|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-K-021-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 27-29 January 2021 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Symptom Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Symptom) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Henry Hoffmann TG-Symptom Driver Ada Health GmbH Germany | | Tel: +49 177 6612889 E-Mail: [henry.hoffmann@ada.com](mailto:henry.hoffmann@ada.com) |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for AI-based symptom assessment. 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. DEL10.14. This draft will be a continuous input- and output document. |

|  |  |
| --- | --- |
| **Change notes:** | **Version 8.0 (submitted as FGAI4-K-021-A01 for the E-meeting K)**   * TODO Migrated document structure to the new TDD template published as FGAI4H-J-10x * TODO added section xxx for   **Version 7.0 (submitted as FGAI4-J-021-A01 for the E-meeting J)**   * Added 5.5 Minimal Minimal Viable Benchmarking - MMVB Version 2.2 * Replaced 4.7 with the work of the breakout group on scores and metrics * Added 2.5.7 Status Update for Meeting J (Online E Meeting) Submission * Added EQA and Barkibu to Appendix A * Renumbered 5.6 🡪 5.7 and 5.5 🡪 5.6 * Updated Appendix B (Glossary) * Updated Appendix E (Meeting list)   **Version 6.0 (submitted as** [**FGAI4-I-021-A01**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0BA86EC9-72BE-4E85-809C-630C495A3728%7D&file=FGAI4H-I-020-A01.docx&action=default) **for the E-meeting I)**   * Added 2.5.6 Status Update for Meeting I (Online E Meeting) Submission * Added 5.4 Minimal Minimal Viable Benchmarking - MMVB Version 2.1 * Renumbered 5.5 🡪 5.6 and 5.4 🡪 5.5 * Added Adam Baker to contributors and conflict of interest. * Added Baidu details to 3.1.1 Topic Group member Systems * Added Baidu details to Appendix A * Updated 4.4 Existing Regulations * Updated 4.7 Scores & Metrics * Updated abstract and small structural details in connection to providing feedback to C105 changes   **Version 5.0 (submitted as** [**FGAI4H-H-021-A01**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-H-021-A01.docx?d=wbd09cb73bf644c7698bd983e7e6b16fd) **for meeting H in Brasilia)**   * Added 2.5.5 Status Update for Meeting H (Brasilia) Submission * Updated 2.5.4 Status Update for Meeting G (Delhi) Submission * Updated 2.2 to the new Focus Group deliverable structure * Added new TG members Buoy, MyDoctor, 1Doc3 and mFine * Added 5.5 on case creation funding considerations * Added image captions and corresponding references * Migrate all meeting minutes and their references to SharePoint * Updated appendix E * Separate authors and contributors according to ITU rules * Added table captions and corresponding references.   **Version 4.0 (submitted as** [**FGAI4H-G-017**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-017.docx?d=w8d301e1e742e4843a66b9091bff1f6e2) **for meeting G in New-Delhi)**   * Updated 1.1 Ethical and cultural considerations * Added 2.5.4 Status Update for Meeting G (Delhi) Submission * Updated 2.6 Next Meetings * Extended 4.2 Clinical Evaluation * Added 5.3 MMVB 2.0 section * Added new TG member Visiba Care * Added Appendix E with a complete list of all TG meetings and related documents * Added Martin Cansdale, Rex Cooper, Tom Neumark, Yura Perov, Sarika Jain, Anastacia Simonchik and Jakub Winter to author list and/or conflict of interest declaration and/or contributors. * Merged meeting F editing by ITU/TSB (Simão Campos)   **Version 3.0 (submitted as** [**FGAI4H-F-017**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-F-017.docx?d=wf55bdbee8809480ca0635f2aaf39624d&Source=https%3A%2F%2Fextranet%2Eitu%2Eint%2Fsites%2Fitu%2Dt%2Ffocusgroups%2Fai4h%2Fdocs%2FForms%2F190903%2Easpx) **for meeting F in Tanzania)**   * Added new TG members Infermedica, Deepcare and Symptify * Added 5.2 section on the MMVB work * Added 2.5.3 Status Update for Meeting F Submission * Updated 2.6 Next Meetings * Refined 3.5 Robustness details * Removed validation outside science   **Version 2.0 (submitted as** [**FGAI4H-E-017**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-017.docx?d=w2ddecda3a8ad48b6870e6c56a7689b1b) **for meeting E in Geneva)**   * Added new TG members Baidu, Isabel and Babylon to header and appendix A. * Added the list of systems that could not be considered in chapter 3 for transparency reasons as Appendix D. * Started a section on scores & metrics. * Refined triage section. * Started the separation into subtopics "Self Assessment" and "Clinical Symptom Assessment". * Refined introduction for better readability. * Added section on benchmarking platforms including AICrowd. * Refined existing benchmarking in science section. * Started section on robustness.   **Version 1.0 (submitted as** [**FGAI4H-D-016**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-016.docx?d=w63899e533e3f41f9a72b0df6ead6a507) **for meeting D in Shanghai)**  This is the initial draft version of the TDD. As a starting point it merges the input documents FGAI4H-A-020, FGAI4H-B-021, FGAI4H-C-019, and FGAI4H-C-025 and fits them to the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6). The focus was especially on the following aspects:   * Introduction to topic and ethical considerations * Workflow proposal for Topic Group * Overview of currently available AI-based symptom assessment applications started * Prior works on benchmarking and scientific approaches including first contributions by experts joining the topic. * Brief overview of different ontologies to describe medical terms and diseases.. |

Contributors

|  |  |
| --- | --- |
| Andreas Kühn Ada Health GmbH Germany |  |
| Jonathon Carr-Brown Your.MD UK | Tel: +44 7900 271580 Email: [jcb@your.md](mailto:jcb@your.md) |
| Matteo Berlucchi Your.MD UK | Tel: +44 7867 788348 Email: [matteo@your.md](mailto:matteo@your.md) |
| Jason Maude Isabel Healthcare UK | Tel: +44 1428 644886 Email: [jason.maude@isabelhealthcare.com](mailto:jason.maude@isabelhealthcare.com) |
| Shubs Upadhyay Ada Health GmbH Germany | Tel: +44 7737 826528 Email: [shubs.upadhyay@ada.com](mailto:shubs.upadhyay@ada.com) |
| Yanwu XU Artificial Intelligence Innovation Business,  Baidu China | Tel: +86 13918541815 Fax: +86 10 59922186 Email: [xuyanwu@baidu.com](mailto:xuyanwu@baidu.com) |
| Ria Vaidya Ada Health GmbH Germany |  |
| Isabel Glusman Ada Health GmbH Germany |  |
| Saurabh Johri  Babylon Health UK | Tel: +44 (0) 7790 601 032 Email: [saurabh.johri@babylonhealth.com](mailto:saurabh.johri@babylonhealth.com) |
| Nathalie Bradley-Schmieg Babylon Health UK | Email: [nathalie.bradley1@babylonhealth.com](mailto:nathalie.bradley1@babylonhealth.com) |
| Piotr Orzechowski Infermedica Poland | Tel: +48 693 861 163 Email: [piotr.orzechowski@infermedica.com](mailto:piotr.orzechowski@infermedica.com) |
| Irv Loh, MD Infermedica USA | Tel: +1 (805) 559-6107 Email: [irv.loh@infermedica.com](mailto:irv.loh@infermedica.com) |
| Jakub Winter Infermedica Poland | Tel: +48 509 546 836 Email: [jakub.winter@infermedica.com](mailto:jakub.winter@infermedica.com) |
| Ally Salim Jr Inspired Ideas Tanzania | Tel: +255 (0) 766439764 Email: [ally@inspiredideas.io](mailto:ally@inspiredideas.io) |
| Megan Allen Inspired Ideas Tanzania | Tel: +255 (0) 626608190 Email: [megan@inspiredideas.io](mailto:megan@inspiredideas.io) |
| Anastacia Simonchik Visiba Group AB Sweden | Tel: +46 735885399 Email: [anastacia.simonchik@visibacare.com](mailto:anastacia.simonchik@visibacare.com) |
| Sarika Jain Ada Health GmbH Germany | Email:[sarika.jain@ada.com](mailto:sarika.jain@ada.com) |
| Yura Perov  EQL UK | Email: [yura@eql.ai](mailto:yura@eql.ai) |
| Tom Neumark University of Oslo Norway | Email: [thomas.neumark@sum.uio.no](mailto:thomas.neumark@sum.uio.no) |
| Rex Cooper Your.MD UK | Email: [rex@your.md](mailto:rex@your.md) |
| Martin Cansdale Your.MD UK | Email: [martin@your.md](mailto:martin@your.md) |
| Martina Fischer Germany |  |
| Lina Elizabeth Porras Santana 1DOC3 Colombia | Email: [linaporras@1doc3.com](mailto:linaporras@1doc3.com) |
| Juan Sebastián Beleño 1DOC3 Colombia | Email: [jbeleno@1doc3.com](mailto:jbeleno@1doc3.com) |
| María Fernanda González Alvarez 1DOC3 Mexico | Email: [mgonzalez@1doc3.com](mailto:mgonzalez@1doc3.com) |
| Adam Baker Babylon Health UK | Email: [adam.baker@babylonhealth.com](mailto:adam.baker@babylonhealth.com) |
| Xingxing Cao Baidu China | Email: [caoxingxing@baidu.com](mailto:caoxingxing@baidu.com) |
| Clemens Schöll Ada Health GmbH Germany |  |
| Audrey Menezes Your.MD UK | Email: [audrey@your.md](mailto:audrey@your.md) |
| Francisco Cheda Barkibu Spain | Email: [fran@barkibu.com](mailto:fran@barkibu.com) |
| Ernesto Hernández Barkibu | Email: [ernesto@barkibu.com](mailto:ernesto@barkibu.com) |

**CONTENTS**

| Page |
| --- |
| [1 Introduction 9](#_Toc62304392)  [2 About the FG-AI4H Topic Group on AI-based symptom assessment 10](#_Toc62304393)  [2.1 Documentation 10](#_Toc62304394)  [2.2 Status of this Topic Group 11](#_Toc62304395)  [2.2.1 Status update for meeting D (Shanghai) 11](#_Toc62304396)  [2.2.2 Status update for meeting E (Geneva) 11](#_Toc62304397)  [2.2.3 Status update for meeting F (Zanzibar) 12](#_Toc62304398)  [2.2.4 Status update for meeting G (Delhi) 13](#_Toc62304399)  [2.2.5 Status update for meeting H (Brasilia) 14](#_Toc62304400)  [2.2.6 Status update for meeting I (Online E Meeting) 16](#_Toc62304401)  [2.2.7 Status update for meeting J (Online E Meeting) 18](#_Toc62304402)  [2.2.8 Status update for meeting K (Online E Meeting) 20](#_Toc62304403)  [2.3 Topic Group participation 22](#_Toc62304404)  [3 Topic description 23](#_Toc62304405)  [3.1 Definition of the AI task 23](#_Toc62304406)  [3.2 Current gold standard 24](#_Toc62304407)  [3.3 Relevance and impact of an AI solution 24](#_Toc62304408)  [3.4 Existing AI solutions 25](#_Toc62304409)  [3.4.1 Topic Group member Systems for AI-based Symptom Assessment 25](#_Toc62304410)  [3.4.2 Other Systems for AI-based Symptom Assessment 27](#_Toc62304411)  [3.4.3 Input Data 29](#_Toc62304412)  [3.4.4 Output Data 31](#_Toc62304413)  [3.4.5 Scope Dimensions 34](#_Toc62304414)  [3.4.6 Additional Relevant Dimensions 35](#_Toc62304415)  [3.4.7 Robustness of systems for AI based Symptom Assessment 36](#_Toc62304416)  [4 Ethical considerations 37](#_Toc62304417)  [5 Existing work on benchmarking 38](#_Toc62304418)  [5.1 Subtopic Self-Assessment 38](#_Toc62304419)  [5.1.1 Publications on benchmarking systems 38](#_Toc62304420)  [5.1.2 Benchmarking by AI developers 38](#_Toc62304421)  [5.1.3 Relevant existing benchmarking frameworks 39](#_Toc62304422)  [5.2 Subtopic [B] 39](#_Toc62304423)  [6 Benchmarking by the Topic Group 39](#_Toc62304424)  [6.1 Benchmarking of the subtopic Self-Assessment 40](#_Toc62304425)  [6.1.1 Benchmarking version MMVB 1.0 41](#_Toc62304426)  [6.1.2 Benchmarking version [X] 56](#_Toc62304427)  [6.2 Subtopic Clinical Symptom Assessment 56](#_Toc62304428)  [7 Overall discussion of the benchmarking 56](#_Toc62304429)  [8 Regulatory considerations 57](#_Toc62304430)  [8.1 Existing applicable regulatory frameworks 57](#_Toc62304431)  [8.2 Regulatory features to be reported by benchmarking participants 58](#_Toc62304432)  [8.3 Regulatory requirements for the benchmarking systems 58](#_Toc62304433)  [8.4 Regulatory approach for the Topic Group 58](#_Toc62304434)  [9 References 59](#_Toc62304435)  [Annex A: Glossary 60](#_Toc62304436)  [Annex B: Declaration of conflict of interests 63](#_Toc62304437) |

**List of Tables**

[Table 1 – Topic Group output documents 11](#_Toc62304240)

[Table 2 – Symptom assessment systems inside the Topic Group 26](#_Toc62304241)

[Table 3 – Symptom assessment systems outside the Topic Group 29](#_Toc62304242)

[Table 4 – Overview symptom assessment system inputs 30](#_Toc62304243)

[Table 5 – Overview symptom assessment system outputs 32](#_Toc62304244)

[Table 6 – Manchester Triage System levels 32](#_Toc62304245)

[Table 7 – Benchmarking iterations 41](#_Toc62304246)

[Table 8 – MMVB input data format 48](#_Toc62304247)

[Table 9 – MMVB 1.0 API output encoding example 49](#_Toc62304248)

[Table 10 – MMVB 1.0 AI output label encoding 50](#_Toc62304249)

[Table 11 – An example of a MMVB 1.0 case-set with a single case. 50](#_Toc62304250)

[Table 12 – Case example for the London Model 53](#_Toc62304251)

**List of Figures**

[Figure 1 - "London Model" used for sampling cases for MMVB 1.0 42](#_Toc62304319)

[Figure 2 - MMVB 1.0 High-level architecture 43](#_Toc62304320)

[Figure 3 - MMVB 1.0 case generation UI 45](#_Toc62304321)

[Figure 4 - MMVB 1.0 screen for running a benchmarking session 46](#_Toc62304322)

[Figure 5 - MMVB 1.0 result screen 46](#_Toc62304323)

FG-AI4H Topic Description Document

Topic Group Symptom Assessment

# Introduction

This topic description document specifies the standardized benchmarking for AI-based symptom assessment systems. It serves as deliverable No. DEL10.14 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

The World Health Organization estimates the shortage of global health workers to increase from 7.2 million in 2013 to 12.9 million by 2035 [WHO2013]. This shortage is driven by several factors including growing population, increasing life expectancy and higher health demands. The 2017 Global Monitoring Report by the WHO and the World Bank reported that half of the world's population lacks access to basic essential health services [WHO/WB2017]. The growing shortage of health workers is likely to further limit access to proper health care, reduce doctor time, and worsen patient journeys to a correct diagnosis and proper treatment.

In recent years, one promising approach to meet the challenging shortage of doctors has been the introduction of AI-based symptom assessment applications that have become widely available. This new class of system provides both consumers and doctors with actionable advice based on symptom constellations, findings and additional contextual information like age, sex and other risk factors. By navigating users to the right care at the right time such systems help using the resources of the health systems more efficient. On the doctors side such systems help to save time by allowing for an automated collection of relevant information before seeing the doctor and to reduce the risk of misdiagnosis.

As an input theses AIs get beside general health profile information the initial presenting complains a user seeks advice for. These systems then follow up with a dialog collecting further evidence on other symptoms the user might have experienced to then present a report providing a general health advice on possible next steps like self-care, to see a pharmacy or seek emergency care, a list of diseases that might have caused the symptoms and explanations on how the symptoms and theses suggestions are related.

While systems for AI-based symptom assessment have great potential to improve health care, the lack of consistent standardisation makes it difficult for organizations like the WHO, governments, and other key players to adopt such applications as part of their solutions to address global health challenges.

The implementation of a standardized benchmarking for AI based symptom assessment applications by the ITU/WHO AI4H Focus Group will therefore be an important step towards closing this gap. Paving the way for the safe and transparent application of AI technology will help improve access to healthcare for many people all over the globe.

# About the FG-AI4H Topic Group on AI-based symptom assessment

The introduction highlights the potential of a standardized benchmarking of AI systems for AI-based symptom assessment to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Symptom at the meeting C in Lausanne, Switzerland, 22-25 January 2019.

It was based on the "symptom checkers" use case, which was accepted at the November 2018 meeting B in New York building on proposals by Ada Health:

* [A-020](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-A-020.docx?d=we280696f99e945f8894a510ff75eeed0): Towards a potential AI4H use case "diagnostic self-assessment apps"
* [B-021](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-B-021-R1.docx?d=w501a8384bf674f8c909d2ab13f52a173): Proposal: Standardized benchmarking of diagnostic self-assessment apps
* [C-019](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-019.docx?d=w0a5639a0e26f474f88c76d7b889dd3eb): Status report on the "Evaluating the accuracy of 'symptom checker' applications" use case

and on a similar initiative by Your.MD:

* [C-025](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-025.docx?d=w6a05e1d093fe4a50915c3f58a299eeb8): Clinical evaluation of AI triage and risk awareness in primary care setting

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, Henry Hoffmann from Ada Health GmbH, Berlin, Germany was nominated as topic driver for the TG-Symptom.

## Documentation

This document is the TDD for the Topic Group on AI-based symptom assessment. 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 symptom assessment. It describes the existing approaches for assessing the quality of such 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 10.14 Symptom assessment (TG-Symptom).” 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-x-021-A01 | Latest update of the Topic Description Document of the TG-Symptom |
| FGAI4H-x-021-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| FGAI4H-x-021-A03 | The presentation summarizing the latest update of the Topic Description Document of the TG-Symptom |

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-Symptom.aspx>

Select the following link:

* <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3B399B54-B2D8-4482-8ED3-44ED80FA22FD%7D&file=FGAI4H-x-021%20-%20TG%20Symptom%20TDD%20draft.docx&action=default>

## Status of this Topic Group

The following subsections describe the update of the collaboration within the TG-Symptom for the official Focus Group meetings.

### Status update for meeting D (Shanghai)

With the publication of the "call for participation" the current Topic Group members, Ada Health and Your.MD, started to share it within their networks of field experts. Some already declared general interest and are expected to join official via input documents at meeting D or E. Before the initial submission of the first draft of this TDD it was jointly edited by the current Topic Group members. Some of the approached experts started working on own contributions that will soon be added to the document. For the missing parts of the TDD where input is needed the Topic Group will reach out to field experts at the upcoming meetings and the in between.

### Status update for meeting E (Geneva)

With Baidu joining at meeting D we introduced the Topic Group differentiation into the subtopics "self-assessment " and "clinical symptom assessment". The corresponding changes to this TDD have been started, however there at the current phase they are still quite close and will mainly differ in the symptom input space and condition output space. Shortly after meeting D Isabel Healthcare, one of the pioneers of the field for diagnostic decision support systems for non-academic use, joined the Topic Group for both subtopics. In the week before meeting E Babylon Health, a large London-based digital health company developing the popular Babylon symptom checker app, joint the Topic Group too.

With more than two participants, the Topic Group on 08.05.2019 started official online meetings. The protocol of the first meeting was distributed through the ai4h email reflector. We will also work on publishing the protocols in the website.

The refinement of the TDD involved primarily:

* adding the new members to the document
* adding the separation into two sub-topics
* the refinement of the triage section
* an improved introduction
* adding a section on benchmarking platforms including AICrowd

The detailed list of the changes is also given in the "change notes" at the beginning of the document.

### Status update for meeting F (Zanzibar)

During meeting E in Geneva, the Topic Group for the first time had a breakout session discussing the specific requirements for benchmarking of AISA systems in person. This meeting can be seen as the starting point for the multilateral work on a standardized benchmarking for this Topic Group.  
  
It was decided that the main objective of the Topic Group for meeting F in Zanzibar was to create a Minimal Minimal Viable Benchmarking (MMVB). The goals of this step as an explicit step before the Minimal Viable Benchmarking (MVB) are:

* show a complete benchmarking pipeline for AISA
* with all parts visible so that we can all understand how to proceed
* get first benchmarking result numbers for Zanzibar
* learn relevant things for MVB that might follow in 1-2 meetings

For discussing the technical details of the MMVB the group held a meeting from 11 - 12 July 2019 in London. A first benchmarking system based on an Orphanet rare disease model was presented and discussed. The main outcomes of this meeting were as follows:

* An agreed-upon set of 11 conditions, 10 symptoms, 1 factor medical model to use for the MMVB.
* To use the pre-clinical triage levels "self-care", "consultation", "emergency", "uncertain" for MMVB
* The data structures to use for the inputs and outputs.
* The agreement on technology agnostic REST API calls for accessing AIs.
* The plan how to work together on drafting a guideline to create/annotate cases for benchmarking.

Based on the meeting outcomes in the following week a second Python based benchmarking framework using the agreed upon data structures and the 11 disease "London" model was implemented and shared via [github](https://github.com/Babylonpartners/itu_who_2019_symptom_assessment_mmv_benchmark).

In addition to the London meeting the group had also 3 other phone calls. The following list shows all meetings together with their respective protocol links:

* 30.5.2019 - Meeting #2 - Meeting E Breakout [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2030-05-2019%20meeting%202%20minutes.docx?d=w6418637cd3e6475f8a5318789527721b)
* 20.06.2019 - Meeting #3 - Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2020-06-2019%20%20meeting%203%20minutes.docx?d=w215896dfe442471cb07f634fbaebe5a6)
* 11-12.7.2019 - Meeting #4 - London Workshop [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-12-2019%20meeting%2014%20minutes.docx?d=w587cebcda07343f0a38184465a1d0f19)
* 15.8.2019 - Meeting #5 - Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2015-08-2019%20meeting%205%20minutes.docx?d=wbe4764613f8f46df905a8efc1a6757fa)
* 23.08.2019 - Meeting #6 - Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2023-08-2019%20meeting%206%20minutes.docx?d=wf9c51f181a9f49258fd28861494de696)

Since the last meeting the Topic Group was joined by Deepcare.io, Infermedica, Symptify and Inspired Ideas. Currently the Topic Group has the following members:

* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay)
* Babylon Health (Saurabh Johri, Yura Perov, Nathalie Bradley-Schmieg)
* Baidu (Yanwu XU)
* Deepcare.io (Hanh Nguyen)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Symptify (Dr Jalil Thurber)
* Your.MD (Jonathon Carr-Brown, Rex Cooper)

At meeting E there was also the agreement that Topic Groups might have their own email reflector. Due to the significant number of members the Topic Group therefore decided to introduce [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) as the groups email reflector.

### Status update for meeting G (Delhi)

At the meeting F in Zanzibar the Topic Group presented a first MMVB - a "minimal minimal viable benchmarking". It showed a first benchmarking pipeline for AI-based symptom assessment systems using synthetic data sampled from a simplistic model and a collection of toy-AI. The main goal of the MMVB was to start learning what benchmarking for this Topic Group could look like. A simple model was chosen to gain insights in the first iteration, onto which more complex layers could be added for subsequent versions. For the latest iteration, the corresponding model and systems are called MMVB 2.0. In general, we expect to continue with further MMVB iterations until all details for implementing the first benchmarking with real data and real AI have been investigated - a version that is then called MVB.

As for the first MMVB iteration we have chosen a workshop format for discussing the technical details of the next benchmarking iteration. The corresponding workshop was held from 10-11.10.2019 in Berlin. As inclusiveness is a key priority for the Focus Group as a whole we also supported remote participation. In the meeting we agreed primarily on:

* Having independent from the MMVB 2 a more cloud based MMVB 1 version benchmarking cloud hosted toy AIs.
* The structure for how to encode attributes of symptoms and findings - a feature that is crucial for benchmarking self-assessment systems.
* A cleaner approach towards factors as the MMVB version.
* An approach how to continue with creation of benchmarking data.
* Exploring whether a 'pruned' subset within SNOMED exists for our use case (to map our symptom ontologies to)

Over the next weeks after the workshop the technical details have then been further refined. All together the have been the following meetings since meeting F:

* 03.08.2019 – Meeting #7 – Meeting F Breakout [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2003-09-2019%20meeting%207%20minutes.docx?d=w32c9a3a90f0645d9bbe335fe88af79de)
* 27.09.2019 – Meeting #8 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2027-09-2019%20meeting%208%20minutes.docx?d=w416a633039c545afa0bc485cba1ffabb)
* 10-11.10.2019 – Meeting #9 – Berlin Workshop [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2010-10-2019-11-10-2019%20meeting%209%20minutes.docx?d=w28f04cfffbe047998a45bc006ac1bd15)
* 17.10.2019 – Meeting #10 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2017-10-2019%20meeting%2010%20minutes.docx?d=w11b253bdb08f4477b39300c857e6ffdb)
* 20.10.2019 – Meeting #11 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2020-10-2019%20meeting%2011%20minutes.docx?d=w43d3b5f99d224d89811c7b48410c8e52)
* 25.10.2019 – Meeting #12 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2025-10-2019%20meeting%2012%20minutes.docx?d=wee119da8d7a64fb796b9a4c182a662e8)
* 30.10.2019 – Meeting #13 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2030-10-2019%20meeting%2013%20minutes.docx?d=w0bfe95524aab406699210287322793a7)

At the time of submission, the MMVB 2 version of the benchmarking software has not been completed yet. The plan is to present a version running on the new MMVB 2 model (also called the "Berlin Model") by the start of meeting G in Delhi.

While the Berlin Model relies on custom symptoms and condition the MVB benchmarking needs to use an ontology all partners can map to. In a teleconference call with SNOMED expert (Ian Arrowsmith) who had, in a prior role, been involved in creating SNOMED findings (minutes in meeting 12 as an addendum), discussion provided some avenues and contacts to help us discover whether it is indeed possible to find a refined subset of SNOMED for our use case to map common symptom and attribute ontologies to.

Beside the work on a MMVB 2 version of model and software we also started to investigate options for funding the independent creation of high-quality benchmarking data. Here we reached out to the Botnar Foundation and the Wellcome trust who have followed and supported the Focus Group since meeting A in Geneva. We expect to integrate their feedback for the funding criteria and requirements in one of the upcoming iterations of this document.

Since meeting F the group was joined by a new company Buoy (Eddie Reyes), mfine (Dr Srinivas Gunda), MyDoctor (Harsha Jayakody), Visiba Care (Anastacia Simonchik). For the first time the group was also joined by the individual experts Muhammad Murhaba (Independent Contributor, NHS Digital) and Thomas Neumark (Independent Contributor, University of Oslo) who supported the group with outreach activities and contributions.

Currently the Topic Group has the following 10 companies and 2 individuals as members:

* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay)
* Babylon Health (Saurabh Johri, Yura Perov, Nathalie Bradley-Schmieg)
* Baidu (Yanwu XU)
* Buoy (Eddie Reyes)
* Deepcare.io (Hanh Nguyen)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Muhammad Murhaba (Independent Contributor)
* MyDoctor (Harsha Jayakody)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale)

The Topic Group email reflector [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) altogether has currently 44 subscribers. The latest Meeting G version of this Topic Description Document lists 20 contributors.

### Status update for meeting H (Brasilia)

Due to limited development resources (vacation, Christmas-break) since the last meeting, our work concentrated on extending the MMVB 1 system. We focused on a feature supporting the benchmarking of the cases defined by our doctors, in addition to the benchmarking with synthetic cases. The updated version has been published to github and deployed to the demo system. The work also included adding another toy AI from the Topic Group member Inspired Ideas.

In the time since the last meeting the Topic Group had primarily one telco for aligning on the steps for meeting H:

* 06.12.2019 – Meeting #14 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-12-2019%20meeting%2014%20minutes.docx?d=w587cebcda07343f0a38184465a1d0f19)
* 06.01.2020 – Meeting #15 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-01-2019%20meeting%2015%20minutes.docx?d=wa35cec942d7b4fcda60b0832e3cf9613&Source=https%3A%2F%2Fextranet%2Eitu%2Eint%2Fsites%2Fitu%2Dt%2Ffocusgroups%2Fai4h%2Ftg%2FSitePages%2FTG%2DSymptom%2Easpx)

In addition to this, our Topic Group also joined with three representatives the workshop of the DAISAM and DASH working groups from 8-9 of January 2020 in Berlin. We contributed there to all tracks and put emphasis on the special requirements of the benchmarking of systems for AI based symptom assessment. The results from these discussions will be reflected in this document over the next versions.

Since the last meeting, the Topic Group approached the Wellcome Trust and the Botnar foundation exploring funding options for the creation of case cards (for more info see 5.5 below). An initial phone call with the Wellcome Trust including Alexandre Cuenat (who previously attended the ITU/WHO AI4H meetings) was arranged. Mr. Cuenat offered to look into opportunities with Wellcome Centres. It was recommended that we look into direct funding options of the Wellcome Innovation stream e.g. applying for an Innovator Award. The Topic Group also received an email from the Botnar foundation, stating that they would get back to us in January. Both opportunities require further exploration in the time after meeting G.

For the Meeting H version of this document we also merged the reformatting done by ITU and revised indexing and descriptions of tables and figures. With the introduction of the new SharePoint folder for all Topic Groups, our Topic Group started migrating all documents to the corresponding TG-Symptom folder <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>. As part of this, the latest TDD draft can always be found under [https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%20TDD%20draft.docx?d=wb569618c24f1445daa93f93aca2bb875](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/OLD%20FGAI4H%20TG%20Symptom%20TDD%20draft%20OLD.docx?d=wb569618c24f1445daa93f93aca2bb875). The protocols of all Topic Group internal meetings have also been uploaded to the folder and the references in this TDD have been updated accordingly.

Since meeting G there has also been some exchange with Baidu, who joined the Topic Group with a focus on the clinical symptom assessment. We are looking forward to integrating material on the benchmarking of AI systems in the clinical context for meeting I.

As our Topic Group is now one of the largest and longest existing ones, we have also been more involved in supporting the onboarding of new Topic Groups. For this we met with members of the newly formed Topic Group Dental Imaging to share insights on starting a Topic Group.

Since the submission for this TDD for meeting G, the Topic Group was joined by 1Doc3, Buoy, mFine and MyDoctor. MyDoctor and mFine joined meeting G and have been onboarded by the group during this meeting. With the new Topic Group members Buoy and 1Doc3 we conducted online onboarding meetings.

Currently the Topic Group has the following 14 companies and 2 individuals as members:

* 1Doc3 (Lina Porras)
* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay, Dr Martina Fischer)
* Babylon Health (Saurabh Johri, Yura Perov, Nathalie Bradley-Schmieg)
* Baidu (Yanwu XU)
* Buoy (Eddie Reyes)
* Deepcare.io (Hanh Nguyen)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Mfine (Dr Srinivas Gunda)
* Muhammad Murhaba (Independent Contributor)
* MyDoctor (Harsha Jayakody)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale)

The Topic Group email reflector [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) altogether has currently 56 subscribers (12 more than for Meeting G). The latest Meeting H version of this Topic Description Document lists 22 (2 more) contributors.

### Status update for meeting I (Online E Meeting)

As the update for meeting H outlined, the work there was focused on extending the current MMVB version to support doctor cases and to connect more toy-AIs. With some new developers joining the Topic Group, since then we could focus more on the next important step of implementing the changes agreed upon at the Berlin workshop in November 2019. Beside a strong focus on the Berlin model extending the London model by symptom attributes and factors this also included more flexible frontend result report drill down, a more refined scoring and metric systems and in general moving the benchmarking system closer to the one needed for the MVB. Given the requirements of the Berlin model it became clear that implementing them would be easier if the software would be separated into dedicated frontend and backend applications, both using tech-stacks allowing to implement more complex features in a more stable and future-proof way. At the time of Meeting I this reimplementation is almost finished. The details of both the new frontend and the new backend are described in “5.4 Minimal Minimal Viable Benchmarking - MMVB Version 2.1”.

At meeting H the Topic Group was also joined by Alejandro Osornio, an expert for ontologies. In the weeks following he proposed a technical solution for how to use SNOMED CTfor encoding the symptoms of the Berlin model. An overview of this work will be outline in section “3.2.2 Ontologies for encoding input data” (not in version yet) and based on this the current implementation work will integrate a mapping to an ontology earlier than expected. Continuing the ontology mapping after meeting I will be one of the priorities.

As suggested in the last meeting the Focus Group started the work on updating the [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6) template for TDDs. Our Topic Group reviewed the draft and contributed the insights from working on this TDD. Once a new version is adopted by the Focus Group we will adjust this TDD to the new structure.

During meeting H the Focus Group discussed the possibility of working on a joint Topic Group overarching tool for creating and annotating benchmarking test data. As part of this discussion our Topic Group also contributed to an initial requirements document. After the meeting this discussion was continued in several online meetings with WG-DASH.

Since the last meeting we also intensified our online collaboration. For coordinating the implementation work we introduced a weekly tech telco. For bringing the clinical discussion on scores and metrics forward the doctors inside the group also started a meeting series. The following list shows all the online meetings since the meeting H:

* 28.03.2020 – Meeting #17 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B24C1610B-A03B-40B8-A47C-9D539E5827CC%7D&file=FGAI4H%20TG%20Symptom%2028-02-2020%20meeting%2017%20minutes.docx&action=default)
* 12.03.2020 – Meeting #18 – Tech Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BDD2C5742-4422-465B-863C-A737B58BAA6D%7D&file=FGAI4H%20TG%20Symptom%2012-03-2020%20meeting%2018%20minutes.docx&action=default)
* 13.03.2020 – Meeting #19 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B18E9043E-AB29-4428-BA7D-251F662C06A7%7D&file=FGAI4H%20TG%20Symptom%2013-03-2020%20meeting%2019%20minutes.docx&action=default)
* 20.03.2020 – Meeting #20 – Tech Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8236F039-4128-484B-8318-2A1A0CA98D5F%7D&file=FGAI4H%20TG%20Symptom%2020-03-2020%20meeting%2020%20minutes.docx&action=default)
* 27.03.2020 – Meeting #21 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0877EB0F-D11F-4D5A-95BE-EFD332F5014A%7D&file=FGAI4H%20TG%20Symptom%2027-03-2020%20meeting%2021%20minutes.docx&action=default)
* 15.04.2020 – Meeting #22 – Tech Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2015-04-2020%20meeting%2022%20minutes.pdf)
* 22.04.2020 – Meeting #23 – Tech Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BFEC194B5-F4CA-4AF4-AF1C-F9FAEA48D339%7D&file=FGAI4H%20TG%20Symptom%2022-04-2020%20meeting%2023%20minutes.docx&action=default)
* 21.04.2020 – Meeting #24 – Clinical Telco (no minutes)
* 24.04.2020 – Meeting #25 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B9072195C-7A17-4DAC-B5CA-ED8768885ECD%7D&file=FGAI4H%20TG%20Symptom%2024-04-2020%20meeting%2025%20minutes.docx&action=default)

All the meetings notes and also be found in the official TG-Symptom SharePoint folder: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

We also started to publish our TG internal Focus Group meeting reports the. The summary of meeting H can be found here:

* [TG-Symptom update on Meeting H](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B518476AA-C7B6-4390-AA3A-5FFD3A95B92B%7D&file=FGAI4H%20TG%20Symptom%20update%20on%20Meeting%20H.docx&action=default)

In addition to the meetings, we also now use the TG slack channel more for ad-hoc communication around technical implementation details and also for the clinical discussion (please reach out to the Topic Driver for details on how to join). Currently it is used by 21 people in the group.

Since Meeting H, we have been joined by three independent contributors, namely Pritesh Mistry, Alejandro Osornio and Salman Razzaki. One company (XUND, represented by Lukas Seper) also joined. In addition, Yura Perov (previously at Babylon) also joined in an independent capacity.

Currently, our Topic Group has the following 15 companies and 6 independent contributors:

* 1Doc3 (Lina Porras and Maria Gonzalez)
* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay, Dr Martina Fischer)
* Alejandro Orsonio (Independent Contributor)
* Babylon Health (Saurabh Johri, Nathalie Bradley-Schmieg, Adam Baker)
* Baidu (Yanwu XU)
* Buoy (Eddie Reyes)
* Deepcare.io (Hanh Nguyen)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Mfine (Dr Srinivas Gunda)
* Muhammad Murhaba (Independent Contributor)
* MyDoctor (Harsha Jayakody)
* Pritesh Mistry (Independent Contributor)
* Dr Salman Razzaki (Independent Contributor)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* XUND (Lukas Seper, Tamás Petrovics, Sophie Pingitzer)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale
* Yura Perov (Independent Contributor)

The Topic Group email reflector [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) altogether has currently 83 subscribers (27 more than for Meeting H). The latest Meeting I version of this Topic Description Document lists 28 (6 more) contributors.

### Status update for meeting J (Online E Meeting)

The work between meeting I and meeting J is divided into two large areas. The first focus was on the finalization of the implementation of the Berlin model. With the separation of the benchmarking system in frontend and backend the implementation was also finished by two teams, one on the backend side. While on both sides the data structures and interface had to be extended to the Berlin models more complex attribute and factor model, the frontend also improved usability and design. The backend had an additional focus to extend the case synthesizer generating the synthetic toy data used for testing the benchmarking system. Building on the new systems the members of the Topic Group started adapting their toy AIs to the new changed backend API interfaces and protocols. At the time of submission of the TDD version for meeting J three toy AIs have be completed with the others to follow in the weeks after meeting J.

With the current version of the software we also introduced the separation between the benchmarking system and the system for annotating/creating new cases by doctors. The corresponding annotation tool was also extended to support the Berlin model. Based on it we expect doctors to start creating benchmarking case vignettes before meeting J and continuing for the weeks after so that we again have the results for both synthetic and real cases. In anticipation of the upcoming next steps on extending the toy model with only 12 diseases and 12 symptoms to a fully condition and symptom space, we have already started to use SNOMED identifiers in the benchmarking system.

The details describing the technical work on frontend, backend and annotation tool can be found in section 5.4 Minimal Minimal Viable Benchmarking - MMVB Version 2.1.

The second large area of work was dedicated to scores and metrics. For driving this forward the doctors inside the Topic Group formed a temporary breakout group working on a document covering all relevant aspects on this topic in full details. The corresponding work was then edited and inserted as new section 4.7 Scores & Metrics into this document.

After meeting I we also continued our contribution to a new template for a topic description documents. The resulting document was submitted as [FGAI4H-J-004](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-J-004.docx) to meeting J.

All the work in the Topic Group was organized online. The following list shows all the online meetings since the since meeting I:

* 29.05.2020 – Meeting #26 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B597F0813-4A2A-479D-AE85-08FEB7304699%7D&file=FGAI4H%20TG%20Symptom%2029-05-2019%20meeting%2026%20minutes.docx&action=default)
* 11.06.2020 – Meeting #27 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5FCF70C5-DDEE-473F-9A14-83DCD71E7307%7D&file=FGAI4H%20TG%20Symptom%2011-06-2019%20meeting%2027%20minutes.docx&action=default)
* 26.06.2020 – Meeting #28 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BFAB2B733-18B9-4EDC-B7A4-9E4B6E305B8D%7D&file=FGAI4H%20TG%20Symptom%2026-06-2019%20meeting%2028%20minutes.docx&action=default)
* 10.07.2020 – Meeting #29 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BAEF6C2DF-2F39-4D04-911C-C8829AADC88E%7D&file=FGAI4H%20TG%20Symptom%2010-07-2019%20meeting%2029%20minutes.docx&action=default)
* 07.08.2020 – Meeting #30 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BD8E9FC2C-0C57-42A7-B206-2531808A130A%7D&file=FGAI4H%20TG%20Symptom%2007-08-2019%20meeting%2030%20minutes.docx&action=default)
* 21.08.2020 – Meeting #31 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC5F08531-5CC4-4DEA-BD39-44948F2A27D7%7D&file=FGAI4H%20TG%20Symptom%2021-08-2019%20meeting%2031%20minutes.docx&action=default)
* 04.09.2020 – Meeting #32 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B81433DBE-97AE-405F-8F8A-84E840E65489%7D&file=FGAI4H%20TG%20Symptom%2004-09-2019%20meeting%2032%20minutes.docx&action=default)
* 18.09.2020 – Meeting #33 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7D2AF17D-BC4B-41DB-809F-E3507CC5B893%7D&file=FGAI4H%20TG%20Symptom%2018-09-2019%20meeting%2033%20minutes.docx&action=default)

For coordinating the implementation work we also continued the weekly tech telco, however having meeting minutes for them proved impracticable. For bringing the clinical discussion on scores and metrics forward the doctors of Topic Group also had additional telcos not listed here.

All the meetings notes and also be found in the official TG-Symptom SharePoint folder: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

We also published a Topic Group internal summary of meeting I that can be found here:

* [TG-Symptom update on Meeting I](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC9A4D585-1897-4FCA-AC4D-F6E9250093D8%7D&file=FGAI4H%20TG%20Symptom%20update%20on%20Meeting%20I.docx&action=default)

Since Meeting I, we have been joined by:

* Barkibu (Ernesto Hernandez and Francisco Cheda)
* EQL (Yura Perov)
* Dr Reza Jarral (Independent contributor)

Currently, our Topic Group has the following 17 companies and 6 independent contributors:

* 1Doc3 (Lina Porras and Maria Gonzalez)
* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay, Ivan Lebovka, Nils Strelow)
* Alejandro Orsonio (Independent Contributor)
* Babylon Health (Saurabh Johri, Adam Baker)
* Baidu (Yanwu XU)
* Barkibu (Ernesto Hernandez)
* Buoy (Eddie Reyes)
* Deepcare.io (Hanh Nguyen)
* EQL (Yura Perov, who moved from Babylon to EQL)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Dr Reza Jarral (Independent contributor)
* Mfine (Dr Srinivas Gunda)
* Muhammad Murhaba (Independent Contributor)
* MyDoctor (Harsha Jayakody)
* Pritesh Mistry (Independent Contributor)
* Dr Salman Razzaki (Independent Contributor)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* XUND (Lukas Seper, Tamás Petrovics, Sophie Pingitzer)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale, Dr Audrey Menezes)

The Topic Group email reflector [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) altogether has currently 99 subscribers (16 more than for Meeting I). The latest meeting I version of this Topic Description Document lists 29 (1 more) contributors.

### Status update for meeting K (Online E Meeting)

With the finalization of the MMVB version implementing the Berlin model and a new benchmarking frontend and backend, the focus of work between Meeting J and Meeting K was one critical task of the Topic Group: agreeing on an ontology and approach for encoding realistic case data for the benchmarking.

As already for the London Model and Berlin Model iterations the work started with organizing the third Topic Group internal workshop from 12.11.2020 – 13.11.2020.

In preparation of the workshop all participants have been asked to prepare answers to the following questions:

1. **Procedure for agreeing on ontologies:** *How would you approach organizing the creation of a joined Snomed-based ontology for symptoms, factors, attributes subset, profile details, expected conditions + all the necessary relations for the benchmarking?*
2. **Available Resources:** *What are the resources you can contribute until the next meeting for technical implementation, working on the joint ontology, creating case data for the benchmarking or updating/migrating the TDD to the new template?*
3. **Next MMVB iteration AIs:** *Under which conditions could you imagine to use already real AIs in the next MMVB iteration? Or should we just stick to toy-AIs for the time being?*
4. **Next MMVB and MVB Disease sets:** *Which set of diseases should we use for the next MMVB version?*
5. **AI metadata:** *What are the relevant metadata-fields that would be needed to describe the context you designed your AI for?*
6. **Benchmarking result sharing:** *How would you like the results of a benchmarking to be shared with the general public, stakeholders, your partners, internally etc.?*
7. **TDD Update work:** *Which sections of the TDD could you imagine to migrate/write/update?*

During the workshop the questions have then been discussed in detail. The main part of the discussion focused on question 1 about the approach for agreeing on a joint ontology. The key points form this have been:

1. We need to try the process of aligning with a few symptoms to see how this works and how to then use this as a blue-print for a the general agreement process
2. We will use the same 11 abdominal related diseases we used in the Berlin Model, but extend the symptom space from the only 11 symptoms in the Berlin Model to all symptoms relevant to these disease
3. As a next step all companies create full detailed case for these disease and/or lookup the symptoms the consider relevant for any of these diseases.
4. Based on these cases the symptoms and their attributes would be grouped/unified to identify the relevant information that needs to be encoded.
5. Based on this symptom/attribute set we would then meet again and try map them to snomed concepts and agree on the level of pre/post-coordination i.e. if it is “pain” + location “right lower quadrant of abdomen” or “abdominal pain” + location “right lower quadrant of abdomen” etc.

Following the workshop the doctors in the Topic Group then created the corresponding case vignettes. The grouping of the symptoms and first steps towards mapping them to SNOMED have then be performed in corresponding follow-up meetings – a process that will continue after the submitting the first draft of the TDD migration work.

Beside the ontology there was also a discussion on point 3 where the consensus was that it should be open to everyone to use their real AIs and whether they do so via the public benchmarking API endpoints or only internally with a local test system. All other points had not been touched in greater detail and the TDD discussion was moved to a dedicated TDD related meeting.

Beside the work on the case description ontology, the Topic Group also contributed to the creation of the new TDD template FGAI4H-J-105 refined since Meeting J. In reflects many of the learnings from writing the earlier versions of this TDD and the way it was necessary to deviate from the original TDD template submitted by this Topic Group as FGAI4H-C-105 to Meeting C.

Based on TDD templated that was then accepted by the Focus Group via the online approval process out Topic Group also started the migration of this TDD document to the new format. Given the vacation, the late final approval and the size of the TG-Symptom TDD (97 pages) the version submitted for Meeting K is a work in progress version. In particular the sections 4 on ethics, and on the theoretical background and the detailed descriptions of the latest benchmarking iterations could not be completed yet. The Topic Group also reviewed the ethics document FGAI4H-K-028, even though this took longer as requested.

The Topic Group also had some contact with the Open-Source initiative, however due to capacity limitations on both sides the original plan to implement a symptom-assessment benchmarking similar to the Berlin Model MMVB version was not realized until Meeting K.

All the work in the Topic Group was organized online. The following list shows all the online meetings since the since meeting J:

* 16.10.2020 – Meeting #34 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BF9F4A45A-7C3B-43B9-BB1C-D2B3BDFE1CD4%7D&file=FGAI4H%20TG%20Symptom%2016-10-2020%20meeting%2034%20minutes.docx&action=default)
* 30.10.2020 – Meeting #35 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB0071FC2-50AC-477A-A879-3571DCD50321%7D&file=FGAI4H%20TG%20Symptom%2030-10-2020%20meeting%2035%20minutes.docx&action=default)
* 12.-13.11.2020 – Meeting #36 – Workshop #3 [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC91B13D2-B7B0-4327-B356-05006134CD83%7D&file=FGAI4H%20TG%20Symptom%2012-11-2020-13-11-2020%20meeting%2036%20minutes.docx&action=default)
* 25.11.2020 – Meeting #37 – Ontology Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BA1A9D5ED-24D7-4455-BFA9-C30937E1DBDF%7D&file=FGAI4H%20TG%20Symptom%2025-11-2020%20meeting%2037%20minutes.docx&action=default)
* 27.11.2020 – Meeting #38 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B3D2D601F-0CF3-4C07-BF47-D2B261E94864%7D&file=FGAI4H%20TG%20Symptom%2027-11-2020%20meeting%2038%20minutes.docx&action=default)
* 11.12.2020 – Meeting #39 – TDD Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B754F2CA4-DC1F-4CE6-BD96-7DABF02D991C%7D&file=FGAI4H%20TG%20Symptom%2011-12-2020%20meeting%2039%20minutes.docx&action=default)
* 14.12.2020 – Meeting #40 – Ontology Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC6C19CDA-34EE-403F-83F8-8BE4A4DEA8C4%7D&file=FGAI4H%20TG%20Symptom%2014-12-2020%20meeting%2040%20minutes.docx&action=default)
* 22.12.2020 – Meeting #41 – Ontology Telco (continued meeting #40 notes)
* 15.01.2020 – Meeting #42 – Telco [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0B660974-351A-4D32-9F0D-02595EE9D18F%7D&file=FGAI4H%20TG%20Symptom%2015-01-2021%20meeting%2042%20minutes.docx&action=default)

All the meetings notes and also be found in the official TG-Symptom SharePoint folder: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

Since Meeting J, we have been joined by:

* PnP (Opeoluwa Ashimi)

Currently, our Topic Group has the following 18 companies and 6 independent contributors:

* 1Doc3 (Lina Porras and Maria Gonzalez)
* Ada Health (Henry Hoffmann, Dr Shubs Upadhyay, Ivan Lebovka, Nils Strelow)
* Alejandro Orsonio (Independent Contributor)
* Babylon Health (Saurabh Johri, Adam Baker)
* Baidu (Yanwu XU)
* Barkibu (Ernesto Hernandez)
* Buoy (Eddie Reyes)
* Deepcare.io (Hanh Nguyen)
* EQL (Yura Perov, who moved from Babylon to EQL)
* Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
* Inspired Ideas (Megan Allen, Ally Salim Jnr)
* Isabel Healthcare (Jason Maude)
* Dr Reza Jarral (Independent contributor)
* Mfine (Dr Srinivas Gunda)
* Muhammad Murhaba (Independent Contributor)
* MyDoctor (Harsha Jayakody)
* PnP (Opeoluwa Ashimi)
* Pritesh Mistry (Independent Contributor)
* Dr Salman Razzaki (Independent Contributor)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* XUND (Lukas Seper, Tamás Petrovics, Sophie Pingitzer)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale, Dr Audrey Menezes)

The Topic Group email reflector [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@lists.itu.int) altogether has currently 104 (duplicates not counted) subscribers (5 more than for Meeting J). The latest meeting I version of this Topic Description Document lists 31 contributors (2 more).

## Topic Group participation

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

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

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

* <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/symptom.aspx>

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

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

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

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the ‘Call for Topic Group participation’ and this link:

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

In addition to the general FG-AI4H mailing list, the Topic Group on AI-based symptom assessment has an *individual mailing list:*

* [fgai4htgsymptom@lists.itu.int](mailto:fgai4htgsymptom@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-based symptom assessment 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 AISA Topic Group originally started without separate subTopic Groups. With Baidu joining in meeting D in Shanghai, the Topic Group was split into the subtopics "self-assessment" and "clinical symptom assessment". The first group addresses the symptom-checker apps used by non-doctors while the second group focuses on symptom-based diagnostic decision support systems for doctors. This document will discuss both sub-topics together.

## 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.

The exact definition of Artificial Intelligence (AI) is controversial. In the context of this document it refers to a field of computer science working on machine learning and knowledge based technology that allows the user to *understand* complex (health related) problems and situations at or above human (doctor) level performance and providing corresponding insights (differential diagnosis, triage) or solutions (next step advice).

The available systems can be divided into consumer facing tools sometimes referred to as "symptom checkers" and professional tools for doctors sometimes described as "diagnostic decision support systems". In general, these systems allow users to state an initial health problem, usually medically termed as the presenting complaint (PC) or chief complaint (CC). Following the collection of PCs, the collection of additional symptoms is performed either proactively - driven by the application using an interactive questioning approach, or passively, allowing the user to enter additional symptoms. Finally, the applications provide an assessment that contains different output components ranging from a general classification of severity (triage), possible differential diagnoses (DD), and advice on what to do next.

## Current gold standard

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

The gold standard for correct differential diagnosis, next step advice and adequate treatment is the evaluation of a medical doctor who is an expert in the respective medical field, which is based on many years of university education and structured training in hospitals and/or in community settings. Depending on context, steps such as triage preceding diagnosis are responsibilities of other health workers. Decision making is often supported by clinical guidelines and protocols or by consulting literature, the internet or other experts.

In recent years, individuals have increasingly begun to use the internet to find advice. Recent publications show that one in four Britons use the web to search their symptoms instead of seeing a doctor [Push Doctor, 2015]. Meanwhile, other studies show that internet self-searches are more likely to incorrectly suggest conditions that may cause inappropriate worry (e.g. cancers for innocuous symptoms).

## 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.

Whilst the shortage of health workers in low- and middle-income countries (LMICs) is worse, in more developed countries health systems face challenges such as increased demand due to increased life expectancy. Additionally, available doctors have to spend considerable amounts of time on patients that do not always need to see a doctor. Up to 90% of people who seek help from primary care have only minor ailments and injuries [Pillay 2010]. The vast majority (>75%) attend primary care because they lack an understanding of the risks they face or the knowledge to care for themselves. In the United Kingdom alone, there are 340 million consultations at the GP every year and the current system is being pushed to do more with fewer resources.

The challenge is to provide high-quality care and prompt, adequate treatment if necessary, develop mechanisms to avoid overdiagnosis and focus the health system resources for the patients in need.

Very few papers on AI-based symptom assessment exist, which are usually based on limited retrospective studies or use case vignettes instead of real cases. Therefore, there is a lack of scientific evidence available that assesses the impact of applying such technologies in a healthcare setting.

AI-powered symptom assessment applications have the potential to improve patient and health worker experience, deliver safer diagnoses, support health management, and save health systems time and money. This could be by empowering people to navigate to the right care, at the right time and in the right place or by enhancing the care that healthcare professionals provide.

Reliable benchmarking of AI solutions will give stakeholders the numbers and metrics required for decision making, building trust, and paving the way for wider adoption of AI-based symptom assessment. This wider adoption could potentially enable outcomes such as earlier diagnosis of conditions, more efficient care-navigation through the health systems and ultimately better health as it is currently pursued by UN's sustainable development goal (SDG) number 3 [UN SDG3]

## Existing AI solutions

This section presents the AI providers currently available and known to the Topic Group. The tables summarize the inputs and outputs relevant for benchmarking. It also presents relevant details concerning the scope of the systems that will affect the definition of categories for benchmarking reports, metrics and scores. Because the field is rapidly changing, this table will be updated before every Focus Group meeting and is currently a draft.

### Topic Group member Systems for AI-based Symptom Assessment

Table 2 provides an overview of the AI systems of the Topic Group members. The initial benchmarking will most likely start with the AI providers that joined the Topic Group and hence focus on the benchmarking relevant technical dimensions found in this group before increasing the complexity to cover all other aspects.

**Table 2 – Symptom assessment systems inside the Topic Group**

| Provider | System | Input | Output | Scope/Comments |
| --- | --- | --- | --- | --- |
| [1DOC3](https://www.1doc3.com/) | 1DOC3 platform | * Age, sex * Free text * Complementary information about signs, symptoms and medication related to the main topic. | * Pre-clinical triage * Possible Causes – differentials. | * Worldwide * Spanish * More than 750 conditions * Web and App for iOS and Android |
| [Ada Health GmbH](https://ada.com/) | Ada app | * Age, sex, risk factors * Free text PC search * Discrete answers to dialog questions for additional symptoms including attribute details like intensity | * Pre-clinical triage * Differentials for PC * Shortcuts in case of immediate danger | * Worldwide * English, German, Spanish, Portuguese, French * Top 1300 conditions * For smartphone users * Android/iOS |
| [Babylon Health](https://www.babylonhealth.com/) | Babylon App | * Age, sex, risk factors, country * Chatbot free text input and free text search (multiple inputs are allowed) * Answers to dialog questions for additional symptoms and risk factors including duration of symptoms, intensity | * Pre-clinical triage * Possible causes ("differentials") * Condition information * Recommendation of appropriate local services and products * Text information about treatments or next steps * Shortcuts in case of immediate danger | * Worldwide * English * 80% of medical conditions * For smartphone/web users * Android/iOS/Web |
| Baidu | Baidu’s Clinical Decision Support System | * Age\*, sex\*, birthplace, occupation, residence, height, weight * Free text of  PC\*, CC\*,  Past Medical History, Family History, Allergic History, Menstrual History\*, Marital and Reproductive History for female * Semi-structure text of medical exam report and test report * \* these details must be provided | * Pre-clinical triage * Diagnosis recommendation with explanation (structure or free text) * Next steps, such as medical exam, test * Treatment recommendation with explanation, such as drug, operations recommendation | * China * Chinese * General practice, 4000+ diagnoses * For Clinicians / Web users * CS SDK / BS SDK / API for HIT Companies integration * Web / mini program apps for Web users |
| Buoy Health |  |  |  |  |
| Deepcare | Deepcare Symptom Checker |  |  | * Users: Doctor and Patient * Platforms: iOS, Android * Language: Vietnamese |
| Infermedica | Infermedica API, Symptomate | * Age, sex * Risk factors * Free text input of multiple symptoms * Region/Travel history * Answers to discrete dialog questions * Lab test results | * Differentials for PC * Pre-clinical triage * Shortcuts in case of immediate danger * Explanation of differentials * Recommended further lab testing | * Worldwide * Top 1000 conditions * 15 language versions * Web, mobile, chatbot, voice |
| Inspired Ideas | Dr. Elsa | * Age, gender * Risk factors * Region/ time of year * Multiple symptoms * Travel history * Answers to discrete dialog questions * Lab test results * Clinicians hypothesis | * List of possible differentials * Condition explanations * Referral & lab test recommendations * Recommended next steps * Clinical triage | * Tanzania, East Africa * Languages: English and Swahili * Android/iOS/Web/API * Users: healthcare workers/ clinicians |
| [Isabel Healthcare](https://www.isabelhealthcare.com/) | Isabel Symptom Checker | * Age * Gender * Pregnancy Status * Region/Travel History * Free text input of multiple symptoms all at once | * List of possible diagnoses * Diagnoses can be sorted by 'common' or 'Red flag' * Each diagnosis linked to multiple reference resources * If triage function selected, patient answers 7 questions to obtain advice on appropriate venue of care | * 6,000 medical conditions covered * Unlimited number of symptoms * Responsive design means website adjusts to all devices * APIs available allowing integration into other systems * Currently English only but professional site available in Spanish and Chinese and model developed to make available in tmost languages |
| mfine |  |  |  |  |
| myDoctor |  |  |  |  |
| [Symptify](https://symptify.com/) | [Symptom Checker](https://symptify.com/) |  |  |  |
| Visiba Group AB | Visiba Care app | * Age * Gender * Chatbot free text input * Region/ time of year * Discrete answers * Lab results, inputs from devices enabled | * List of possible diagnoses * pre-clinical triage including format of meeting (digital or physical) * Next-step advice * condition information | * Language: Swedish * Android/iOS/Web * Users: Doctor and Patient |
| [XUND Solutions](http://xund.ai) | XUND App | * Age * Gender * Risk factors * Guided dialogue * Standardised answers | * Pre-clinical triage * In-depth explanations * Recommendations * Navigation within healthcare system | * Europe (CEE & CIS) * Primary healthcare (350 conditions); up to 500 planned * German, English, Hungarian * Patient-centered * Mobile & API |
| [Your.MD Ltd](https://www.your.md/) | Your.MD app | * Age, sex, medical risk factors, * Chatbot free text input * User consultation output (report) | * Differentials for PC * Pre-clinical triage * Shortcuts in case of immediate danger * Condition information * Recommendation of appropriate local services and products * Medical factors | * Worldwide * English, * Top 370 conditions (building to 500). * For smartphone users Android /iOS and web and messaging groups Skype etc |

### Other Systems for AI-based Symptom Assessment

Table 3 lists the providers of AI symptom assessment systems who have not joined the Topic Group yet. The list is most likely incomplete and suggestions for systems to add are appreciated. The list is limited to systems that actually have some kind of AI that could be benchmarked. Systems that e.g. show a static list of conditions for a given finding or pure tele-health services have not been included.

**Table 3 – Symptom assessment systems outside the Topic Group**

| Provider | System | Input | Output | Scope/Comments |
| --- | --- | --- | --- | --- |
| [Aetna](https://www.aetna.com/individuals-families.html) | [Symptom checker](https://www.healthwise.net/aetna/Content/CustDocument.aspx?XML=STUB.XML&XSL=CD.FRONTPAGE.XSL) |  |  |  |
| AHEAD Research | [Symcat](http://www.symcat.com/) |  |  |  |
| [Curai](https://www.curai.com/) | [Patient-facing DDSS](https://medium.com/curai/using-ai-ml-to-scale-the-worlds-best-healthcare-to-every-human-being-8cbc56df21d6) / [Chatbot](https://medium.com/@xamat/curai-6408bbc78b87) |  |  |  |
| [DocResponse](https://www.docresponse.com/) | [DocResponse](https://www.docresponse.com/) |  |  | * for doctors |
|  | Doctor Diagnose |  |  | * Android |
| [Drugs.com](https://www.drugs.com) | [Symptom Checker](https://www.drugs.com/symptom-checker/) |  |  | * Triage * Note: Harvard Health decision guide used |
| [EarlyDoc](https://www.earlydoc.com/) |  |  |  | * Web |
| [FamilyDoctor.org](https://familydoctor.org/) | [Symptom Checker](https://familydoctor.org/your-health-resources/health-tools/symptom-checker/) |  |  | * Web |
| [Healthline](https://www.healthline.com/) | [Symptom Checker](https://www.healthline.com/symptom-checker) |  |  |  |
| [Healthtap](https://www.healthtap.com/) | [Symptom Checker](https://www.healthtap.com/member/login) (for members) |  |  |  |
| [Isabel Healthcare](https://symptomchecker.isabelhealthcare.com/) | [Isabel Symptom Checker](https://symptomchecker.isabelhealthcare.com/suggest_diagnoses_advanced/landing_page) |  |  |  |
| [K Health](https://www.khealth.ai/) | K app chatbot |  |  |  |
| [Mayo Clinic](https://www.mayoclinic.org/) | [Symptom Checker](https://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075) |  |  |  |
| [MDLive](https://www.mdlive.com/) | Symptom checker on MDLive app |  |  |  |
| [MEDoctor](https://www.medoctor.io/) | [Symptom Checker](https://www.medoctor.com/Freemium/interview) |  |  |  |
| [Mediktor](https://www.mediktor.com/en) | [Web-based symptom checker,](https://www.mediktor.com/en) or Mediktor app |  |  |  |
| [NetDoktor](https://www.netdoktor.de/) | [Symptom Checker](https://www.netdoktor.de/symptom-checker/) |  |  |  |
| [PingAn](http://www.pingan.cn/en/index.shtml) | Good Doctor app |  |  |  |
| [Sharecare, Inc.](https://www.sharecare.com/) | [AskMD](https://www.sharecare.com/askmd/get-started) |  |  |  |
| [WebMD](https://www.webmd.com/) | Symptom checker | * Age, Gender, Zip code * Multiple presenting symptoms * Answers to discrete dialog questions | * List of possible differentials * Explanation of differentials * Possible treatment options |  |

### Input Data

AI systems in general are often described as functions mapping an input space to an output space. To define a widely accepted benchmarking it is important to collect the different input and output types relevant for symptom assessment systems.

#### Input Types

Table 4 gives an overview of the different input types used by the AI systems listed in Table 2.

**Table 4 – Overview symptom assessment system inputs**

| Input Type | Short Description | Number of Systems |
| --- | --- | --- |
| General Profile Information | General information about the user/patient like age, sex, ethnics and general risk factors. |  |
| Presenting Complaints | The health problems the users seeks advice for. Usually entered in search as free text. |  |
| Additional Symptoms | Additional symptoms answered by the use if asked. |  |
| Lab Results | Available results from lab tests that the user could enter if asked. |  |
| Imaging Data (MRI, etc.) | Available imaging data that the use could upload if available digitally. |  |
| Photos | Photos of e.g. skin lesions. |  |
| Sensor Data | Data from self tracking sensor devices like scales, fitness trackers, 1-channel ECG |  |
| Genomics | Genetic profiling information from sources like 23andMe. |  |
| ... |  |  |

#### Ontologies for encoding input data

For benchmarking the different input types need to be encoded in a way that allows each AI to "understand" its meaning. Since natural language is intrinsically ambiguous, this is achieved by using a terminology or ontology defining concepts like symptoms, findings and risk factors with a unique identifier, the most commonly used names in selected languages and often a set of relations describing e.g. the hierarchical dependencies of "pain at the left hand" and "pain in the left arm".

There is a large number of ontologies available (e.g. at <https://bioportal.bioontology.org/>). However most ontologies are specific for a small domain, not well maintained, or have grown to a size where they are not consistent enough for describing case data in a precise way. The most relevant input space ontologies for symptom assessment are described in the following sub sections

##### SNOMED Clinical Terms

SNOMED CT (<http://www.snomed.org/>) describes itself with the following five statements:

* Is the most comprehensive, multilingual clinical healthcare terminology in the world
* Is a resource with comprehensive, scientifically validated clinical content
* Enables consistent representation of clinical content in electronic health records
* Is mapped to other international standards
* Is in use in more than eighty countries

Maintenance and distribution is organized by the SNOMED International (trading name for the International Health Terminology Standards Development Organisation). SNOMED CT is seen to date as the most complete and detailed classification for all medical terms. SNOMED CT is only free of charge in member countries. In non-member countries the fees are prohibitive. While being among the largest and best maintained ontologies, it is partially not precise enough for encoding symptoms, findings and their details in a unified unambiguous way. Especially for phenotyping rare disease cases it does not yet have high enough resolution (e.g. Achromatopsia and Monochromatism are not separated, or "Increased VLDL cholesterol concentration" is not as explicit as e.g. "increased muscle tone"). SNOMED CT is also currently adapted to fit the needs of ICD-11 to link both classification systems (see below).

##### Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) (www.human-phenotype-ontology.org) is an ontology focused on phenotyping patients especially in context of hereditary diseases, containing more than 13,000 terms. In context of rare disease it is the most commonly used ontology and was adopted by OrphanNet for encoding the conditions in their rare disease database. Other examples are the 100K Genomes UK, NIH UDP, Genetic and Rare Diseases Information Center (GARD). The HPO is part of the Monarch Initiative, an NIH-supported international consortium dedicated to semantic integration of biomedical and model organism data with the ultimate goal of improving biomedical research (HPO) [[1]](#footnote-1).

##### Logical Observation Identifiers Names and Codes (LOINC)

LOINC is a standardized description of both, clinical and laboratory terms. It embodies a structure / ontology, linking related laboratory tests / clinical assessments with each other. It is maintained by the Regenstrief Institute. LOINC covers the domain of clinical observations, it can be used for symptoms, scales and specially results from clinical studies and procedures.

##### Unified Medical Language System (UMLS)

The UMLS, which is maintained by the US National Library of Medicine, brings together different classification systems / biomedical libraries including SNOMED CT, ICD, DSM and HPO and links these systems creating an ontology of medical terms. UMLS contains very useful lexical resources, useful to develop NLP tools. Very rarely used for clinical coding.

### Output Data

Beside the inputs, the outputs need to be specified in a precise and unambiguous way too. For every test case the output needs to be clear so that the scores and metrics can assess the distance between the expected results and the actual output of the different AI systems.

#### Output Types

As for the input types, Table 5 lists the different output types that the systems listed in 3.1.1 and 3.1.2 generate.

**Table 5 – Overview symptom assessment system outputs**

| Output Type | Short Description | Number of Systems |
| --- | --- | --- |
| Clinical Triage | Initial classification/prioritization of a patient on arrival in a hospital / emergency department. |  |
| Pre-Clinical Triage | A general advice of the severity of the problem and on how urgent actions need to be taken ranging from e.g. "self-care" over "see doctors within 2 days" to "call an ambulance right now" |  |
| Differential Diagnosis | A list of diseases that might cause the presenting complaints, usually ranked by some score like probability. |  |
| Next Step Advice | A more concrete advice suggesting doctors or institutions that can help with the specific problem. |  |
| Treatment Advice | Concrete suggestions of how to treat the problem e.g. with exercises, maneuvers, self medication etc. |  |
| ... |  |  |

The different output types will be explained in detail in the following section:

##### Clinical Triage

The most simple output of symptom based DDSS is a pre-clinical triage. Triage is a term commonly used in clinical context to describe the classification and prioritization of patients based on their symptoms. Most hospitals use some kind of triage systems in their emergency department for deciding how long a patient can wait so that people with severe injuries are treated with higher priority than stable patients with minor symptoms. One triage system commonly used is the Manchester Triage System (MTS) which defines the levels shown in Table 6.

**Table 6 – Manchester Triage System levels**

| Level | Status | Colour | Time to Assessment |
| --- | --- | --- | --- |
| 1 | Immediate | Red | 0 min |
| 2 | Very urgent | Orange | 10 min |
| 3 | Urgent | Yellow | 60 min |
| 4 | Standard | Green | 120 min |
| 5 | Non urgent | Blue | 240 min |

The triage is usually performed by a nurse for every incoming patient in a triage room equipped with devices of measuring the vital signs. While there are some guidelines clinics report a high variance in the classification between different nurses and on different days.

##### Pre-Clinical Triage

As triage helps with the prioritization of patients in an emergency setting, the pre-clinical triage helps users of self-assessment applications independent of a diagnosis to help decide when and where to seek care. In contrast to the clinical triage where there are several methods known, pre-clinical triage is not standardized. Different companies use different in-house classifications. Inside the Topic Group for instance the following classifications are used.

**1DOC3**

* No need for any other medical attention
* Should have a medical appointment in a few weeks or months
* Should have a medical appointment in a few days
* Should have a medical appointment in a few hours
* Should have a medical attention immediately

**Ada Health Pre-Clinical Triage Levels**

* Self-care
* Self-care Pharma
* Primary care 2-3 weeks
* Primary care 2-3 days
* Primary care same day
* Primary care 4 hours
* Emergency care
* Call ambulance

**Babylon Pre-Clinical Triage Levels**

Generally:

* Self-care
* Pharmacy
* Primary care, 1-2 weeks
* Primary care, same day urgently
* Emergency care (usually transport arranged by patient, including taxi)
* Emergency care with ambulance

With additional information provided per condition.

**Deepcare Triage Levels**

* Self-care
* Medical appointment (as soon as possible)
* Medical appointment same day urgently
* Instant medical appointment (Teleconsultation)
* Emergency care
* Call ambulance

**Infermedica Triage Levels**

* Self-care
* Medical appointment
* Medical appointment within 24 hours
* Emergency care / Hospital urgency
* Emergency care with ambulance

On top of that the system provides information on whether remote care is feasible (e.g. teleconsultation). Additional information provided per condition (e.g. doctor's specialty in case of medical appointments).

**Inspired Ideas Triage Levels**

* Self-care
* Admit patient / in-patient
* Refer patients to higher level care (District Hospital)
* Emergency Services

Triage is completed by a community health worker/ clinician, typically at a lower level health institution such as a village dispensary.

**Isabel Pre-Clinical Triage Levels**

* Level 1 (Green): Walk in Clinic/Telemedicine/Pharmacy
* Level 2 (Yellow): Family Physician/Urgent Care Clinic/Minor Injuries Unit
* Level 3 (Red): Emergency Services

Isabel does not advocate self care and assumes the patient has decided they want to seek care now but just need help on deciding on which venue of care.

**Symptify Pre-Clinical Triage Levels**

**Visiba Care Pre-Clinical Triage Levels**

* Self-care
* Medical appointment - digital - same day
* Medical appointment - digital - 1-2 weeks
* Medical appointment - physical primary care
* Emergency services

Depending on the condition additional adjustments possible.

**Your.MD Pre-Clinical Triage Levels**

* Self-care
* Primary care 2 weeks
* Primary care 2 days
* Primary care same day
* Emergency care

For a standardized benchmarking the Topic Group has to agree on a subset or superset for annotating test cases and for computing the benchmarking scores.

* existing pre-clinical triage scales
  + scales used by health systems e.g. NHS
* discussion tradeoff between number of different values and inter-annotator-agreement
* discussion tradeoff between number of different values and helpfulness for the user
* discuss challenge to define an objective ground truth for benchmarking
* available studies, e.g. on the spread among triage nurses

##### Differential Diagnosis

Using SNOMED CT for representing differential diagnosis provides a clinical level of detail, with very specific diagnosis concept and terms, and is automatically multi-lingual. Using a classification like ICD has the limitation of using broad categories, with a valuable epidemiologic meaning but too general for clinical use.

The hierarchies in SNOMED CT support for the selection of the appropriate level of detail for each differential diagnosis, i.e. ranging from “Autoimmune disease”, to “Rheumatoid arthritis” or “Rheumatoid arthritis of distal radioulnar joint”.

##### Next Step Advice

* to be written

##### Treatment Advice

* to be written

### Scope Dimensions

The table of existing solutions also lists the scope of the intended application of these systems. Analysing them suggests the following dimensions should be considered as part of the benchmarking:

Regional Scope

Some systems focus on a regional condition distribution and symptom interpretation, whereas others don't use the regional information. As this is an important distinction between the systems, the benchmark may need to present the results by region as well as the overall results. Since the granularity varies, starting at continent-level but also going down to the neighbourhood-level. The reporting most likely needs to support a hierarchical or multi-hierarchical structure.

Condition Set

With subtypes there are many thousands of known conditions. The systems differ in the range as well in depth of condition they support. Most systems focus on the top 300 to top 1500 conditions while others also include the 6000-8000 rare diseases. Other systems with a narrower intended focus e.g. tropical diseases or single disease only. The benchmarking therefore needs to be categorized by different condition sets to account for the different system capabilities.

Age Range

Most systems are created for the (younger) adult range and highly based on these conditions. Only few are explicitly created for pediatrics, especially very young children and some try to cover the whole lifespan of humans. The benchmarking therefore needs to be categorized into different age ranges.

Languages

Though there are some systems covering more than one language, common systems are created mostly in English. As it is essential for patient-facing applications to provide low-thresholds for everyone to access this medical information, this dimension may be taken into account as well - especially if at some point the quality of natural language understanding of entered symptoms is assessed.

### Additional Relevant Dimensions

Besides scope, technology and structure, the analysis of the different applications revealed several additional aspects that need to be considered to define the benchmarking:

Dealing with "No-Answers" / missing information

Some systems are not able to deal with missing information as they require always a "yes" or "no" answer when asking patients. This may be a challenge for testing with e.g. case vignettes as it won't be possible to describe the complete health state of an individual with every detail that is imaginable.

Dialog Engines

More modern systems are designed as chatbots engaging in a dialog with the user. The number of questions asked is crucial for the system performance and might be relevant for benchmarking. Furthermore dialog based systems proactively asking for symptoms are challenging if case vignettes are used for benchmarking since the dialog might not ask for the symptoms in the vignettes. Later iterations of the benchmarking might explicitly conduct a dialog to include the performance of the dialog, while first iterations might provide the AIs with complete cases.

Number of Presenting Complaints

The systems differ in the number of presenting complaints the user can enter. This might influence the cases used for benchmarking e.g. by starting with cases having only one presenting complaint.

Multimorbidity

Most systems don't support the possibility that a combination of multiple conditions is responsible for the users presenting complaints (multi-morbidity). The benchmarking therefore should mark multi-morbid and mono-morbid cases and differentiate the reported performance accordingly. The initial benchmarking might also be restricted to mono-morbid cases.

Symptom Search

Most systems allow to search for the initial presenting complaints. The performance of the search and if the application is able to provide the correct finding given the terms entered by users is also crucial for the system performance and could be benchmarked.

Natural Language Processing

Some of the systems support full natural language process for both the presenting complaints the dialog in general. While these systems are usually restricted to few languages, they provide a more natural experience and possible more complete collection of the relevant evidence. Testing the natural language understanding of symptoms might therefore be another dimension to consider in the benchmarking.

Seasonality

Some systems take into account seasonal dynamics in certain conditions. For example, during springtime there can be a spike in allergies and, hence, relevant conditions may be more probable than during other periods. Other examples include influenza spikes in winter or malaria in rainy seasons.

### Robustness of systems for AI based Symptom Assessment

As meeting D underlined with the introduction of a corresponding ad-hoc group, robustness is an important aspect for AI systems in general. Especially in recent years it could be shown that systems performing well on a reasonable benchmarking test set completely fail if adding some noise or a slight valid but unexpected transformation to the input data. For instance traffic signs might not be recognized any more if a slight modification like a sticker is added that a human driver would hardly notice. Based on the knowledge of such behaviours, the results of AI systems could be deliberately compromised e.g. to get more money from the health insurance for a more expensive disease, or faster appointments.

A viable benchmarking should therefore assess also the robustness. While for e.g. deep learning based image processing technologies robustness is a more important issue, also symptom based assessment can compromised. The reminder of this section gives an overview of the most relevant robustness and stability issues that should be assessed as part of the benchmarking.

Memory Stability & Reproducibility

An aspect of robustness is also the stability of the results. For instance a technology might use data structures like hash maps that depend on the current operating systems memory layout. In this case running the AI on the same case after restart again might lead to slightly different, possibly worse results.

Empty case response

AI should respond correctly to empty cases e.g. with an agreed-upon error message or some "uncertain" expressing that the given evidence is insufficient for a viable assessment.

Negative evidence only response

Systems should have no problems with cases containing only negative additional evidence besides the presenting complaints.

All symptoms response

Systems should respond correctly to requests giving evidence to all i.e. several thousand symptoms rather than e.g. crashing.

Duplicate symptom response

The systems should be able to deal with requests containing duplicates e.g. multiple times with the same symptom - possibly even with contradicting evidence. This might include cases where a presenting complaint is mentioned in the additional evidence again. A proper error message pointing on the invalid case would be considered as correctly dealing with duplicate symptoms.

Wrong symptom response

Systems should respond properly to unknown symptoms.

Symptom with wrong attributes response

Systems should respond properly to symptoms with wrong/incorrect attributes.

Symptom without mandatory attribute response

Systems should respond properly to symptoms with missing but mandatory attributes.

# 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-Symptom.

* 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 for assessing the quality of AI-based symptom-assessment systems. 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 Self-Assessment

*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 AI-based symptom-assessment 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”*.

* 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 [YOUR TOPIC] 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.

* 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).

* 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 symptom assessment 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).

## Benchmarking of the subtopic Self-Assessment

The benchmarking of AI-based symptom assessment 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.

The main goal of the benchmarking framework for AI-based symptom assessment is to enable the participation of as many existing systems as possible. In contrast to for instance most image processing AI systems where the input is already provided in a standardized format, for symptom assessment there is no such standard yet and needs to be developed as part of the benchmarking.

The first version where the group expects this preparation work to be finished and then real AIs to be benchmarked with real data to produce results that allow stakeholders to make decisions for the first time was named “minimal viable benchmarking” MVB.

The Topic Group agreed that in preparation of building this minimal viable benchmarking, we need to work on a benchmarking iteration where every detail is visible for analysis and optimization. Since this can be seen as "minimal" version of the MVB this version was given the name MMVB. All versions before the MVB are handled as MMVB x.y versions.

**Table 7 – Benchmarking iterations**

| Short Name | Name | Focus/Goals |
| --- | --- | --- |
| MMVB 1.0 | Minimal Minimal Viable Benchmarking | * show a complete benchmarking pipeline including case generation, AI, metrics, reports * with all parts visible to everyone so that we can all understand how to proceed with relevant details for MVB * learn about the needed data structures and scores * write/test some first case annotations guidelines * learn about the cooperation on both software and annotation guidelines * have a foundation for further discussions on if an own benchmarking software is needed or crowdAI could be used * Target: meeting F Zanzibar |
| MMVB 2.0 | Minimal Minimal Viable Benchmarking Version 2 | * extend the MMVB model to attributes * refine the MMVB factor model * switch to cloud-based toy AI hosting * test one-case-at-a-time testing |
| MMVB 2.1 | Minimal Minimal Viable Benchmarking Version 2.1 | * a new dedicated benchmarking frontend * a new backend infrastructure * a first simple case annotation tool |
| MMVB 2.2 | Minimal Minimal Viable Benchmarking Version 2.2 | * full implementation of the Berlin model in frontend, backend and annotation tool * improve AI error handling / health check * improved usability of the frontend |
| *MMVB 3.0* | *Minimal Minimal Viable Benchmarking Version 3.3* | * a first benchmarking using a snomed-based ontology for encoding case symptoms and their attributes * Target: Q4 of 2021 |
| *MVB* | *Minimal Viable Benchmarking* | * *first benchmarking with real AI and real data* * *Target: end of 2021* |
| *Vx.0* | *TG Symptom Benchmarking Vx.0* | * *the regular e.g. quarterly benchmarking for this Topic Group* * *continuous integration of new features* |

The latest version of the benchmarking is MMVB 2.2. It allows to benchmark toy-AIs using synthetic toy data sampled from an agreed upon simple model with 11 conditions and 12 symptoms including attribute details for the symptoms and a model where factors are handled as distributions modifying the prior probability of conditions.

The next benchmarking iteration MMVB 3.0 is expected to switch to a model using symptoms and attributes described by a SNOMED based ontology. Agreeing on this ontology is the most complex sub-task in the Topic Groups work and therefore we currently don’t expect a new benchmarking version until last quarter of 2021.

### Benchmarking version MMVB 1.0

This section includes all technological and operational details of the benchmarking process for the benchmarking version MMVB 1.0.

#### Overview

This section provides an overview of the key aspects of this benchmarking iteration, version MMVB 1.0. The main goal of it was to see a first working benchmarking pipeline for symptom assessment systems. The technical requirements have been discussed by the Topic Group in the first Topic Group workshop 11.-12.7.2019 in London. The MMVB 1.0 benchmarking software was then implemented based on the outcomes of this meeting in the following weeks.

Since a central part of a standardized benchmarking is agreeing on inputs and outputs of the AI systems, the work was started by defining a simple medical domain model containing hand selected conditions, symptoms, factors and profile information. Based on this domain model then the structure of inputs, outputs and the encoding of the expected outputs was defined. We refer to this model as the "London-model".

As shows, the model consists of 11 conditions from the field of abdominal pain together with 10 symptoms and one factor. The model states also the expected triage level which can be primary care (PC), self-care (SC) or emergency care (EC). The Topic Group also decided to use symptoms with all attributes “baked” into them (sometimes call pre-coordinated symptoms) and to leave the more complex explicit modelling of attributes to the next MMVB iterations.

Calendar

Description automatically generated

**Figure 1 - "London Model" used for sampling cases for MMVB 1.0**

#### Benchmarking methods

This section provides details about the methods of the benchmarking version MMVV 1.0. 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 this very first MMVB 1.0 version the benchmarking software was implemented as a python backend application proving all the benchmarking functionality via REST APIs to an HTML5+JS frontend for performing the benchmarking and displaying the results. The components are:

Case Generator

The case generator reads the London Model and provides a service for sampling test cases from it that can be used for the benchmarking. The generated case-sets are stored in the Case Storage.

Evaluator

The evaluator is the core of the benchmarking pipeline feeding all benchmarking cases of a case-set in the Case Storage to all the toy-AIs. This includes both several trivial toy-AIs directly implemented in the benchmarking backend, as well the actual toy-AIs hosted by the benchmarking participants in their own data centres. The remove toy-AIs expose all a REST API endpoint that is called by the evaluator. The results of each AI are persisted in the Results Storage.

Metrics Calculator

As the report displayed by the web interface is dynamically filtered and aggregated the Metrics Calculator is called directly by the frontend application to compute the scores for all benchmarking metrics.

Domain Model

The domain model i.e., the London Model, is the medical model describing the 11 diseases with their 11 symptoms the doctors of the Topic Group created for the purpose of benchmarking. It is manually exported from the google spreadsheet as CSV file that is then pre-processed into a JSON file which is then used by the Case Generator.

Case Storage / Result Storage

Both the generated cases as well as the results collected from the different AIs are persistent as JSON files in the filesystem. At this early stage it was decided that a proper database was not needed yet.

An architecture overview can be seen in ***Figure 2***. While every participant had their own instance of the benchmarking system running for implementing their toy-AI, there was also central system hosted by Babylon Health setup with all the API endpoints of the participants. Every participant hosted its toy-AI in their data centre using a technology of their choice.

Diagram

Description automatically generated

**Figure 2 - MMVB 1.0 High-level architecture**

##### Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture. In the MMVB 1.0 there are the following relevant data flows:

Model Generation

* The medical domain model is defined by the doctors direct in a google spreadsheet.
* From there it is exported as CSV file
* The CSV is then pre-processed and converted into a JSON file by python script.
* This JSON model is then used by the benchmarking as Domain Model.

Synthetic Case Generation

* The user triggers the creation of a new case-set in the web-interface.
* As result the case-set is stored to the Case Storage

Manual Case Generation

* The doctors created a set of 12 manual cased directly in the same spreadsheet as the London Model based on a template structure
* The cases have been exported as CSV file
* The CSV was then transformed into the same case format used by the Case Generator and store as case-set in the Case Storage where it can be use as any other case-set by the benchmarking

Benchmarking

* The Evaluator reads a selected case-set from the Case Storage and sends it to the AIs
* The AIs respond with their result which are then stored by the evaluator to in the Result Storage.
* The web-app then uses the Metrics Calculator to compute the metrics for based on the results stored in the Results Storage as well as the corresponding cases stored in the Case Storage
* The computed results are then finally display by the web-application

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

In contrast to the later MVB, all MMVB iterations of the benchmarking are designed to facilitate the development of the benchmarking for AI-based symptom assessment systems. They use only toy-data and toy-AIs, hence safe and secure system operation have not been explicitly considered. The benchmarking system was hosted by Babylon Health in their infrastructure applying their standards. All the toy-AIs that participated in the MMVB 1.0 benchmarking have been hosted by the individual companies following their own standards for safe and secure operation. The only security consideration applied was that only Babylon, hosting the benchmarking system, had access to it and all the data stored including the REST API endpoints of all toy-AIs.

The data used for the benchmarking was generated with a case synthesizer running on the benchmarking system. All data sets and all results have been stored into the file system and a simple database with no further protection against data-loss or manipulation.

The benchmarking system persisted all results from all AIs, including any timeouts and errors. All results have been displayed by the benchmarking frontend application that was freely accessible in the web – including any issues with the AIs so that the AI developers could use this for debugging their toy-AIs. The benchmarking system was not part of any automated monitoring and needed to be restarted on demand.

##### Benchmarking process

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

The focus of MMVB 1.0 was to develop and test a very first symptom-assessment benchmarking pipeline. The process for this covered the step 1) case set generation, 2) running the benchmarking for a selected dataset against all AI systems 3) computing & showing the results. For performing the different steps, the benchmarking system offered a simple web-based user interface focused on the task rather than on user experience considerations. The user interface was public with no password protection so that all developers and interested people from the Focus Group could explore it.

For creating the benchmarking case-set the UI provided the screen shown in ***Figure 2***. The only relevant parameter was here the number of cases to generate. It was also possible to select an existing case set by entering its identifier.

Graphical user interface

Description automatically generated with low confidence

**Figure 3 - MMVB 1.0 case generation UI**

The user was then able to trigger the execution of the benchmarking in the screen shown ***Figure 3*** . The first version did not support further selection of the AIs to run the benchmarking. Once the benchmarking was started a real-time log was displayed informing the user about the status.

Text, letter

Description automatically generated

**Figure 4 - MMVB 1.0 screen for running a benchmarking session**

Running the benchmarking did not include the computation of the scores for the metrics. This was then triggered by clicking the “Calculate & Evaluate” button in ***Figure 4***. As result the report table show in this figure was generated.

A picture containing table

Description automatically generated

**Figure 5 - MMVB 1.0 result screen**

For this first MMVB 1.0 version there was no scheduled benchmarking. Every developer could run a benchmarking at any time for building their own toy-AI which helped both the development of the pipeline and the toy-AIs.

For participating in the benchmarking with a toy-AI the process was to send a corresponding email with the API endpoint plus a name of the toy-AI to Yura Perov who was Babylon Health scientist responsible for the benchmarking system instance. For building the toy-AI all participants had access to the git repository of the benchmarking system which contained the code for some toy-AIs. Inside the Topic Group there was also an invitation mail shared with the request and response objects for implementing the API endpoint. The participants then improved and tested their AI using the benchmarking system. If AIs were broken the developers of the benchmarking system and the AI sorted the issues out by email or the issue tracking in git.

#### 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-based symptom assessment. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking.

The MMVB 1.0 uses as input for the AIs a simplistic user profile, explicit presenting/chief complaints (PC/CC), and additional features. The additional features might also contain risk factors. Table 5 shows the concrete fields with corresponding examples.

**Table 8 – MMVB input data format**

| Field name | Example | Description |
| --- | --- | --- |
| profileInformation | "profileInformation": {  "age": 38,  "biologicalSex": "male"  } | * General information about the patient * Age is unrestricted, however for the case creation it was agreed to focus on 18-99 years. * As sex we started with the biological sex "male" and "female" only |
| presentingComplaints | "presentingComplaints": [  {  "id": "c643bff833aaa9a47e3421a",  "name": "Vomiting",  "state": "present"  }  ] | * The complaints the user seeks and explanation/advice for * Always present * A list, but for the MMVB always with exactly one entry |
| otherFeatures | "otherFeatures": [  {  "id": "e5bcdaa4cf15318b6f021da",  "name": "Increased Urination Freq.",  "state": "absent"  },  {  "id": "c643bff833aaa9a47e3421a",  "name": "Vomiting",  "state": "unsure"  }  ], | * Additional symptoms and factors available * Might include "absent", "present" and "unsure" symptoms/factors * Might be empty |

As the London Model is not defining any identifiers the benchmarking system generated new ones using a hash function on the name. The different toy-AI systems used the same hash function to identify the given symptoms in the London Model again - a design detail to be addressed in the next MMVB versions.

#### 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.

For the MMVB 1.0 the AI systems are supposed to respond with a JSON object encoding the conditions that might have caused the symptoms in the given input object as well as their triage result. While every case has only on correct condition the AI systems are expected to generate a list of possible explanations that is sorted by descending likelihood. The group decided to not include an explicit score yet since the semantics of the scores of the group members is different and not comparable. The list of conditions might be empty, and if so, it means that with the given evidence no conclusive differential result was possible. For the benchmarking only the id of the condition was used. The name was added for improved readability by the developers.

In addition to the triage levels defined by the underlying London Model, for triage the AI might response with "UNCERTAIN" to declare that with the given evidence no conclusive triage result was possible.

shows an example of the data expected from an MMVB 1.0 toy-AI as response. Anything that could not be parsed into this structure was logged as an error.

**Table 9 – MMVB 1.0 API output encoding example**

| Field name | Example | Description |
| --- | --- | --- |
| conditions | "conditions": [  {  "id": "ed9e333b5cf04cb91068bbcde643",  "name": "GERD"  }  ] | * The conditions the AI considers best explaining the presenting complaints. * Ordered by relevance descending |
| triage | "triage": "EC" | * The triage level the AI considers adequate for the given evidence * Uses the same abbreviations defined by the London-model EC, PC, SC, UNCERTAIN |

#### Test data label/annotation structure

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.

For the MMVB 1.0 benchmarking iteration consists of the one and only condition a case was generated for and its corresponding triage level. With the difference that the “correct” condition is only one rather than a list, the structure of the expected output is similar to the structure of the AI output described in the previous section. Again, it contains beside the necessary condition identifier also the human readable form. Shows an example of how the expected output is encoded.

**Table 10 – MMVB 1.0 AI output label encoding**

| Field name | Example | Description |
| --- | --- | --- |
| condition | "condition": [  {  "id": "85473ef69bd60889a208bc1a6",  "name": "simple UTI"  }  ] | * The conditions expected/accepted as top result for explaining the presenting complaints based on the given evidence. * A list, but only one entry for mono-morbid cases as it is the case for MMVB |
| expectedTriageLevel | "expectedTriageLevel": "PC" | * The expected triage level |

For the MMVB 1.0 benchmarking system case data is organized in case sets. Each case sets is encoded as JSON file with an array for the cases. The cases then contain the actual case data shared with the AI and separate section with the label/annotations to predict. Shows an example of a complete case set structure combining profile information, presenting complaints, other features and the expected values to predict.

**Table 11 – An example of a MMVB 1.0 case-set with a single case.**

|  |
| --- |
| {  "cases": [  {  "caseData": {  "caseId": "case\_mmvb\_0\_0\_1\_a\_13588414",  "metaData": {  "description": "a synthetic case for the MMVB"  },  "otherFeatures": [  {  "id": "6e16a75aff90a62324940175453741f1",  "name": "Diarrhoea",  "state": "absent"  },  {  "id": "7a3094ceceac3afdae15243c11031588",  "name": "sharp lower quadrant pain",  "state": "present"  }  ],  "presentingComplaints": [  {  "id": "bcdc01d83bfb31c85ec47efc0642304e",  "name": "Weight Loss",  "state": "present"  }  ],  "profileInformation": {  "age": 64,  "biologicalSex": "female"  }  },  "valuesToPredict": {  "condition": {  "id": "42e009a4e3d8c8a17a29b4c57311e9cf",  "name": "IBD (first presentation non flare)"  },  "expectedTriageLevel": "PC"  }  }  ]  } |

#### Scores and metrics

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.

As the MMVB benchmarking iterations only use toy-AIs and toy-data the focus for the scores and metrics was to have metrics for implementing the benchmarking at all. For this purpose the Topic Group decided to use the classic top-n metrics top-1, top-3 and top-10 defining if the correct diseases is withing the first n suggested conditions. The score is used internally by most Topic Group members and can also be found the few existing papers on benchmarking systems for AI-based symptom assessment. For the triage the standard accuracy was used as well as the triage similarity score. The similarity score was defined as distance between the correct triage and the expected triage along the SC, PC, EC scale normalized by 2. For an UNCERTAIN triage level, the metric used 0.2 as soft triage level. The details for this soft triage match have not been discussed in the group as the goal was only to have some second more soft triage metric. In this first iteration of the MMVB no robustness metrics have been implemented. As a first non-medical performance metric the number of successfully processed cases as computed.

#### Test dataset acquisition

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

The primary data generation strategy for the MMVB 1.0 was to use the London-model and to sample cases from it. Sampling is done in several steps.

As first step the profile information is sampled. Here age is sampled form an equal distribution [18, 80] years. Sex is sampled between with 0.5 probability from (male, female). As next step the condition is sampled from prior probability distribution of the conditions for the sex sampled before. For this the number of “x” in the prior cells of the model () ranging from “x” to “xxx” is interpreted as prior probability between 0.3 and 0.9. For each case then the symptoms are sampled according to their condition probability following the same scale from 0.3 to 0.9. However, for each symptom there is only a 0.8 probability to include the symptom in the case and of these the symptoms are marked as “unsure” with probability 0.1.

Even if synthetic data will play an important role especially for benchmarking robustness, the Topic Group agrees that the MVB benchmarking always must contain real cases as well as designed case vignettes. This case data needs to be of exceptionally high quality since it is used to potentially influence business relevant stakeholder decisions. At the same time, it must be systematically ruled out that any Topic Group member can access the case data before the benchmarking, effectively ruling out that the Topic Group can check the quality of the benchmarking data. This is an important point to maintain trust and credibility.

For creating the benchmarking data therefore, a process is needed that blindly creates with reliably reproducible high-quality benchmarking data that all the Topic Group members can trust to be fair for testing their AI systems. With the growing number of Topic Group members form the industry it also becomes more and more clear that "submitting an AI" to a benchmarking platform e.g., as a docker container containing all the companies IP is not feasible, and hence the process does not only to guarantee high quality by also high efficiency and scalability.

One way to approach this is to define a methodology, processes and structures that allows clinicians all around the world in parallel to create the benchmarking cases.

As part of this methodology annotation guidelines are a key element. The aim is that these could be given to any clinician tasked with creating synthetic or labelling real world cases, and if the guidelines are correctly adhered to, will facilitate the creation of high quality, structured cases that are "ready to use" in the right format for benchmarking. The process would also include an n-fold peer reviewing processes.

There will be two broad sections of the guideline:

1. **Test Case Corpus Annotation Guideline** - this is the wider, large document that contains the information on context, case requirements, case mix, numbers, funding, process, review. It is addressed to institutions like hospitals that would participate in the creation of benchmarking data.
2. **Case Creation Guideline** - the specific guidelines for clinicians creating individual cases.

As part of MMVB 1.0 the Topic Group decided to start the work on some first annotation guidelines and test them with real doctors. Due to the specific nature of the London Model the MMVB 1.0 is based on, a first, very specific annotation guideline was drafted to explore this topic and learn from the process. The aim was to:

* create some clinically sound cases for MMVB 1.0 within a small "sandbox" of symptoms and conditions that were mapped by the clinicians in the group.
* explore what issues/challenges will need to be considered for a broader context

A more detailed description of the approach and methodology will be outlined in the [MMVB guideline](https://docs.google.com/document/d/1SLc8yrNr5s1RQQVy5rD4BdG4jvmFOg3rIwbwKminU68/edit) itself, but broadly followed the following process:

* Symptoms and conditions mapped by TG clinicians within sandbox of GI/Urology/Gynaecology conditions
* Alignment on case structure and metrics being measured.

The bulk of this activity was carried out in a face to face meeting in London, telcos and also through working on shared documents.

**Table 12 – Case example for the London Model**

|  |  |
| --- | --- |
| **Age**  18-99 | 25 |
| **Gender**  Biological, only male or female | male |
|  | |
| **Presenting Complaint** (from symptom template) | vomiting |
| **Other positive features** (from symptom template) | abdominal pain central crampy "present",  sharp lower quadrant pain 1 day "absent"  diarrhoea "present"  fever 'absent" |
| **Risk factors** | n/a |
|  | |
| **Expected Triage/Advice Level**  What is the most appropriate advice level based on this symptom constellation | self-care |
| **Expected Conditions** (from condition template) | viral gastroenteritis |
| **Other Relevant Differentials** (from condition template)  What other conditions is it relevant to have on a list based on the history. | irritable bowel syndrome |
| **Impossible Conditions** (from condition template) (are there any conditions, based on the above info, including demographics, where it is not possible\* for a condition to be displayed) – e.g. endometriosis in a male | ectopic pregnancy |
| **Correct conditions** (from condition template) | appendicitis |

The instructions (with an example) were shared with clinicians in the Topic Group companies and some cases were created for use by the MMVB 1.0. Feedback was collected on the quality of guidelines and process. As part of the work for meeting H, the MMVB was extended by supporting benchmarking based on the 12 cases manually created by our doctors.

Both the synthetic data and the cases created by the doctors served as expected their purpose for allowing to build and test a first version of a benchmarking so that we will continue this approach for the next MMVB iterations.

#### 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*).

For the MMVB 1.0 iteration only synthetic cases and 12 cases created by the doctors in the Topic Group have been used. The cases are highly specific of this minimalistic benchmarking iteration and are not based on real cases. Hence, the data is freely accessible under the following URL:

https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=1175944267

#### 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.

For the MMVB 1.0 assessing any baseline was out of scope.

#### Reporting methodology

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

As the MMVB 1.0 uses toy AIs and toy data, the only stakeholder interested in results was the Focus Group itself. The results have been documented in this TDD document and presented at the Focus Group meeting in Zanzibar, 2-5 September 2019.

In this early development phase, all AI developers had always full transparent access to all results of all AI systems by using the screen shown in ***Figure 4***.

The future reporting methodology was discussed in during the Topic Group workshop planning the MMVB 1.0 benchmarking iteration. From the discussion was clear that in contrast to other Topic Groups, that a single leaderboard is not sufficient for the benchmarking systems for AI-based symptom-assessment. There is the need for numerous dimensions to group and filter the results by in order to answer questions reflecting the full range of possible use cases (narrow and wide) e.g. the questions which systems are viable choices in Swahili speaking, offline scenarios with a strong focus on pregnant women vs. a general use symptom-assessment tool.

As first step in this direction for MMVB 1.0, a simple interactive table was implemented to show that it is possible to filter results. For the illustrative purposes of the MMVB 1.0, three simple groups are introduced that filter the results by the age of case patients.

From the workshop it became also clear that for this Topic Group it is unlikely that all results for all AI can always be publicly shared. The thinking was going the direction of opting-in/out for result publication in combination with means for allowing participants to share access to their own results with their own stakeholders.

#### Result

This section gives an overview of the results from runs of this benchmarking version of your topic.

As this benchmarking iterations was only an intermediate development step, no final result was recorded.

#### 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).

As intended, the MMVB 1.0 reached a point where first results from a new build benchmarking pipeline for AI-based symptom-assessment can be seen. The first minimalistic user interfaces allowed to create case sets and run benchmarking against toy-AIs partially hosted in the cloud by the different participants. Building this iteration also successfully established the cooperation on the technical implementation inside the Topic Group.

While the MMVB 1.0 provides a good starting point, we collected the following learnings for the next MMVB iterations until the work on the MVB can start:

Adding symptom attributes

Using symptoms with the attributes already embedded like in “sharp, lower right quadrat abdominal pain” are too simplistic and need to be replaced in the next iteration.

Adding more factors

The London Model contains the factor “females only” as comment and the factor “missed period” only as binary flag. For the next iteration modelling factors as probability distribution is needed.

Adding “dimensions” and using them in a more interactive reporting

For exploring the necessary drill-down reporting features needed to provide stakeholder with the answers from the benchmarking for their decision-making, we need to introduce the annotation for both data and AI with additional flexible metadata like their offline capabilities.

Implementation of some robustness scores

In the next iterations we need to see how to integrate the medical performance metrics with non-medical ones and the once for robustness as the technically work in a different way than only comparing AI results with expected results.

Better support for “unsure” / “unknown” AI answers

In the current iteration answering with a dangerously wrong or misleading answers is counted in the same way as stating that no reliable answer could be computed. As this is a feature some of the real AIs have, we need to reflect this in the MVB metrics.

Scores dealing with AI errors

While for the internal benchmarking of participants error play no important role as final numbers are only taken after all bugs have been removed, for the benchmarking by this Focus Group we have to expect some AIs to fail with an error on certain cases which needs to be reflected in some of the metrics.

Dynamic AI self-registration through the web-interface

The development of the MMVB 1.0 has shown that it would be more practicable if participants could register their different toy-AIs themselves without changing the codebase of the benchmarking system.

Running the benchmarking by case rather than by AI

In the current implementation performs the benchmarking AI by AI collecting all results from one AI before collecting it form the next AI. As part of increasing the resilience of the benchmarking against manipulation the next iteration should all AIs for the result of a case in parallel in combination with a short timeout so that side-channel communication between AIs would not provide any advantage.

Agreeing on how to encode test data is the core task of this Topic Group

The work on the first workshop also underlined that the most important and most complex unsolved task of the Topic Group on AI-based symptom assessment is agreeing on a joint input ontology for encoding factors, symptoms and their attributes in a way that can be interpreted by all AIs.

The need for a sub-topic taking care of NLP dialogs

The workshop for MMVB 1.0 has raised the point that we at some point will need a sub-group taking care of benchmarking the conversational NLP part of the self-assessment dialog some of the participants support with their systems.

A case set statistics analysis is needed

The work on the pipeline has shown that future version would need tools for checking the statistics of the benchmarking data to make sure that there are for instance no issues with the case synthesizer.

#### Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed.

As the MMVB 1.0 was an intermediate development step the benchmarking system and the corresponding toy-AI endpoints have been already retired. While the code of the benchmarking system is still in GitHub it was up to the participants how they handle the retirements of their endpoints and their source code. The synthetic test data was not archived. Both the London Model used for generating the synthetic test data as well as the 12 cases manually created by the doctors of the Topic Group are still available at:

<https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=575520860>

### 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 Clinical Symptom Assessment

In the current phase of specifying the benchmarking the difference between self-assessment and clinical symptom assessment are not relevant. Therefore, it was decided by the Topic Group to start the specification of the benchmarking for clinical symptom assessment only after at least the minimal viable benchmarking (MVB) version for self-assessment was completed or a new joining company has the capacity to drive this sub-topic.

# 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-Symptom.

## 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 symptom-assessment.

* 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

[HPO]https://hpo.jax.org/app/

[Pillay2010].Pillay N. The economic burden of minor ailments on the national health service in the UK. SelfCare 2010; 1:105-116

[Push Doctor 2015]. PushDoctor. UK Digital Health Report 2015. <https://www.pushdoctor.co.uk/digital-health-report>

[UN SDG] United Nations. Goal 3: Ensure healthy lives and promote well-being for all ages. <https://www.un.org/sustainabledevelopment/health/>

[WHO2013] *Global health workforce shortage to reach 12.9 million in coming decades.* Von WHO: <https://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en/> abgerufen

[WHO/WB2017] *Tracking Universal Health Coverage: 2017 Global Monitoring Report.* World Health Organization and International Bank for Reconstruction and Development / The World Bank. 2017. <http://pubdocs.worldbank.org/en/193371513169798347/2017-global-monitoring-report.pdf>.

Annex A:  
Glossary

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

|  |  |  |
| --- | --- | --- |
| Acronym/Term | Expansion | Comment |
| AI | [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) | While the exact definition is highly controversial, in context of this document it refers to a field of computer science working on machine learning and knowledge based technology that allows to *understand* complex (health related) problems and situations at or above human (doctor) level performance and providing corresponding insights (differential diagnosis) or solutions (next step advice, triage). |
| AI-MD | AI based medical device |  |
| AI4H | Artificial intelligence for health |  |
| AISA | AI-based symptom assessment | The abbreviation for the topic of this Topic Group. |
| API | [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) | the software interface systems communicate through. |
| AuI | Augmented Intelligence |  |
| CC | Chief Complaint | See "Presenting Complaint". |
| CfTGP | Call for Topic Group participation |  |
| CONSORT-AI | Consolidated Standards of Reporting Trials |  |
| DD | Differential Diagnosis |  |
| DEL | Deliverable |  |
| FDA | Food and Drug administration |  |
| FG | [Focus Group](https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx) | An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas. |
| FGAI4H | Focus Group on AI for Health |  |
| GDP | Gross domestic product |  |
| GDPR | General Data Protection Regulation |  |
| IIC | International Computing Centre | The United Nations data center that will host the benchmarking infrastructure. |
| IMDRF | International Medical Device Regulators Forum |  |
| IP | Intellectual property |  |
| ISO | International Standardization Organization |  |
| ITU | [International Telecommunication Union](https://www.itu.int) | The United Nations specialized agency for information and communication technologies – ICTs. |
| LMIC | Low-and middle-income countries |  |
| MDR | Medical Device Regulation |  |
| MMVB | Minimal minimal viable benchmarking | A simple benchmarking sandbox for understanding and testing the requirement for implementing the MVB. See chapter 5.2 for details. |
| MRCGP | [Membership of the Royal College of General Practitioners](https://en.wikipedia.org/wiki/Membership_of_the_Royal_College_of_General_Practitioners) | A postgraduate medical qualification in the United Kingdom run by the Royal College of General Practitioners. |
| MTS | Manchester Triage System | A commonly used systems for the initial assessment of patients e.g. in emergency departments. |
| MVB | minimal viable benchmarking |  |
| NGO | [Non Governmental Organization](https://en.wikipedia.org/wiki/Non-governmental_organization) | NGOs are usually non-profit and sometimes international organizations independent of governments and international governmental organizations that are active in humanitarian, educational, health care, public policy, social, human rights, environmental, and other areas to affect changes according to their objectives. (from Wikipedia.en) |
| PC | [Presenting Complaint](https://en.wikipedia.org/wiki/Presenting_problem) | The health problems the user of an symptom assessment systems seeks help for. |
| PC | Primary Care | A pre-clinical triage level suggested by many symptom-checkers. |
| PII | Personal identifiable information |  |
| PMCF | Post Market Clinical Follow Up | A requirement by regulators for Software as a medical device. This refers to clinical studies of the product in the real world that serve to show evidence of the claimed benefits of a medical device. |
| PROMs | Patient Reported Outcome Measures | This are outcomes reported by patients (usually through questionnaires) about their quality of life |
| SaMD | Software as a medical device |  |
| SDG | [Sustainable Development Goals](https://www.un.org/sustainabledevelopment/) | The United Nations Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. Currently there are 17 goals defined. SDG 3 is to "Ensure healthy lives and promote well-being for all at all ages" and is therefore the goal that will benefit from the AI4H Focus Groups work the most. |
| 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 TG-Symptom |
| TG | Topic Group |  |
| Triage |  | A [medical term](https://en.wikipedia.org/wiki/Triage) describing a heuristic scheme and process for classifying patients based on the severity of their symptoms. It is primarily used in emergency settings to prioritize patients and to determine the maximum acceptable waiting time until actions need to be taken. |
| WG | Working Group |  |
| WHO | World Health Organization |  |

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).

1DOC3

[1DOC3](https://www.1doc3.com/) is a digital health startup based in Colombia and Mexico, was founded in 2014 and provide the first layer of access to affordable healthcare for spanish speaking people on their phone. 1DOC3 has developed a Medical Knowledge graph in Spanish and a proprietary AI assisted technology to improve user experience by effectively symptom checking, triaging and pre diagnosing, **optimizing doctors’ time** allowing 1DOC3 to serve 350K consultations a month.

People actively involved: Lina Porras ([linaporras@1doc3.com](mailto:linaporras@1doc3.com)), Juan Beleño ([jbeleno@1doc3.com](mailto:jbeleno@1doc3.com)) and María Fernanda González ([mgonzalez@1doc3.com](mailto:mgonzalez@1doc3.com))

Ada Health GmbH

[Ada Health GmbH](https://ada.com/) is a digital health company based in Berlin, Germany, developing diagnostic decision support systems since 2011. In 2016 Ada launched the Ada-App, a DSAA for smartphone users, that since then has been used by more than 5 million users for about 10 million health assessments (beginning of 2019). The app is currently available in 6 languages and available worldwide. At the same time, Ada is also working on Ada-Dx, an application providing health professionals with diagnostic decision support, especially for complex cases. While Ada has many users in US, UK and Germany, it also launched a Global Health Initiative focusing on impact in LMIC where it partners with governments and NGOs to improve people's health.

People actively involved: Henry Hoffmann ([henry.hoffmann@ada.com](mailto:henry.hoffmann@ada.com)), Shubhanan Upadhyay ([shubs.upadhyay@ada.com](mailto:shubs.upadhyay@ada.com)),

Further contributions to this document: Andreas Kühn, Clemens Schöll, Johannes Schröder, Sarika Jain, Isabel Glusman, Ria Vaidya ([ria.vaidya@ada.com](mailto:ria.vaidya@ada.com)), Martina Fischer

Babylon Health

Babylon Health is a London-based digital health company which was founded in 2013. Leveraging the increasing penetration of mobile phones, Babylon has developed a comprehensive, high-quality, digital-first health service. Users are able to access Babylon health services via three main routes: i) Artificial Intelligence (AI) services, via our chatbot, ii) "Virtual" telemedicine services and iii) physical consultations with Babylon's doctors (only available in the UK as part of our partnership with the NHS). Babylon currently operates in the U.K., Rwanda and Canada, serving approximately 4 million registered users. Babylon's AI services will be expanding to Asia and opportunities in various LMICs are currently being explored to bring accessible healthcare to where it is needed the most.

People actively involved: Saurabh Johri ([saurabh.johri@babylonhealth.com](mailto:saurabh.johri@babylonhealth.com)), Nathalie Bradley-Schmieg ([nathalie.bradley1@babylonhealth.com](mailto:nathalie.bradley1@babylonhealth.com)), Adam Baker ([adam.baker@babylonhealth.com](mailto:adam.baker@babylonhealth.com))

Involved earlier from Babylon Health: Yura Perov (involved now too; with EQL)

Baidu

Baidu is an international company with leading AI technology and platforms. After years of commercial exploration, Baidu has formed a comprehensive AI ecosystem and is now at the forefront of the AI industry in terms of fundamental technological capability, speed of productization and commercialization, and “open” strategy. Baidu Intelligent Healthcare—an AI health-specialized division established in 2018—is seeking to harness Baidu's core technology assets to use evidence-based AI to empower primary health care. The division’s technology development strategy was developed in collaboration with the Chinese government and industry thought leaders. It's building capacity in China’s public health-care facilities at a grassroots level through the development of its Clinical Decision Support System (CDSS), an AI software tool for primary health-care providers built upon medical natural language understanding and knowledge graph technology. By providing explainable suggestions, CDSS guides physicians through the clinical decision-making process like diagnosis, treatment plans, and risk alert. In the future, Baidu will continue to enhance user experience and accelerate the development of AI applications through the strategy of “strengthening the mobile foundation and leading in AI”.

People actively involved: Yanwu XU (xuyanwu@baidu.com), Xingxing Cao (caoxingxing@baidu.com)

**Barkibu**

[Barkibu](https://www.barkibu.com/) is a pet health care and insurance company based in Coruña, Spain and founded in 2015.  
Through the Barkibu app, pet parents can get assistance on how to take care of their pets, check their symptoms and get immediate triage, talk to a live vet or find the best suited clinic for their pet’s problem. We do this through a combination of an AI powered vet assistant that runs our proprietary algorithms fed with real case data, a chat & video telehealth platform and a comprehensive insurance coverage policy.  
People actively involved: Francisco Cheda Pérez ([fran@barkibu.com](mailto:fran@barkibu.com)), Ernesto Hernández Cura ([ernesto@barkibu.com](mailto:ernesto@barkibu.com))

Deepcare

Deepcare is a Vietnam based medtech company. Founded in 2018 by three co-founders. Actually, we provide a Teleconsultation system for vietnamese market. AI-based symptom checker is our core product. It actually is available only in vietnamese language.

People actively involved: Hanh Nguyen ([hanhnv@deepcare.io](mailto:hanhnv@deepcare.io)), Hoan Dinh ([hoan.dinh@deepcare.io](mailto:hoan.dinh@deepcare.io)), Anh Phan ([anhpt@deepcare.io](mailto:anhpt@deepcare.io))

**EQL**

EQL is a digital health-tech organisation based in London, UK, which focuses on MSK conditions and physiotherapy. EQL’s product, Phio Access, provides a conversational AI-enabled digital solution to support triage for MSK conditions. Phio Access is currently available to 9.5 million people in the UK and in active use by several major healthcare providers, including Circle, BMI, Connect Health, Healthshare. EQL is currently working on its next-generation products, with the extended application of AI and ML technology for MSK medicine and physiotherapy.

People actively involved: Yura Perov ([yura@eql.ai](mailto:yura@eql.ai)).

Infermedica

[Infermedica](http://infermedica.com/), Inc. is a US and Polish based health IT company which was founded in 2012. The company provides customizable white-label tools for patient triage and preliminary medical diagnosis to B2B clients, mainly health insurance companies and health systems. Infermedica is available in 15 language versions and offered products include Symptom Checker, Call Center Triage and Infermedica API. To date the company's solutions provided over 3.5 million health assessments worldwide.

People actively involved: Dr. Irv Loh ([irv.loh@infermedica.com](mailto:irv.loh@infermedica.com)), Piotr Orzechowski ([piotr.orzechowski@infermedica.com](mailto:piotr.orzechowski@infermedica.com)), Jakub Winter ([jakub.winter@infermedica.com](mailto:jakub.winter@infermedica.com)), Michał Kurtys ([michal.kurtys@infermedica.com](mailto:michal.kurtys@infermedica.com))

Inspired Ideas

[Inspired Ideas](http://inspiredideas.io/) is a technology company in Tanzania that believes in using technology to solve the biggest challenges across the African continent. Their intelligent Health Assistant, [Dr. Elsa,](https://drelsa.xyz/) is powered by data and artificial intelligence and supports healthcare workers in rural areas through symptom assessment, diagnostic decision support, next step recommendations, and predicting disease outbreaks. The Health Assistant augments the capacity and expertise of healthcare providers, empowering them to make more accurate decisions about their patients' health, as well as analyzes existing health data to predict infectious disease outbreaks six months in advance. Inspired Ideas envisions building a complete end-to-end intelligent health system by putting digital tools in the hands of clinicians all over the African continent to connect providers, improve health outcomes, and support decision making within the health infrastructure that already exists.

People actively involved: Ally Salim Jr ([ally@inspiredideas.io](mailto:ally@inspiredideas.io)), Megan Allen (megan@inspiredideas.io)

Isabel Healthcare

[Isabel Healthcare](http://www.isabelhealthcare.com) is a social enterprise based in the UK. Founded in 2000 after the near fatal misdiagnosis of the co-founder's daughter, the company develops and markets machine learning based diagnosis decision support systems to clinicians, patients and medical students. The Isabel DDx Generator has been used by healthcare institutions since 2001.Its main user base is in the USA with over 160 leading institutions but also has institutional users around the world, including emerging economies such as Bangladesh, Guatemala and Somalia . The DDx Generator is also available in Spanish and Chinese. The Isabel Symptom Checker and Triage system has been available since 2012. This system is freely available to patients and currently receives traffic from 142 countries. The company makes its APIs available so EMR vendors, health information and telehealth companies can integrate Isabel into their own systems. The Isabel system has been robustly validated since 2002 with several articles in peer reviewed publications.

People actively involved: Jason Maude ([jason.maude@isabelhealthcare.com](mailto:jason.maude@isabelhealthcare.com))

Tom Neumark

I am a postdoctoral research fellow, trained in social anthropology, employed by the University of Oslo. My qualitative and ethnographic research concerns the role of digital technologies and data in improving healthcare outcomes in East Africa. This research is part of a European Research Council funded project, based at the University of Oslo, titled 'Universal Health Coverage and the Public Good in Africa'. It has ethical approval from the NSD (Norway) and NIMR (Tanzania); in accordance with this, the following applies: Personal information (names and identifiers) will be anonymized unless the participant explicitly wishes to be named. No unauthorized persons will have access to the research data. Measures will be taken to ensure confidentiality and anonymity. More information available on request.

Visiba Group AB

Visiba Care supplies and develops a software solution that enables healthcare providers to run own-brand digital practices. The company offers a scalable and flexible platform with facilities such as video meetings, secure messaging, drop-ins and booking appointments. Visiba Care enables larger healthcare organisations to implement digital healthcare on a large scale, and include multiple practices with unique patient offers in parallel. The solution can be integrated with existing tools and healthcare information systems. Facilities and flows can be added and customised as needed.

Visiba Care was founded in 2014 to make healthcare more accessible, efficient and equal. In a short time, Visiba Care has been established as a market-leading provider of technology and services in Sweden, enabling existing healthcare to digitalise their care flows. Through its innovative product offering and the value it creates for both healthcare providers and patients, Visiba Care has been a driving force in the digitalisation of existing healthcare. Through our platform, thousands of patients today can choose to meet their healthcare provider digitally. As of today, Visiba Care is active in 4 markets (Sweden, Finland, Norway and UK) with more than 70 customers and has helped facilitate more than 130.000 consultations. Most customers are present in Sweden today, and our largest client is the Västra Götaland region with 1.6 million patients.

We have been working specifically with AI-based symptom assessment and automated triage for 2 years now, and this becomes a natural step to expand our solution and improve patient onboarding within the digi-physical careflow.

People actively involved: Anastacia Simonchik ([anastacia.simonchik@visibacare.com](mailto:anastacia.simonchik@visibacare.com))

Your.MD Ltd

[Your.MD](https://www.your.md/) is a Norwegian company based in London. We have four years' experience in the field, a team of 50 people and currently delivers next steps health advice based on symptoms and personal factors to 650,000 people a month. Your.MD is currently working with Leeds University's eHealth Department and NHS England to scope a benchmarking approach that can be adopted by organisations like the National Institute of Clinical Excellence to assess AI self-assessment tools. We are keen to link all these initiatives together to create a globally recognised benchmarking standard.

People actively involved: Jonathon Carr-Brown ([jcb@your.md](mailto:jcb@your.md)), Matteo Berlucchi ([matteo@your.md](mailto:matteo@your.md)), Rex Cooper ([rex@your.md](mailto:rex@your.md)), Martin Cansdale ([martin@your.md](mailto:martin@your.md)), Audrey Menezes ([audrey@your.md](mailto:audrey@your.md))

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

1. [↑](#footnote-ref-1)