|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-N-025-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 15-17 February 2022 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Endoscopy Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Endoscopy) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Jianrong Wu TG-Endoscopy Tencent Healthcare (Shenzhen) Co.,Ltd China | | E-Mail: edwinjrwu@tencent.com |
| **Contact:** | | Shan Xu TG-Endoscopy CAICT China | | E-Mail: [xushan@caict.ac.cn](mailto:xushan@caict.ac.cn) |

|  |  |
| --- | --- |
| **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.   *Version 2* (submitted as FGAI4H-K-025-A01-R01 to meeting K (e-meeting), 27 – 29 Jan 2021)   * Arranged and aligned the TDD following the new template J-105.   *Version 3* (submitted as FGAI4H-M-025-A01-R01 to meeting M (e-meeting), 28 – 30 Sept 2021)   * Introduced 3 new participants, and 1 new subtopic of endoscopic ultrasound. * Modified the chapter of Method, including additional content to chapter of Input data structure, new chapter of Output data structure, new chapter of Annotation procedure combining content from previous chapters of Information requirements of annotation, Annotation of detection, Annotation of classification, Annotation of segmentation * Deleted chapter of Existing work on benchmarking, move corresponding content to the chapter of Evaluation score and metrics.   *Version 4* (submitted as FGAI4H-N-025-A01-R01 to meeting N (e-meeting), 15 – 17 Feb 2022)   * Introduced 1 new participants * Modified the structure of the TDD, such as adding new chapter of Definition of the AI task, Current gold standard, Relevance and impact of an AI solution * Updated content of chapter Status of this topic group, Definition of the AI task, Current gold standard, Relevance and impact of an AI solution, Existing AI solutions, Annotation of classification and Annotation of segmentation |

**Contributors**

|  |  |
| --- | --- |
| Yajun Zhang Tencent Technology (Shenzhen) Co.,Ltd China  Yi Cai Olympus Medical Systems Corp.  Japan  Junbo Li Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences China  Yanchun Zhu China Unicom (Guangdong) Industrial Internet Co., Ltd China | Email: [yajunzhang@tencent.com](mailto:yajunzhang@tencent.com)  Email: [i.sai@olympus.com](mailto:i.sai@olympus.com)  Email: [Lijb@sibet.ac.cn](mailto:Lijb@sibet.ac.cn)  Email: [zhuyc82@chinaunicom.cn](mailto:zhuyc82@chinaunicom.cn) |

**CONTENTS**

[1 Introduction 5](#_Toc95748523)

[1.1 About the FG-AI4H topic group on Endoscopy 5](#_Toc95748524)

[1. Documentation 5](#_Toc95748525)

[2. Status of this topic group 6](#_Toc95748526)

[9. Topic group participation 7](#_Toc95748527)

[10. Topic description 7](#_Toc95748528)

[11. Subtopics 7](#_Toc95748529)

[17. Definition of the AI task 8](#_Toc95748530)

[21. Current gold standard 9](#_Toc95748531)

[24. Relevance and impact of an AI solution 9](#_Toc95748532)

[27. Ethical considerations 10](#_Toc95748533)

[28. Regulatory considerations 10](#_Toc95748534)

[2 Existing AI solutions 10](#_Toc95748535)

[3 Existing work on benchmarking 11](#_Toc95748536)

[29. Publications on benchmarking systems for Endoscopy 12](#_Toc95748537)

[30. Benchmarking by AI developers 12](#_Toc95748538)

[31. Relevant existing benchmarking frameworks 12](#_Toc95748539)

[4 Benchmarking by the topic group 12](#_Toc95748540)

[4.1 Method 13](#_Toc95748541)

[4.1.1 Input data structure 13](#_Toc95748542)

[4.1.2 Output data structure 13](#_Toc95748543)

[4.1.3 Annotation procedure 14](#_Toc95748544)

[4.2 Score and metrics 18](#_Toc95748545)

[4.2.1 Evaluation score and metrics 18](#_Toc95748546)

[4.2.2 Other quality score and metrics criteria 19](#_Toc95748547)

[5 Overall discussion of the benchmarking 20](#_Toc95748548)

[6 Regulatory considerations 20](#_Toc95748549)

[6.1 Existing applicable regulatory frameworks 21](#_Toc95748550)

[6.2 Regulatory features to be reported by benchmarking participants 21](#_Toc95748551)

[6.3 Regulatory requirements for the benchmarking systems 21](#_Toc95748552)

[6.4 Regulatory approach for the topic group 21](#_Toc95748553)

[7 References 21](#_Toc95748554)

[Annex A: Glossary 23](#_Toc95748555)

[Annex B: Declaration of conflict of interests 24](#_Toc95748556)

**List of Tables**

[Table 1: Topic group output documents 6](#_Toc95748467)

[Table 2: Benchmarking processes 19](#_Toc95748468)

**List of Figures**

[Figure 1: Illustration of annotation procedure for detection 17](#_Toc95388875)

[Figure 2: Illustration of annotation procedure for classification 19](#_Toc95388876)

**FG-AI4H Topic Description Document**

**Topic group - AI for Endoscopy**

1. **Introduction**

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 diagnoses.

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).

* 1. **About the FG-AI4H topic group on Endoscopy**

The introduction highlights the potential of a standardized benchmarking of AI systems for Endoscopy to help solve 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 the FG-AI4H meeting I e-meeting, 7-8 May 2020​​, Dr Jianrong Wu from Tencent Healthcare was nominated as the topic driver for the TG-Endoscopy.

**Documentation**

This document is the TDD for the TG-Endoscopy. 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 10.20 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:

* https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/\_layouts/15/WopiFrame.aspx?sourcedoc=%7BA115EA26-A356-4C3D-8F86-6E88EC7A9DF3%7D&file=TDD-[TG-Endoscopy]-MeetingJ.docx&action=default

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 Endoscopy.
* **Section 3** describes the existing approaches for assessing the quality of AI-based Endoscopy 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.
* **Section 5** summarizes the results of the topic group’s benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

**Status of this topic group**

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

**Status update for the meeting I**

* Discussed the proposal from Tencent
* Approved AI for Endoscopy as a use case for FG-AI4H
* Established the topic group at Meeting I (online, 7-8 May 2020)
* Nominated the topic group driver

**Status update for meeting J**

* Start the draft of TDD
* Start the draft of the call for participation
* Present the initial documents for AI for endoscopy (TG-Endoscopy)

**Status update for meeting K**

* Updates the initial documents of AI for endoscopy
* Reach out for new participants.

**Status update for meeting L**

* Transform the initial document of TG-Endoscopy into TDD template format
* Reach out for new participants.

**Status update for meeting M**

* Update the TDD for AI for endoscopy
* Add new subtopic as endoscopic ultrasound
* Invite new Participants.

**Status update for meeting N**

* Update the TDD for AI for endoscopy
* Modified the structure of chapters
* Invite new participants

**Topic group participation**

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](mailto: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 for Endoscopy and how this can help to solve a relevant ‘real-world’ problem. This topic groups summarize uses cases of a certain health topic or problem and similar AI benchmarking requirements for Endoscopy.

**Subtopics**

Currently, TG-Endoscopy started without separate subtopic groups, however, it is possible that future subtopics for gastroscopy, colonoscopy, laryngoscopy, rhinoscopy, endoscopic ultrasound might be introduced to pursue different topic-specific specializations, and the topic group driver will moderate the activities within the topic group.

**Gastroscopy**

[TBC] [call for contribution]

**Colonoscopy**

Colorectal cancers (CRC) are the third most prevalent cancer and the second-highest cause of cancer deaths worldwide. Colonoscopy is the gold standard for CRC screening, but cannot detect all colonic polyps, some of which may be 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 demonstrate that computer assisted-detection systems (CADe) might have the potential to serve as real-time ‘experts’ to improve the quality of colonoscopies.

**Laryngoscopy**

[TBC] [call for contribution]

**Rhinoscopy**

[TBC] [call for contribution]

**Endoscopic Ultrasound**

Endoscopic ultrasound (EUS) is a minimally invasive procedure in which endoscopy is combined with ultrasound to obtain images of the internal organs. EUS has emerged as an important imaging modality for the diagnosis and staging of benign and malignant lesions in the upper digestive tract and the respiratory system. EUS is most commonly used for staging of GI malignancies, evaluating pancreaticobiliary disease, evaluating subepithelial abnormalities, evaluating extraluminal abnormalities, staging of lung cancer and image guidance for therapeutic procedures [1]. Research has shown remarkable improvement of sensitivity and specificity in various tumor diagnoses using EUS. For instance, EUS is capable to identify small pancreatic tumours with a staging sensitivity greater than 90% [2]. Endobronchial ultrasound (EBUS) has been used for lung cancer staging with a diagnostic accuracy of 90–100% [3].

Research on artificial intelligence (AI) in EUS is still limited [4]. Only a handful of reports were published based on limited clinical data through retrospective or prospective studies, with a focus on pancreatic diseases. Currently, there’s no commercial AI product for EUS on the market.

**Definition of the AI task**

The application of AI in the field of endoscopy varies according to different clinical goals. In general, it is mainly divided into the following three categories: classification, detection, and segmentation.

**Classification**

Classification is a machine learning task for determining which classes are in an image, video or other types of data. It refers to training machine learning models with the intent of finding out which classes are present. In practical clinical applications, it is possible to classify with/without lesions in endoscopic entire image data from a patient. It is also possible to evaluate the image quality of all endoscopic images of the patient, from which the quality of the image meets the diagnostic quality requirements. A separate task from classification is localization, which is essentially a classification task at the pixel level of an image.

**Detection**

Object detection combines classification and localization to determine what objects are in the image or video and specify where they are in the image. It applies classification to distinct objects and uses bounding boxes. Object detection is useful in identifying objects in an image or video. In endoscopy, detection can effectively reduce the missed rate of screening. For example, in colorectal examination, the detection rate of polyps can be improved.

**Segmentation**

Image segmentation separates an image into regions in pixel level, with particular shape and border, delineating potentially meaningful areas for further processing, such as measurement, classification and object detection. The regions may not take up the entire image, but the goal of image segmentation is to highlight foreground elements and make it easier to be evaluate. Image segmentation provides pixel-by-pixel details of an object, making it different from classification and object detection. For example, in endoscopy, the tumor size and area can be automatically calculated based on the image segmentation results.

**Current gold standard**

[TBC]

**Colonoscopy**

[TBC]

**Endoscopic Ultrasound**

Endoscopic ultrasound (EUS) has become a widely used diagnostic and therapeutic procedure for various types of pancreatic diseases throughout the world. Clinical evidences have shown the benefits of EUS over the potential adverse events (AEs) and clinical guidelines have been published and continuously updated to ensure the safely use of the procedures [5][6][7][8]. European Society of Gastrointestinal Endoscopy (ESGE) has suggested EUS for pancreatic cancer screening in selected high-risk patients [5], recommended EUS-guided sampling for pancreatic solid lesions as first line procedure and EUS-guided sampling for biochemical analysis plus cytopathologic examination for pancreatic cystic lesions, etc. [6], and recommended EUS as therapeutic procedures over various types of diseases, including percutaneous transhepatic biliary drainage (PTBD), pancreatic duct (PD) drainage, etc [7].

**Relevance and impact of an AI solution**

**Colonoscopy**

[TBC]

**Endoscopic Ultrasound**

EUS has been proven to be an effective imaging modality for local-reginal staging of gastrointestinal tumors. The diagnostic ability of EUS is higher than that of computed tomography (CT), transabdominal ultrasonography, and magnetic resonance imaging (MRI) [9][10][11]. It has also proved to be a useful alternative therapeutic modality in surgery. However, EUS may be less accurate for early staging of esophageal cancer. According to a meta-analysis by Puli et al., the diagnostic accuracy of EUS was higher for T3-T4 lesions (> 90%) than T1-T2 (65%) [11]. It’s also shown low accuracy of using EUS for differentiating benign and malignant rectal cancer after treatment. Another limitation for EUS (as well as other ultrasonography procedures) is its operator-dependency. The performance of EUS improves with experiences. High inter-observer variability (61%-77%) has been reported and a wide range of overall accuracy for tumor staging could be found between different studies (63% to 95%) [9].

AI is believed to play important role in EUS procedures, not only to detect anatomical features, differentiate benign and malignant lesions, delineate lesion contours, but more important to reduce learning time for junior endoscopists, decrease work load and standardize the overall quality of EUS procedures.

**Ethical considerations**

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises several 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 concerns should be taken into careful consideration and relevant regulations should be followed. Otherwise, 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.

[TBC]

**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 the 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 related to TG- Endoscopy.

[TBC]

1. **Existing AI solutions**

Tencent Healthcare and Changhai Hospital developed a CADe system in 2021 that was built based on the You Only Look Once v2 deep learning framework. The system detects potential polyp and presents an alert rectangle surrounding polyps on a second monitor for colonoscopists. Colonoscopists detect more polyps and adenomas with the aid of computer-aided detection (CADe) system, particularly polyps < 5 mm and flat polyps. However, high efficacy is not realized in colonoscopies with inadequate bowel preparation and withdrawal time. [12]

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 the 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.[13]

Wision AI and Sichuan Provincial People’s Hospital developed a real-time automatic polyp detection system in 2018 that detects colorectal polyps during an endoscopic examination using deep learning. The detection algorithm is a deep CNN based on SegNet architecture. If any polyp is detected by the system, a hollow tracking box around would be shown on the monitor. As a conclusion, in a low prevalent ADR population, an automatic polyp detection system during colonoscopy resulted in a significant increase in the number of diminutive adenomas detected, as well as an increase in the rate of hyperplastic polyps. the cost–benefit ratio of such effects has to be determined further [14][15].

National Chiao Tung University and Tri-Service General Hospital developed a computer-aided diagnosis (CADx) system in 2018 with a deep neural network to analyze narrow-band images of diminutive colorectal polyps. The system could classify the polyps in narrow-band images as neoplastic or hyperplastic. [16]

Sun Yat-sen University developed a CADe system in 2018 with deep learning to detect upper gastrointestinal cancers by endoscopy that is named as Gastrointestinal Artificial Intelligence Diagnostic System (GRAIDS). It is the first real-time AI-aided image recognition system that has been implemented in clinical practice for detecting upper gastrointestinal cancers during endoscopy. [17]

Zhongshan Hospital and University of California developed an artificial intelligence–based CNN-CAD system through transfer learning leveraging a state-of-the-art pretrained CNN architecture, ResNet50. The system is used to determine the invasion depth of the gastric cancer and screen patients for endoscopic resection. This system distinguished early gastric cancer from deeper submucosal invasion and minimized overestimation of invasion depth, which could reduce unnecessary gastrectomy.[18]

Cancer Institute Hospital Ariake, AI Medical Service and Tada Tomohiro Institute of Gastroenterology and Proctology developed a CNN-based diagnostic system based on Single Shot MultiBox Detector architecture to detect gastric cancer in endoscopic images. This constructed CNN system for detecting gastric cancer could process numerous stored endoscopic images in a very short time with a clinically relevant diagnostic ability. [19]

Renmin Hospital of Wuhan University developed a system using a novel deep convolution neural network (DCNN) to detect early gastric cancer (EGC) without blind spots during esophagogastroduodenoscopy (EGD). This system could identify EGC from non-malignancy and classify gastric location into 10 or 26 parts with high accuracy. [20]

Kindai University developed a system in 2017 that could diagnose colon polyps as adenomatous or non-adenomatous using a simple CNN. [21]

1. **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.

**Publications on benchmarking systems for Endoscopy**

While a representative comparable benchmarking for TG-Endoscopy 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”*.

[TBC]

**Benchmarking by AI developers**

All developers of AI solutions for TG-Endoscopy 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.

[TBC]

**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).

[TBC]

1. **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).

* 1. **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 challenging 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.

* + 1. **Input data structure**

Ultrasound and endoscopic images or videos captured with gastroscope, colonoscope, laryngoscope, laryngoscope or EUS, should be submitted as separate files in the following format:

1. Endoscopic data

* Image file format: JPEG format, PNG format or BMP format.
* Image file names: be unique in the dataset and anonymize the personal information of the patient.
* Image resolution: original resolution as captured with endoscopic device, including gastroscope, colonoscope, laryngoscope, laryngoscope and EUS.
* Video file format: AVI format or MPEG-4 format.
* Video file names: be unique in the dataset and anonymize the personal information of the patient.
* Video resolution: original resolution as captured with endoscopic device, including gastroscope, colonoscope, laryngoscope, laryngoscope and EUS.

1. Ultrasound data

* Image file format: JPEG format, PNG format, BMP format or DICOM format.
* Image file names: be unique in the dataset and anonymize the personal information of the patient. For the DICOM file, anonymizing should be performed to remove sensitive information in the DICOM tag.
* Image resolution: original resolution as captured with EUS.
* Video file format: AVI format or MPEG-4 format or DICOM format.
* Video file names: be unique in the dataset and anonymize the personal information of the patient. For the DICOM file, anonymizing should be performed to remove sensitive information in the DICOM tag.

Video resolution: original resolution as captured with EUS.

* + 1. **Output data structure**

The output of the algorithm should be documented in an arranged and clear way, like a CSV, XML or JSON file with the following information.

* Info of data (name, format, etc).
* Result of the data. It would depend upon the specific condition and the type of task that is being benchmarked.
  + - 1. **Detection**
* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Location info: coordinates of a specific point (left-top or center of the bounding box) in the image. For video data, the slice index should be recorded.
* Size info: height and width in pixels.
* Task info(optional): task ID, task name, task type, etc.
  + - 1. **Classification**
* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Task info(optional): task ID, task name, task type, etc.
  + - 1. **Segmentation**
* Data info: data name, data format, etc.
* Result info:
* Category info: the types would depend on the task.
* Path of segmentation file: the stored path of the segmentation file.
* Segmentation border info(optional): coordinates of points of the segmentation mask.
* Task info(optional): task ID, task name, task type, etc.
  + 1. **Annotation procedure**

Guarantying the quality of data annotation and reducing individual differences among doctors, it is recommended that the annotation process should include multiple doctors’ independent annotation, cross-annotation, arbitration, and review.

If appropriate, a corresponding clinical diagnosis report or pathological report would be recommended for reference even the gold standard in data annotation.

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

* + - 1. **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 a rectangular box which is called a bounding box.

1. **Independent annotation**

Independent annotation by 2 or more doctors to confirm whether the endoscopic image/video contains lesions or intended objects and if so, mark the location and size of the lesion or intended objects with a bounding box. All the marked bounding boxes should be documented in a clear way, like a CSV file. Independent annotation requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

Annotated results (bounding box like [*x, y, w, h, s*], where *s* is the slice index in video and equal to 0 in image), annotator information, annotation procedure information, the date, annotation serial number.

1. **Cross-annotation**

The independent annotations by different annotators are crossed evaluated to identify the relationship between each other by calculating the similarity, like IoU (Intersection over Union)[22]. If the similarity satisfies pre-set requirements, the independent annotations would be merged to the gold standard candidate in a specific manner, like average. Crossed annotation requirements include:

* Non-annotating information (optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information (mandatory):

Cross-annotated results (bounding box like [*x, y, w, h, s*], if the pre-set requirements are not satisfied, the bounding box should be [0, 0, 0, 0, 0]), annotation serial numbers for merging, merge manner, annotation procedure information, the date, annotation serial number.

1. **Arbitration**

If the similarity calculated in the cross-annotation step does not satisfy the pre-set requirements, the corresponding data will be transferred to the arbitration expert to review and re-annotate as a gold standard candidate. Otherwise, the gold standard candidate in step Cross-annotation would be transferred to the review expert. Arbitration requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

The arbitrated results (bounding box like [*x, y, w, h, s*]), annotation serial numbers for arbitration, arbitration expert information, annotation procedure information, the date, annotation serial number.

1. **Review**

The gold standard candidates would be confirmed by the review experts one by one. The data approved by the review expert would be marked as the gold standard. Otherwise, the data without review approval would be sent back to the arbitration procedure or modified by the review expert to generate the gold standard. Review requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)

* Annotating information(mandatory):

The review results (gold standard or sent back to arbitration), serial number for review, review expert information, annotation procedure information, the date, annotation serial number.

**图形用户界面, 文本, 聊天或短信

描述已自动生成****Figure 1: Illustration of annotation procedure for detection**

* + - 1. **Annotation of classification**

Annotation of classification means arranging a category for 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 objective evidence, like pathological results.

In the subjective annotation procedure, the annotation would be made manually without objective evidence.

1. **Independent annotation**

Independent annotation of classification by 2 or more doctors to confirm which category the data should be arranged. All the annotated results should be documented in a clear way, like a CSV file. Independent annotation requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

Annotated results, annotator information, annotation procedure information, the date, annotation serial number.

1. **Cross-annotation**

The independent annotations by different annotators are crossed evaluated to identify the relationship between each other by calculating the level of consistency. If the level of consistency satisfies pre-set requirements, the independent annotations would be merged to the gold standard candidate in a specific manner, like majority rule. Crossed annotation requirements include:

* Non-annotating information (optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information (mandatory):

Cross-annotated results, annotation serial numbers for merging, merge manner, annotation procedure information, the date, annotation serial number.

1. **Arbitration**

If the level of consistency of independent annotations does not satisfy the pre-set requirements, the corresponding data will be transferred to the arbitration expert to review and re-annotate as a gold standard candidate. Otherwise, the gold standard candidate in step Cross-annotation would be transferred to the review expert. Arbitration requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race).

* Annotating information(mandatory):

The arbitrated results, annotation serial numbers for arbitration, arbitration expert information, annotation procedure information, the date, annotation serial number.

1. **Review**

The gold standard candidates would be confirmed by the review experts one by one. The data approved by the review expert would be marked as the gold standard. Otherwise, the data without review approval would be sent back to the arbitration procedure or modified by the review expert to generate the gold standard. Review requirements include:

* Non-annotating information(optional):

Image/video name, image/video identification code, collection device model, collection date, image size, collection frame rate, hospital, patient information (age, gender, race)

* Annotating information(mandatory):

The review results (gold standard or sent back to arbitration), serial number for review, review expert information, annotation procedure information, the date, annotation serial number.

图片包含 图示

描述已自动生成

**Figure 2: Illustration of annotation procedure for classification**

* + - 1. **Annotation of segmentation**

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

Segmentation annotations are usually done on the basis of detection annotations. Usually, the segmentation annotation is independently outlined by three doctors, and the overlapping area of more than two doctors is taken as the gold standard segmentation annotations. Because segmentation annotation is very time-consuming, especially if there are multiple images in a subject that need to be annotated. In order to improve the efficiency of segmentation annotation, it is sometimes allowed to sketch the contour of the object by low-level doctors, and confirm or modify by high-level experts to form the gold standard segmentation annotations.

[TBC] [call for contribution]

* 1. **Score and metrics**
     1. **Evaluation score and metrics**

This section focuses on the benchmarking processes in the context of AI for endoscopy for quality assessment. It addresses different aspects of benchmarking in 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: Benchmarking processes**

| **Methodology** | **Description** |
| --- | --- |
| True positive (TP) | The number of correctly identified samples. The number of frames with an endoscopic finding 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 erroneously are identified as a frame with an endoscopic finding. |
| Recall (REC) or  Sensitivity (SENS) | 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. |
| DICE coefficient (DICE) | This metric measures the similarity between two sets of data and is most broadly used in the validation of image [segmentation](https://radiopaedia.org/articles/segmentation?lang=us). It equals twice the number of elements common to both sets divided by the sum of the number of elements in each set. |
| Jaccard coefficient or  Intersection over Union (IoU) | This metric measures the similarity between two sets of data and is most broadly used in the validation of object detection and image segmentation. It equals the number of elements common to both sets divided by the sum of the number of unique elements in each set. |
| Polyp detection rate (PDR) | This metric is the percentage of patients undergoing screening endoscopy who have one or more polyp detected. |
| Adenoma detection rate (ADR) | This metric is the percentage of patients undergoing screening endoscopy who have one or more conventional adenomas detected. |

* + 1. **Other quality score and metrics criteria**

[TBC] [call for contribution]

* + - 1. **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. |

* + - 1. **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*).

[TBC]

* + - 1. **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.

[TBC]

* + - 1. **Reporting methodology**

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

[TBC]

* + - 1. **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.

[TBC]

* + - 1. **Discussion of the benchmarking**

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

[TBC]

* + - 1. **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).

[TBC]

1. **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 4.2.2.6).

1. **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 the 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 related to TG-Endoscopy.

* 1. **Existing applicable regulatory frameworks**

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as a 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.

* 1. **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 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).

* 1. **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.

* 1. **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 to be compliant. The discussion here focuses on the guidance and best practices 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.”*

1. **References**
2. Reddy, Yogananda, and Robert P. Willert. "Endoscopic ultrasound: what is it and when should it be used?." Clinical medicine 9.6 (2009): 539.
3. Mertz, Howard R., et al. "EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma." Gastrointestinal endoscopy 52.3 (2000): 367-371.
4. Wallace, Michael B., et al. "Minimally invasive endoscopic staging of suspected lung cancer." Jama 299.5 (2008): 540-546.
5. Kuwahara, Takamichi, et al. "Current status of artificial intelligence analysis for endoscopic ultrasonography." Digestive Endoscopy 33.2 (2021): 298-305.
6. Săftoiu, Adrian, et al. "Role of gastrointestinal endoscopy in the screening of digestive tract cancers in Europe: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement." Endoscopy 52.04 (2020): 293-304.
7. Dumonceau, J-M., et al. "Indications, results, and clinical impact of endoscopic ultrasound (EUS)-guided sampling in gastroenterology: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline." Endoscopy 43.10 (2011): 897-912.;
8. Van Der Merwe, Schalk W., et al. "Therapeutic endoscopic ultrasound: European Society of Gastrointestinal Endoscopy (ESGE) Guideline." Endoscopy (2021).
9. Forbes, Nauzer, et al. "Adverse events associated with EUS and EUS-guided procedures." Gastrointestinal endoscopy 95.1 (2022): 16-26.
10. Valero, Manuel, and Carlos Robles-Medranda. "Endoscopic ultrasound in oncology: An update of clinical applications in the gastrointestinal tract." World journal of gastrointestinal endoscopy 9.6 (2017): 243.
11. Takamichi Kuwahara,[Kazuo Hara](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Hara%2C+Kazuo),[Nobumasa Mizuno](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Mizuno%2C+Nobumasa),[Shin Haba](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Haba%2C+Shin),[Nozomi Okuno](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Okuno%2C+Nozomi),[Hiroki Koda](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Koda%2C+Hiroki),[Akira Miyano](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Miyano%2C+Akira),[Daiki Fumihara](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Fumihara%2C+Daiki), Current status of artificial intelligence analysis for endoscopic ultrasonography, Diagnostics 2021, 11(1), 18;
12. Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J, Gathercole L, Smith MA. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. Gut 2001; 49: 534-539;
13. Zhao, Sheng-Bing, et al. "Establishment and validation of a computer-assisted colonic polyp localization system based on deep learning." World Journal of Gastroenterology 27.31 (2021): 5232.
14. Japan Science and Technology Agency, Immediate detection of colorectal cancer with AI. (<https://www.jst.go.jp/EN/achievements/research/bt2019-07.html>, accessed on Jan 13, 2022)
15. Wang, Pu, et al. "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study." Gut 68.10 (2019): 1813-1819..
16. Wang, Pu, et al. "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy." Nature biomedical engineering 2.10 (2018): 741-748..
17. Chen, Peng-Jen, et al. "Accurate classification of diminutive colorectal polyps using computer-aided analysis." Gastroenterology 154.3 (2018): 568-575..
18. Luo, Huiyan, et al. "Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study." The Lancet Oncology 20.12 (2019): 1645-1654..
19. Zhu, Yan, et al. "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy." Gastrointestinal endoscopy 89.4 (2019): 806-815..
20. Hirasawa, Toshiaki, et al. "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images." Gastric Cancer 21.4 (2018): 653-660..
21. Wu, Lianlian, et al. "A deep neural network improves endoscopic detection of early gastric cancer without blind spots." Endoscopy 51.06 (2019): 522-531..
22. Komeda, Yoriaki, et al. "Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience." Oncology 93.Suppl. 1 (2017): 30-34.
23. Rosebrock, Adrian. "Intersection over Union (IoU) for object detection, 2016." URL https://www. pyimagesearch. com/2016/11/07/intersection-over-unioniou-for-object-detection (2005).

**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 | A 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 | Personally identifiable information |  |
| […] |  |  |

**Annex B:  
Declaration of conflict of interests**

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

**Tencent Healthcare (Shenzhen) Co., Ltd**

Tencent Healthcare is committed to play a good role as a digital assistant in healthcare scenario, by integrating the digital capabilities of Tencent and its partners to build series of practical products, including Tencent AIMIS, medical insurance payment by Wechat and electronic health card. By opening up Tencent's advanced technologies and services such as artificial intelligence, big data, cloud computing, etc., to realize the combination of advanced technology and healthcare.

**China Academy of Information and Communications Technology**

[TBC]

**Olympus Medical Systems Corp.**

[TBC]

**Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences**

[TBC]

**China Unicom (Guangdong) Industrial Internet Co., Ltd**

[TBC]

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