|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-G-207 | | |
| **ITU-T Focus Group on AI for Health** | | |
| **Original: English** | | |
| **WG(s):** | | Plenary | | | New Delhi, 13-15 November 2019 |
| **DOCUMENT** | | | | | |
| **Source:** | | Editor | | | |
| **Title:** | | DEL07: AI4H evaluation specification (Outline) | | | |
| **Purpose:** | | Discussion | | | |
| **Contact:** | | Markus Wenzel WG-O & Fraunhofer HHI Germany | | Email: [markus.wenzel@hhi.fraunhofer.de](mailto:markus.wenzel@hhi.fraunhofer.de) | |
|  | | Secretariat of the Focus Group | | Email: [tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int) | |

|  |  |
| --- | --- |
| **Abstract:** | The present document proposes an outline for the future deliverable "AI4H Evaluation Specification". Background: The ITU/WHO Focus Group on Artificial Intelligence for Health (AI4H) has proposed a list of deliverables at meeting "G" in New Delhi in November 2019, including this "AI4H Evaluation Specification", which combines a set of four deliverables as umbrella. |

Call for Participation

If you are interested in contributing to the *AI4H Evaluation Specification*, please contact the editor of this document (Markus Wenzel, [markus.wenzel@hhi.fraunhofer.de](mailto:markus.wenzel@hhi.fraunhofer.de)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) with "ITU/WHO Deliverable - AI4H Evaluation Specification" as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Table of Contents

[Objectives 3](#_Toc25854674)

[1 AI4H Evaluation Process Description 4](#_Toc25854675)

[1.1 Objectives 4](#_Toc25854676)

[1.2 … 4](#_Toc25854677)

[1.3 … 4](#_Toc25854678)

[1.4 Bibliography 4](#_Toc25854679)

[2 AI Test Specification 5](#_Toc25854680)

[2.1 Objectives 5](#_Toc25854681)

[2.2 … 5](#_Toc25854682)

[2.3 … 5](#_Toc25854683)

[2.4 Bibliography 5](#_Toc25854684)

[3 AI Test Metric Specification 6](#_Toc25854685)

[3.1 Objectives 6](#_Toc25854686)

[3.2 … 6](#_Toc25854687)

[3.3 … 6](#_Toc25854688)

[3.4 Bibliography 6](#_Toc25854689)

[4 Clinical Validation 7](#_Toc25854690)

[4.1 Objectives 7](#_Toc25854691)

[4.2 … 7](#_Toc25854692)

[4.3 … 7](#_Toc25854693)

[4.4 Bibliography 7](#_Toc25854694)

AI4H Evaluation Specification (Outline)

# Objectives

The ITU/WHO Focus Group on Artificial Intelligence for Health (AI4H) has proposed a list of deliverables at meeting "G" in New Delhi in November 2019. The present document proposes an outline for the future deliverable "AI4H Evaluation Specification". This deliverable combines a set of four separate deliverables as umbrella, which address four important aspects related to the evaluation of artificial intelligence (AI) and machine learning (ML) models/methods for health purposes (cf. Table I).

Table 1 – Proposed deliverable titles with initial editors and document numbers

| Deliverable title | Initial editors | Number |
| --- | --- | --- |
| AI4H Evaluation Specification | Markus Wenzel | G-207 |
| AI4H Evaluation Process Description | Sheng Wu | G-207-A01 |
| AI Test Specification | Auss Abbood | G-207-A02 |
| AI Test Metric Specification | Luis Oala | G-207-A03 |
| Clinical Validation | Naomi Lee, Manjula Singh, Rupa Sarkar | G-207-A04 |

Each editor will propose an initial outline (=Table of Contents), define the objectives of the future deliverable, and collect a bibliography of existing literature and material relevant for the development of the respective document. A short call for participation, the expertise profile of potential contributors, a time plan, and a brief characterisation of the target audience serve as preface.

* Call for Participation
* Expertise profile of potential contributors
* Time plan (first draft, release 1)
* Target audience
* Initial outline
* Objectives
* …
* Bibliography

*The remainder of this document suggests a coherent structure for the deliverables A01-A04,   
which the respective editors may use - if they think it is helpful.*

# AI4H Evaluation Process Description

Call for Participation

If you are interested in contributing to the *AI4H Evaluation Process Description*, please contact the editor of this document (Sheng Wu; [wus@who.int](mailto:wus@who.int)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) using "ITU/WHO Deliverable - AI4H Evaluation Process Description" as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Expertise profile of potential contributors

To be defined (TBD).

Time plan (first draft, release 1)

TBD.

Target audience

TBD.

Initial table of contents

TBD.

## Objectives

***[Suggestion:]*** The AI4H evaluation process description serves as overview. This process description may include, among other aspects, disentangling *in silico* tests of the AI/ML model, and health-related evaluation aspects, which may require a clinical validation, e.g. in randomized controlled trials. Furthermore, novel/specific aspects will be identified that should be considered in the case of AI. Naturally, the evaluation process description will consider the requirements to achieve regulatory clearance for the health intervention containing an AI software component - see deliverable XXX. Etc.

## …

…

## …

…

## Bibliography

*TBD. Format: APA bibliography style with DOI links. References like [Name, Year].*

# AI Test Specification

Call for Participation

If you are interested in contributing to the *AI4H Test Specification*, please contact the editor of this document (Auss Abbood, [AbboodA@rki.de](mailto:AbboodA@rki.de)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) using "ITU/WHO Deliverable - AI4H Test Specification" as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Expertise profile of potential contributors

To be defined (TBD).

Time plan (first draft, release 1)

TBD.

Target audience

TBD.

Initial table of contents

TBD.

## Objectives

***[Suggestion:]*** This document specifies how an AI can and should be tested *in silico*. Among other aspects, best practices for test procedures known from machine learning challenges will be reviewed in this document. Etc.

## …

…

## …

…

## Bibliography

*TBD. Format: APA bibliography style with DOI links. References like [Name, Year].*

Suggested reading:

* Anderson‐Cook et al. (2019). How to Host An Effective Data Competition: Statistical Advice for Competition Design and Analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(4), 271-289. [[doi](https://doi.org/10.1002/sam.11404)]
* Reinke et al. (2018). How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 388-395). Springer, Cham. [[doi](https://doi.org/10.1007/978-3-030-00937-3_45)]

# AI Test Metric Specification

Call for Participation

If you are interested in contributing to the *AI4H Test Metric Specification*, please contact the editor of this document (Luis Oala, [luis.oala@hhi.fraunhofer.de](mailto:luis.oala@hhi.fraunhofer.de)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) using "ITU/WHO Deliverable - AI4H Test Metric Specification" as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Expertise profile of potential contributors

To be defined (TBD).

Time plan (first draft, release 1)

TBD.

Target audience

TBD.

Initial table of contents

TBD.

## Objectives

***[Suggestion:]*** Review of test metrics and criteria that should be considered (performance measures, robustness, transparency/explainable machine learning, uncertainty quantification etc.). Etc.

## …

…

## …

…

## Bibliography

*TBD. Format: APA bibliography style with DOI links. References like [Name, Year].*

Suggested reading:

* Maier-Hein et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1), 5217. [[doi](https://doi.org/10.1038/s41467-018-07619-7)]
* Michaelis et al. (2019). Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv preprint*. [[doi](https://arxiv.org/abs/1907.07484)]

# Clinical Validation

*Call for Participation*

If you are interested in contributing to the *Clinical Validation* document, please contact the editors of this document (Naomi Lee, [naomi.lee@lancet.com](mailto:naomi.lee@lancet.com); Manjula Singh, [singhmanjula.hq@icmr.gov.in](mailto:singhmanjula.hq@icmr.gov.in); Rupa Sarkar, [r.sarkar@lancet.com](mailto:r.sarkar@lancet.com)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) using "ITU/WHO Deliverable - Clinical Validation" as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Expertise profile of potential contributors

To be defined (TBD).

Time plan (first draft, release 1)

TBD.

Target audience

TBD.

Initial table of contents

TBD.

## Objectives

*TBD*

## …

*…*

## …

*…*

## Bibliography

*TBD. Format: APA bibliography style with DOI links. References like [Name, Year].*

Suggested reading:

Relevant literature includes work from the EQUATOR NETWORK (such as TRIPOD-ML, CONSORT-AI, SPIRIT-AI), from Cochrane, and NICE, and further work related to evidence-based medicine in general and, in particular, with a focus on AI and ML.

TRIPOD ML

* Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, *393*(10181), 1577-1579. [[doi](https://doi.org/10.1016/S0140-6736(19)30037-6)]
* Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, *162*(1), W1-W73. [[doi](https://doi.org/10.7326/M14-0698)]

CONSORT-AI & SPIRIT-AI

* Liu, X., Faes, L., Calvert, M. J., & Denniston, A. K. (2019). Extension of the CONSORT and SPIRIT statements. *The Lancet*, *394*(10205), 1225. [[doi](https://doi.org/10.1016/S0140-6736(19)31819-7)]
* The CONSORT-AI Extension: Reporting Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (registered on 8th of May, 2019). [[link](http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials/#AI)]
* Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, *8*(1), 18. [[doi](https://doi.org/10.1186/1741-7015-8-18)]
* Protocol Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (SPIRIT-AI Extension) (registered 21 June 2019). [[link](http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials-protocols/#AI)]

COCHRANE & NICE

* Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.
* The English National Institute for Health and Care Excellence (2019). Evidence Standards Framework for Digital Health Technologies. [[link](https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf)]

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