcome of this benchmarking iteration?
* How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
* Are there real benchmarking results and interesting insights from this data?
	+ How was the performance of the AI system compared to the baseline?
	+ How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
	+ How did the results change in comparison to the last benchmarking iteration?
* Are there any technical lessons?
	+ Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
	+ How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
* Are there any lessons concerning data acquisition?
	+ Was it possible to collect enough data?
	+ Were the data as representative as needed and expected?
	+ How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
	+ Was it possible to find annotators?
	+ Was there any relevant feedback from the annotators?
	+ How long did it take to create the dataset?
* Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
	+ Are metrics missing?
	+ Do the stakeholders need different reports or additional metadata (e.g., do they need the “offline capability” included in the AI metadata so that they can have a report on the best offline system for a certain task)?
* Are there insights on the benchmarking process?
	+ How was the interest in participation?
	+ Are there reasons that someone could not join the benchmarking?
	+ What was the feedback of participants on the benchmarking processes?
	+ How did the participants learn about the benchmarking?

#### Subtopic [X] details

*Topic driver: If you have insights specific to the subtopics, describe them in dedicated subsections.*

### Retirement

*Topic driver: describe what happens to the benchmarking data and the submitted AI models after the benchmarking.*

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section [DEL04](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC68833D1-9B31-4E8E-8A4A-3939D7DEA56F%7D&file=DEL04.docx&action=default) “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

* What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency, or published)?
* What happens to the submitted AI models after the benchmarking?
* Could the results be reproduced?
* Are there legal or compliance requirements to respond to data deletion requests?

# Discussion

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 4.2.11).

* What is the overall outcome of the benchmarking thus far?
* Have there been important lessons?
* Are there any field implementation success stories?
* Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
* Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
* Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
	+ Did the AI system perform as predicted relative to the baselines?
	+ Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?
* Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
* What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

|  |  |  |
| --- | --- | --- |
| A |  | B |

Captions for figures use WinWord style "Figure\_No & title"

Figure 1: Example of a figure

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

|  |  |  |
| --- | --- | --- |
| Acronym/Term | Expansion | Comment |
| TDD | Topic Description Document | Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group [YOUR TOPIC GROUP] |
| TG | Topic Group |  |
| WG | Working Group |  |
| FGAI4H | Focus Group on AI for Health |  |
| AI | Artificial intelligence |  |
| ITU | International Telecommunication Union |  |
| WHO | World Health Organization |  |
| DEL | Deliverable  |  |
| CfTGP | Call for topic group participation |  |
| AI4H  | Artificial intelligence for health |  |
| IMDRF | International Medical Device Regulators Forum |  |
| MDR | Medical Device Regulation |  |
| ISO | International Standardization Organization |  |
| GDPR | General Data Protection Regulation |  |
| FDA | Food and Drug administration |  |
| SaMD | Software as a medical device |  |
| AI-MD | AI based medical device |  |
| LMIC | Low-and middle-income countries |  |
| GDP | Gross domestic product |  |
| API | Application programming interface |  |
| IP | Intellectual property |  |
| PII | Personal identifiable information |  |
| […] |  |  |

Annex B:
Declaration of conflict of interests

In accordance with the ITU transparency rules, this section lists the conflict of interest declarations for everyone who contributed to this document. Please see the guidelines in FGAI4H-F-105 "ToRs for the WG-Experts and call for experts" and the respective forms (Application form & Conflict of interest form).

Company/Institution/Individual XYZ

A short explanation of the company’s area of activity and how the work on this document might benefit the company and/or harm competitors. A list of all people who contributed to this document on behalf of this company and any personal interest in this company (e.g., shares).

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