|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-J-042 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 30 September – 2 October 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | Editors | | |
| **Title:** | | Draft description of the peer review process for FG-AI4H deliverables | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Monique Kuglitsch Fraunhofer HHI Germany | | Email: [monique.kuglitsch@hhi.fraunhofer.de](mailto:monique.kuglitsch@hhi.fraunhofer.de) |
| **Contact:** | | Eva Weicken Fraunhofer HHI Germany | | Email: [eva.weicken@hhi.fraunhofer.de](mailto:eva.weicken@hhi.fraunhofer.de) |
| **Contact:** | | Naomi Lee The Lancet United Kingdom | | Email: [naomi.lee@lancet.com](mailto:naomi.lee@lancet.com) |

|  |  |
| --- | --- |
| **Abstract:** | To ensure that the WHO/ITU FG-AI4H deliverables—a key contribution of our activities—achieve the maximum level of quality and offer value for stakeholders, this document proposes a two-part (internal and external) peer review process. |

**CONTENTS**

| Page |
| --- |
| [1 Introduction 3](#_Toc52139093)  [2 Why do we need a review process for deliverables? 3](#_Toc52139094)  [3 Who should do the review process? 3](#_Toc52139095)  [4 What is the workflow for the review process? 3](#_Toc52139096)  [5 How do we find (and convince) reviewers? 5](#_Toc52139097)  [6 What does a review entail? 6](#_Toc52139098)  [7 Responding to reviewers’ comments 8](#_Toc52139099)  [Annex A: Peer review of the TDDs 9](#_Toc52139100) |

**List of Tables**

| Page |
| --- |
| [Table A.1 – Checklist for peer review of Topic Description Documents (TDDs) 9](#_Toc52139101) |

**List of Figures**

| Page |
| --- |
| [Figure 1 – Workflow of the review process. 5](#_Toc52139102) |

Description for the peer review process for FG-AI4H deliverables

# Introduction

A main goal of the International Telecommunication Union/World Health Organization Focus Group on Artificial Intelligence for Health (FG-AI4H) is to document guidelines for an internationally applicable standardized benchmarking framework for artificial intelligence (AI) in health. The FG-AI4H deliverables are output documents that cover all aspects that should be considered when benchmarking AI in health. Working groups produce deliverables that address all regulatory, ethical, data-specific, and evaluation requirements. The requirements defined by the working groups are subsequently addressed in the deliverables of the topic groups [called topic description documents (TDD)] for specific health use cases (e.g., AI-based image recognition in radiology). To ensure that the deliverables achieve the maximum level of quality and have value for stakeholders, FG-AI4H has established an internal and external peer review process. This document sets out the details for such a peer review process.

# Why do we need a review process for deliverables?

Reviews can serve multiple purposes. They can be used:

* to support authors to produce the highest quality deliverable document by providing critical evaluation and constructive comments;
* to gauge the state of the deliverable (and the maturity of the associated topic group or working group), to generate constructive feedback (particularly at earlier draft stages), and to bring the deliverable closer to the final version (or to provide a confirmation that the final version has been achieved);
* to ensure that each deliverable achieves a consistent level of quality: in technical/scientific terms, in style (format and grammar), in originality (and acknowledgement of previous work), and in completeness;
* to ensure the value of the deliverable (i.e., that it is useful for the target audience); and
* to foster comparison with other guidelines (e.g., standardized benchmarking frameworks developed for other fields or by other organizations) with a similar scope.

# Who should do the review process?

There should be an internal review process (by at least two FG-AI4H members) followed by an external review process (by at least two non-FG-AI4H members). Depending on the deliverable, the reviewers should cover the range of relevant expertise needed to thoroughly understand and evaluate the content (i.e., medical *and* artificial intelligence/machine learning content). In some instances, one reviewer may be deemed responsible for the “medical” content and a second reviewer may be deemed responsible for the “artificial intelligence/machine learning content”; in other cases, both reviewers may have sufficient experience in both fields to evaluate both aspects. Relevant expertise can be acquired through a higher education in a relevant field, demonstrable practical experience (e.g., conducting research or development for a university or company) in a relevant field, and a publication record in a relevant field. For instance, when seeking reviewers for the TDD on “Malaria Detection,” it would be recommended to select reviewers with relevant expertise in a communications field (e.g., machine learning/artificial intelligence, electrical engineering, and computer vision) and in a relevant medical/health field (e.g., tropical medicine, epidemiology, or public health). It would also be valuable if either (or both) reviewers could provide feedback on ethical or regulatory aspects. If this is not possible, it might be necessary to seek a third (or fourth) reviewer.

# What is the workflow for the review process?

The workflow for the review process has multiple steps as shown in Figure 1.

* **Submission** of a deliverable once the document is deemed ready by the authors. This is accompanied by a list of suggested external reviewers (see details in Section 5).
* **Round 1 of the review process** is by at least two FG-AI4H members who were not involved in preparation of the deliverable, do not have a conflict of interest, and have sufficient expertise to provide feedback.[[1]](#footnote-1) Reviewers are given two weeks to complete a review. The review should result in a clear suggestion of acceptance, minor revision, major revision, or rejection; accompanied by overall feedback, section-wise feedback and, for TDDs, a completed checklist. The review is submitted as a report to the *Committee of Deliverable Editors*.[[2]](#footnote-2)
* **Appraisal of the round 1 review** is by the *Committee of Deliverable Editors*. They determine the next steps for the given document: whether the reviews seem reasonable and whether the deliverable should undergo another round 1 review (i.e., if the review resulted in a major revision or rejection) or is approved for the round 2 review (i.e., if the review resulted in a minor revision or acceptance). At this point, the reviews and the decision of the *Committee of Deliverable Editors* are shared with the authors of the deliverable. If the reviewers indicated that they prefer to remain anonymous, this must be respected. Otherwise, the names of the reviewers are disclosed at this stage.
* **Round 2 of the review process** is by at least two external experts.[[3]](#footnote-3) Reviewers are given two weeks to complete a review. The review should result in a clear suggestion of acceptance, minor revision, major revision, or rejection; accompanied by overall feedback, sectionwise feedback and, for TDDs, a completed checklist. The review is submitted as a report to the *Committee of Deliverable Editors*.
* **Appraisal of the round 2 review** is by the *Committee of Deliverable Editors*. They determine the next steps for the given document: whether the reviews seem reasonable and whether the deliverable should undergo another round 1 review (i.e., if the review resulted in a major revision or rejection), whether the deliverable requires minor modifications but then will be finalized (i.e., if the review resulted in a minor revision), or whether the deliverable is finalized in the current state (i.e., if the review resulted in acceptance). At this point, the reviews and the decision of the *Committee of Deliverable Editors* are shared with the authors of the deliverable. If the reviewers indicated that they prefer to remain anonymous, this must be respected. Otherwise, the names of the reviewers are disclosed at this stage.
* **Finalized** deliverables are marked as *Approved by Peer Review Process* and can officially be used for publication.

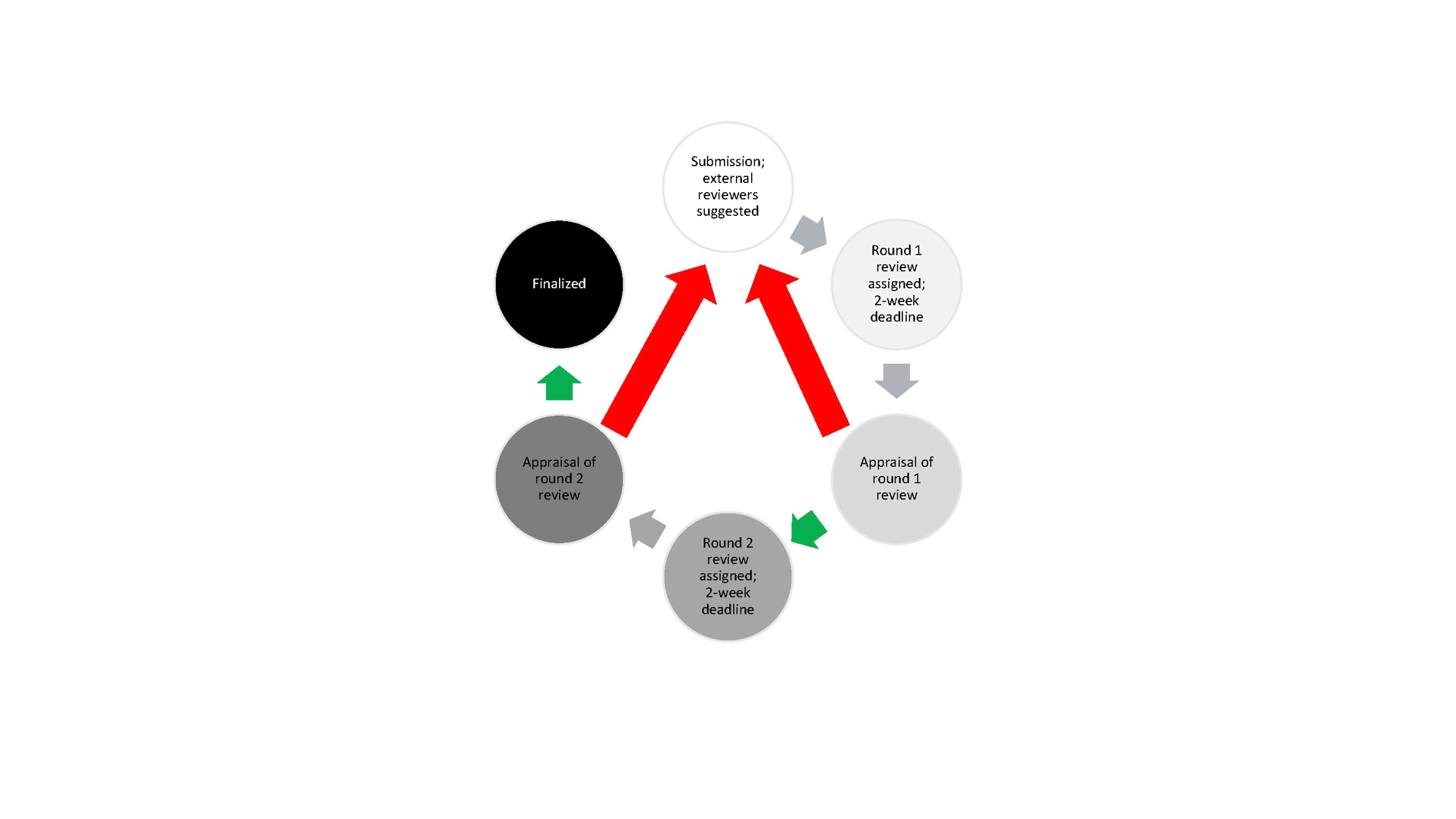


Figure 1 – Workflow of the review process.

# How do we find (and convince) reviewers?

Internal reviewers will benefit from the quid-pro-quo principle. Because each topic group/working group is preparing at least one deliverable document (and will eventually require its own review), through offering to review the deliverable in another topic/working group, there is a better chance of the favour being returned. Internal reviewers may also learn from reading other deliverables and apply these lessons to their own document.

All FG-AI4H members who are willing to serve as a reviewer should complete the *Internal Reviewer Volunteer Form*[[4]](#footnote-4) where they indicate their name, affiliation, contact details, role(s) in FG-AI4H, area(s) of expertise and justification, and whether they prefer to remain anonymous. This information will be stored in a database that is maintained by the *Committee of Deliverable Editors[[5]](#footnote-5)* and used to assist in the selection of internal reviewers. The *Committee of Deliverable Editors* makes the final selection and contacts the reviewers using the *Review Request Template: Internal*[[6]](#footnote-6) to inform them of their assignment to review a given document.

External reviewers should have sufficient expertise in the subject of interest to provide valuable feedback. However, as “external” indicates, they cannot be part of FG-AI4H or directly involved in (or benefit from) FG-AI4H activities. It is the responsibility of the deliverable authors to recommend at least five possible external reviewers [providing their name, affiliation, contact details, area(s) of expertise and justification] when they submit their deliverable for review. Depending on the subject of interest, these reviewers can be found:

* within professional societies and organizations,
* within the author list of published articles with a common or complementary theme, or
* within one’s professional network (current and past colleagues, current and past co-authors, or current and past collaborators where a conflict of interest in not perceived to exist and the reviewer can give independent feedback).

External reviewers can benefit from reviewing deliverables through:

* learning about the newest activities in AI/health standardization,
* exposure to FG-AI4H activities and connection with members, and
* pride of knowing that they are furthering FG-AI4H activities (and supporting the safe delivery of global healthcare).

As for the internal reviewers, the name, affiliation, contact details, and expertise (with justification) of all potential external reviewers should be stored in the database that is maintained by *Committee of Deliverable Editors*. The *Committee of Deliverable Editors* makes the final selection of external reviewers and uses the *Review Request Template: External[[7]](#footnote-7)* to request that they review a given document. According to the information provided in the returned template, the database should be updated.

# What does a review entail?

As indicated in Section 2, both internal and external reviewers are provided a two-week period during which to conduct the review. Reviewer instructions (e.g., from *Wiley*, *Elsevier*, and *Nature Research*) offer slight variations of the same steps. In many cases, however, these steps closely follow the structure of a scientific journal manuscript (e.g., abstract, introduction, methods, etc.) rather than an FG-AI4H deliverable. Therefore, we kindly request that the reviewer use personal judgement whether these steps are relevant. The review process is single-blind, whereby the deliverable authors are publicly known and the reviewers are anonymous. If the reviewers waive their right to anonymity, the process becomes fully open. The latter is encouraged as it can facilitate an open dialogue which is constructive and supportive. When reviewing a TDD, reviewers are requested to consider the additional criteria provided in the annex.

**Overall feedback**

Overall summary

* Reviewers should provide a short message to the authors and *Committee of Deliverable Editors* that provides an overall impression of the deliverable.

Recommendation

* Reviewers should provide a recommendation for acceptance, minor revision, major revision, or rejection of the deliverable.

First impressions

* What are the outstanding features of the deliverable?[[8]](#footnote-8)
* What is the main message being sent by the deliverable? Is this relevant and of value to the field?[[9]](#footnote-9)
* Does the deliverable present research/outcomes that are original (if not, please provide relevant references)?[[10]](#footnote-10)
* Are there any flaws to the deliverable that should prevent it from proceeding to publication (if so, please provide details)?[[11]](#footnote-11)
* What suggested improvements (e.g., additional experiments; additional sources of data) should be considered during a revision?[[12]](#footnote-12)

Format

* Is the deliverable logically structured, well written (e.g., in terms of language/grammar), and supported by the appropriate use of figures/tables?[[13]](#footnote-13)
* Is the deliverable of a reasonable length or should it be condensed (please provide more detail)?
* Does the deliverable use full sentences and paragraph form (as opposed to bullets=?

**Section-wise feedback**

Abstract

* Does the abstract provide a clear and concise summary (with a reasonable word count) and highlight key findings?[[14]](#footnote-14)

Introduction

* Is it logically organized? If not, how can it be improved?[[15]](#footnote-15)
* Does it clearly introduce (and contextualize) the text that follows?[[16]](#footnote-16)
* Does it summarize recent research and highlight gaps in knowledge? Does it justify the need for the current study? Does it suggest the target readership and application?[[17]](#footnote-17)

Methods

* Are the methods applied in the deliverable controversial/discredited (if so, please provide details and relevant references) or valid (if so, please confirm)?[[18]](#footnote-18)
* Are the methods correctly applied and clearly explained? Would it be possible to reproduce these methods and generate similar results?[[19]](#footnote-19)
* Were best practices used (e.g., in terms of ethical considerations)?[[20]](#footnote-20)
* Is there a sufficient quantity and quality of data to derive outcomes of statistical significance? Are the sources (and handling) of data clearly indicated? Are uncertainties correctly managed?[[21]](#footnote-21)

Results and discussion

* Is there a complete and coherent description of the findings (i.e., what the data show)? Are figures and/or tables provided?[[22]](#footnote-22)
* What was discovered in the data?[[23]](#footnote-23) Is there a logic in how the data are used and interpreted?[[24]](#footnote-24) Is this supported by the quantitative (statistical; e.g., significance or goodness of fit) or qualitative evidence?[[25]](#footnote-25)
* What trends are observed in the data and how do they compare to published research?[[26]](#footnote-26)
* Would additional experiments add weight to the results? If so, please provide details.[[27]](#footnote-27)
* Are open questions (areas of further research) acknowledged and discussed?[[28]](#footnote-28)

Conclusion

* Is it clearly stated whether the objectives of the study were achieved?[[29]](#footnote-29)
* Is the content consistent with the evidence provided in the preceding sections?

References

* Are all sources of information clearly cited and referenced?
* Is a consistent referencing style used throughout the deliverable?
* Should additional references be included?
* Are references relevant (e.g., recent and from reputable sources)?

# Responding to reviewers’ comments

In certain instances (e.g., in the case of minor revision or major revision in either round of review), it might be appropriate for the authors to respond directly to the comments of the reviewers. In this situation, the deliverable authors are requested to provide:

* A tracked changes version and a clean version of the deliverable and
* A point-by-point response to the reviewers comments, indicating if appropriate, where in the text (line number) changes were made.

# Annex A: Peer review of the TDDs

Topic group (TG) deliverables (i.e., TDDs) should follow the structure and content defined in document [FG-AI4H-J-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-105.docx). Therefore, in addition to considering the general features described in the main text of this document, round-1 reviewers of TDDs (No. 10 deliverables) should consider the specific features listed in FG-AI4H-J-105. A checklist is provided below. However, the round-1 review should consult FG-AI4H-J-105 for greater details.

Table A.1 – Checklist for peer review of Topic Description Documents (TDDs)

| Checklist item | Complete | Comments |
| --- | --- | --- |
| **Title page**  Is there a title page that contains the required components (e.g., document number, location/date, source, title, and contact/contributor details)? |  |  |
| **Abstract**  Is there an abstract, based on the draft provided in FG-AI4H-J-105, updated for the TG? |  |  |
| **Change notes**  Are all edits made to past versions of the TDD listed as change notes? Is all relevant information provided: the version, the previous document name (FG-AI4H-X-#), the meeting (name and location) at which the previous document was submitted, and the specific additions/updates? Have regular updates of the TDD been provided to each FG-AI4H meeting? Does the most-recent update coincide with the last FG-AI4H meeting? |  |  |
| **Table of contents**  Is the table of contents accurate (e.g., do the section titles and page numbers listed in the table of contents reflect the actual section titles and page numbers)? |  |  |
| **Introduction**  Does the introduction open with a short (half-page) general overview of the addressed health topic (including basic information about the AI task) and highlight the expected impact of the benchmarking?  Does subsection 1.1 provide information about the topic group (when/where it was established, the topic group driver, and the contributors)? Are topic group output documents listed and described in a table (1.1.1)? Is the status of the topic group (e.g., progress since the last FG-AI4H meeting) provided (1.1.2)? Are instructions on how to participate in the topic group provided (e.g., a link to the “Call for Topic Group Participation” and instructions for joining the mailing list) (1.1.3)?  Does subsection 1.1.4 describe the health topic? Are subtopics—including their focus—introduced (1.1.5)? Note, if subtopics have not yet been established, a few sentences about potentially relevant future subtopics could appear here. Are AI tasks defined (1.1.6)? Specifically, are the “AI requirements specifications” in [DEL3](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7997F2C1-5A1D-4409-B2A0-CBC4E9CE8CDA%7D&file=DEL03.docx&action=default) (i.e., the informational, functional, behavioural, and operational aspects of an AI system; e.g., what is the AI doing? What input data are fed into the AI? What output is generated?) outlined? Is the current gold standard described (1.1.7)? For instance, how is the task currently solved and what are the limitations? Is the relevance of an AI solution demonstrated (1.1.8)? Here, the document should highlight the expected impacts of deploying AI for the intended task (e.g., with regards to the public health system).  Does subsection 1.2 adequately address ethical considerations (highlighted in the output document of WG-Ethics; [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))?  Does subsection 1.3 adequately address regulatory considerations (highlighted in the output document of WG-RC; [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))? Specifically, are existing applicable regulatory frameworks (1.3.1), regulatory features to be reported by benchmarking participants (1.3.2), regulatory requirements for the benchmarking systems (1.3.3), and a regulatory approach for the topic group (1.3.4; i.e., documentation and transparency, risk management and lifecycle approach, data quality, intended use and analytical and clinical validation, data protection and information privacy, and engagement and collaboration) discussed? |  |  |
| **Existing AI solutions**  Does section 2 provide an overview (e.g., operations, limitations, robustness, and scope) of existing AI solutions (in operation) that address the given health topic? |  |  |
| **Existing work on benchmarking**  Does section 3 address existing work on benchmarking of AI systems relevant to this topic?  Does subsection 3.1 review relevant literature on the topic (please consult [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), [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), [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), [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), 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) for guidance). For instance, are scores/metrics, test data acquisition, and AI system performance (both clinical evaluation and intra-AI system comparison) discussed?  Does subsection 3.2 provide an overview of internal benchmarking systems used by AI developers to assess performance? In particular, what can be learned from these approaches (e.g., with regard to scores/metrics, test data acquisition, etc.)?  Does subsection 3.3 introduce and describe relevant benchmarking frameworks (or platforms) currently in operation? |  |  |
| **Benchmarking**  Does section 4 provide the technical and operational details on benchmarking of this topic? Does it consider [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), [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), [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), [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), [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), [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), [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), [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), [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), [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), [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), [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), [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), and [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), [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), [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), [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)?  Does subsection 4.1 provide details on the continuous improvement of benchmarking through past (and anticipated) iterations?  Does subsection 4.2 give the technical and operational details of specific benchmarking versions? Specifically, does it provide an overview of the specific benchmarking version (4.2.1; this can include the overall scope and new features), the methodology and architecture (4.2.2; e.g., the most relevant components, their interaction, and underlying technologies/frameworks; the benchmarking system data flow; safe/secure operation and hosting; and details about the benchmarking), AI input data structure for benchmarking (4.2.3; e.g., the general types of data that are fed in the AI and the encoding process), the AI output data structure (4.2.4; e.g., the nature of the output, the encoding method, and expected error types), the test data label/annotation structure (4.2.5; e.g., the general label types, the encoding method, the handling of metadata, and the embedding of labels in the input data set), the scores and metrics (4.2.6; e.g., the anticipated use of the scores and metrics by various stakeholders to make decisions; and the scores and metrics selected for various uses including medical—or non-medical—performance, robustness, and explainability), the test data acquisition (4.2.7; e.g., quality control of the data set, control mechanisms, data sources, and data storage), baseline acquisition (4.2.8; e.g., a comparison of the performance of an AI system relative to healthcare workers doing the same task), reporting methodology (4.2.9; e.g., the release of benchmarking results to participants, the publishing of results, and the existence of feedback channels), results (4.2.10; e.g., the details about the benchmarking run: when it was conducted, who participated, and the overall performance of AI systems), discussion of the benchmarking (4.2.11; e.g., insights gained from the benchmarking run), and retirement (4.2.12; e.g., the next steps for the AI system and data once the benchmarking run is completed). |  |  |
| **Discussion**  Does section 5 provide a discussion of all benchmarking activities conducted in the topic group to date? This can include: lessons learned through benchmarking, lessons learned through field implementation, lessons learned about the impacts of the AI system on healthcare, updates made to the benchmarking process and methods (regarding technical architecture, data acquisition, etc.). |  |  |
| **Annex A**  Does the glossary provide useful and accurate information that facilitates reading of the document? |  |  |
| **Annex B**  Is the declaration of conflict of interest complete for all contributors to the document? |  |  |

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

1. Selection and assignment of the round 1 reviewers and the criteria used for the review are provided in Section 5. [↑](#footnote-ref-1)
2. A small group of FG-AI4H members (selected by the FG-AI4H chair) with experience in the editorial process and with representation from FG-AI4H leadership, ITU, and WHO. Their names are public and they stay in the *Committee* for a pre-determined period of time. If the *Committee* loses a member, a suitable replacement should be found. [↑](#footnote-ref-2)
3. Selection and assignment of the round 2 reviewers and the criteria used for the review are provided in Section 5. [↑](#footnote-ref-3)
4. Form can be found on the FG-AI4H collaboration site ([FGAI4H-J-042-A01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-042-A01.docx)). [↑](#footnote-ref-4)
5. The database (e.g., Excel spreadsheet) should include the following columns: name of reviewer, affiliation, contact details, relevant expertise and justification, and request for (or waiver of) anonymity. One page of the database should be reserved for internal reviewers and another page should be used for external reviewers. [↑](#footnote-ref-5)
6. Template can be found on the FG-AI4H collaboration site ([FGAI4H-J-042-A02](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-042-A02.docx)). [↑](#footnote-ref-6)
7. Template can be found on the FG-AI4H collaboration site ([FGAI4H-J-042-A03](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-042-A03.docx)). [↑](#footnote-ref-7)
8. <https://www.nature.com/nature-research/editorial-policies/peer-review#the-review-process> [↑](#footnote-ref-8)
9. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-9)
10. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-10)
11. <https://www.nature.com/nature-research/editorial-policies/peer-review#the-review-process> [↑](#footnote-ref-11)
12. <https://www.nature.com/nature-research/editorial-policies/peer-review#the-review-process> [↑](#footnote-ref-12)
13. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-13)
14. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-14)
15. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-15)
16. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-16)
17. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-17)
18. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-18)
19. <https://www.nature.com/nature-research/editorial-policies/peer-review#the-review-process> [↑](#footnote-ref-19)
20. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-20)
21. <https://www.nature.com/nature-research/editorial-policies/peer-review#the-review-process> [↑](#footnote-ref-21)
22. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-22)
23. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-23)
24. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-24)
25. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-25)
26. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-26)
27. <https://www.elsevier.com/reviewers/how-to-review> [↑](#footnote-ref-27)
28. <https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html> [↑](#footnote-ref-28)
29. https://authorservices.wiley.com/Reviewers/journal-reviewers/how-to-perform-a-peer-review/step-by-step-guide-to-reviewing-a-manuscript.html [↑](#footnote-ref-29)