|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-K-025-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 27-29 January 2021 |
| **DOCUMENT** |
| **Source:** | TG-Endoscopy Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Endoscopy) |
| **Purpose:** | Discussion |
| **Contact:** | Jianrong WuTG-EndoscopyTencent Healthcare (Shenzhen) Co.,LtdChina | E-Mail: edwinjrwu@tencent.com |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based Endoscopy. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking (and follows the template structure defined in document FGAI4H-J-105). 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. DEL 10.20. This draft will be a continuous input and output document. |
| **Change notes:** | Version 1 (submitted as FGAI4H-J-025-A01-R01 to meeting J (e-meeting), 30 Sept – 2 Oct 2020)* Drafted this TDD.
 |

**Contributors**

|  |  |
| --- | --- |
| Yajun ZhangTencent Technology (Shenzhen) Co.,LtdChina | Email: yajunzhang@tencent.com  |

**CONTENTS**

|  Page |
| --- |
| [1 Introduction 5](#_Toc62227638)[2 About the FG-AI4H topic group on Endoscopy 5](#_Toc62227639)[2.1 Documentation 5](#_Toc62227640)[2.2 Status of this topic group 6](#_Toc62227641)[2.2.1 Status update for meeting J 6](#_Toc62227642)[2.3 Topic group participation 6](#_Toc62227643)[3 Topic description 7](#_Toc62227644)[3.1 Gastroscopy 7](#_Toc62227645)[3.2 Colonoscopy 7](#_Toc62227646)[3.3 Laryngoscopy 7](#_Toc62227647)[3.4 Rhinoscopy 7](#_Toc62227648)[3.5 Colposcopy 7](#_Toc62227649)[3.6 Existing AI solutions 8](#_Toc62227650)[4 Ethical considerations 8](#_Toc62227651)[5 Existing work on benchmarking 8](#_Toc62227652)[6 Benchmarking by the topic group 9](#_Toc62227653)[6.1 Method 9](#_Toc62227654)[6.1.1 Inputs data structure 10](#_Toc62227655)[6.1.2 Data laundry 10](#_Toc62227656)[6.1.3 Screening process 10](#_Toc62227657)[6.1.4 Annotation of Detection 11](#_Toc62227658)[6.1.5 Annotation of classification 11](#_Toc62227659)[6.1.6 Annotation of segmentation 12](#_Toc62227660)[6.1.7 Information requirements of annotation 12](#_Toc62227661)[6.2 Score and metrics 13](#_Toc62227662)[6.2.1 Annotation quality score and metrics criteria 13](#_Toc62227663)[6.2.2 Other quality score and metrics criteria 14](#_Toc62227664)[7 Overall discussion of the benchmarking 15](#_Toc62227665)[8 Regulatory considerations 15](#_Toc62227666)[8.1 Existing applicable regulatory frameworks 15](#_Toc62227667)[8.2 Regulatory features to be reported by benchmarking participants 16](#_Toc62227668)[8.3 Regulatory requirements for the benchmarking systems 16](#_Toc62227669)[8.4 Regulatory approach for the topic group 16](#_Toc62227670)[9 References 16](#_Toc62227671)[Annex A: Glossary 17](#_Toc62227672)[Annex B: Declaration of conflict of interests 18](#_Toc62227673) |

**List of Tables**

[Table 1: Topic group output documents 6](#_Toc62227674)

[Table 2: Existing benchmarking processes 8](#_Toc62227675)

**List of Figures**

[Figure 1: Screening procedure 10](#_Toc62227676)

[Figure 2: Annotation procedure for a detection 11](#_Toc62227677)

[Figure 3: Illustration of intersection over union (IoU) 14](#_Toc62227678)

FG-AI4H Topic Description Document

Topic group-Endoscopy

# Introduction

This topic description document specifies the standardized benchmarking for Endoscopy systems. It serves as deliverable No. DEL 10.20 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

Endoscopy is the core technical means for early diagnosis and screening of digestive cancer. Implementing endoscopic screening for digestive cancer can detect and treat precancerous lesions, which can drastically reduce the incidence and mortality of digestive cancer. Due to factors such as the endoscopic doctor's operating, the ability to identify lesions, and visual fatigue, a considerable proportion of lesions in clinical diagnosis, including even advanced and precancerous lesions, may be missed by the endoscopic doctor.

In recent years, with the breakthrough of the new generation of artificial intelligence technology represented by deep learning, revolutionary progress has been made in the field of automatic recognition of medical images. The real-time assistance of artificial intelligence to detect and classify gastrointestinal lesions is expected to help clinicians improve their examination quality and reduction of missed diagnosis.

The document aims at developing a standardised benchmarking approach for AI-based applications for endoscopy.

# About the FG-AI4H topic group on Endoscopy

The introduction highlights the potential of a standardized benchmarking of AI systems for Endoscopy 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-Endoscopy at the meeting I e-meeting, 7-8 May 2020​​.

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 I e-meeting, 7-8 May 2020​​, Dr. Jianrong Wu from Tencent Healthcare was nominated as topic driver for the TG-Endoscopy.

## Documentation

This document is the TDD for the TG-Endoscopy. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for Endoscopy. It describes the existing approaches for assessing the quality of Endoscopy systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It 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. Finally, it summarizes the results of the topic group’s benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD 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 10.20 Endoscopy (TG- Endoscopy).” 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-J-025-A01-R01 | Latest update of the Topic Description Document of the TG-Endoscopy  |
| FGAI4H-J-025-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| N/A | The presentation summarizing the latest update of the Topic Description Document of the TG-Endoscopy |

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

* https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Endoscopy.aspx

Select the following link:

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

## Status of this topic group

The following subsections describe the update of the collaboration within the TG-Endoscopy for the official focus group meetings.

The Endoscopy Topic group was approved and established at the FG-AI4H Meeting I (online, 7-8 May 2020).

The group has a Topic Group Driver who will moderate the activities of the Topic Group. Topic Groups summarize uses cases of a certain health topic or problem and similar AI benchmarking requirements. However, inside a Topic Group different Sub-topic Groups can be established to pursue different topic-specific specializations. TG-Endoscopy will start without separate subtopic Groups. However, it is possible that during the process subtopics will be introduced as contributors will deem fit.

### Status update for meeting J

* Work on this document
* Work on the benchmarking software
* Progress with data laundry, data acquisition, annotation, screening process etc.

## Topic group participation

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

* <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Endoscopy.pdf>

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

* <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Endoscopy.aspx>

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>

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/go/fgai4h>

# Topic description

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

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-Endoscopy currently has no subtopics. Future subtopics for gastroscopy, colonoscopy, laryngoscopy, rhinoscopy, colposcopy might be introduced.

The performance of the endoscope-assisted diagnostic system relies on massive and high-quality annotated data to fully learn the features of the lesion. Effective collection of large samples of high-quality annotated data has become the primary basis and prerequisite for endoscopic assisted diagnosis systems.

Additional diseases and conditions that are relevant to this Topic Group may be added in the future.

## Gastroscopy

[TBC] [call for contribution]

## Colonoscopy

Colorectal cancers (CRC) is the third most prevalent cancer and the second highest cause of cancer deaths worldwide. Colonoscopy is considered to be the gold standard for CRC screening, but cannot detect all colonic polyps, some of which maybe neoplasms. Colonoscopy has been reported to miss 17-48% of adenomas which are considered to be 50%-60% causes of interval cancers.

Over the last two decades, computer-assisted polyp detection has been actively explored to improve inspecting quality and reduce adenoma miss rates (AMR). Recently, artificial intelligence (AI) has made remarkable breakthroughs in medical fields with deep learning and convolutional neural networks (CNNs). With enough qualified learning materials, CNNs can reach even higher real-time detecting accuracy than human experts, which suggests that computer assisted-detection systems (CADe) might serve as real-time ‘expert’ to improve the quality of colonoscopists.

## Laryngoscopy

[TBC] [call for contribution]

## Rhinoscopy

[TBC] [call for contribution]

## Colposcopy

[TBC] [call for contribution]

## Existing AI solutions

National Cancer Center Hospital, and NEC Japan successfully developed a system in 2017 that immediately detects colorectal cancer and ulcerative colon polyps, a precursor to cancer, during an endoscopic examination using artificial intelligence (AI). It automatically detects colorectal cancer and polyps from images and videos taken during an endoscopic examination of the colon, and aids in discovery of lesions by endoscopists. It improves polyp detection, which was an issue during such exams, and increases the detection rate. In this manner, it greatly contributes to the prevention and early detection of colorectal cancer.[1]

# 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-Endoscopy.

It is necessary to collect massive data for AI solution development, however, ethical considerations such as patient privacy concern should be taken into careful consideration and relevant regulations should be followed. Otherwhile, the privacy of patients must be protected in the process of data collection, transmission and utility. If the data contains patient private information or identified codes, data desensitization must be performed.

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and Endoscopy 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.

Table 2: Existing benchmarking processes

| Methodology | Description |
| --- | --- |
| True positive (TP) | The number of correctly identified samples. The number of frames with an endoscopic finding which correctly is identified as a frame with an endoscopic finding. |
| True negative (TN) | The number of correctly identified negative samples, i.e., frames without an endoscopic finding which correctly is identified as a frame without an endoscopic finding. |
| False positive (FP) | The number of wrongly identified samples, i.e., a commonly called a "false alarm". Frames without an endoscopic finding which is erroneously identified as a frame with an endoscopic finding. |
| False negative (FN) | The number of wrongly identified negative samples. Frames without an endoscopic finding which erroneously is identified as a frame with an endoscopic finding. |
| Recall (REC) | This metric is also frequently called sensitivity, probability of detection and true positive rate, and it is the ratio of samples that are correctly identified as positive among all existing positive samples. |
| Precision (PREC) | This metric is also frequently called the positive predictive value. It shows the ratio of samples that are correctly identified as positive among the returned samples (the fraction of retrieved samples that are relevant). |
| Specificity (SPEC) | This metric is frequently called the true negative rate. It shows the ratio of negatives that are correctly identified as such (e.g., the fraction of frames without an endoscopic finding are correctly identified as a negative result). |
| Accuracy (ACC) | The percentage of correctly identified true and false samples. |
| Matthews correlation coefficient (MCC) | MCC takes into account true and false positives and negatives. It is a balanced measure even if the classes are of very different sizes. |
| F1 score (F1) | A measure of a test's accuracy by calculating the harmonic mean of the precision and recall. |

# Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the Endoscopy 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”*, [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”*, [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 the validation and evaluation of AI for health models, 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).

## Method

The method for AI benchmarking includes data format requirement of input data and output data, training and testing data annotation quality control as well as testing metrologies and scoring matrixes.

Data annotation is currently a spontaneous non-standard process. It is a challenge task to guarantee the accuracy and representativeness of learning materials without the standardized data annotation quality control measures which are widely recognized by the industry. What’s more, this may also bring a greater risk of erroneous judgment for endoscopic assisted diagnosis.

Note: the current version focused more on the quality control of digestive endoscopic data annotation and the method of annotating the lesion location.

### Inputs data structure

* Image file format: JPEG format;
* Image file names: anonymized;
* Video file format: TBC;
* Video file names: anonymized;
* Macroscopic (without magnification);
* Image size: TBC.

### Data laundry

The data needs preliminary filtering and laundering to eliminate completely worthless data, such as: missing data, image parameter mismatch, non-inspection site image, foreign matter in the image, image artifacts, image quality cannot satisfy the diagnostic requirements.

### Screening process

The screening process is illustrated in Figure 1.



**Figure 1: Screening procedure**

In order to reduce the differences between observers, all conclusions about confirming the existence of polyp lesions should be based on at least three experienced endoscopic doctors (˃3000 operation cases). At least two doctors should confirm the existence of polyp lesions in a blind method.

Guarantying the quality of data annotation and reduce individual differences among doctors, it is recommended that the annotation process should include multiple doctors' independent annotation, cross-annotation, arbitration, and review those four procedures. The annotation process is illustrated in Figure 2.

### Annotation of Detection

The annotation of detection includes localizing the object inside the data and categorizing it. The bounding box is usually used to localize the object with rectangular box.

#### Independent annotation

Independent detection by 3 or more doctors to confirm whether the endoscopic image contains lesions, and if so, mark the location and size of the polyp lesion in a clear way.

#### Cross-annotation

The annotated results are crossed to identify the relationship between each other by calculating the IoU (Intersection over Union).

#### Arbitration

In the cross-annotation step, some images' data marked results do not satisfy requirements that cannot be go to the review procedure will be transferred to the arbitration expert to review and re-annotate as a gold standard candidate.

#### Review

The data for completed gold standard candidate, which will be confirmed by the review experts one by one. Review experts approved candidate data will be marked as gold standard. Review experts do not approve of the candidate data which will be sent back to arbitration procedure and arbitrated by another arbitration expert.



**Figure 2: Annotation procedure for a detection**

### Annotation of classification

Annotation of Classification means arranging a category to the data. For example, the decision of the category might be made subjectively, based on the manual observing of features in the entire or part of data. Also, the category might be made objectively, based on the corresponding pathological results.

Independent annotation of classification by 3 or more doctors to confirm which category the data should be arranged. If all independent annotation are completely consistent, the annotation is regarded as gold standard. Otherwise, if there is inconsistent annotation, this case should be transferred to the arbitration expert for re-annotation. All the annotation are checked by review experts one by one, while the approvals are regarded as gold standard and the rest will be sent back to arbitration experts for another round of annotation. If the corresponding pathological report is available, the category classification should be made following the report.

[TBC]

[call for contribution]

### Annotation of segmentation

Annotation of segmentation means the annotation of every pixel in object within a data. Practically, a polygon is used to annotate the contour of the object, which will be filled for a mask.

[TBC]

[call for contribution]

### Information requirements of annotation

#### Independent annotation requirements

Independent annotation requirements include:

1. Non-annotating information:
* Image name, image identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
1. Annotating information:
* Independently label the results (x, y, w, h), the doctor information, the date, the result serial number

#### Cross-annotation requirements

Cross-annotation requirements include:

1. Non-annotating information:
* Image name, image identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
1. Annotating information:
* The results after cross label, the serial number of independently label for merging, merge type, procedure type (arbitration or review), and the serial number after merge

#### Arbitration requirements

Arbitration requirements include:

1. Non-annotating information:
* Image name, image identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
1. Annotating information:
* The arbitration results, the serial number after merging for arbitration, arbitration expert information, the date, and the arbitration results serial number

#### Review requirements

Review requirements include:

1. Non-annotating information:
* Image name, image identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)
1. Annotating information:
* The review results (gold standard or sent back to arbitration), the serial number of arbitrations, review expert information, the date, and the serial number of audit result

## Score and metrics

### Annotation quality score and metrics criteria

Intersection over Union (IoU) [2] is an evaluation metric used to measure the accuracy of an object detector on a particular dataset. Intersection over Union is simply an evaluation metric. Any algorithm that provides predicted bounding boxes as output can be evaluated using IoU.

More formally, in order to apply Intersection over Union to evaluate an (arbitrary) object detector we need:

1. The ground-truth bounding boxes (i.e., the hand labelled bounding boxes from the testing set that specify where in the image our object is).
2. The predicted bounding boxes from our model.

Computing Intersection over Union can therefore be determined via:



In the numerator we compute the area of overlap between the predicted bounding box and the ground-truth bounding box.

The denominator is the area of union, or more simply, the area encompassed by both the predicted bounding box and the ground-truth bounding box.

Dividing the area of overlap by the area of union yields our final score — the Intersection over Union.

An Intersection over Union score > 0.5 is normally considered a "good" prediction (see Figure 3).



**Figure 3: Illustration of intersection over union (IoU)**

### Other quality score and metrics criteria

[TBC] [call for contribution]

#### Available public data and undisclosed test data set collection

| Dataset Name | Description |
| --- | --- |
| CVC-ClinicDB | CVC-ClinicDB is a database of frames extracted from colonoscopy videos. These frames contain several examples of polyps. In addition to the frames, we provide the ground truth for the polyps. This ground truth consists of a mask corresponding to the region covered by the polyp in the image. |
| Kvasir | The Kvasir dataset consists of images, annotated and verified by medical doctors (experienced endoscopists), including several classes showing anatomical landmarks, pathological findings or endoscopic procedures in the GI tract, i.e., hundreds of images for each class..  |

#### 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*).

#### 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.

#### Reporting methodology

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

#### 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.

#### 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).

#### Retirement

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).

# Overall discussion of the benchmarking

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).

# Regulatory considerations

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-Endoscopy.

## 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 endoscopy.

## 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).

## 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.

## Regulatory approach for the topic group

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.”*

# References

[1] <https://www.jst.go.jp/EN/achievements/research/bt2019-07.html>

[2] <https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>

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

Company/Institution/Individual XYZ

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