|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-N-006-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 15-17 February 2022 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Cardio Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Cardio) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Benjamin R. Henri Muthambi, DrPH, MPH TG-Cardio Topic Driver Institutes of Institutes of Epidemiology & Public Health, Inc. South Africa & USA | | E-Mail: [brm5-@-caa.columbia.edu](mailto:brm5-@-caa.columbia.edu) |
| **Contact:** | | Vacant [Recruiting] TG-Cardio Co-Topic Driver | | E-Mail: |
| **In Memoriam:** (Served: 11/2018 – 1/2022) | | Nao Sipula, MBChB, DPH(Global Health) TG-Cardio Co-Topic Driver  WatifHealth Inc., South Africa | | E-Mail: |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based Cardiovascular Disease (CVD) Management. 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. 10.01. This draft will be a continuous input- and output document. |

|  |  |
| --- | --- |
| **Change notes:** | *Topic Driver: Please list the changes of the current TDD version in comparison to earlier versions*. *This can include content updates in specific sections, additional or completed sections, updates on subtopics, etc.*  **Version 5 (submitted as FGAI4H-N-006 to meeting N (2/2022 – online mtng)**   * Further updates to conform with new consolidated final TDD template for multiple subtopics issued 12/2020;   **Version 4 (submitted as FGAI4H-L-006 to meeting L (5/2021 – online mtng)**   * Updates to conform with new consolidated final TDD template for multiple subtopics issued 12/2020;   **Version 3 (submitted as FGAI4H-K-006 to meeting K (1/2021 – online meeting)**   * Updates to conform with new consolidated final TDD template for multiple subtopics issued 12/2020;   **Version 2 (submitted as FGAI4H-H-006-A01 to meeting H (1/2020 - in Brasilia, Brazil)**   * Subtopic-specific TDD update limited to TG-Cardio Subtopic: AI-mediated Cardiovascular Disease (CVD) risk prediction; also clarifying that other subtopics would submit separate subtopic specific TDDs;   **Version 1 (submitted as FGAI4H-B-102 to meeting B (1/2019 in Lausanne, Switzerland)**   * Pre-TDD Use case proposal adopted (revision of proposal submitted in NY, 11/2018) |

Contributors (&/or advisory members of subtopic group)

[List is in development based on participants leading sub-topics and/or currently enrolled in the mailing list of TG-Cardio Subtopic A - TG-Cardio Subtopic: AI-mediated Cardiovascular Disease (CVD) risk prediction, and other subtopic participants.]

|  |  |
| --- | --- |
| Name Benjamin R. Henri Muthambi, DrPH, MPH Company/Institute: Institutes of Institutes of Epidemiology & Public Health Country: South Africa (& USA) | Tel: +1-xx-xxx-xxxx Email: [brm5-@-caa.columbia.edu](mailto:brm5-@-caa.columbia.edu) |
| Name Nao Sipula, MBChB, DPH(Global Health) Company/Institute: WatifHealth, Inc. Country: South Africa | Tel: +27-xx-xxx-xxxx Email: [Nao-@-watifhealth.com](mailto:Nao-@-watifhealth.com) |
| Name Nathan Guo, PhD Company/Institute: Shukun Country: China | Tel: +10-89735152 Email: [guoning@shukun.net](mailto:guoning@shukun.net) |
|  |  |

*The table of contents/list of tables/ list of figures are generated automatically by MS Word provided that the correct WinWord styles are used (Heading 1, Heading 2, etc.). They can be updated with a right click when the cursor is over the table/list, then “Update table”. Please familiarize yourself on how to properly use MS word formatting with “apply styles”, in particular when adding new (sub)-sections, figures and tables.*

**CONTENTS**

| Page |
| --- |
| [1 Introduction 6](#_Toc95829327)  [2 About the FG-AI4H topic group AI-based Cardiovascular Disease (CVD) Management 6](#_Toc95829328)  [2.1 Documentation 6](#_Toc95829329)  [2.2 Status of this topic group 7](#_Toc95829330)  [2.2.1 Status update for: Meeting N (15 – 17 Feb 2022) 7](#_Toc95829331)  [2.2.2 Status updates for most recent 3 meetings: Meeting L (19 – 21 May 2021), Meetings K (27-29 January 2021) & J (2 October 2020) 8](#_Toc95829332)  [2.3 Topic group participation 9](#_Toc95829333)  [3 Topic description 10](#_Toc95829334)  [3.1 Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction 10](#_Toc95829335)  [3.1.1 Definition of the AI task 10](#_Toc95829336)  [3.1.2 Current gold standard 12](#_Toc95829337)  [3.1.3 Relevance and impact of an AI solution 12](#_Toc95829338)  [3.1.4 Existing AI solutions 13](#_Toc95829339)  [3.2 Subtopic [B] 14](#_Toc95829340)  [4 Ethical considerations 14](#_Toc95829341)  [5 Existing work on benchmarking 15](#_Toc95829342)  [5.1 Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction 15](#_Toc95829343)  [5.1.1 Publications on benchmarking systems 15](#_Toc95829344)  [5.1.2 Benchmarking by AI developers 18](#_Toc95829345)  [5.1.3 Relevant existing benchmarking frameworks 18](#_Toc95829346)  [5.2 Subtopic [B] 18](#_Toc95829347)  [6 Benchmarking by the topic group 18](#_Toc95829348)  [6.1 Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction 19](#_Toc95829349)  [6.1.1 Benchmarking version [Y] 19](#_Toc95829350)  [6.1.2 Benchmarking version [X] 29](#_Toc95829351)  [6.2 Subtopic [B] 29](#_Toc95829352)  [7 Overall discussion of the benchmarking 29](#_Toc95829353)  [8 Regulatory considerations 30](#_Toc95829354)  [8.1 Existing applicable regulatory frameworks 30](#_Toc95829355)  [8.2 Regulatory features to be reported by benchmarking participants 30](#_Toc95829356)  [8.3 Regulatory requirements for the benchmarking systems 31](#_Toc95829357)  [8.4 Regulatory approach for the topic group 31](#_Toc95829358)  [9 References 32](#_Toc95829359)  [Annex A: Glossary 33](#_Toc95829360)  [Annex B: Declaration of conflict of interests 34](#_Toc95829361) |

**List of Tables**

| Page |
| --- |
| [Table 1: Topic group output documents 7](#_Toc95829225) |

**List of Figures**

| Page |
| --- |
| [Figure 1: Example of a figure 32](#_Toc95829226) |

FG-AI4H Topic Description Document

Topic group - TG-Cardio

# Introduction

This topic description document specifies the standardized benchmarking for AI-based Cardiovascular Disease (CVD) Management systems. It serves as deliverable No. FG-AI4H-H-006-A01 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

Cardiovascular diseases (CVD) is a major threat to human health and the leading cause of death worldwide (WHO, 2014; ADA, 2019). Certain subgroups including diabetics have higher CVD risk (ADA, 2019), hence improved CVD risk prediction is also critical for better diabetes management and reducing mortality. The project aims to evaluate accuracy of various AI-based Cardiovascular Disease (CVD) Management systems (for more specifics, see TG Cardio subtopics later in this document).

# About the FG-AI4H topic group AI-based Cardiovascular Disease (CVD) Management

The introduction highlights the potential of a standardized benchmarking of AI systems for AI-based Cardiovascular Disease (CVD) Management 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-Cardio at the meeting C in Lausanne, Switzerland, 22-25 January 2019.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD.During FG-AI4H meeting C in Lausanne, Switzerland, 22-25 January 2019., Benjamin R. Henri Muthambi, DrPH, MPH from Institutes of Epidemiology & Public Health, Inc. (South Africa & USA) was nominated as topic driver for the TG-Cardio.

## Documentation

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

This document is the TDD for the TG-Cardio. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for AI-based Cardiovascular Disease (CVD) Management. It describes the existing approaches for assessing the quality of AI-based Cardiovascular Disease (CVD) Management systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group’s benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable “DEL 10.006 AI-based Cardiovascular Disease (CVD) Management (TG-Cardio).” 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-N-006-A01 | Latest update of the Topic Description Document of the TG-Cardio, AI-based Cardiovascular Disease (CVD) Management |
| FGAI4H-L-006-A01 | Latest update of the Topic Description Document of the TG-Cardio, AI-based Cardiovascular Disease (CVD) Management |
| FGAI4H-L-006-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| FGAI4H-L-006-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-Cardio, AI-based Cardiovascular Disease (CVD) Management |

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-Cardio.aspx>

Select the following link:

* + TDD for TG-Cardio (new version consolidating all subtopics: due at Meeting N: 15 – 17 Feb 2022) to be posted here: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Cardio.aspx>
  + TDD for TG-Cardio (new version consolidating all subtopics: due at Meeting L: 19 – 21 May 2021) to be posted here: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Cardio.aspx>
  + TDD for TG-Cardio | Subtopic: Cardiovascular Disease (CVD) risk prediction: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-H-006-A01.docx> (this version is getting phased out; to be replaced with new version consolidating all subtopics: due at Meeting K: 27-29 January 2021)

## Status of this topic group

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

### Status update for: Meeting N (15 – 17 Feb 2022)

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

* *Work on this document*: Completed 3rd draft of new consolidated TDD (version issued 12/2020) for most of Sections 1 through 5.1.1. including relevant sections from previous TDD update (19 – 21 May 2021) pertaining to TG-Cardio topic: AI-based Cardiovascular Disease (CVD) Management.
  + <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B87B490D1-D44E-455B-ABE3-A9CB4EC966A9%7D&file=1-FGAI4H-J-105(2021-TDD-TG-CARDIO).docx&action=default&CT=1611918628555&OR=DocLibClassicUI>
* *Work on the benchmarking software:*
  + TG-Cardio expects to use FGAI4H-sponsored benchmarking software in development; and
  + Various elements of benchmarking software to be developed during implementation of preliminary training of algorithms, and subsequent pilot evaluation studies of CVD risk prediction accuracy using academic/educational-use data already procured;
* *Progress with data acquisition, annotation, etc*: A wide range of *potential/not yet accessed* secondary data have been identified from real world data (RWD) sources already collected/archived in repositories such as electronic medical record systems, registries, and secondary data from previous research studies conducted for different objectives) in Western Europe, North America, South America, South Africa, India and China. In addition, academic/educational datasets have been procured for use in preliminary training of algorithms, and subsequent pilot evaluation studies of CVD risk prediction accuracy.
* *Overview of the online meetings including links to meeting minutes:* With recent availability of the hopefully final TDD template issued 12/2020, and completion of initial and current draft of most of sections 1 through 5.1.1. therein, TG-Cardio anticipates future online meetings focusing on completion of the TDD deliverable subsequent to FGAI4H meeting N (Feb 2022);
* *Relevant insights from interactions with other working groups or topic groups*: While TG-Cardio was largely in hiatus awaiting the updated TDD template developed throughout 2020 and issued 12/2020, TG-Cardio convenors have worked on transfer of prior content to the initial and current drafts based on the above referenced updated TDD template, and various topic group members have been encouraged and reported participation in other technical initiatives to gain insights which can be brought to bear on TG-Cardio work. Specific insights gained are expected to be incorporated into further TDD development.
* *Partners joining the topic group*: Quite a number of members have joined the mailing list of TG-Cardio Subtopic A - TG-Cardio Subtopic: AI-mediated Cardiovascular Disease (CVD) risk prediction.
* *List of current partners:* List is in development based on participants currently enrolled in the mailing list of TG-Cardio Subtopic A - TG-Cardio Subtopic: AI-mediated Cardiovascular Disease (CVD) risk prediction.
* *Relevant next steps*:
  + Continuation of development of outstanding sections of the TDD has been ongoing after Meeting K (Jan 2021) – including ongoing recruitment efforts for a co-chair of the topic group to share workload of coordination topic group activity, in conjunction with use case proposers for the other envisaged/identified subtopics in TG-Cardio.
  + Using academic/educational datasets that have already been procured, a parallel/concurrent initiative has begun preparations for preliminary training of algorithms, to be followed by subsequent pilot evaluation studies of CVD risk prediction accuracy. This is an e-course-based research initiative sponsored and spearheaded on the project-based e-learning platform of IEPH/Institutes of Epidemiology & Public Health, Inc.

### Status updates for most recent 3 meetings: Meeting L (19 – 21 May 2021), Meetings K (27-29 January 2021) & J (2 October 2020)

* **Meeting L (19 – 21 May 2021)**

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B049A6BAB-00A9-42E9-BC25-B4DD0EC64DD9%7D&file=FGAI4H-J-006-A03.pdf&action=default>

* **Meeting K (27-29 January 2021)**

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B049A6BAB-00A9-42E9-BC25-B4DD0EC64DD9%7D&file=FGAI4H-J-006-A03.pdf&action=default>

* **Meeting J (2 October 2020)**

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B049A6BAB-00A9-42E9-BC25-B4DD0EC64DD9%7D&file=FGAI4H-J-006-A03.pdf&action=default>

## Topic group participation

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

* [The publicly-accessible copy of CfTGP is outdated. To access current version, participants must register first following info below (1st URL below), in order to access CfTGP (2nd URL below)]
* Registration https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/\_layouts/15/WopiFrame.aspx?sourcedoc=%7B87B490D1-D44E-455B-ABE3-A9CB4EC966A9%7D&file=1-FGAI4H-J-105(2021-TDD-TG-CARDIO).docx&action=default&CT=1611918628555&OR=DocLibClassicUI: <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/reg2.aspx>
  + Current version of CfTGP: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-H-006-A02.docx>

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-Cardio-Risk-Prediction.aspx>

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

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

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

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

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

* TG-Cardio | Subtopic: AI-mediated Cardiovascular Disease (CVD) risk prediction:
  + [fgai4htgcardiocp@lists.itu.int](mailto:fgai4htgcardiocp@lists.itu.int)
* TG-Cardio | Subtopic: AI-mediated Cardiac Image Analyses
  + [fgai4htgcardiocia@lists.itu.int](mailto:fgai4htgcardiocia@lists.itu.int)

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

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

# Topic description

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

Topic groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Cardio has identified four (4) potential subtopics so far. The 4 TG-Cardio subtopic categories with brief information on their focus/mandate as broadly classified by [Yan et al (2019)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748906/pdf/jgc-16-08-585.pdf) are listed below, including their activation status:

[X] A. **CLINICAL/RISK PREDICTIONS -** ***Cardiovascular disease (CVD) Risk Prediction***. (Subtopic sections included in this TDD document – Subtopic Driver(s) Muthambi et al.)

[ ] B. **CARDIAC IMAGE ANALYSES – *Coronary CT Image Processing/Image Recognition for Coronary CT angiography (CCTA) in coronary artery disease (CAD) diagnosis***. (Subtopic sections to be included in this TDD document - Subtopic Driver(s) Guo et al.)

[ ] C. **INTELLIGENT ROBOTS** – ***Surgical Robot Technologies incl. AI-assisted Minimally Invasive Cardiac Surgery*** (Subtopic not activated/proposed)

[ ] D. **PRECISION MEDICINE** – ***AI-assisted Individualized Medicine and healthcare customized for each patient***. (Subtopic not activated/proposed)

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

## Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction

### Definition of the AI task

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

More specifically, the aims of the TG Cardio subtopic on CVD Risk Prediction are to:

1. Assess:
2. CVD risk prediction accuracy of various machine learning (ML) methods benchmarked against CVD risk based on actually observed occurrence of first CVD event (truth) documented in diverse cohorts/populations data, and
3. Replicability/reproducibility of ML prediction of CVD risk using 'external data' from diverse populations meeting prescribed criteria but 'not previously accessed' (undisclosed) data to the ML algorithms under evaluation;
4. Compare CVD risk prediction accuracy of several ML algorithms [referenced above under (a)] to:
5. Several routine clinical-use CVD risk scoring tools/calculators, and
6. Traditional multivariate statistical methods (in collaboration with other co-investigators who recently undertook similar risk prediction accuracy studies);
7. Determine which methods, if any; consistently show better predictive accuracy across diverse populations. Using the above-referenced methods, benchmarking, anticipated findings and peer-review thereof, the project expects to establish an evidence-based standards-setting blueprint.

Additional elucidation of what this subtopic seeks to accomplish is outlined below:

* *What the AI is doing*: In this use case, AI is used to predict risk of CVD outcomes such as myocardial infarction (MI), stroke or death.
* *The kind of AI task is implemented (e.g., classification, prediction, clustering, or segmentation task)*: More specifically, various machine learning (ML) risk prediction algorithms are used to improve CVD risk prediction over CVD risk scoring tools/calculators that are used as the standard of practice.
* *Input data fed into the AI model*: Systematic reviews show the 7 core risk factors taken into account among categories of predictors mainly used in clinical CVD risk scoring tools/prediction calculators, namely demographics (such as sex, age, race), physical examination (incl. BMI), systolic blood pressure, lipid levels & other blood variables, comorbidities (incl. history of diabetes), lifestyle (incl. smoking status), antihypertensive treatment, family history, and genetics [Dahagam et al, 2016; Alaa et al, 2019]. Lipid lowering agents, such as statins have not been historically included among predictors in these CVD risk predictions. Beyond the 7 core risk factors widely used, risk predictions using ML algorithms entail computational complexity arising from exponentially increasing the number of predictor variables to more than 400 [Alaa et al, 2019]. To distinguish the CVD risk prediction accuracy gain derived from using ML risk prediction algorithms from that derived from just using more variables, the more complex ML risk prediction using more variables can be compared to a simpler ML risk prediction using the same 7 core predictors typically used by CVD risk scoring tools/prediction calculators.
* *Output generated*: The main outcome conditions of interest are first fatal or non-fatal CVD events, defined by any of these ICD-10 diagnoses codes: I20-I25 (coronary/ischaemic heart disease), I50 (heart failure events, including acute and chronic systolic heart failure), I60-I69 (cerebrovascular disease), and F01 (vascular dementia), or any of these ICD-9 codes: 410-414 (ischemic heart disease), 436-438, and 430-434 (cerebrovascular disease). CVD risk prediction models including AI models proposed are generally designed to generate results assigning predicted levels of risk stratifications/categories to each individual whose risk of an adverse CVD outcome is being predicted (viz. risk of deaths due to myocardial infarctions, MI). Risk strata/categories are primarily represented as gradations of progressive levels of severity of predicted risk of a given adverse CVD outcome, such as:
  + Low risk
  + Normal risk
  + Intermediate risk
  + Very high risk

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

### Current gold standard

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

* *How the task is currently solved without AI*: While varying by country or region of the world, prevailing standards-of-care for CDV risk prediction include use of a variety of clinical CVD risk scoring tools/calculators (WHO, 2019) which incorporate several factors with well-established etiological associations with CVD such as age, sex, BMI, systolic blood pressure, smoking, A1C, lipid levels, age at diagnosis &/or onset of diabetes, diabetes duration, and antihypertensive and lipid-reducing drugs, but do not necessarily include a comparable set of predictors.
* *Issues which occur with the current gold standard, and limitations thereof*: These risk prediction methods often have low accuracy as they often fail to identify many people who would benefit from preventive treatment (i.e. low sensitivity resulting in higher false negative rate), while others receive unnecessary interventions (i.e. low specificity resulting in higher false positive rate).
* *Estimated performance of the current state of the art methods*: Prior studies suggest that only about ~50% of myocardial infarctions (MIs) and strokes occur among persons predicted to be at risk of CVD, with the use of clinical CVD risk scoring tools/calculator methods that are the prevailing standard of care (gold standard), [Ridker et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18997196/) .

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

### Relevance and impact of an AI solution

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

* *Why solving the addressed task with AI is relevant*: Several studies have demonstrated that certain ML algorithms can improve CVD risk prediction accuracy over CVD risk scoring tools/calculators currently used in the standard of practice in many countries. Although improvement of accuracy of CVD risk prediction was demonstrated in studies undertaken in a number of different populations, these studies used disparate study populations; either traditional multivariate statistical methods or disparate AI algorithms, specifically ML algorithms. Further, these studies examined incomparable sets of predictors, and often did not consider the broader range of potential predictor data made possible by mining electronic health records, big data aggregation, or accounting for complex interactions that ML can handle more easily. Different metrics/measures of predictive accuracy & related methodologies were also used. Notwithstanding these challenges, promising results from prior studies highlight the need to advance standards-setting and robust systematic evaluation to facilitate adoption of AI-mediated CV risk prediction and incorporation in user-friendly AI-assisted clinical Decision Support Systems (DSS).
* *Impact of deploying envisaged* *AI-assisted clinical Decision Support Systems (DSS)*: Deployment of AI/ML algorithms is expected to help improve CVD risk prediction accuracy over CVD risk scoring tools/calculators. In turn, this will help better identify patients at particular risk levels for adverse outcomes (or those with complex health needs), and enable clinicians to provide more appropriate interventions, i.e. identification of more people who would benefit from preventive treatment, while reducing the number of those who receive unnecessary interventions, due to use of less accurate CVD risk scoring tools/calculators.
* *Why benchmarking for this topic is important*: If AI/ML algorithms are shown to improve CVD risk prediction accuracy over CVD risk scoring tools/calculators currently used in the standard of practice, the evidence from benchmarking will facilitate recommendations for adoption of more accurate ML-assisted CVD risk prediction as a new standard of practice which can be incorporated into user-friendly AI-assisted clinical Decision Support Systems (DSS). Wider use of more accurate risk CVD risk prediction methods will result in more people who need indicated CVD interventions receiving them, with fewer who do not need such interventions receiving them, and tremendous efficiencies and cost-savings. Improved outcomes will in turn include reduced CVD-related morbidity and mortality

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

### Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.

* *General status and the maturity of AI systems for the TG-Cardio subtopic of CVD risk prediction*: The number of CVD risk prediction studies demonstrating potential AI/machine learning solutions is growing rapidly, and a number of health organizations are developing, piloting and implementing their own proprietary AI/ML-based clinical decision support sub-systems embedded in real world health system settings. These predictive algorithms are thus used to help identify patients at particular risk levels for adverse outcomes (or those with complex health needs), including use as adjuncts to existing standard-of-care CVD risk scoring tools/prediction calculators. Highlighting the need for standards-setting and robust evaluation before adoption of AI in health, prior systematic studies of CVD risk prediction accuracy often used: disparate study populations; either traditional multivariate statistical methods or disparate ML algorithms; incomparable sets of predictors often not considering the broader range of potential predictor data made possible by mining electronic health records, big data aggregation, or accounting for complex interactions that ML can handle more easily; and also used different metrics/measures of predictive accuracy.

Of note, none of the CVD risk prediction algorithms studied are known to have yet been approved by FDA or other countries' regulatory authorities for use as the standard-of-care for clinical decision support in patient care/individualized healthcare.

* *Currently known AI systems applicable to this use case of CVD risk prediction studies:* Studies of AI-mediated CVD risk prediction various ML algorithms, including public domain ML algorithms studied for disease risk prediction accuracy, which can be loosely categorized as:
  + Simple linear (Linear Discriminant Analysis/LDA),
  + Nonlinear (Classification and Regression Trees/CART; K-Nearest Neighbors/kNN; & gradient boosting classifier/GBC), &
  + Complex nonlinear methods (Support Vector Machines/SVM; Random Forest/RF; & Artificial Neural Networks/ANNs).

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

## Subtopic [B]

*Topic driver: If you have subtopics in your topic group, describe how the existing AI solutions in the second subtopic [B] deviate from the description in the previous section. Please use the same subsection structure as above for the first subtopic [A]. If there are no subtopics in your topic group, you can remove the “Subtopic” outline level, but - of course - you need to keep the subsections! In this case, please adapt the lower outline levels accordingly (section numbering).*

# Ethical considerations

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

* *Ethical implications and real-world consequences of biases in algorithms which highlight the need for introducing benchmarking*: Several challenges have plagued the plethora of emerging AI/ML-based clinical decision support sub-systems. Highlighting the need for standardization of evaluation of predictive fidelity of these algorithms is, most notably, racial bias recently revealed in an evaluation of an algorithm which is live and deployed at scale in the management of the health of large populations across the United States [Obermeyer et al., 2019; Paul, 2019]. This algorithm is reportedly one of the most widely used among typical examples of a class of commercial risk-prediction tools that is, by industry estimates, said to be applied, each year, to nearly 200 million people across the United States. Unaware of racial biases recently uncovered in this algorithm, which could be adversely impacting millions of African Americans as reported in the above-referenced evaluation study, large health systems and health insurance claims payers depend on this algorithm to target care management programs designed to patients predicted to be at "high-risk" of various adverse outcomes (or those with complex health needs).

**All Subtopic Groups**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

# Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and AI-based Cardiovascular Disease (CVD) Management 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.

## Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction

*Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the first subtopic [A] in this section. If there are no sub-topics, you can remove the “Subtopic” outline level, but - of course - you need to keep the subsections below!*

### Publications on benchmarking systems

While a representative comparable benchmarking for Cardiovascular Disease (CVD) risk prediction 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”*.

* *Most relevant peer-reviewed scientific publications in the literature on: benchmarking or objectively measuring the performance of systems, including most relevant approaches compared using retrospective study designs*: Three (3) domains of risk prediction approaches can be identified from relevant peer-reviewed publications in the literature that feature comparisons of CVD risk prediction methods (whereby the main methods used in the current standards-of-practice are the gold standard). Although prior studies evaluating AI/ML improvement in accuracy of CVD risk prediction often only entailed comparisons of a particular standard-of-practice clinical CVD risk scoring tool/calculator versus a given multivariate statistical risk prediction method or a given AI/ML algorithm, these comparisons can be aggregated into the above-referenced three (3) domains of CVD risk prediction methods that are listed below:

1. *Methods used as the* *current standard-of practice/gold standard*: Clinical CVD risk scoring tools/calculators are the main methods in most countries’ current standards-of-practice. These include risk calculators generating the following risk scores: ACC/AHA ([Goff et al, 2014](https://www.ncbi.nlm.nih.gov/pubmed/24222018/)), QRISK2 ([Hippisley-Cox et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18573856/)), Framingham ([D’Agostino et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18212285/)), SCORE ([Conroy et al, 2003](https://doi.org/10.1016/S0195-668X(03)00114-3)), DECODE ([Balkau et al, 2004](https://doi.org/10.1007/s00125-004-1574-5)), Reynolds Risk Score ([Ridker et al, 2007](https://www.ncbi.nlm.nih.gov/pubmed/17299196/)); UKPDS ([Simmons et al, 2009](http://care.diabetesjournals.org/content/32/4/708) ; [UKPDS risk engine](https://www.dtu.ox.ac.uk/riskengine/) ), Swedish NDR 5-yr risk equation ([Cederholm et al, 2008](http://care.diabetesjournals.org/content/31/10/2038#ref-6); [Jackson R, 2008](http://dx.doi.org/10.1136/hrt.2007.138040)), & WatifHealth algorithms ([Sipula N, 2018](http://diabetescare.africa/));
2. *Multivariate statistical risk prediction methods* incl. Cox Proportional Hazards and Multiple Logistic Regression.
3. *AI/ML algorithms previously used &/or assessed for disease risk prediction accuracy*: are loosely categorized as simple linear (Linear Discriminant Analysis/LDA), nonlinear (Classification and Regression Trees/CART; K-Nearest Neighbors/kNN; & gradient boosting classifier/GBC) & complex nonlinear methods (Support Vector Machines/SVM; Random Forest/RF; & Artificial Neural Networks/ANNs). Relevant studies/assessments of disease risk prediction accuracy of ML include [Narain et al, 2016](https://doi.org/10.2147/PPA.S108203) (FHS-USA); [Fox et al, 2016](https://doi.org/10.1001/jamacardio.2015.0300) (JHS-USA); [Ambale-Venkatesh et al, 2017](https://doi.org/10.1161/CIRCRESAHA.117.311312) (MESA-USA); [Weng et al, 2017](https://dx.doi.org/10.1371%2Fjournal.pone.0174944) (NHS-CPRD-UK); [Unnikrishnan et al, 2016](https://dx.doi.org/10.1155%2F2016%2F3016245) (BMES-AUS); & related methodology: [Rahimian et al, 2018](https://dx.doi.org/10.1371%2Fjournal.pmed.1002695) ; [Luo et al, 2016](https://dx.doi.org/10.2196%2Fjmir.5870); [Bal et al, 2014](https://dx.doi.org/10.1155%2F2014%2F137896) .

* *Scores and metrics that have been used*: As indicated in earlier sections, studies seeking to identify methods with better CVD risk prediction accuracy for first CVD event after 5 – 10 yr follow-up often entailed comparisons of a particular standard-of-practice clinical CVD risk scoring tool/calculator versus a given multivariate statistical risk prediction method or a given AI/ML algorithm. Measures of risk prediction accuracy in such comparisons mainly consisted of the following metrics:
  1. Two simple metrics to be compared are:
     + Accuracy of each risk prediction method (defined as the number of correctly predicted CVD cases divided by the total number of actually observed CVD diagnoses in each retrospective cohort used; multiplication by 100 gives a percentage, e.g. 95% accurate), and ii)
     + The degree of agreement between each risk prediction method vs. observed CVD events in each retrospective cohort used (Kappa statistic);
  2. Somewhat more advanced metrics to be compared include:
     + Area under the curve/AUC (area under the receiver operating characteristic, AUROC),
     + Sensitivity,
     + Specificity,
     + Positive predictive value (PPV),
     + Negative predictive value (NPV).
* *How test data were collected*: The Retrospective cohorts used for CVD risk prediction evaluation studies are often sourced from real-life/real-world secondary data that’s previously collected from patient electronic medical records (EMR) systems, registries or collected through previous systematic cohort studies conducted for different research objectives, whereby data sources identified already contained data on the relevant predictor and outcome variables.
* *How the AI system performed and how it compared to the current gold standard*: In general, CVD risk score calculators used as the standard-of-practice had lower accuracy in CVD risk prediction, and thus often failed to identify many people who would benefit from preventive treatment, while others receive unnecessary interventions. For example, ~50% of myocardial infarctions (MIs) and strokes occurred among persons predicted to be at risk of CVD ([Ridker et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18997196/)).
* *How the utility of the AI system can be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)*: The primary epidemiologic study design which is usually used in most CVD risk prediction accuracy evaluation studies assembles retrospective cohorts from real-life/real-world secondary data that’s sourced from electronic medical record systems, registries or previous systematic cohort studies conducted for different research objectives whereby such data already contained:
  + The relevant predictor variable data typically used for CVD risk score calculation in standard-of-practice tools/calculators, and other variables of interest for use in the AI/ML system evaluated; and
  + A 5-10-year length of follow-back to identify pre-CVD patients, thus allowing sufficient follow-up of such patients till time of occurrence of diagnoses of CVD or a censored follow-up time of 5-10 years.
* *Clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings*: Given the existence of risk score calculators already in use as the standard-of-practice for CVD risk prediction, attempts at undertaking placebo trials are likely to encounter hard-to-overcome ethics challenges in most settings. Further, it would not be scientifically valid to undertake a a 2-arm trial using the standard-of-care in one arm while the 2nd arm uses the standard of care plus AI, in the way drug trials can be done with such an approach. Even with a placebo trial, patient recruitment challenges would most likely make it harder to achieve sufficient sample size within a reasonable time. Given these challenges, the most feasible, quick and cheap studies to evaluate CVD risk prediction accuracy are still retrospective cohort studies; however, to address regulatory concerns, further review and consideration may be warranted for non-randomized real-world evidence studies which emulate RCTs using retrospective cohort data from the above-referenced real-world data sources ([Franklin et al, 2020](https://doi.org/10.1161/CIRCULATIONAHA.120.051718)).
* *Most relevant gaps in the literature (what is missing concerning AI benchmarking)*:

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

### Benchmarking by AI developers

All developers of AI solutions for Cardiovascular Disease (CVD) risk prediction 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.

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

### Relevant existing benchmarking frameworks

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

**Subtopic Group Task**: Review above section for responsiveness to the questions below, and suggest appropriate updates to this section as needed:

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

## Subtopic [B]

*Topic driver: If there are subtopics in your topic group, describe the existing work on benchmarking for the second subtopic [B] in this section using the same subsection structure as above. (If there are no sub-topics, you can remove the “Subtopic” outline level.)*

# Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the AI-based Cardiovascular Disease (CVD) Management 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).

## Subtopic A: AI-mediated Cardiovascular Disease (CVD) risk prediction

*Topic driver: Please refer to the above comments concerning subtopics.*

The benchmarking of AI-mediated Cardiovascular Disease (CVD) risk prediction is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

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

### Benchmarking version [Y]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [Y] (latest version, chronologically reversed order).

#### Overview

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

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

#### Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

##### Benchmarking system architecture

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

* How does the architecture look?
* What are the most relevant components and what are they doing?
* How do the components interact on a high level?
* What underlying technologies and frameworks have been used?
* How does the hosted AI model get the required environment to execute correctly? What is the technology used (e.g., Docker/Kubernetes)?

##### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

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

##### Safe and secure system operation and hosting

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

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

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

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

##### Benchmarking process

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

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

#### AI input data structure for the benchmarking

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

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

#### AI output data structure

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

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

#### Test data label/annotation structure

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

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

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

#### Scores and metrics

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

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

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

#### Test dataset acquisition

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

* How does the overall dataset acquisition and annotation process look?
* How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
* Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
* How was the dataset documented and which metadata were collected?
  + Where were the data acquired?
  + Were they collected in an ethical-conform way?
  + Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
  + Do the data contain ‘sensitive information’ (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
  + What kind of data anonymization or deidentification has been applied?
  + Are the data self-contained (i.e., independent from externally linked datasets)?
  + How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
  + What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
* Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
* Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
* How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
* Specific data governance derived by the general data governance document (currently [F-103](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FGAI4H-F-103-DataPolicy.pdf) and the deliverables beginning with [DEL05](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B2012357A-941E-44BD-B965-370D7829F52C%7D&file=DEL05.docx&action=default))
* How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
* Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
* How was the annotation organized?
  + How many annotators/peer reviewers were engaged?
  + Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
  + How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
  + If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
  + How was the annotation of each case documented?
  + Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
* Were data/label update/amendment policies and/or criteria in place?
* How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
* How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
* Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
* What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

#### Data sharing policies

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

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

#### Baseline acquisition

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

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

#### Reporting methodology

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

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

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

#### Result

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

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

#### Discussion of the benchmarking

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

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

#### Retirement

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

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

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

### Benchmarking version [X]

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

*Topic driver: Provide details of previous benchmarking versions here using the same subsection structure as above.*

## Subtopic [B]

*Topic driver: If there are subtopics in your topic group, please provide the details about the benchmarking of the second subtopic [B] here using the same subsection structure as above (please refer to earlier comments – in red fonts - concerning subtopics).*

# Overall discussion of the benchmarking

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

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

# Regulatory considerations

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

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

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

## Existing applicable regulatory frameworks

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

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

## Regulatory features to be reported by benchmarking participants

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

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

## Regulatory requirements for the benchmarking systems

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

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

## Regulatory approach for the topic group

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

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

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

# References

*Topic driver: Add the bibliography here.*

*Topic driver: If you include figures in this document, please use the following MS Word format/style (otherwise the figure won’t be included in the table of figures).*

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

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

Figure 1: Example of a figure

Annex A:  
Glossary

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

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

Annex B:  
Declaration of conflict of interests

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

Company/Institution/Individual XYZ

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

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