



WG(s): Plenary Helsinki, 20-22 September 2022

DOCUMENT

Source: TG-Symptom Topic Driver
Title: Att.1 – TDD update (TG-Symptom)
Purpose: Discussion

Contact: Henry Hoffmann Tel: +49 177 6612889
TG-Symptom Email: henry.hoffmann@ada.com
Ada Health GmbH
Germany

Contact: Martin Cansdale Email: martin@livehealthily.com
Healthily
UK

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 13.0 (submitted as FGAI4-P-021-A01 for Meeting P in Helsinki)

- Added 2.2.13 Status Update for Meeting P (Helsinki) Submission

Version 12.0 (submitted as [FGAI4-O-021-A01](#) for Meeting O in Berlin)

- Added 2.2.12 Status Update for Meeting O (Berlin) Submission
- Added Martin Cansdale as topic driver / contact

Version 11.0 (submitted as [FGAI4-N-021-A01](#) for the E-meeting N)

- Added 2.2.11 Status Update for Meeting N (Online E Meeting) Submission
- Added Flo and Kahun to 3.4.1 Topic group member Systems for AI-based Symptom Assessment
- Updated contributors list
- Added Flo and Kahun to Annex B

Version 10.0 (submitted as [FGAI4-M-021-A01](#) for the E-meeting M)

- Added 2.2.10 Status Update for Meeting M (Online E Meeting) Submission
- Added 5.1.1.2.9 "CONSORT-AI and SPIRIT-AI reporting guidelines"
- Updated 3.4.4.1.2 Your.MD triage levels
- Updated 3.4.1 Your.MD system table row
- Added 3.4.4.1.4 *Next Step Advice*
- Added 3.4.4.1.5 *Treatment Advice*
- Update Annex B

Version 9.0 (submitted as [FGAI4-L-021-A01](#) for the E-meeting L)

- Added 2.2.9 Status Update for Meeting L (Online E Meeting) Submission
- Migrated 6.1.2 Minimal Minimal Viable Benchmarking - MMVB Version 2.2
- Migrated Chapter 4 – Ethical considerations
- Migrated Chapter 5 – Existing work on benchmarking
- Update Annex B

Version 8.0 (submitted as [FGAI4-K-021-A01](#) for the E-meeting K)

- Migrated document structure to the new TDD template published as FGAI4H-J-105
- Added 2.2.8 Status Update for Meeting K (Online E Meeting) Submission

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](#) 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](#) 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](#) 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](#) 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](#) 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](#) 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](#). 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

Matteo Berlucchi
Your.MD
UK

Tel: +44 7867 788348
Email: matteo@your.md

Jason Maude
Isabel Healthcare
UK

Tel: +44 1428 644886
Email: jason.maude@isabelhealthcare.com

Shubhanan Upadhyay
Ada Health GmbH
Germany

Tel: +44 7737 826528
Email: shubs.upadhyay@ada.com

Yanwu XU
Artificial Intelligence Innovation Business,
Baidu
China

Tel: +86 13918541815
Fax: +86 10 59922186
Email: 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

Nathalie Bradley-Schmieg
Babylon Health
UK

Email: nathalie.bradley1@babylonhealth.com

Piotr Orzechowski
Infermedica
Poland

Tel: +48 693 861 163
Email: piotr.orzechowski@infermedica.com

Irv Loh, MD
Infermedica
USA

Tel: +1 (805) 559-6107
Email: irv.loh@infermedica.com

Jakub Winter
Infermedica
Poland

Tel: +48 509 546 836
Email: jakub.winter@infermedica.com

Ally Salim Jr
Inspired Ideas
Tanzania

Tel: +255 (0) 766439764
Email: ally@inspiredideas.io

Megan Allen Inspired Ideas Tanzania	Tel: +255 (0) 626608190 Email: megan@inspiredideas.io
Anastacia Simonchik Visiba Group AB Sweden	Tel: +46 735885399 Email: anastacia.simonchik@visibacare.com
Sarika Jain Ada Health GmbH Germany	Email: sarika.jain@ada.com
Yura Perov Independent contributor UK	Email: yura.perov@gmail.com
Tom Neumark University of Oslo Norway	Email: thomas.neumark@sum.uio.no
Rex Cooper Your.MD UK	
Martina Fischer Germany	
Lina Elizabeth Porras Santana 1DOC3 Colombia	Email: linaporras@1doc3.com
Juan Sebastián Beleño 1DOC3 Colombia	Email: jbeleno@1doc3.com
María Fernanda González Alvarez 1DOC3 Mexico	Email: mgonzalez@1doc3.com
Adam Baker Babylon Health UK	Email: adam.baker@babylonhealth.com
Xingxing Cao Baidu China	Email: caoxingxing@baidu.com
Clemens Schöll Ada Health GmbH Germany	
Audrey Menezes Your.MD UK	Email: audrey@your.md
Francisco Cheda Barkibu Spain	Email: fran@barkibu.com

Ernesto Hernández
Barkibu

Email: ernesto@barkibu.com

Saddif Ahmed
Flo Health
London

Email: s_ahmed@flo.health

Anna Klepchkova
Flo Health
London/Minsk

Email: a_klepchkova@flo.health

Michal Tzuchman Katz
Kahun
Israel, Tel Aviv

Email: michal@kahun.com

CONTENTS

1	Introduction.....	12
2	About the FG-AI4H topic group on AI-based symptom assessment	12
2.1	Documentation	13
2.2	Status of this topic group.....	14
2.2.1	Status update for meeting D (Shanghai)	14
2.2.2	Status update for meeting E (Geneva).....	14
2.2.3	Status update for meeting F (Zanzibar).....	14
2.2.4	Status update for meeting G (Delhi)	16
2.2.5	Status update for meeting H (Brasilia).....	17
2.2.6	Status update for meeting I (Online E Meeting)	19
2.2.7	Status update for meeting J (Online E Meeting).....	21
2.2.8	Status update for meeting K (Online E Meeting).....	23
2.2.9	Status update for meeting L (Online E Meeting)	26
2.2.10	Status update for meeting M (Online E Meeting).....	31
2.2.11	Status update for meeting N (Online E Meeting).....	35
2.2.12	Status update for meeting O (Berlin)	39
2.2.13	Status update for meeting P (Helsinki)	43
2.3	Topic group participation	45
3	Topic description	45
3.1	Definition of the AI task.....	46
3.2	Current gold standard	46
3.3	Relevance and impact of an AI solution.....	46
3.4	Existing AI solutions	47
3.4.1	Topic group member Systems for AI-based Symptom Assessment.....	47
3.4.2	Other Systems for AI-based Symptom Assessment.....	50
3.4.3	Input Data.....	51
3.4.4	Output Data	53
3.4.5	Scope Dimensions	57
3.4.6	Additional Relevant Dimensions	58
3.4.7	Robustness of systems for AI based Symptom Assessment	58
4	Ethical considerations	59
4.1	The ethical implications of applying the AI model in real world scenarios.....	60
4.2	The ethical implications of introducing benchmarking.....	61
4.3	The ethical implications of collecting the data for benchmarking	61
4.4	Risks facing individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground	61
4.5	How the privacy of personal health information protected	62

5	Existing work on benchmarking.....	63
5.1	Subtopic Self-Assessment	63
5.1.1	Publications on benchmarking systems.....	63
5.1.2	Benchmarking by AI developers.....	67
5.1.3	Relevant existing benchmarking frameworks	68
5.1.4	Scores & Metrics	70
5.1.5	Metrics for Symptom Assessment.....	76
5.1.6	Performance and Accuracy	77
5.1.7	Putting it all together for Clinicians	84
5.1.8	Additional clinical considerations and limitations	86
5.1.9	Conclusion.....	87
5.2	Subtopic [B].....	87
6	Benchmarking by the topic group.....	87
6.1	Benchmarking of the subtopic Self-Assessment	88
6.1.1	Benchmarking version MMVB 1.0.....	89
6.1.2	Benchmarking version MMVB 2.0 - 2.2.....	103
6.2	Subtopic Clinical Symptom Assessment.....	130
7	Overall discussion of the benchmarking.....	130
8	Regulatory considerations	131
8.1	Existing applicable regulatory frameworks.....	131
8.2	Regulatory features to be reported by benchmarking participants	132
8.3	Regulatory requirements for the benchmarking systems	132
8.4	Regulatory approach for the topic group.....	133
9	References.....	134

List of Tables

Table 1	– topic group output documents.....	13
Table 2	– Symptom assessment systems inside the topic group.....	47
Table 3	– Symptom assessment systems outside the topic group.....	50
Table 4	– Overview symptom assessment system inputs	51
Table 5	– Overview symptom assessment system outputs	53
Table 6	– Manchester Triage System levels	53
Table 7	- Ground truth approaches with their problems	72
Table 8	– Overview patient case metrics	79
Table 9	– Benchmarking iterations	88
Table 10	– MMVB 1.0 input data format	94

Table 11 – MMVB 1.0 API output encoding example	96
Table 12 – MMVB 1.0 AI output label encoding	96
Table 13 – An example of a MMVB 1.0 case-set with a single case.	97
Table 14 – Case example for the London Model.....	99
Table 15 – MMVB 2.2 input data format	120
Table 16 – MMVB 2.2 AI output structure	122
Table 17 – MMVB 2.2 case with labels included.....	123
Table 18 - MMVB 2.2 overall case-set structure.....	123

List of Figures

	Page
Figure 1 – Some of the case vignettes created by the doctors after workshop #3	26
Figure 2 – Attributes of symptom "abdominal pain" collected from the workshop #3 case vignettes	27
Figure 3 – Example of workshop #3 symptoms phrases mapped to SNOMED CT (ignoring attributes).....	28
Figure 4 – Experimental first simple SNOMED CT based case creation tool.....	29
Figure 5 – Case symptoms separated by category	32
Figure 6 – SNOMED search results for "headache" (left side) and the ancestors and children for the selected "Headache (finding)" concept.	33
Figure 7 – Editor that describes attributes severity, clinical course and finding site of an abdominal pain finding.	36
Figure 8 – SNOMED concept browser with the new feature showing how often concepts have been used in cases and if they are appropriate for use in case vignettes.	37
Figure 9 – Screenshot of the first benchmarking results in the audit benchmarking system.....	40
Figure 10 – Main roadmap items for the remaining time of the topic group.....	41
Figure 11 – Example “Model Facts” label for sepsis machine learning model from Sendak et al, 2020. (Nature)	85
Figure 12 – "London Model" used for sampling cases for MMVB 1.0.....	90
Figure 13 – MMVB 1.0 High-level architecture.....	91
Figure 14 – MMVB 1.0 case generation UI.....	93
Figure 15 – MMVB 1.0 screen for running a benchmarking session	93
Figure 16 – MMVB 1.0 result screen.....	94
Figure 17 – Abdominal Pain symptom with attributes inside the Berlin Model.....	104
Figure 18 – Factors with attribute details inside the Berlin Model.....	104
Figure 19 – Refined factor distributions for ectopic pregnancy inside the Berlin Model.....	104
Figure 20 – MMVB 2.2 High-level architecture.....	106
Figure 21 – MMVB 2.2 ai-implementations API	107

	Page
Figure 22 – MMVB 2.2 cases and case-sets API.....	108
Figure 23 – MMVB 2.2 benchmarking-sessions API.....	109
Figure 24 – MMVB 2.2 metrics API	109
Figure 25 – 2.2 Version of the Benchmarking start page	112
Figure 26 – The AI implementations list now featuring the ability of adding new AIs and editing existing ones.....	113
Figure 27 – MMVB 2.2 case sets overview page	114
Figure 28 – MMVB 2.2 case set creation page	115
Figure 29 – MMVB 2.2 benchmarking sessions overview page	116
Figure 30 – MMVB 2.2 benchmarking session creation page.....	117
Figure 31 – MMVB 2.2 benchmarking session runner.....	118
Figure 32 – MMVB 2.2 benchmarking result page	119
Figure 33 – MMVB 2.2 General input case structure	120
Figure 34 – MMVB 2.2 case-set raw viewer	125
Figure 35 – MMVB 2.2 case-set statistics view.....	127
Figure 36 – MMVB 2.2 example of a case defined by a doctors using the case annotation tool	128

FG-AI4H Topic Description Document

Topic Group Symptom Assessment

1 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 these 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 these 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.

2 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](#): Towards a potential AI4H use case "diagnostic self-assessment apps"

- [B-021](#): Proposal: Standardized benchmarking of diagnostic self-assessment apps
- [C-019](#): Status report on the "Evaluating the accuracy of 'symptom checker' applications" use case

and on a similar initiative by Your.MD:

- [C-025](#): 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.

2.1 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

2.2 Status of this topic group

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

2.2.1 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.

2.2.2 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 listed in the "change notes" at the beginning of the document.

2.2.3 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.

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](#)
- 20.06.2019 - Meeting #3 - Telco [Minutes](#)
- 11-12.7.2019 - Meeting #4 - London Workshop [Minutes](#)
- 15.8.2019 - Meeting #5 - Telco [Minutes](#)
- 23.08.2019 - Meeting #6 - Telco [Minutes](#)

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 Shubhanan 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 as the groups email reflector.

2.2.4 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](#)
- 27.09.2019 – Meeting #8 – Telco [Minutes](#)
- 10-11.10.2019 – Meeting #9 – Berlin Workshop [Minutes](#)
- 17.10.2019 – Meeting #10 – Telco [Minutes](#)
- 20.10.2019 – Meeting #11 – Telco [Minutes](#)
- 25.10.2019 – Meeting #12 – Telco [Minutes](#)
- 30.10.2019 – Meeting #13 – Telco [Minutes](#)

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 altogether has currently 44 subscribers. The latest Meeting G version of this Topic Description Document lists 20 contributors.

2.2.5 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](#)
- 06.01.2020 – Meeting #15 – Telco [Minutes](#)

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>. 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 Shubhanan Upadhyay, Dr Martina Fischer)
- Babylon Health (Saurabh Johri, Yura Perov, Nathalie Bradley-Schmiegl)

- 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 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.

2.2.6 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 reimplementations is almost finished.

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 CT for encoding the symptoms of the Berlin model. An overview of this work will be outline in section “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](#) 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](#)
- 12.03.2020 – Meeting #18 – Tech Telco [Minutes](#)
- 13.03.2020 – Meeting #19 – Telco [Minutes](#)
- 20.03.2020 – Meeting #20 – Tech Telco [Minutes](#)
- 27.03.2020 – Meeting #21 – Telco [Minutes](#)
- 15.04.2020 – Meeting #22 – Tech Telco [Minutes](#)
- 22.04.2020 – Meeting #23 – Tech Telco [Minutes](#)
- 21.04.2020 – Meeting #24 – Clinical Telco (no minutes)
- 24.04.2020 – Meeting #25 – Telco [Minutes](#)

All the meeting notes can 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 summary of meeting H can be found here:

- [TG-Symptom update on Meeting H](#)

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 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.

2.2.7 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 been 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 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.

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](#) 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](#)
- 11.06.2020 – Meeting #27 – Telco [Minutes](#)
- 26.06.2020 – Meeting #28 – Telco [Minutes](#)
- 10.07.2020 – Meeting #29 – Telco [Minutes](#)
- 07.08.2020 – Meeting #30 – Telco [Minutes](#)
- 21.08.2020 – Meeting #31 – Telco [Minutes](#)
- 04.09.2020 – Meeting #32 – Telco [Minutes](#)
- 18.09.2020 – Meeting #33 – Telco [Minutes](#)

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 meeting notes can 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](#)

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 Shubhanan 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 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.

2.2.8 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 from 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 blueprint for 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 diseases 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.

In reporting period 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](#)
- 30.10.2020 – Meeting #35 – Telco [Minutes](#)

- 12.-13.11.2020 – Meeting #36 – Workshop #3 [Minutes](#)
- 25.11.2020 – Meeting #37 – Ontology Telco [Minutes](#)
- 27.11.2020 – Meeting #38 – Telco [Minutes](#)
- 11.12.2020 – Meeting #39 – TDD Telco [Minutes](#)
- 14.12.2020 – Meeting #40 – Ontology Telco [Minutes](#)
- 22.12.2020 – Meeting #41 – Ontology Telco (continued meeting #40 notes)
- 15.01.2020 – Meeting #42 – Telco [Minutes](#)

All the meeting notes can 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 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).

2.2.9 Status update for meeting L (Online E Meeting)

Following Meeting K the topic group continued the work on defining a symptom ontology which had started back in workshop #3 from 12.11.2020 – 13.11.2020. After the doctors from the topic group created cases containing all symptoms (**Figure 1** shows some of these cases) and after the symptom's different presentations had been clustered (See **Figure 2** for an example), the group met again to review the results and discuss next steps.

ID	Company	Clinician	True Condition	Triage Level	Typicality	Biological sex	Age	Presenting Complaint(s)	Additional Evidence
1	Ada	Shubs	Apendicitis	EC	Typical	F		Sharp right lower quadrant pain for since yesterday (12 21 hours or so), getting progressively worse.	Vomited once this morning, no blood. Currently feels nauseated. No diarrhoea or constipation. Felt feverish since last night Temperature was 38.2C. No pv bleeding. Drinking fluids, urine output ok at present. Mouth feels dry. No dysuria. Feels generally unwell. Last menstrual period was 2 weeks ago. Denies recent unprotected sexual intercourse. Smoker. Nil other PMHx.
6	Infermedica	Michal	Irritable Bowel Syndrome	SC	Typical	F	25	abdominal pain,	abdominal pain lasting for 10 months, gets better on the weekends, increases in stressful situations, gets better after defecation rectal pressure flatulence constipations and diarrheas no vomiting no weight loss no anorexia no gastrointestinal bleeding no fever fatigue
12	Independent	Reza	Acute Pyelonephritis	EC	Atypical	F	33	Fever and vomiting	2 days of progressive loin pain Sweaty Temperature was 39.2C Nausea and vomiting Anorexia Negative pregnancy test Some dysuria in the past week Lethargic and bed bound No frequency
25	PnP	PnP clinician	Viral Gastroenteritis	SC	Typical	M		Watery stool (3 episodes) 2 Vomiting (5 episodes) Symptoms started 48 hrs. ago	Temperature >39 o C Symptoms persistent despite the use of antimalaria and antibiotics by the mother, mild abdominal discomfort Stool watery, non-bloody, non-mucoid Child taken care of by maid when mother goes to work, attends kindergarten where one other kid in his play group has similar symptom No abdominal distention, no dysuria, no history of food allergy or intolerance, no previous history of abdominal surgery
29	Independent	Eva	Acute Pyelonephritis	EC	Typical	F	25	Increasing flank pain and fever	Fever (39 °C) and chills for 1 day Increasing flank pain since morning Costovertebral angle tenderness Dysuria, frequent urination and urgency for 2 days Feels generally unwell and weak nausea, vomiting denies pregnancy
32	1Doc3	Maria/Lina	Acute cholecystitis		Typical	F	42	6 hour of moderate abdominal pain and right subcostal tenderness, asociated to vomiting.	Feels feverish Is overweight Has polycystic ovary syndrome Is taking oral contraceptives Smoking (+) Nauseas (+)

Figure 1 – Some of the case vignettes created by the doctors after workshop #3

```
abdominal pain
  finding site
    left
    right lower quadrant
    flank
    periumbilical
  quality|
    cramping
    sharp
  intensity
    moderate
    intensity (8/10),
  time since onset / duration
    Over the last week
    for since yesterday (12 hours or so)
    Over the last week
    8 hours of
  progression / clinical course
    getting progressively worse
    increasing
    progressive
```

Figure 2 – Attributes of symptom "abdominal pain" collected from the workshop #3 case vignettes

One of the outcomes of this meeting was that as the next step the doctors would explore expressing the cases using SNOMED CT concepts. The focus was here on the symptoms and the attributes have therefore been ignored. **Figure 3** shows an example of how this manual mapping looked. The work on this step showed that in general symptom mapping is feasible. It was also noted that SNOMED CT has a very strong bias towards professionally used clinical findings and not all lay use patient-reported details would be available at the level of detail supported by symptom assessment applications.

Case Snippets	Symptom	Snomed Name	Snomed Id	Mapping Quality
No diarrhoea	diarrhoea	Diarrhea (finding)	62315008	neg.
no diarrhoea				neg.
5 weeks of on-off diarrhoea with mucus				attributes missing
Intermittent diarrhoea for the last 2 months, no blood no mucus.				attributes missing
Chronic Diarrhoea "Poorly formed predominantly type 6 stool for 1 year Worse after heavy drinking Worse with work stress				attributes missing
diarrhea				perfect
Watery stool (3 episodes)				attributes missing
Stool watery, non-bloody, non-mucoid				attributes missing
diarrheas				attributes missing
no ... constipation	constipation	Constipation (finding)	14760008	neg.
constipations				attributes missing
Felt feverish since last night Temperature was 38.2C.	fever	Fever (finding)	386661006	attributes missing
no fever				neg.
NO fever				neg.
Fever				perfect
Temperature was 39.2C				attributes missing
Fever of 38C				attributes missing

Figure 3 – Example of workshop #3 symptoms phrases mapped to SNOMED CT (ignoring attributes)

In a subsequent meeting the topic group then discussed how to approach the mapping of attributes. During the discussion it became clear that the effort of defining which attributes are allowed for which symptoms and which attribute states are valid for an attribute in context of a given symptom will be a lot of work and that keeping this mapping up to date would cause a lot of maintenance work too. For this reason, we decided to first test if it would be feasible to not create such an attribute mapping and rely only on the SNOMED CT findings “as is”, possibly extended by only the most common attributes “severity” and parts of “clinical course”.

Based on this idea the topic group implemented the minimalistic SNOMED CT case creation tool depicted in **Figure 4**. Its main purpose was to allow the doctors in the group to see how well mapping the symptoms mentioned in the workshop #3 cases only using a search function. For this task the tool provided some metadata fields for author, expected disease, the factors age and sex, a field for the case vignette text and a comment field for providing feedback on how well it worked. The main feature was a search field that allowed searching findings in SNOMED CT and to add them as “presenting complaint”, “present” or “not present”. For this feature it was agreed to restrict the search to the finding subtree. In addition to this the doctors in the group reviewed the tree and excluded selected sub-trees and findings matching certain rules to narrow the tree down to the symptoms relevant to modelling self-assessment cases mainly consisting of patient reportable findings.

ITU/WHO - FG AI4H - TG-Symptom SNOMED CT case sketching tool V0.1

Case List

Expected Disease: test	Author: ivan		
Expected Disease: Acute cholecystitis	Author: Henry		
Expected Disease: Inflammatory Bowel disease (first pres - UC)	Author: Shubs		
Expected Disease: Appendicitis	Author: Milan		
Expected Disease: Simple UTI	Author: Henry		
Expected Disease: Appendicitis	Author: Henry		
Expected Disease: Irritable Bowel Syndrome	Author: Milan		

Edit Case

Author	Expected Disease	Age	Sex
Henry	Appendicitis	25	Female

Comment

- At least for the tool we would need negation group macros - ideally via snomed
- can' find Urinary symptoms (finding) SCTID: 249274008

Description

Case 30

8 hours of right lower quadrant abdominal pain, initially periumbilical with progressive intensity (8/10), associated with three vomiting episodes.

"Dehydrated
Pregnancy test (-)
Taking oral contraceptives
No urinary symptoms
Fever of 38.9°C"

Presenting Complaint

Abdominal pain (finding) x Vomiting symptom (finding) x

Present Symptoms

Fever (finding) x Pregnancy test negative (finding) x Oral contraception (finding) x

Absent Symptoms

Dysuria (finding) x

SUBMIT

Figure 4 – Experimental first simple SNOMED CT based case creation tool

The tool was then tested by the doctors in the group. While this testing is still ongoing some of the insights so far are:

- For a realistic assessment of the approach, we need to integrate a real search engine supporting synonyms, ids, fuzzy search, shorter-match-preference, start-of-word-preference, perfect-match-preference etc. Even if fuzzy search is not supported the search provided by Snowstorm servers will likely provide a good starting point for this.
- We need to add a visualization of the hierarchies around the search concept to enable the selection of the right post-coordination level – especially for adding attribute details.
- We need to replace the plain tag-lists for the symptoms with a UI that allows to specify a few hand-picked attributes that apply to almost all findings.
- This might include a mechanism to visualize/edit the findings with attribute post-coordination of the same base finding together.
- We need to encourage the user to add “not present” symptoms as high as possible in the hierarchy.

- We need to provide easy means to negate common finding groups like “Urinary symptoms” which can consist of multiple concepts.
- We need to provide a built-in way to mark suggested findings as unappropriated for creating symptom-assessment cases for further improving the case creation tool.

Beside the work on the ontology workstream the group also continued the migration and extension of contents from our original TDD version into this new J-105 format. The main points added in in the submission for meeting L are:

- The migration and extension of the MMVB 2.0-2.2 descriptions summarized as the new section 6.1.2
- The migration and extension of the chapter on ethical considerations as chapter 4
- The migration of the existing work on benchmarking as chapter 5

Since meeting K there has also been some more exchange with the open-source initiative of the focus group, but so far the test-wise integration of symptom-assessment as demo use-case has not been started.

All the work in the topic group was organized online. The following list shows all the online meetings since the since meeting K:

- 19.01.2020 – Meeting #43 – Telco [Minutes](#)
- 12.03.2020 – Meeting #44 – Telco [Minutes](#)
- 26.03.2020 – Meeting #45 – Telco [Minutes](#)
- 16.04.2020 – Meeting #46 – Telco [Minutes](#)
- 23.04.2020 – Meeting #47 – Telco [Minutes](#)
- 07.05.2020 – Meeting #48 – Telco [Minutes](#)

All the meeting notes can 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 K, we have been joined by:

- Nivi (David Tresner-Kirsch)

Yura Perov/EQL changed his role to “independent contributors”.

Currently, our topic group has the following 18 companies and 7 independent contributors:

- 1Doc3 (Lina Porras and Maria Gonzalez)
- Ada Health (Henry Hoffmann, Dr Shubhanan Upadhyay, Ivan Lebovka)
- Alejandro Orsonio (Independent Contributor)
- Babylon Health (Saurabh Johri, Adam Baker)

- Baidu (Yanwu XU)
- Barkibu (Ernesto Hernandez)
- 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)
- Dr Reza Jarral (Independent contributor)
- Mfine (Dr Srinivas Gunda)
- Muhammad Murhaba (Independent Contributor)
- MyDoctor (Harsha Jayakody)
- Nivi (David Tresner-Kirsch)
- 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)
- Yura Perov (Babylon and EQL before)

The topic group email reflector fgai4htgsymptom@lists.itu.int altogether has currently 109 (duplicates not counted) subscribers (5 more than for Meeting K). The latest meeting L version of this Topic Description Document lists 31 contributors.

2.2.10 Status update for meeting M (Online E Meeting)

Following meeting L, the work on the SNOMED annotation tool continued, starting with a meeting that structured the required next steps to address the learnings from meeting L. The focus was to improve the symptom search because it plays a central role in encoding test cases for benchmarking using SNOMED. To date, the search was implemented as a simple infix search over the normalized symptom name space. Testing this approach led to the conclusion that we need to improve the search by:

- supporting synonyms
- have a fuzzier search tolerant (also against typos)
- allowing search for identifiers
- shorter-match-preference
- start-of-word-preference
- perfect-match-preference

The group agreed that even if fuzzy search is not supported, the best course of action would be to update the annotation tool to use Snowstorm – the most frequently used server for hosting a SNOMED instance that provides a flexible API for querying all relevant concepts. (The swagger API specification of a Snowstorm instance can be found here: <https://snowstorm-training.snomedtools.org/snowstorm/snomed-ct/swagger-ui.htm>).

API access to the Snowstorm server has been implemented in the annotation tool backend. The corresponding search functionality was then provided via a new search API to the frontend application.

In the previous version of the case editor, the search was integrated as one tag lists for each group of symptoms: presenting complaints, present symptoms absent symptoms. While this was sufficient to test adding symptoms “as is” from SNOMED, it became clear that in addition we will need to explicitly support editing symptom details (such as attribute expression), meaning that the tag-list approach would be too simplistic. Together with the transition to the new search API, the UI was therefore separated into individual lists for each of the symptom categories and a dedicated section for the symptom search so that the search results could be added to any of the categories (see Figure 5)

Presenting Complaint	Present Symptoms	Absent Symptoms
Diarrhea symptom (finding) ⊖	Fever symptoms (finding) ⊖	Urinary symptom change (finding) ⊖
Mucus in stool (finding) ⊖	Tired (finding) ⊖	Complaining of a rash (finding) ⊖
Feces: fresh blood present (finding) ⊖	Weight loss (finding) ⊖	Dizziness (finding) ⊖
Left sided abdominal pain (finding) ⊖		No symptom relieving factor (finding) ⊖
Cramping pain (finding) ⊖		Foreign travel history finding (finding) ⊖
Moderate pain (finding) ⊖		

Figure 5 – Case symptoms separated by category

The design was aligned with the official SNOMED browser, separating the search result window and a window that shows the details for the selected search results (in particular the concept hierarchy around it as in many cases the desired concept is one of parents or children of the selected search result). Search, search results and the ancestors and children for the selected search result can be seen in Figure 6.

The screenshot displays the 'Snomed Concept Browser' interface. On the left, a search bar contains the text 'headache'. Below it is a table of search results with columns 'SnomedId' and 'Name'. The first row is selected and highlighted in blue: SnomedId: 25064002, Name: Headache (finding). Other rows include 'Headache site (finding)', 'Sick headache (disorder)', 'Migraine without aura (disorder)', 'Viral headache (finding)', 'Daily headache (disorder)', and 'Acute headache (finding)'. At the bottom of the table, it shows '1 row selected' and 'Rows per page: 100'. On the right side, there are three buttons: 'SET AS PRESENTING COMPLAINT', 'ADD AS PRESENT', and 'ADD AS ABSENT'. Below these buttons is the 'Symptom Details' section for SnomedId: 25064002, FSN: Headache (finding). The 'Ancestors' section shows a hierarchy of concepts: Finding of sensation by site, Head finding, Clinical finding, Finding of body region, Pain finding at anatomical site, Pain / sensation finding, SNOMED CT Concept, Finding of head and neck region, Finding by site, Sensory nervous system finding, Neurological finding, and Pain. The 'Children' section lists various subtypes: Headache caused by drug, Acute headache, Short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing syndrome, Frequent headache, Intermittent headache, Headache due to reversible cerebral vasoconstriction syndrome, Medication overuse headache, Orthostatic headache, Migraine variant with headache, Headache character - finding, Cervicogenic headache, Frontal headache, and Headache associated with substance abuse or withdrawal.

Figure 6 – SNOMED search results for "headache" (left side) and the ancestors and children for the selected "Headache (finding)" concept.

We moved the source code of the annotation tool to GitHub to work on the software in a cooperative way:

https://github.com/FG-AI4H-TG-Symptom/annotation_tool

The topic group also started to use the GitHub ticket system to organize the tasks:

https://github.com/FG-AI4H-TG-Symptom/annotation_tool/issues

Following meeting M, the next steps are to:

- implement an ECL based pre-filtering to filter better to concepts relevant for creating or annotating cases.
- add an attribute editor for the symptoms.
- add mechanisms marking and visualizing findings that have been used in other cases and/or should not be used.
- refine the hierarchy view.
- implement/integrate case and case set handling from previous versions.
- create realistic cases using the tool.

The next steps also include the support of cases created using the annotation tool in the MMVB benchmarking system implemented by the group. In addition to updating some of the toy-AIs, this includes the company internal implementation of a first mapping from the SNOMED symptom space to their native ontologies to confirm that the approach will generally work.

The end of this reporting period represented a milestone for the cooperation with the open-source initiative of the Focus Group. Even if TG-Symptom is not ML centric like most of the other topic groups, we have been chosen as one of the test use-cases. As part of this process, a group has been formed that will guide the implementation of the TG-symptom benchmarking in this framework. The current structure of this group can be found in this team allocation matrix:

https://docs.google.com/spreadsheets/d/17gDoEVA8qe_SBPYMIddl0dVzyTc29Y5O/edit#gid=1198500177

- The first TG meeting #55 was already joined by a regulatory representative of this group (Carolin Prabhu) and further alignment meetings will likely take place soon after submitting this TDD iteration.

All the work in the topic group was organized online. The following list shows all the online meetings since the since meeting L:

- 11.06.2021 – Meeting #49 – Telco [Minutes](#)
- 15.07.2021 – Meeting #50 – Telco [Minutes](#)
- 30.07.2021 – Meeting #51 – Telco [Minutes](#)
- 20.08.2021 – Meeting #52 – Telco [Minutes](#)
- 23.08.2021 – Meeting #53 – Telco [Minutes](#)
- 03.09.2021 – Meeting #54 – Telco [Minutes](#)
- 17.09.2021 – Meeting #55 – Telco [Minutes](#)

All the meeting notes can 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 L there have been no new joiners. Within the group, Alejandro Orsonio now works for SNOMED International which will help the topic group's work, but in the meantime will continue his role as an Independent Contributor. Buoy's representative changed from Eddie Reyes to Sarah Hassonjee.

Currently, our topic group has the following 18 companies and 7 independent contributors:

- 1Doc3 (Lina Porras and Maria Gonzalez)
- Ada Health (Henry Hoffmann, Dr Shubhanan Upadhyay, Ivan Lebovka, Milan Jovanovic)
- Alejandro Orsonio (Independent Contributor)
- Babylon Health (Saurabh Johri, Adam Baker)
- Baidu (Yanwu XU)
- Barkibu (Ernesto Hernandez)
- Buoy (Sarah Hassonjee)
- 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)

- Dr Reza Jarral (Independent contributor)
- Mfine (Dr Srinivas Gunda)
- Muhammad Murhaba (Independent Contributor)
- MyDoctor (Harsha Jayakody)
- Nivi (David Tresner-Kirsch)
- 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, Martin Cansdale, Dr Audrey Menezes)
- Yura Perov (Babylon)

The topic group email reflector fgai4htgsymptom@lists.itu.int altogether has currently 109 (duplicates not counted). The latest meeting M version of this Topic Description Document lists 31 contributors.

2.2.11 Status update for meeting N (Online E Meeting)

The work following meeting M was split into five workstreams. The first workstream focused on continuing the work on the annotation tool with the goal of finding a way to create annotated cases for benchmarking AI-based symptom assessment systems. Following meeting M, the tool supported symptom and finding search using the SNOMED ontology. Building on top of this, additional work was done to support the annotation of symptom attributes (e.g. intensity or finding site). While SNOMED's support for findings is adequate, the precision and quality of attributes is not suboptimal to use for case encoding. Based on the approach [outlined](#) by our topic group's SNOMED expert, we implemented the following changes:

- Refactoring the case editor's tag-like symptom lists into tables to show more details for each symptom.
- Implementing a dedicated modal symptom editor to annotate attributes supported the symptoms/findings according to the SNOMED ontology.
- Implementing a sub-component to select the attributes states still available for post-coordinating the symptoms/findings (e.g. a sub-structure of the abdomen if "abdominal pain" was selected).
- An additional comment field allowing our doctors to provide feedback on how well the attribute encoding worked.
- Implementing the necessary backend SnowStorm server API calls to query the pre-coordinated, post-coordinated attribute states and the attributes supported by symptoms/findings.

Figure 7 shows the modal finding editor with the new attribute selection feature.

Create New Case

Author: Dr. Smith | Expected Disease: Appendicitis | Age: 43 | Sex: Male

Edit Attributes

SnomedCT Id: 21522001 | Name: Abdominal pain (finding)

Snomed Id	Name	Postcoordinated State
246112005	severity	Select state: Severities (qualifier value)
263502005	clinical course	Select state: Acute onset (qualifier value)
363698007	finding site	Select state: [Open dropdown menu]

Comment:

Presenting Complaint

SnomedId: 21522001

Present Symptoms

SnomedId:

Postcoordinated State Dropdown:

- Structure of blood vessel in pericolic tissue (body structure)
- Entire colonic submucosa and colonic muscularis propria (body structure)
- Structure of right adrenal cortex (body structure)
- Structure of left adrenal cortex (body structure)
- Structure of right calyx (body structure)
- Structure of left calyx (body structure)
- Part of left kidney (body structure)
- Part of right kidney (body structure)

Figure 7 – Editor that describes attributes severity, clinical course and finding site of an abdominal pain finding.

We also implemented a feature that makes it easier for our doctors to identify the symptoms and findings that they most likely want to use in cases by highlighting the entities that have been used in other case vignettes. The feature also allows the explicitly marking of findings as inappropriate if they should be explicitly avoided (see **Figure 8**).

The screenshot displays the SNOMED Concept Browser interface. At the top, it shows the 'Finding Name' with three status indicators: 'used in presenting complaint' (blue), 'used in present' (purple), and 'used in absent' (orange). Below this is a search bar with the text 'abdominal'. A table lists search results with columns for 'SnomedId', 'Name', and 'Inapt'. The row for 'Abdominal heart (disorder)' (SnomedId: 14886009) is highlighted, and its 'Inapt' status is marked with a red exclamation point. To the right of the table, there are three buttons: 'SET AS PRESENTING COMPLAINT', 'ADD AS PRESENT', and 'ADD AS ABSENT'. Below these buttons, the 'Symptom Details' section shows the SnomedId (14886009) and FSN (Abdominal heart (disorder)). The 'Ancestors' section lists various related concepts in blue rounded rectangles, including 'Congenital cardiovascular disorder', 'Finding of upper trunk', 'Disorder of thoracic segment of trunk', 'Viscus structure finding', 'Clinical finding', 'Congenital anomaly of upper trunk', 'Congenital anomaly of thorax', 'Congenital anomaly of cardiovascular structure of trunk', 'Disorder of body system', 'Finding of trunk structure', 'Mediastinal finding', 'Cardiac finding', 'Finding of region of thorax', 'Congenital malformation', 'SNOMED CT Concept', 'Structural disorder of heart', 'Disorder of trunk', 'Disorder of thorax', 'Cardiovascular finding', and 'Congenital anomaly of trunk'.

Figure 8 – SNOMED concept browser with the new feature showing how often concepts have been used in cases and if they are appropriate for use in case vignettes.

The technical work on the annotation tool was complemented by extending the clinical vignettes database that will be used for testing the annotation tool and performing a first benchmarking with the real AIs. The medical doctors of the topic group used two approaches to further enrich the database. First they reached out to the medical community to ask for support to create new clinical vignettes. The second approach consisted of internet search for high quality clinical vignettes that are freely available.

In the third workstream we explored the feasibility of integrating the TG-Symptom annotation tool with the recent developments on the “annotation package” developed by the focus group’s open-code initiative. The topic group reached out to the group and discussed how the annotation tool could be registered and called as an external case editor. This included investigating how annotated TG-Symptom cases could be stored in the annotation package system. In preparation of calling our editor “stand-alone”, the annotation tool was refactored to separate the case editor components from the case list and to accept case data via URL encoded parameters.

Shortly before meeting M, the TG-Symptom agreed to engage in the audit trial initiative as another showcase. While it was clearly stated that we will not be able to follow the proposed timeline and are unlikely to have the evaluation ready for publication before meeting N, we worked together with the corresponding TG-Symptom audit group on both an audit benchmarking script and on the audit process itself. As part of the work to implement an evaluation script we created an initial version published in the GitHub-repo that the audit group created for us:

- <https://github.com/aiaudit-org/trial-audits-team-a-tg-symptoms>

Any real audit trial would require the work on SNOMED-based encoding of cases to be completed. Here we focused on applying the audit framework on the MMVB 2.2 system using synthetic cases sampled from the Berlin model with 11 abdominal conditions and corresponding toy-AIs. Since these toy-AIs are hosted in the cloud to use them directly, the docker-based benchmarking approach + internet access would be needed, which was not available when the development was started.

Therefore, as an intermediate step, we implemented benchmarking based on the submission of solution files. To facilitate this approach, we created a helper script taking a dataset exported from the [MMVB 2.2. system's case-set page](#) and converting into audit annotation files and AI input-files. We also implemented a script iterating all cloud hosted AI-systems and recording their response to the AI input cases in audit submission files that could then be used for manual solution upload and local offline testing. To evaluate the solutions, we extended the evaluation script to use the same TG-Symptom specific metrics also supported by the MMVB 2.2 system. The work reached the point where the scripts successfully run locally. We expect to present the benchmarking results in the audit system during meeting N. The general background and usage are described in detail in the corresponding [readme file](#).

Independent of the technical audit work, we also worked with the TG-Symptom audit group on the formal side of the audit benchmark. We focused on the audit-checklist. Given that the TG-Symptom AIs largely apply non-ML AI-techniques, adapting the checklist to TG-Symptom needs was more challenging than expected. The current master is expected to be merged the week before meeting N. It is a combination of a more specific version of the original default checklist and technical points chosen from a list of TG-Symptom specific aspects relevant to understand the performance of systems for AI-based symptom assessment. The latest version of the merged checklist can be found [here](#). The list of TG-Symptom specific technical details can be found [here](#).

All the work in the topic group was organized online. The following list shows all the online meetings since the since meeting M:

- 12.10.2021 – Meeting #56 – Telco [Minutes](#)
- 15.10.2021 – Meeting #57 – Telco [Minutes](#)
- 29.10.2021 – Meeting #58 – Telco [Minutes](#)
- 03.11.2021 – Meeting #59 – Telco [Minutes](#)
- 12.11.2021 – Meeting #60 – Telco [Minutes](#)
- 26.11.2021 – Meeting #61 – Telco [Minutes](#)
- 02.12.2021 – Meeting #62 – Telco [Minutes](#)
- 10.12.2021 – Meeting #63 – Telco [Minutes](#)
- 21.01.2022 – Meeting #64 – Telco [Minutes](#)
- 04.02.2022 – Meeting #65 – Telco [Minutes](#)

All the meeting notes can also be found in the official TG-Symptom SharePoint folder:

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

In addition to the regular bi-weekly TG-Symptom meetings there have been weekly informal developer stand-ups to sync on technical details as well as informal management syncs. TG-Symptom also participated in several meetings of the TG-Symptom audit group.

Since meeting M, the topic group has been joined by:

- Flo (Anna Klepchukova, Saddif Ahmed)
- Kahun (Michal Tzuchman Katz)

Currently, our topic group has the following 20 (+2) companies and 7 independent contributors:

- 1Doc3 (Lina Porras and Maria Gonzalez)
- Ada Health (Henry Hoffmann, Dr Shubhanan Upadhyay, Ivan Lebovka, Milan Jovanovic)
- Alejandro Orsonio (Independent Contributor)
- Babylon Health (Saurabh Johri, Adam Baker)
- Baidu (Yanwu XU)
- Barkibu (Ernesto Hernandez)
- Buoy (Sarah Hassonjee)
- Deepcare.io (Hanh Nguyen)
- Flo (Anna Klepchkova, Saddif Ahmed)
- Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
- Inspired Ideas (Megan Allen, Ally Salim Jnr)
- Isabel Healthcare (Jason Maude)
- Kahun (Michal Tzuchman Katz)
- Dr Reza Jarral (Independent contributor)
- Mfine (Dr Srinivas Gunda)
- Muhammad Murhaba (Independent Contributor)
- MyDoctor (Harsha Jayakody)
- Nivi (David Tresner-Kirsch)
- 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, Martin Cansdale, Dr Audrey Menezes)
- Yura Perov (Babylon)

The topic group email reflector fgai4htgsymptom@lists.itu.int altogether has currently 117 (+8) (duplicates not counted) subscribers. The latest meeting N version of this Topic Description Document lists 34 (+3) contributors.

2.2.12 Status update for meeting O (Berlin)

In the reporting period between meeting N and meeting O we continued the work on the five active workstreams introduced in the previous report. After adding the new features for SNOMED-based symptom attribute specification, the work in the first workstream focused on testing these new features. For this the topic group doctors encoded about 30 cases and collected all observed bugs and usability issues, which then have been translated into corresponding tickets and discussed with the engineer responsible for the annotation tool. So far 50% of the issued have been resolved. The work is expected to be finished shortly after meeting O.

In preparation of the upcoming test-encoding of benchmarking cases, the topic group doctors also started to reach out to other doctors interested in contributing. This included both English native speakers as well as non-English native speakers to assess the robustness of the case annotation and process in this respect. In preparation of this work, we also started the revision of the necessary annotation guidelines.

In workstream three we continued the work on integrating TG-Symptom annotation tool with the annotation package developed by the focus group's open-code initiative. The topic group's project plan now includes migrating the annotation tool unchanged to the infrastructure of the open code initiative, integrating with the annotation package API, and migrating storage of cases to the annotation package. Remaining tasks include specifying and implementing user roles, case creation and testing workflows, and reporting of statistics.



Figure 9 – Screenshot of the first benchmarking results in the audit benchmarking system

In the time since meeting N we also continued the cooperation with the audit trial initiative and the corresponding TG-Symptom audit group. As part of this work, we could see first benchmarking results for the challenge we setup inside the audit benchmarking platform. Figure 9 shows the results with the expected performance scores already measured by the MMVB 2.0 benchmarking system previously developed by the topic group.

The cooperation with the audit group also covered the finalization of the audit questionnaire for TG-Symptom systems. Until early June the questionnaire will be implemented in the audit platform. In parallel we now investigate the scores and metrics for both, the qualitative results from the questionnaire and the quantitative results from the benchmarking. In contrast to the other topic groups participating in the audit trial we do not plan to publish any paper before the actual benchmarking mid 2023.

From 7.4.2022 to 8.4.2022 the topic group held its fourth workshop. The focus was here to plan out the remaining time until the final document submission deadline set by the focus group to mid 2023. As output of the workshop, we created a roadmap outlining all relevant foreseeable tasks. The main points are shown in Figure 10.



Figure 10 – Main roadmap items for the remaining time of the topic group.

The tickets identified during the workshop have been added to a new Jira instance setup for this purpose. With this transition we will stop using the github issue tracking that did not meet all requirements. According to the plan, in meeting #69 we also started reviewing the AI benchmarking interface. Different than expected we decided to investigate FHIR as format for encoding the benchmarking cases to send to the participating AIs which will increase the interoperability of the benchmarking cases.

During workshop #4 the topic group also agreed to nominate Martin Cansdale from Healthily (former Your.MD) as co-driver of the topic group to facilitate the implementation of the outlined roadmap.

All the work in the topic group was organized online. The following list shows all the online meetings since the since meeting N:

- 18.02.2022 – Meeting #66 – Telco [Minutes](#)
- 01.04.2022 – Meeting #67 – Telco [Minutes](#)
- 07.04.2022 – 08.04.2022 – Meeting #68 – Workshop #4 [Minutes](#)
- 20.05.2022 – Meeting #69 – Telco [Minutes](#)

All the meeting notes can also be found in the official TG-Symptom SharePoint folder:

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

In addition to the regular bi-weekly TG-Symptom meetings there have been weekly informal developer stand-ups to sync on technical details as well as informal management syncs. TG-Symptom also participated in several meetings of the TG-Symptom audit group. We also introduced a weekly management call.

Currently, our topic group has the following 22 companies and 7 independent contributors:

- 1Doc3 (Lina Porras and Maria Gonzalez)
- Ada Health (Henry Hoffmann, Dr Shubhanan Upadhyay, Ivan Lebovka, Milan Jovanovic)
- Alejandro Orsonio (Independent Contributor)
- Babylon Health (Saurabh Johri, Adam Baker)
- Baidu (Yanwu XU)
- Barkibu (Ernesto Hernandez)
- Buoy (Sarah Hassonjee)
- Deepcare.io (Hanh Nguyen)
- Flo (Anna Klepchukova, Saddif Ahmed)
- Healthily (Jonathon Carr-Brown, Martin Cansdale, Dr Audrey Menezes)
- Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys)
- Inspired Ideas (Megan Allen, Ally Salim Jnr)
- Isabel Healthcare (Jason Maude)
- Kahun (Michal Tzuchman Katz)
- Dr Reza Jarral (Independent contributor)
- Mfine (Dr Srinivas Gunda)
- Muhammad Murhaba (Independent Contributor)
- MyDoctor (Harsha Jayakody)
- Nivi (David Tresner-Kirsch)
- 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)
- Yura Perov (Babylon)

The topic group email reflector fgai4htgsymptom@lists.itu.int altogether has currently 121 (+4) (duplicates not counted) subscribers. The latest meeting O version of this Topic Description Document lists 37 contributors.

2.2.13 Status update for meeting P (Helsinki)

According to the plan outlined in the previous report, the work after meeting O continued with implementing the necessary annotation tool changes requested by the Topic Group doctors, followed by an additional round of re-testing.

Based on the Annotation Tool the doctors started to revise and extend the step-by-step annotation guidelines on how to use the tool to create cases. The work showed how crucial the quality of this document will be for the quality of the datasets for the benchmarking. Accordingly, we will give it more priority.

In parallel to writing the guidelines the doctors involved in the annotation tool development reached out to other doctors among the Topic Group members to ask for additional feedback on the usability of the annotation tool and in preparation to the upcoming first round of benchmarking test case creation. As part of this we had several meetings where we introduced doctors to the status of the Topic Group as well as the details on testing the annotation tool and creating benchmarking cases.

Beside finishing the annotation tool, the technical work since the last meeting focused on continuing the evaluation of FHIR as format for storing benchmarking cases. In several workshops we analysed the different structures provided by the FHIR specification and agreed on a first version of a FHIR encoded benchmarking case using SNOMED as the reference ontology. To make sure that the specified FHIR documents are valid, we implemented test code for generating cases by using the FHIR HAPI via Kotlin and the Python FHIR resources project. The current fully self-contained FHIR case use a top-level bundle containing observation resources for all the symptoms as well as the expected conditions and expected triage result. The format specifies means to represent present, non-present and skipped symptoms, attribute expressions and the representation of age and sex (which will explicitly not use a patient resource). We also planned out the next technical steps which will involve:

- Finalisation of the FHIR python script generating a case according to the agreed upon schema (done)
- Test-wise automated conversion of existing MMVB 2.2 cases to FHIR
- Implementation of AI endpoints accepting FHIR cases by TG members
- Update and test of the audit benchmark script to use the FHIR endpoints and FHIR cases
- Update AIs to use FHIR as output
- Update and test the audit benchmark to use the FHIR outputs
- Update the annotation tool to directly read/write FHIR cases

In this past reporting period, the Topic Group also continued the cooperation with the TG-Symptom audit group. The work focused here on further refining the audit questionnaire and test wise filling of it by Ada and Healthily. The filling was conducted in documents shared only between each company and the audit-group so that for instance the criticality of certain IP details can be discussed without already sharing them outside the audit-group. As next steps the questionnaires will now be evaluated as soon as the audio group agreed on the scoring metrics. Filling the questionnaires already showed that some of the questions likely need to be rephrased and some of the answers that might be necessary to interpret the benchmarking results are difficult to share because of business constraints and IP considerations.

All the work in the topic group was organized online. For the past reporting period we switched from having bi-weekly meetings to shorter but weekly calls to coordinate the work in the different workstreams. The notes of the meetings of the corresponding 8 online-meetings notes have been published as one single document:

- 07.06.2022 – 06.09.2022 – Telco [Minutes](#)

All the meeting notes can be found in the official TG-Symptom SharePoint folder:

<https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Symptom.aspx>

TG-Symptom also participated in several meetings of the TG-Symptom audit group.

Currently, our topic group has the following 22 companies and 7 independent contributors (new contributors inside the companies are marked bold):

- 1Doc3 (Lina Porras, Maria Gonzalez, **Jhon Lince**)
- Ada Health (Henry Hoffmann, Dr Shubhanan Upadhyay, Ivan Lebovka, Milan Jovanovic)
- Alejandro Orsonio (Independent Contributor)
- Babylon Health (Saurabh Johri, Adam Baker)
- Baidu (Yanwu XU)
- Barkibu (Ernesto Hernandez)
- Buoy (Sarah Hassonjee)
- Deepcare.io (Hanh Nguyen)
- Flo (Anna Klepchukova, Saddif Ahmed)
- Healthily (Jonathon Carr-Brown, Martin Cansdale, Dr Audrey Menezes)
- Infermedica (Piotr Orzechowski, Dr Irv Loh, Jakub Winter, Michal Kurtys, **Mateusz Palczewski, Jakub Jaszczak**)
- Inspired Ideas (Megan Allen, Ally Salim Jnr)
- Isabel Healthcare (Jason Maude)
- Kahun (Michal Tzuchman Katz)
- Dr Reza Jarral (Independent contributor)
- Mfine (Dr Srinivas Gunda)
- Muhammad Murhaba (Independent Contributor)
- MyDoctor (Harsha Jayakody)
- Nivi (David Tresner-Kirsch)
- 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)
- Yura Perov (Babylon)

The topic group email reflector fgai4htgsymptom@lists.itu.int altogether has currently 126 (+5) (duplicates not counted) subscribers. The latest meeting P version of this Topic Description Document lists 37 contributors.

2.3 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.

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

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>

3 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.

3.1 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](#) "*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.

3.2 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).

3.3 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 [Pillay2010]. 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 SDG 3].

3.4 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.

3.4.1 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
IDOC3	IDOC3 platform	<ul style="list-style-type: none"> Age, sex Free text Complementary information about signs, symptoms and medication related to the main topic. 	<ul style="list-style-type: none"> Pre-clinical triage Possible Causes – differentials. 	<ul style="list-style-type: none"> Worldwide Spanish More than 750 conditions Web and App for iOS and Android
Ada Health GmbH	Ada app	<ul style="list-style-type: none"> Age, sex, risk factors Free text PC search 	<ul style="list-style-type: none"> Pre-clinical triage Differentials for PC 	<ul style="list-style-type: none"> Worldwide

		<ul style="list-style-type: none"> Discrete answers to dialog questions for additional symptoms including attribute details like intensity 	<ul style="list-style-type: none"> Shortcuts in case of immediate danger 	<ul style="list-style-type: none"> English (US/UK), German, Spanish, Portuguese, French, Swahili, Romania Top 1300 conditions For smartphone users Android/iOS
Babylon Health	Babylon App	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> Worldwide English 80% of medical conditions For smartphone/web users Android/iOS/Web
Baidu	Baidu's Clinical Decision Support System	<ul style="list-style-type: none"> 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 <p>* these details must be provided</p>	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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
Deepcare	Deepcare Symptom Checker			<ul style="list-style-type: none"> Users: Doctor and Patient Platforms: iOS, Android Language: Vietnamese
Flo Health	Flo App	<ul style="list-style-type: none"> Age, assumes female sex Symptoms, risk factors, menstrual cycle history Answers to discrete dialog questions 	<ul style="list-style-type: none"> Differential suggestions with explanation 	<ul style="list-style-type: none"> Worldwide - women only English and 21 other languages Women only App (iOS and Android) and web based
Infermedica	Infermedica API, Symptomate	<ul style="list-style-type: none"> Age, sex Risk factors Free text input of multiple symptoms Region/Travel history 	<ul style="list-style-type: none"> Differentials for PC Pre-clinical triage Shortcuts in case of immediate danger 	<ul style="list-style-type: none"> Worldwide Top 1000 conditions 15 language versions Web, mobile, chatbot, voice

		<ul style="list-style-type: none"> Answers to discrete dialog questions Lab test results 	<ul style="list-style-type: none"> Explanation of differentials Recommended further lab testing 	
Inspired Ideas	Dr. Elsa	<ul style="list-style-type: none"> Age, gender Risk factors Region/ time of year Multiple symptoms Travel history Answers to discrete dialog questions Lab test results Clinicians hypothesis 	<ul style="list-style-type: none"> List of possible differentials Condition explanations Referral & lab test recommendations Recommended next steps Clinical triage 	<ul style="list-style-type: none"> Tanzania, East Africa Languages: English and Swahili Android/iOS/Web/API Users: healthcare workers/ clinicians
Isabel Healthcare	Isabel Symptom Checker	<ul style="list-style-type: none"> Age Gender Pregnancy Status Region/Travel History Free text input of multiple symptoms all at once 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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 most languages
Kahun	Kahun decision support, Patient1st	<ul style="list-style-type: none"> Age, sex Risk factors Pregnancy Status Answers to discrete dialog questions Lab test results 	<ul style="list-style-type: none"> Pre-clinical triage Differentials for PC Shortcuts in case of immediate danger 	<ul style="list-style-type: none"> Worldwide Multilingual Over 6000 medical conditions Android/iOS/Web/API
Visiba Group AB	Visiba Care app	<ul style="list-style-type: none"> Age Gender Chatbot free text input Region/ time of year Discrete answers Lab results, inputs from devices enabled 	<ul style="list-style-type: none"> List of possible diagnoses pre-clinical triage including format of meeting (digital or physical) Next-step advice condition information 	<ul style="list-style-type: none"> Language: Swedish Android/iOS/Web Users: Doctor and Patient
XUND Solutions	XUND App	<ul style="list-style-type: none"> Age Gender Risk factors Guided dialogue Standardised answers 	<ul style="list-style-type: none"> Pre-clinical triage In-depth explanations Recommendations Navigation within healthcare system 	<ul style="list-style-type: none"> Europe (CEE & CIS) Primary healthcare (350 conditions); up to 500 planned German, English, Hungarian Patient-centered Mobile & API
Your.MD Ltd	Your.MD app	<ul style="list-style-type: none"> Age, sex, medical risk factors, 	<ul style="list-style-type: none"> Differentials for PC Pre-clinical triage 	<ul style="list-style-type: none"> Worldwide English,

		<ul style="list-style-type: none"> • Chatbot free text input • User consultation output (report) 	<ul style="list-style-type: none"> • Shortcuts in case of immediate danger • Condition information • Recommendation of appropriate local services and products • Medical factors 	<ul style="list-style-type: none"> • >630 conditions • For smartphone users Android /iOS and web and messaging groups Skype etc
--	--	--	--	--

3.4.2 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	Symptom checker	•	•	•
AHEAD Research	Symcat	•	•	•
Curai	Patient-facing DDSS / Chatbot	•	•	•
DocResponse	DocResponse	•	•	• for doctors
	Doctor Diagnose	•	•	• Android
Drugs.com	Symptom Checker	•	•	• Triage • Note: Harvard Health decision guide used
EarlyDoc		•	•	• Web
FamilyDoctor.org	Symptom Checker	•	•	• Web
Healthline	Symptom Checker	•	•	•
Healthtap	Symptom Checker (for members)	•	•	•
K Health	K app chatbot	•	•	•
Mayo Clinic	Symptom Checker	•	•	•
MDLive	Symptom checker on MDLive app	•	•	•
MEDoctor	Symptom Checker	•	•	•

Mediktor	Web-based symptom checker , or Mediktor app	•	•	•
NetDoktor	Symptom Checker	•	•	•
PingAn	Good Doctor app	•	•	•
Sharecare, Inc.	AskMD	•	•	•
WebMD	Symptom checker	<ul style="list-style-type: none"> • Age, Gender, Zip code • Multiple presenting symptoms • Answers to discrete dialog questions 	<ul style="list-style-type: none"> • List of possible differentials • Explanation of differentials • Possible treatment options 	•

3.4.3 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.

3.4.3.1 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
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.
...	

3.4.3.2 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

3.4.3.2.1 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).

3.4.3.2.2 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].

3.4.3.2.3 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.

3.4.3.2.4 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.

3.4.4 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.

3.4.4.1 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
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:

3.4.4.1.1 Clinical Triage

The simplest 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.

3.4.4.1.2 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.

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-limiting
- Self-care
- Primary care 2 weeks
- Primary care 2 days
- Primary care same day
- Emergency A&E
- Emergency ambulance

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 trade-off between number of different values and inter-annotator-agreement
- discussion trade-off 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

3.4.4.1.3 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”.

3.4.4.1.4 Next Step Advice

Next Steps Advice exist to guide the user towards a suggested action to take, and are often based on pre-clinical triage. Next Steps vary in their granularity across different tools. They can guide the

user towards specific services or institutions in order to further assess or treat their symptoms or conditions that may be compatible with their symptoms. There is sometimes a brief explanation of the type of action the recommended service might take. Examples of services or institutions that are recommended across the various tools include primary care doctors, genitourinary medicine clinics, pharmacists, dietitians, and psychologists. These services and institutions tend to be localized to country-specific guidelines or recommendations. Specific named services and institutions may also be recommended in line with commercial partnerships. Some Next Steps Advice sections also provide information about what action the user should take if their symptoms change, worsen, or do not improve, as a form of ‘safety netting.’

3.4.4.1.5 Treatment Advice

Treatment Advice provides the user with advice as to how manage the user’s symptoms or condition compatible with their symptoms. This can be in the form of possible treatment options that might be suggested by a recommended service or institution outlined in the Next Steps Advice (e.g. medicine that might be prescribed), or may be concrete suggestions of how to treat the problem if it is a condition or symptom that can be managed with self-care (e.g. simple painkillers, exercises). Treatment Advice is often generic and common across countries based on the evidence base for the condition or symptom in question, but may on occasion be informed by local guidelines or recommendations; further information can also be provided in the form of links to medically validated health information sources.

3.4.5 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 paediatrics, 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.

3.4.6 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.

3.4.7 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 be compromised. The remainder 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.

4 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 "*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.

Across the world, people are increasingly making use of digital infrastructures, such as dedicated health websites, wearable technologies and AISAs, in order to improve and maintain their health. A UK survey found that a third of the population uses internet search engines for health advice. This digitally mediated self-diagnosis is also taking place in countries in the global South, where access

to healthcare is often limited, but where mobile and internet penetration over the last decade has increased rapidly.

This section widens the lens in considering the ethical and cultural dimensions and implications of AISAs beyond their technical accuracy. It considers that humans, and their diverse social and cultural environments, should be central at all stages of the product's life-cycle. This means recognizing both people's formal rights and obligations but also the substantive conditions that allow them to achieve and fulfil them. This means considering the economic and social inequalities at a societal and global level that shape AISAs and their deployment.

4.1 The ethical implications of applying the AI model in real world scenarios

Effect on decision-making in health

AISAs will shape how individuals seek care within a healthcare system which have important ethical implications. They may influence users to act when there is no need –or stop them from acting, by not seeing a doctor when they ought to. Healthcare workers using AISAs may come to rely heavily upon them reducing their own independent decision-making, a phenomena termed 'automation bias'. These behaviours will vary depending on the healthcare system, such as the availability of healthcare workers, drugs and diagnostic tests. For instance, if the AISA makes suggestions for next steps that are unavailable or inaccessible to users, they may choose not to utilise the AISA, turning instead to alternative forms of medical advice and treatment. The individual health-seeking behaviour can also be shaped by existing hierarchies. For instance, a healthcare worker may feel undermined if a patient ignores their medical advice in favour of that given by the AISA, potentially hindering the patient's access to healthcare.

Accountability

AISAs raise serious ethical questions around accountability. Some of these are designed to be answered through the benchmarking process, but others might not have clear-cut answers. As the UK Academy of Medical Royal Colleges has suggested, while accountability should lie largely with those who designed the AI system (when used correctly), what happens when a clinician or patient comes to trust the system to such an extent that they 'rubber stamp' its decisions? It is also worth noting that there is evidence from the global South that AISAs, and related technologies, are currently being used not only by health professionals and patients, but also by intermediates with little healthcare training.

Technical robustness, safety and accuracy

AISAs must be technically robust and safe to use. They must continue working in the contexts they were designed for, but must also anticipate potential changes to those contexts. AISAs may be maliciously attacked or may break down, which can cause a problem when they are relied upon, necessitating contingency measures to be built into the design.

Wider potential societal effects of AISAs

There may also be long-term effects of AISAs on the public healthcare system. For instance, they may discourage policy makers from investing in human resources. This may adversely affect more vulnerable, marginalised or remote populations who are unable to use AISAs due to factors including a lack of adequate digital data infrastructures and digital illiteracy. This could exacerbate an existing 'digital divide'. Furthermore, in the case of clinician-facing AISAs, consideration would need to be put to re-skilling health workers, many of whom are increasingly required to utilise in their working lives various other digital diagnosis and health information systems.

AISAs will also rely upon existing digital infrastructures that consume resources in their design, production, deployment and utilization. Responsibility around this digital infrastructure is dispersed across many bodies, but the group should at least be aware of the harms that may exist to the environment along the supply chain, including the disposal of outdated or non-functioning hardware.

4.2 The ethical implications of introducing benchmarking

Setting up benchmarking of AISAs will help assess their accuracy, a vital dimension of their quality. This will be important in considering the ethical and cultural dimensions and implications of using AISAs compared to other options, which include not only the aforementioned digital-based solutions but most significantly human experts – with variable levels of expertise, accessibility and supportive infrastructures, such as diagnostic tests and drugs.

Benchmarking could:

- TODO give the illusion of safety
- TODO risk replacing real-world studies of the technology that would assess the wider impact and context of them

The quality of a technology should be assessed in multifaceted ways that go further than benchmarking via independently curated datasets 1. This means benchmarking should not aim to replace real world studies in clinical settings. The metrics and results should be interpreted with a recognition of limitations that these datasets might have. Whilst there would be a drive to ensure global representation of relevant data, this is itself inherently prone to bias from those who curate the data. If the data is being curated from electronic records, for example, this may mean that underrepresented communities or settings that use paper records are not represented in the dataset. As such, benchmarking should not be seen as a replacement for prospective studies in clinical settings. The ‘users’ of the outputs of benchmarking (e.g. governments, health systems, regulators or clinicians), should take these limitations into account, and ensure benchmarking alone is not seen as a guarantee of safety or efficacy of AISAs, and rather as one part of a more holistic evaluation.

- TODO Create burdens that thereby put better-resourced actors (including companies and governments) at an advantage

4.3 The ethical implications of collecting the data for benchmarking

An important question around fairness concerns the data collected for the training of the AISAs. How has authority been established for the ownership, use and transfer of this data? There may be important inequalities at different scales, from individual to larger entities such as governments and corporations, that need to be considered. Glossing over exchanges between these actors as mutually beneficial or egalitarian may obscure these inequalities. For instance, an actor may agree to provide health data in exchange for better trained models or even for a subsidised or free service, but in the process may lose control over how that data is subsequently used.

4.4 Risks facing individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground

There is a potential for AISAs to produce biased advice: systematically incorrect advice, resulting from AI systems trained on data that is not representative of populations that will use or be impacted by these systems. There is a particular danger of biased advice affecting vulnerable or marginalised populations. Dangers and risks include:

- An individual may receive an erroneous diagnosis that could lead to them seeking the wrong kind of treatment either within or outside the healthcare setting.
- If an algorithm/technology is scaled more widely it has the potential to affect many more people than, for instance, an individual clinician.
- Misguided decisions on healthcare expenditures and coverage, whether public or private
- Erroneous decisions concerning the disease burden for particular populations

4.5 How the privacy of personal health information protected

AISAs must adhere to strict standards around data governance, privacy and quality. This also applies to the benchmarking process of AISAs, which requires labelled test data. Depending on the approach for creating the data set this might involve real anonymized patient cases, in which case privacy and protection is crucial. Given the importance of this issue, the Focus Group actively works on ensuring that the used data meets high standards for ethics and protection of personal data. There are a number of regulations that can be drawn upon including the European Union General Data Protection Regulation and the US Health Insurance Portability and Accountability Act. National laws also exist, or are being developed, in a number of countries.

How is it 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)?

The design of the AISA should consider fairness. Issues, shaped by social, political, economic and cultural factors, such as access to, and ability to use, the AISA are important - including access to appropriate smartphone devices, language, and digital literacy. For instance, it has been shown that in Sierra Leone, AI tools designed to predict the mobility during the Ebola outbreak by tracking mobile phone data failed because they did not consider how mobile phones were often shared among friends, neighbours and family.⁸ The group should also consider how wider infrastructures, such as electricity and internet, interact with a particular AISA to shape fairness.

Compared to medical professional assessment and conventional diagnostics, an AI system should lead to an increase in both specificity and sensitivity in the context of diagnosis. In certain contexts, a trade-off of specificity against sensitivity is possible. This context must be made clear before establishing a benchmark. For example, in emergency settings it might be advisable to increase sensitivity even if specificity is slightly reduced. An effective benchmark will be adapted to these settings. In order to be judged "better" than conventional diagnostics, an AI system (or medical professionals using this system) must prove superiority to the prior gold standard.

Explicability - The current benchmarking process is intended to evaluate the accuracy of an AISA's prediction. However, the importance of explaining and communicating such predictions, and the potential trade-offs in accuracy, should be considered by the group. Such explicability should also be considered in regard to the expected users of the AISA, from physicians to community health workers to the public.

- (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 are your experiences and learnings from addressing ethics in your TG?

5 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.

5.1 Subtopic Self-Assessment

5.1.1 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](#) “AI4H evaluation process description,” [DEL07_2](#) “AI technical test specification,” [DEL07_3](#) “Data and artificial intelligence assessment methods (DAISAM),” and [DEL07_4](#) “Clinical Evaluation of AI for health”.

To establish a standardized benchmarking for AI-based symptom assessment systems, it is valuable to analyse previous benchmarking work in this field. So far, little work has been performed, which is also a reason that the introduction of a standardized benchmarking framework is important. The current work falls into several subcategories that will be discussed in their own subsections.

5.1.1.1 Scientific Publications on Benchmarking AI-based Symptom Assessment Applications

Whilst rare, a few publications exist that worked on assessing the performance of AI-based symptom assessment systems. For reviewing, the details of the different approaches and their relevance for setting up a standardized benchmarking the most relevant publications will be discussed in the subsequent sections:

5.1.1.1.1 “ISABEL: Accuracy of a Machine Learning Based Ddx Generator“ [rx13]

563 cases of diagnostic error were collected over a period of 2 years from case reports, journals and detailed press articles. The cases covered 300 diagnoses and 27 specialties and, on average, contained 6 clinical features each. The free text case presentations were entered into Isabel DDx Generator and the position of the known final diagnosis within the tool’s list of ranked possible diagnoses recorded. Results: In 74% of the cases the final diagnosis was in the top 3 suggestions. In 87% of cases the final diagnosis was in the top 5 suggestions and in 98% of cases the final diagnosis was in the top 10 suggestions.

5.1.1.1.2 „Asking ISABEL for diagnostic dilemmas in pediatrics: How does a web based diagnostic checklist perform?“ [rx14]

Using a case-based textbook, 10 participants selected keywords from 25 cases; each read the same cases and were blinded to the diagnosis. Age, gender, and keywords were entered into Isabel. The primary outcome measure was Isabel’s inclusion of the diagnosis on 'Page 1' (top 10 diagnoses) and 'View all' (top 30 diagnoses). The secondary outcome measure was the impact of level of training on Isabel's success rate. Lower level of training (LLT) was defined as medical student and resident. Higher level of training (HLT) was defined as junior and senior faculty.

Isabel’s performance with published cases:

- Isabel included the diagnosis in 60% (149/248) of cases on 'Page 1', and 81% (202/248) on 'View all' (p=0.001).
- With LLT users, Isabel included the diagnosis in 55% (55/100) of cases on 'Page 1' compared to 64% (94/148) with HLT (p=0.18).
- With LLT users, Isabel included the diagnosis in 78% (78/100) of cases on 'View All' compared to 84% (124/148) with HLT (p=0.25).

5.1.1.1.3 Performance of a Web-Based Clinical Diagnosis Support System for Internists [rx15]

We tested 50 consecutive Internal Medicine case records published in the New England Journal of Medicine. We first either manually entered 3 to 6 key clinical findings from the case (recommended approach) or pasted in the entire case history. The investigator entering key words was aware of the correct diagnosis. We then determined how often the correct diagnosis was suggested in the list of 30 differential diagnoses generated by the clinical decision support system. We also evaluated the speed of data entry and results recovery.

The clinical decision support system suggested the correct diagnosis in 48 of 50 cases (96%) with key findings entry, and in 37 of the 50 cases (74%) if the entire case history was pasted in. Pasting took seconds, manual entry less than a minute, and results were provided within 2–3 seconds with either approach.

5.1.1.1.4 "Comparison of physician and computer diagnostic accuracy." [rx16]

Semigran et al. expounded on their 2015 systematic assessment of online symptom checkers by comparing checker performance—the previous 45 vignettes—to physician (n=234) diagnoses. Physicians reported the correct diagnosis 38.1% more often symptom checkers (72.1% vs. 34.0%), additionally outperforming in the top three diagnoses listed (84.3% vs. 51.2%). Physicians were also more likely to list the correct diagnoses for high-acuity and uncommon vignettes, while symptom checkers were more likely to list the correct diagnosis for low-acuity and common vignettes. While the study is limited by physician selection bias, the significance of the results lies in the vast outperformance of physician diagnoses.

5.1.1.1.5 "A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom checkers." [rx17]

Unique algorithms (n=4) from the top 20 web-based symptom checkers were evaluated for their ability to diagnose degenerative cervical myelopathy (DCM): WebMD, Healthline, Healthtools.AARP, and NetDoctor. A single case vignette of up to 31 DCM symptoms derived from 4 review articles was entered into each symptom checker. Only 45% of the 31 DCM symptoms were associated with DCM as a differential by the symptom checkers, and in these cases a majority 79% ranked DCM in the bottom two-thirds of differentials. Insofar as web-based symptom checkers are able to detect symptoms of degenerative disorder, the authors conclude their is technological potential, but an overall lack of acuity.

5.1.1.1.6 "Evaluation of symptom checkers for self diagnosis and triage" [rx10]

TODO

5.1.1.1.7 "ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation" [rx11]

TODO

5.1.1.1.8 "Safety of patient-facing digital symptom checkers." [rx12]

TODO

5.1.1.2 Clinical Evaluation of AI-based Symptom Assessment

While there is currently a stronger focus on patient-facing symptom assessment systems, some work has also been done on assessing the performance of similar systems in a clinical context. The relevant publications are discussed in the following sub sections.

5.1.1.2.1 "A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. Emergencias" [rx18]

One report was published in 2017 assessing a single AI-based symptom assessment in a Spanish Emergency Setting.¹⁵ The tool was used for non-urgent emergency cases and users were included who were above 18 years, willing to participate, had a diagnosis after leaving the emergency department and if this diagnosis was part of the Mediktor dictionary at this time. With this setting, the symptom assessment reached an F1 Score of 42.9%, and F3 score of 75.4% and F10 score of 91.3% for a total of 622 cases.

5.1.1.2.2 "Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department" [rx19]

The results of a subsequent prospective study to the Moreno et al. (2017) evaluation of Mediktor were published in 2019. This study was also conducted in an emergency room setting in Spain and consisted of a sample of 219 patients. With this setting, the symptom assessment reached an F1 Score of 37.9%, and F3 score of 65.4% and F10 score of 76.5%. It was further determined that Mediktor's triage levels do not significantly correlate with the Manchester Triage System for emergency care, or with hospital admissions, hospital readmissions and emergency screenings at 30 days.

5.1.1.2.3 "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence." [rx20]

Recently, a study by Liang et. al. showed a proof of concept of a diagnostic decision support system for (common) pediatric conditions based on a natural language processing approach of EHR. The F1 Score was overall between junior and senior physicians group with an average F1 score of 0.885 for the covered conditions.

5.1.1.2.4 "Evaluating the potential impact of Ada DX in a retrospective study." [rx21]

A retrospective study evaluated the diagnostic decision support system Ada DX in 93 cases of confirmed rare inflammatory systemic diseases. Information from patients' health records was entered in Ada DX in the cases' course over time. The system's disease suggestions were evaluated with regard to the confirmed diagnosis. The system's potential to provide correct rare disease suggestions early in the course of cases was investigated. Correct suggestions were provided earlier than the time of clinical diagnosis in 53.8% of cases (F5) and 37.6% (F1) respectively. At the time of clinical diagnosis the F1 score was 89.3%.

5.1.1.2.5 "Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain." [rx22]

The results of a prospective observational study were published in 2016 in which researchers evaluated the accuracy of a web-based symptom checker for ambulatory patients with knee pain in the United States. The symptom checker had the ability to provide a differential diagnosis for 26 common knee-related conditions. In a sample size of 259 patients aged above 18 years, the symptom assessment reached an F10 score of 89%.

5.1.1.2.6 "How Accurate Are Patients at Diagnosing the Cause of Their Knee Pain With the Help of a Web-based Symptom Checker?" [rx23]

In a follow up to the Blisson et al. (2014) study investigating the accuracy of a web-based symptom checker for knee pain, a prospective study was conducted across 7 sports medicine clinics to evaluate patient's ability to self-diagnose their knee pain with the help of the same symptom checker within a cohort of 328 patients aged 18–76 years. Patients were allowed to use the symptom checker, which generated a list of potential diagnoses after patients had entered their symptoms. Each diagnosis was linked to informative content. Patients then self-diagnosed the cause of their knee pain based on the information from the symptom checker. In 58% of cases, one of the patients' self-diagnoses matched the physician diagnosis. Patients had up to 9 self-diagnoses.

5.1.1.2.7 "Are online symptoms checkers useful for patients with inflammatory arthritis?" [rx24]

A prospective study in secondary care in the United Kingdom evaluated the NHS Symptom Checker for triage accuracy and Boots WebMD for diagnostic accuracy against physician diagnosis of inflammatory arthritis: rheumatoid arthritis (n = 13), psoriatic arthritis (n = 4), unclassified arthritis (n = 4) and inflammatory arthralgia (n = 13). The study aimed to expand literature into the effectiveness of online symptom checkers in real patients in relation to how the internet is used to search for health information. 56% of patients were suggested the appropriate level of care by the NHS Symptom Checker, while 69% of rheumatoid arthritis patients and 75% of psoriatic arthritis patients had their diagnosis listed amongst the top five differential diagnoses by WebMD. Low triage accuracy led the authors to predict an inappropriate use of healthcare resources as a result of these web-based checkers.

5.1.1.2.8 "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis" [rx25]

In the study it was hypothesised that an artificial intelligence (AI) powered triage and diagnostic system would compare favourably with human doctors with respect to triage and diagnostic accuracy. A prospective validation study of the accuracy and safety of an AI powered triage and diagnostic system was performed. Identical cases were evaluated by both an AI system and human doctors. Differential diagnoses and triage outcomes were evaluated by an independent judge, who was blinded from knowing the source (AI system or human doctor) of the outcomes. Independently of these cases, vignettes from publicly available resources were also assessed to provide a benchmark to previous studies and the diagnostic component of the Membership of the Royal College of General Practitioners (MRCGP) exam. Overall, it was found that the Babylon AI powered Triage and Diagnostic System was able to identify the condition modelled by a clinical vignette with accuracy comparable to human doctors (in terms of precision and recall). In addition, it was found that the triage advice recommended by the AI System was, on average, safer than that of human doctors, when compared to the ranges of acceptable triage provided by independent expert judges, with only a minimal reduction in appropriateness.

5.1.1.2.9 "CONSORT-AI and SPIRIT-AI reporting guidelines" [rx28]

Adapted from traditional clinical trials guidelines, the CONSORT-AI and SPIRIT-AI reporting guidelines were published in September 2020 after a staged Delphi consensus process of academics, AI experts and clinicians. They comprise of a checklist for anyone reporting the results of a trial that includes an AI intervention, and publishing its protocol. The checklist aims to standardise the sharing of information from the trial. The CONSORT-AI and SPIRIT-AI reporting guidelines are a significant and formative part of the evolution of clinical AI research, as they encourage research teams to conduct and report trials in a trusted way. If AI in healthcare is to be trusted by clinical stakeholders and decision makers, then transparent reporting, all the way from data sourcing, to clinical outcomes, is key. These reporting guidelines are a first step in establishing a minimum standard of accountability in AI interventions.

5.1.1.3 Benchmarking Publications outside Science

In addition to scientific benchmarking attempts, there are several newspaper articles reporting tests of primarily user-facing symptom assessment applications. Since these articles have not been peer reviewed and are not always follow following scientific standards, they will not be discussed in this TDD.

5.1.2 Benchmarking by AI developers

All developers of AI solutions for TG Symptom 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.

Probably the most sophisticated systems for benchmarking symptom assessment systems are the ones created by the different companies developing such systems for internal testing and quality control. While most of the details are unlikely to be shared by the companies, this section points out insights relevant for creating a standardized benchmarking.

Dataset Shift

In most test sets the distribution of conditions is not the same as the distribution found in the real world. There are usually a few cases for even the rarest conditions while at the same time the number of common cold cases is limited. This gives rare diseases a much higher weight in the aggregation of the total scores. While this is desirable to make sure that all disease models perform well, in some cases it is more important to measure the net performance of systems in real world scenarios. In this case the aggregation function needs to scale the individual cases results with its expected top match prior probability in order to get the mathematically correct expectation-value for the score. For example, errors on common-cold cases need to be punished harder than errors on cases of rare diseases that only a few people suffer from. The benchmarking should include results with and without correction of this effect.

Medical distance of the top matching diseases to the expected ones

In case the expected top match is not the first position and the listed conditions are not in e.g. a set of "expected other conditions", the medical distance between the expected conditions and actual conditions could be included in the measure.

The rank positions

In case the expected top-match is not in the first position, the actual position might be part of the scoring. This could include the probability integral of all higher-ranking conditions or the difference between the top scores and the score of the expected disease.

The role of the secondary matches

Since AISA systems usually present multiple possible conditions, even if the top match is correct the qualities of the other matches need to be considered as well. For example, the highly relevant differentials that should be ruled out are much better secondary diagnoses than random diseases.

- Which scores and metrics have been used?

- How did they approach the acquisition of test data?

5.1.3 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](#) “*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).

Document [C031](#) provides a list of the available platforms. While not specific for symptom assessment it provides important examples for many aspects of benchmarking ranging from operational details, over scores & metrics, leader boards, reports to the overall architecture. Due to high numbers of participants and the prestige associated with a top rank, the platforms have also substantial experience in designing the benchmarking in a way that is hard or impossible to manipulate.

In response to the call for benchmarking platforms ([FGAI4H-C-106](#)), in meeting D in Shanghai [FGAI4H-D-011](#) suggested the use of AICrowd. As discussed in meeting D, the topic group had a look at AICrowd to get a first overview if it could be an option for benchmarking the AI systems in this topic group.

5.1.3.1 Features requiring special attention in TG Symptom

While many AI benchmarks also involve tasks in health, the benchmarking for this topic group has some specific requirements that will be discussed in this section.

Technology Independence

The AI systems that are a part of this topic group run on complex architecture and use a multitude of technologies. In contrast, most benchmarking platforms have been primarily designed for use with Python based machine learning prototypes. One important requirement is therefore that the benchmarking platform is completely technology agnostic e.g. by supporting AIs submitted as docker containers with a specified interface.

Custom Scores & Metrics

For the tasks benchmarked by the common benchmarking platforms the focus is on only a small number of scores. In many cases it is even possible to use common ready-made built-in scores. For benchmarking the performance in our topic group, we will need to implement a multitude of new scores and metrics to reflect the different aspects of the quality and performance of self-assessment systems. It is therefore important that the benchmarking platform allows to define and add new custom scores - ideally by configuration rather than changing the platform code, to compute them as part of the benchmarking and automatically add them to the generated reports.

Custom Reports & Additional Reporting Dimensions

Together with the many more scores, the platform also needs to support the generation of reports that include all the scores in a readable way. Beside the scores there are also many dimensions to organize the reports by so that it is clear which technologies fit the needs of specific use cases.

Interactive Reports & Data Export

Since the number of dimensions and score will grow fast, it will not always be possible to automatically provide the reports answering all the details for all possible use cases. For this case the platform needs to either provide interactive navigation and filtering of the benchmarking result data or at least an easy way to export the data for further processing e.g. in tools like Tableau.

Support for Interactive Testing

Whilst for the first benchmarking iterations providing cases with all the evidence at once might suffice, later iterations will probably also test the quality of the dialog between the system and the user e.g. only answering questions the AI systems explicitly ask for. The platform should allow a way to implement this dialog simulation.

Stability & Robustness & Performance & Errors

Beside benchmarking using the test data as-is, we also need to assess the stability of the results given a changed symptom order or in a second run. We also need to record the run time for every case or possible error codes, hanging AIs and crashes without itself being compromised. Recording these details in a reliable and transparent way requires the benchmarking platform to perform a case-by-case testing rather than e.g. letting the AI batch-process a directory of input files.

Sand-Boxing

Not specific for this topic group but of utmost importance is that the platform is absolutely safe with regard to blocking any access of the AI on anything outside its sandbox. It must not be possible to access the filesystem of the benchmarking machine, databases, network etc. The AI must not be able to leak the test data to the outside world, nor see the correct labels, nor manipulate the recorded benchmarking results, nor access other AIs or their results. The experience with protecting all kinds of manipulation attempts is the biggest advantage that using a ready-made benchmarking platform could provide.

Online-Mode

Beside the sand-boxed mode for the actual official benchmarking it would simplify the implementation of the benchmarking wrapper if there would also be a way to submit a hosted version of the AI. This way the developers could test run the benchmarking on some public e.g. synthetic dataset get some preliminary results.

Due to the challenges and complexities inherent to this modality, we are developing our own minimum viable benchmarking.

5.1.3.2 AICrowd

The general preliminary assessment is that AICrowd has the potential to serve as a benchmarking platform software for the first iteration of the benchmarking in our topic group. However, benchmarking and reporting is designed for one primary and one secondary score. Adding the high-dimensional scoring systems with reporting organized by a multitude of additional dimensions is not yet supported and needs to be implemented. This also applies to the automatic stability and

robustness testing. The interactive dialog simulation needed for future benchmarking implementations would need to be implemented from scratch. In general, we found that the documentation for installing the software, the development process and for extending it is not as detailed and up to date as needed and the necessary changes would probably require close cooperation with the developers of the platform.

The topic group will discuss if the strong experience in water-tight sandboxing and the design of the platform itself outweighs the work of changing an existing platform to the topic group's needs, compared to implementing a new specialized solution. The final decision was to develop a custom benchmarking tool through a range of iterations.

Features requiring special attention in TG Symptom

There are several features unique to TG Symptom that require attention and discussion.

5.1.4 Scores & Metrics

At the core of every AI benchmarking there are scores and metrics that assess the output of the different systems. In context of the topic group the scores have to be chosen in a way that can facilitate decision making when it comes to deciding on possible solutions for a given health task in a given context.

5.1.4.1 Outline of clinical considerations for scores and metrics

The aim of this section is to outline the perspectives that must be considered for the development of clinically relevant metrics for benchmarking. It is crucial for clinical stakeholders that the outputs of benchmarking are able to answer questions that are relevant to clinical outcomes for decision makers.

A benchmarking framework should aim to assess some of these outcomes, and this section will aim to identify:

- which aspects are important to evaluate AI systems in a health setting
- which of these are feasible to capture in benchmarking
- which may not, and would be required to be captured in some other way (e.g. clinical studies)

Examples of stakeholders that might require clinically relevant metrics:

- Health system and governmental decision makers (including procurement, tendering, commissioning)
- Healthcare payers and providers
- Clinicians interacting with or using symptom assessment tools (either via patient facing self-assessment tools or clinician facing decision support tools)
- Regulators and notified bodies

It must be noted that clinical evaluation of AI tools in healthcare is already performed through clinical trials. The special, nuanced considerations over and above clinical studies for medical hardware, medications or surgical interventions have been carefully addressed in the CONSORT-AI reporting guidelines released in September 2020. In addition, as part of Software as a Medical Device Regulations (e.g. the FDA, EU MDR), it is a requirement that companies building these

tools carry out Post Market Clinical Follow Up (PMCF) in order to demonstrate claimed clinical benefits really do occur in the real world.

Bearing this in mind, it is pertinent to discuss where the ITU/WHO benchmarking fits within the overall terrain of clinical evaluation. This is currently under detailed discussion among various clinical and academic experts in the ITU/WHO Clinical Evaluation Working Group. Further details about the consensus and outcomes from this work will be updated here in this document.

Oversight

Whilst the benchmarking framework and clinical metrics are created as a result of collaborative efforts between various companies, there is also crucial involvement from a diverse range of independent stakeholders across the world (including practising clinicians, academics, technology experts and ethics experts).

This process has oversight and input from the WHO and ITU via an independent Clinical Evaluation Working Group and Regulatory Working Group.

5.1.4.2 What do we mean by clinical metrics?

These refer to measurements relevant to the stakeholders outlined above and could be split into:

- Performance and Accuracy measures
- Safety measures
- Clinical outcome measures

This list is not exhaustive, and later other possible measures specific to AI systems in this modality will be discussed. All of these should be addressed in the clinical evaluation of an AI system deployed into a healthcare system.

The detail of exactly which metrics to be considered will be discussed further in this section, but it is important to outline the following key determinants:

- Intended use of the device
- Intended users
- Risk classification
- Point of Information and comparison to Standard of Care
- Intended/stated benefits to the user, clinical workflow and health system

For the purposes of benchmarking AI powered symptom assessment tools, it is important to consider clinical metrics within the context of the above categories, and subsequently discuss which of these are then possible to actually measure through an independently curated, representative and high quality test data set.

In modalities such as image classification, the inputs and outputs are quite clear compared to symptom assessment. The metrics to be considered for these tasks will be very different to those for symptom assessment tools. The next section outlines some of the key differences and special considerations for this modality.

5.1.4.3 More complex than image classification

Image classification of, say, a histological slide for identifying potentially cancerous cells will have a clear set of inputs and outputs. Inputs will be image pixel data, and the outputs (cancerous vs non-cancerous, for instance) can be benchmarked compared to gold standards.

In comparison, symptom assessment tools will have a large range of information to convert to inputs. Examples might be:

- Symptoms - even the way these are captured might vary. Some may use structured text, others free text, others may also additionally have other augmenting ways to capture symptoms.
- Attributes (such as onset, character, location, intensity)
- Risk factors
- Medication
- Demographic information
- Location, region
- Seasonality (e.g. hayfever in summer, winter for certain respiratory viruses, malaria in rainy season)
- Increasingly other data points can be included as part of AI powered assessment tools (examples include wearables data,
- Also, Clinical Decision Support Tools (covered later on in the process), could additionally take in clinical signs, bedside tests, lab tests and imaging data.

Additionally, for an image classification task, the AI tool is provided with the image data it needs to perform the task on. In other words, it is given all the relevant information it needs to get to its output. In contrast, consider an AI based symptom assessment tool. The performance of the task will be greatly affected by another task - how effectively it collects and comprehends the information. As an example, it may be important to collect the critical information that somebody is pregnant. If the tool has not elicited this information, it may not consider this when providing condition suggestions. Clearly a balance must be reached - there could also be situations where too many questions are asked, resulting in a user not completing their assessment.

5.1.4.4 Important Considerations

Before discussing metrics, there are some key nuances to consider:

The Ground truth problem

In order to derive metrics around performance, ground truth, or gold standards must be established.

There are different approaches to this, but examples, as well as their problems are outlined in the table below:

Table 7 - Ground truth approaches with their problems

Gold Standard Case Vignettes	Creation of vignette cases by clinicians. In these, the gold standard conditions and triage level can be defined by the author	These might be based on clinician experience or real cases they have seen, and as such are subject to the clinicians own heuristics and biases.
-------------------------------------	--	---

	<p>Examples in literature include Semigran et al paper</p>	<p>Require quality and peer review. Consensus on gold standards and disagreement/variability in opinion between clinicians is common.</p> <p>Clinician opinion is of lower certainty compared to a definitive imaging, lab or histopathology report (we have learned from our colleagues in these topic groups that even for histopath/imaging reports, there can be issues with gold standards owing to intra and inter-observer variability)</p>
<p>EHR records (retrospective)</p>	<p>Converting anonymised/pseudonymised electronic health record episodes into cases, with the coded condition as Gold standard.</p>	<p>There are many issues with using EHRs as ‘gold standard’.</p> <p>Examples:</p> <ul style="list-style-type: none"> • Coding issues (usually optimised for billing) • • Depend on whether the clinician actually documented info (for example important negative information) • • May not the same point of information (outlined below) • • The final diagnosis may occur later down the line (after the evolution of symptoms and time, as well as added data points e.g. labs imaging) • • Geographic variability <p>Lack of standard structure</p>

Defining metrics in the context of point of information

Scope of data

To create a safe and effective symptom checker, the datasets used in each symptom checker need to have defined and standardised limits with respect to information that can be asked of the user, and to conditions that are provided to the user as possible outcomes. Symptoms are the most logical

inclusion in a symptom checker, as the name would suggest; as features that can be described by the user in a typical medical history-taking exercise, they reasonably easily translate to a symptom-checking environment (notwithstanding the nuances of symptom descriptions). But while there has historically been great emphasis on the power of a well-taken medical history in determining diagnoses [rx26], other elements of a medical assessment add critical data that enhance the diagnostic process. To this end, the criteria for inclusion becomes more complex.

Elements of the past medical history and social history that are sought depend on the scope of the symptom checker. It is once again logical to ask for information that can affect the probability of possible outcomes, such as a history of smoking or co-morbid hypertension in a middle-aged male user using the symptom checker to assess his chest pain. Beyond this, however, there is significant uncertainty in the boundaries of information that can or should be asked. Should all users be asked, for example, if they live at home alone, or if they live with somebody who can drive? This can influence whether a user is advised to call an ambulance or to go to the emergency department if they are having sudden-onset visual impairment. Similarly, should users be asked about the number of people that live in their household, or more information about their socioeconomic status? This can influence whether or not an outcome of scabies is given to a user with itchy hands. Furthermore, specific information may be necessary depending on the features the symptom checker offers. For example, if a symptom checker offers social care services alongside their outcomes and/or triages, a more thorough social history would be relevant.

The place of physical examination in a symptom checker is also ambiguous. Should signs or physical examination findings be asked, and if so, which ones? Ultimately, these elements of the medical assessment need to be reasonable with respect to what the user can find or self-evaluate with an untrained eye. For example, asking the user if they have pain on their ribs only with pressing the area in order to elicit the sign of rib tenderness is reasonable; asking them to perform the maneuver required to elicit Murphy's sign, however, is not. With the increasing use of wearables, it is also important to consider whether or what elements of self-monitoring should be included, such as temperature, blood pressure measurements, or blood glucose readings. In spite of the limitations of the wearables themselves, the use of this data could make outcomes more accurate, if available.

The question of which conditions are to be included in a symptom checker is somewhat more straightforward. The suite of conditions needs to be limited to conditions that could reasonably be diagnosed or suspected with the information that can be provided by a user through a symptom-checking tool. This entails the exclusion of those conditions which require clinician face-to-face contact, laboratory testing, or clinical procedure to be diagnosed or suspected. While no condition can be diagnosed with 100% probability without a full and comprehensive medical assessment (and sometimes, not even then), many self-care conditions (e.g. viral colds, sinusitis, constipation) can, with the appropriate inclusion and exclusion of purely symptoms, be comfortably 'diagnosed' through a symptom checker - that is to say, suspected with a very high probability. With more complex conditions, particularly conditions requiring emergency treatment, uncommon or rare conditions, or conditions requiring histological diagnosis or specialist management, the specification of 'suspicion' in a symptom checker becomes an important one. The inclusion of such conditions, ones that can be suspected but not diagnosed, is necessary for engendering trust in a symptom checker. An application which considers only self-care conditions or conditions requiring routine review is not particularly helpful. A symptom checker needs to have the ability to consider or rule out emergency or urgent conditions. Appendicitis, for example, can only be diagnosed with certainty through surgical exploration. However, every general practitioner would be expected to consider or suspect this diagnosis in a user with acute right lower quadrant abdominal pain. Alongside trust, the value of including such conditions is in directing users to appropriate actions and focusing clinical contact time. If appendicitis is considered in a consultation, depending on the likelihood of this condition based on the user's symptoms, this suspicion may come with a recommendation to attend the emergency department in the first instance to assess, via face-to-face

contact with a clinician, whether surgical exploration is necessary, regardless of whether or not dysmenorrhoea is also a likely diagnosis.

Defining the scope

When considering which information is considered appropriate for a symptom checker, it is also important to consider who defines which information is considered appropriate. As there is no established gold standard, the most appropriate method of defining the boundaries of these metrics is to use a panel of clinicians to reach consensus. This is a method also used in clinical medicine for identifying diagnostic reference standards in the absence of gold standard tests [rx27]. The use of ‘expert panels’ for diagnosis is not without issues, such as intra- and inter-observer variability. It is therefore important to define a clear methodology for how consensus will be reached to minimise this variability and increase reproducibility as much as possible. Furthermore, the make-up of members on the panel is dependent on the focus of the symptom checker, so it needs to be considered whether the panel is limited to general practitioners, or a mix of general practitioners and specialists, or whether there are multiple panels appropriate for different conditions or presentations. Numbers of panel members also need to be decided (an odd number is most practical), as does the criteria for inclusion on a panel (e.g. area of expertise, number of years of experience). There is also the consideration of whether a panel of patients is necessary, particularly for defining the validity of user symptoms and the finer details relating to language and presentation of information.

Getting the comparison right

A useful source of data or point of comparison for this process of defining appropriate clinical metrics may be the use of telehealth consultations. As medical assessments which are undertaken with the barrier of a phone or screen, they, like symptom checker consultations, lack access to physical examination and investigations. They may therefore offer the most like-to-like comparison for defining clinical metrics in a symptom checker, though they are still limited by the variability of individual clinician choices of what to ask and record in a consultation. Audio or telephone rather than video consultations are likely the most useful of telehealth consultations, as they do not have the added visual information that video consultations can provide. However, there are still issues with telehealth consults which affect the reliability of their data. The ability of the clinician to hear the voice of the patient in the consultation provides some degree of clinical information not available in a symptom checker. A telehealth consultation is also more dynamic than a machine-powered consultation, allowing the user to spontaneously change their history, clarify mutual understanding or the clinician to switch to video as required. The datasets gleaned from telehealth consultations are also affected by specific patient populations that utilise telehealth more frequently (e.g. rural/remote populations), or contexts in which telehealth is used (e.g. the 2020 coronavirus pandemic).

Another consideration for a source of datasets or a point of comparison is electronic health record (EHR) data. The use of these datasets, however, is rife with issues. Information recorded in EHRs is incomplete and variable. It is also limited by what the clinician or writer has chosen to gather and document, or believes is relevant, and may not include all questions asked of, and all information given by, a patient in order to develop a complete picture of the possible diagnoses. The language used in EHRs is often heavily medical in nature, acting as ‘translations’ of patient histories, and are also reflective of the writer’s training, ethnicity, coding requirements, and/or practices of the institution in which they work. In addition, the “true condition” taken as gold standard might have

been arrived at through a combination of taking a history, clinical exam, bedside tests and potentially lots of other tests, so as discussed, it does not serve as a perfect like for like comparison to a patient facing symptom assessment tool.

Performance does not equal Utility

Whilst it is important to consider the performance of AI based symptom checkers as part of evaluation, another key aspect in clinical evaluation is that of utility or impact. An AI tool may claim to have a 99% sensitivity, but clinical stakeholders are also considered with more. In particular, what the impact is in a clinical setting. These clinical outcomes might be considered at patient level, clinician level (how does it impact clinical workflow) or health system level.

Different metrics for different use cases/contexts

Symptom assessment tools are not one homogenous group of tools. There are numerous variations on intended use, intended users, locations, populations, etc. Some might be specific for a certain region or population, others more general. They may also perform varying tasks (e.g. taking structured inputs vs free text), and be geared optimally towards their particular intended use case.

With this in mind, there need to be context relevant metrics. Current discussions in the topic group centre around developing the ability to drill down into different contexts and use cases. For example, a stakeholder might be a health ministry in a certain region using the ITU/WHO benchmarking metrics to assess potential partners. They may be more interested in viewing specific metrics (or metrics from a specific sample of the independent data set) that are relevant to the setting in which they want to deploy an AI symptom assessment tool.

Some examples of contextual variations include:

- Geography/region/seasonality (important to also note that there are large demographic variances within countries, and within cities)
- Population demographics – e.g. age groups, subpopulations, biological sex, language
- Health literacy
- Digital literacy
- Focus (via intended use) on specific medical specialties (e.g. Pediatrics or Musculoskeletal medicine)

5.1.5 Metrics for Symptom Assessment

As indicated previously, these can be looked at with respect to:

- Performance and Safety measures (How accurate is this device? How safe is this device?)
- Clinical outcome and income measures (How does it impact clinical practice - for patients, clinician workflow, or health systems)

Currently, the key outputs of patient-facing symptom assessment are:

- **Condition suggestions.** Some tools provide an ordered list of possible conditions (with varying nomenclature, the most common being differential diagnosis), others have an

unordered list of possibilities. Another variable between tools¹ is the indication of probabilities for each suggestion. Sometimes informally any of this is called a “[differential] diagnosis” but it is only an informal name due to the fact that symptom assessment apps generally don’t have enough accuracy (yet); also due to liability considerations and due to important safety & regulatory reasons.

- **Pre-clinical triage.** This gives advice about what level of care the user should seek. There is a large variance between tools in the exact levels, which can range from ‘self-care’ to “call ambulance”. The triage advice might be given as direct advice or just as “information” (the choice depends on the accuracy of a tool and the estimation of that accuracy by an app provider; by the amount of the liability that an app provider would like to take; on regulations requirements in a particular country; etc.).

There are other tasks that require their own metrics:

- Quality of information gathering (how well does the tool collect the required information)
- Safety of information gathering (does the tool consider relevant serious symptoms, and ask them)
- Tools that assimilate free text also require
- (To be completed)

5.1.6 Performance and Accuracy

5.1.6.1 Condition Suggestions

When measuring performance, traditional metrics in healthcare are focused around diagnostic accuracy. “How did the test predict the diagnosis of a condition compared to the Gold standard?”

At present, published clinical studies of AI based symptom checkers focus on assessing this by comparing the following to the Gold standard.

- Top matching condition (i.e. did the top condition suggested by the AI tool match to the gold standard?)
- Top 3 or 5 matching condition (i.e. was the condition defined in the gold standard present in the top 3 or top 5 of the list of suggestions of the AI tool?)

Whilst these measures are a good starting point, some important nuance must be considered.

- The top match condition metric assumes that it is always possible to get to the “final diagnosis” indicated in the gold standard from the inputs available to a symptom assessment tool. In other words, it assumes that taking the information that is available from a medical history will be enough to accurately predict what the patient has in future. This has been discussed further in the Point of Information section. Much depends on the stated intended use of the tool, but in general, at the time of writing, it is not a goal (and it simply can’t be a goal) to give a diagnosis. Aside from the most clear-cut cases, the process of diagnosis requires so much more information: observing the evolution over time, clinical examination, bedside tests, lab and imaging tests. (In future, we will be able to come back to this when benchmarking clinician facing decision support tools)

¹ Through this section, the terms tool and app are used interchangeably, and refer to patient facing symptom assessment tools.

- The top 3 or top 5 condition matching metric is useful as it starts looking at the list of conditions suggested by the AI tool. However, there is no measure of how good the other suggestions on the list are.

This leads us to discuss the merits of metrics that aim to measure the quality of the entire list of differentials provided by an AI tool:

- (e.g. 1-2 conditions that a patient has) but rather a differential diagnosis, since even the most astute clinicians can't always give a certain, precise diagnosis for a patient without further tests (e.g. X-rays, blood tests, etc.) in many cases. This means that when we measure accuracy of the "diagnostic" capabilities of symptom assessment apps, we always need to keep in mind that we are always evaluating their results for a particular patient case against a "ground truth" (i.e. the real, objective one) of differential diagnosis distribution of possible conditions. For example, for a relatively young patient who has a heart attack their differential diagnosis might look as follows: 15% heart attack, 75% panic attack, and around 10% for other conditions. Note that when collecting this information from clinicians, we most likely need to approximate those probabilities or even just consider orders of likelihoods without assigning numerical estimates.

For any metric evaluation, it is also important to define what exactly is being evaluated. For example, what exactly are we trying to assess: Is it a new symptomatic condition(s)? A new symptomatic or asymptomatic condition(s)? Should it include flares of existing conditions? Should it include acute presentations of chronic conditions? Etc.

5.1.6.2 The presence of more than one condition (multi-morbidity)

When we mentioned above a distribution over the differential diagnosis for a particular patient, we often assume that a patient has generally only one of those conditions (i.e. a condition "of interest" or the "ground truth" condition). However, while less likely it is often possible that a patient has multiple conditions "of interest" (or "ground truth" conditions) at the same time (e.g. two conditions which are both new), and this affects the shape of the "ground truth" differential diagnosis distribution for the patient (e.g., for a patient who has a whiplash and a dislocation of shoulder after a traumatic car accident, the differential diagnosis distribution might look like this: 85% whiplash; 90% dislocation of shoulder; and some other conditions with other probabilities that are similar to whiplash or/and dislocation of shoulder).

5.1.6.3 Similar presentations, varying outcomes

It is possible that very similar constellations of symptoms have variable outcomes. As a very simplified illustrative example, if there is a cohort of 100 people (female) aged 25-30 who present to a GP in a very similar way, say, right lower abdominal pain, fever and mild dysuria for 3 days, and followed them up 1 month later there would be a natural variation in what they ended up having. X% might have a urinary tract infection, Y% might have appendicitis, Z% might have pyelonephritis, an even smaller proportion might have an ectopic pregnancy. This happens despite the fact that the "inputs" captured at that first point of information were very similar. If a symptom assessment tool is providing condition outputs accompanied by probabilities for their differential diagnosis lists/suggestions, it may be an important consideration to include metrics that assess whether these distributions match real world/gold standards.

Some conditions have pathognomonic symptoms or signs - i.e. very strong indicators of a particular condition. But in reality, for most conditions it is not so simple - there is much more uncertainty. One of the main roles of those in primary or emergency care is to manage this uncertainty.

Also, note that a symptom assessment tool might predict a chance/probability for many different conditions (often for hundreds of them), and that is quite different from the settings of “diagnostic tests” which often just need to determine whether a patient has or does not have one or few particular conditions.

5.1.6.4 Basic Metrics for Differential Diagnosis

There may be several ways to calculate these metrics for symptom assessment tools. Other approaches and descriptions are welcomed.

Let’s say a patient has a specific presentation (including: symptoms, if any; their medical history; etc.) and let’s say that there is the “ground truth” that he/she has e.g. new presentations of conditions $\{X_1, \dots, X_n\}$ (where n is likely to be equal to or less than 1), and does not have any other conditions.

Let’s say that a symptom assessment tool, after collecting all information it could collect from a patient, has identified that a patient might have some conditions $\{Z_1, \dots, Z_m\}$ (for simplicity, let’s assume that the app just says whether each condition is present or not; generally, it might return some likelihood/probability of it, or some other degree of certainty).

Any metric we calculate might be influenced by the outcome types of the tool. As mentioned, some assessment tools return ordered lists of conditions with probabilities; and others might just return ordered lists or even unordered lists. This means that only the top matching condition (top-N) metric might be calculated only for some of the tools.

The following metrics can be calculated per patient case (and then aggregated later) for a specific symptom assessment tool/app:

Table 8 – Overview patient case metrics

<p>Recall (also called: true positive rate, sensitivity)</p>	<p>It is the ratio of conditions “of interest” that the patient has and which were identified by the app (i.e. presumed by the app to be happening to the patient) to the number of the conditions that the patient has.</p> <p>Note that in the case of a “simple” “one-condition diagnostic test” for a specific condition, the recall is much “simpler” and is usually calculated in an aggregated way across all patient cases: it is the ratio of sick people (i.e. people with that specific condition) correctly identified as sick (i.e. presumed to have that specific condition) to the number of sick people (i.e. having that specific condition).</p>
<p>Precision (also called: positive predictive value)</p>	<p>It is the ratio of conditions that the patient “of interest” has and which were identified by the app to the number of the conditions that the app has identified.</p>
<p>F1-score and Fn-score</p>	<p>It is the harmonic mean of precision and recall. In Fn-score, recall is considered n times as important as precision ($n > 0$).</p>
<p>Specificity (also called: selectivity; true negative rate)</p>	<p>It is the ratio of the conditions “of interest” that the patient does not have and which were not flagged (i.e. were not highlighted as present/”likely”) by the app to the conditions which the patient does not have.</p> <p>Note:</p> <ul style="list-style-type: none"> False positive rate (also called: fall-out or false alarm ratio) can be calculated as follows: $1.0 - \text{specificity}$.

	<ul style="list-style-type: none"> • Since there are quite a lot of conditions that a patient might have in general, and since symptom assessment apps generally can rule out most of conditions that a patient does not have, specificity might often be close to 1.0. <p>Because of that, a receiver operating characteristic curve (ROC curve) (calculated over many cases, not just for one case) that is created by plotting the recall (sensitivity) against the false positive rate (i.e. 1.0 - specificity) might look almost “trivial” in many cases since the false positive rate might often be close to 0.0.</p>
Accuracy	<p>Different things might be meant by “accuracy” and there are multiple ways to define “accuracy”. Informally, different metrics can be called “accuracy”. Because of this, it is recommended to always clarify what is meant by “accuracy” if this term is used.</p>
Top-N (Top matching condition)	<p>One of the simple ways to calculate it: for each case top-N is equal to 1.0 if an app’s output (i.e. a “differential diagnosis”) contains the condition of “interest” in its top N conditions in the output (assuming that the app returns ordered conditions). If there are M conditions of “interest” for the case, and K of them are in the top-N conditions from an app’s output, then top-N for the case for the app is equal to K/M. Note that an app might return less than N conditions in its output for a case: in this scenario, if an app returns $J < N$ conditions, it might be assumed that top-N is equal to top-J for this particular case for this app (note that J might be equal to 0).</p> <p>Note that if an app provides an unordered list of conditions, then it is not possible/trivial to calculate Top-N.</p>

There are at least 3 approaches to elicit conditions of “interest” for a case:

- 1) For each case condition(s) of “interest” is/are provided by a creator of the case, or by an independent clinician (or a panel), or it comes from an EHR where we are certain about the diagnosis (i.e. exact condition(s)) of a patient.
- 2) Another option is to ask a clinician or a panel of clinics to provide a “differential diagnosis” (with multiple conditions) in some form, to which apps’ outcomes will be compared.
- 3) Ultimately, we might want to naturally obtain a distribution over a differential diagnosis for very similar patients with very similar symptoms and risk factors. This is exactly the sought differential diagnosis distribution, to which we can compare apps’ outcomes. However, this requires a lot of data and to the best of our knowledge many existing EHR systems might not be suitable for this due to many factors (due to their size/record format/“accuracy”/etc.). Also, it might be very hard to scale this approach internationally for different regions (e.g. due to the variability in how EHR systems are implemented and used).

(Note that in the 2nd and 3rd approach we receive a distribution over condition(s) of “interest”.)

It should also be noted that there are always special cases: some patients might be healthy or they might have a condition that is not known by an app, and so for the purpose of the metric calculation there might be zero conditions of “interest” that a patient has. Such situations might cause recall to

be ill-defined. Also, an app might return an outcome with no conditions at all (i.e. it might assume that a patient does not have any conditions at all, at least of those that it is aware of), and in this case the precision might be ill-defined. Special rules must be applied for such cases as appropriate.

The metrics mentioned above might be calculated per each patient case. They can then be aggregated, e.g. by taking an average. A weighted average can be used (e.g. by weights associated with the severity of conditions, by epidemiological rates associated with the condition or by some other weights to balance patient cases and/or avoid bias; etc.). Also, note that another alternative is to treat all patients and all conditions as separate (but correlated) random variables such that e.g. each pair of a patient-condition (e.g. a patient X has a condition Y) is a separate variable (e.g. a Boolean one), and then the metrics might be calculated for all of them at the same time in one batch. One more alternative is to treat each patient-condition pair as a separate variable but instead of treating them as one batch, consider patients for each disease independently (in some sense, this is equivalent to treating a symptom assessment app as a set of independent one-disease “diagnostic tests”). There are other alternatives as well.

A curve of recall and precision could then be plotted over multiple test cases. (The area under such a curve might also be calculated.) There are multiple ways to calculate such a curve, similarly to how there are multiple ways (as discussed above) to calculate e.g. aggregate metrics for recall and precision.

Metrics such as precision and recall measure presence in the differential diagnosis list, but they cannot tell if differentials are returned with appropriate likelihood nor if they are returned in the right order (for instance with decreasing level of confidence).

To address this, Normalised Discounted Cumulative Gain (NDCG) is an example of a metric that gives a measure of ranking quality. Each item (that is a disease) has an assigned relevance. In the setting of a differential diagnosis list, relevance should be proportional to confidence in disease - more probable diseases should have higher confidence.

If items with high relevance are in the top of ranking the score will be high.

In the prior mentioned example of a young man with chest pain:

- 15% heart attack - relevance “medium”
- 75% panic attack - relevance “high”

(and around 10% for other conditions - we might give some conditions label “low”)

Note that this proposition measures only ability to return diseases from the most probable one to least probable. However, from a clinical perspective returning “heart attack” might be as relevant (and important) as “panic attack” because of its potential seriousness.

5.1.6.5 Basic metrics for triage

For triage, it might be of interest to measure whether the triage recommendation returned by an app is suitable for a particular patient case. There are two approaches to consider triage in this context. The first is to consider triage outcomes based on the seriousness of conditions that are present within the possible conditions.. The other approach is the presence of serious symptoms within an assessment. In reality clinicians will use a combination of both of these to settle on the most appropriate outcome. To benchmark this, it may be necessary to have an externally defined set of serious conditions and red flag symptoms.

One way to measure triage performance is to have one “perfect”, “ground truth” triage option for each patient case (provided by an expert clinician/panel of experts), and to match it with a tool’s

triage, such that there is a match or no match. This can be averaged (and weighted if appropriate similarly to the differential diagnosis as described above) across multiple patient cases. This way we capture the “accuracy” of triage.

However, this approach has limitations because:

- a. there might be multiple appropriate triages for the same patient (especially if different experts believe some similar but different triage options are appropriate), and
- b. some triages are safe but overcautious e.g. if a patient needs to see a primary care doctor but he/she is directed by an app to a hospital urgently, then the tool’s decision is safe but not accurate. This also has the potential to overburden health systems inappropriately.

Hence, the ground truth for a particular patient case might consist of a set (or a range if we assume (partial) ordering of triages) of appropriate triages rather than just one triage option. This set can also be separated into at least two subsets:

1. a subset of triage options that are safe and non-overcautious and
2. a subset of triage options that are safe but overcautious.

If an assessment tool returns a triage outcome from any of two subsets, it could be deemed a safe decision. However, if the tool returns a triage outcome that it is in the second subset, then it could be deemed a safe but overcautious triage. For example, for a condition for which it is okay (safe) to see a primary care doctor in a few weeks, it is most probably safe but overcautious to go to a hospital or even go there by an ambulance.

With this separation, triage outcome can be measured with e.g. two metrics: (a) how safe it is; (b) how safe and non-overcautious it is? The latter might be called “accuracy” of triage. There are obviously other variations over these metrics.

Note that if there is some full or partial ordering of triage outcomes, then the two subsets mentioned above might be simplified: e.g. in the case of a full ordering, two threshold triage outcomes might be needed: e.g. one to define the “minimum safe triage outcome” and the “minimum safe not overcautious triage outcome”.

A special case scenario for an assessment tool might be when it does not return any triage outcomes or any explicit recommendation. This is a special type of a triage outcome, and depending on the particular message that is returned by an app in such cases, it either should be treated as one of the default outcomes (e.g. if anyone is advised to see a doctor “quite urgently” in such cases by an app, such an outcome could be mapped to an urgent primary care appointment) or it should be treated as a separate category of triage outcomes for the benchmark purposes (and hence probably reported and analysed semi-independently which might involve additional complexity for report generation especially when they are aggregated for different apps).

It is a significant challenge to standardise and map triage outcomes (of different tools) to one particular set, especially internationally. This is because of local/contextual variation. Options for care in a remote village in one country are very different to those in an affluent neighbourhood in a large city in the same or other country, for instance.

It also needs to be considered that there are apps that return only information and do not provide any explicit triage outcomes. In addition overall triage metrics, e.g. triage safety, might be ultimately not that useful. For example, if a patient who needs to see a primary care doctor urgently is triaged by an app to see a primary care doctor non-urgently, this is not safe but it is probably still less “unsafe” than advising that patient to self-care (e.g. “stay at home and keep hydrated”, even without going to a pharmacist who has medical knowledge).

Hence, some additional stratification/analysis of triage outcomes is important. One more metric can be relative/absolute confusion matrices where e.g. one dimension contains “expected” triages and another dimension contains triages provided by an app. In addition to this, different triage combinations (e.g. pairs of an expected and provided triage) can have different weight.

5.1.6.6 About statistical significance of metrics

[To be added]

5.1.6.7 “Ground truth” for differential diagnosis/triage

[To be added:]

- How to estimate it?,
- Should it be perfect?
- How to average if we have multiple approximations of ground truth from different experts.
- How to use an expert panel’s opinion?

5.1.6.8 On balancing/biasing/weighting/stratification

[To be added]

5.1.6.9 Other performance measures

The parameters discussed so far relate to the performance of the outputs of symptom assessment tools.

Clinical stakeholders are also concerned with the performance of how data is collected and interpreted by the tool.

With this in mind, other aspects to include in overall performance metrics would be:

- Quality of question flow: How much of the relevant positive and negative evidence was actually elicited? Were there any key serious symptoms (red flags) not elicited by the tool?
- Do users actually understand/comprehend the questions? (this may be difficult to assess within benchmarking. It might be captured within usability testing)
- Measuring the task of converting the lived experience of the user, into the tool. Different tools use different methods for this - some use Natural Language Processing, for example. It is important to have metrics on the performance/accuracy of this task.
- Metrics to consider that there are certain conditions that may be explicitly ruled out due to the age or the biological sex of the user. An example of this would be pregnancy related conditions (e.g. pre-eclampsia) should usually not be included in the list of differentials for someone whose biological sex is male²)

Clinical Outcomes/Impact

Metrics measured here will relate to the stated clinical benefits of the AI tool. Mostly these are best measured with a study in a clinical setting.

² It is also important to consider this within the context of people who have congenital sex differentiation disorders, where this notion may not be so clear cut for some users.

Examples of clinical outcomes and impact measurements could be:

Patient Journey:

- Effect on patient journey - satisfaction with navigating health services after being given triage advice
- Increase/decrease in time spent waiting for appointments for conditions that can be helped with self-care or other healthcare pathway (e.g. pharmacy)
- Effect on waiting times for appointments
- PROMs (Patient Reported Outcome Measures)

Clinical Workflow

- Effect on consultation times between patient and HCP
- Effect on clinician caseload management

Health System

- Effect on demand on emergency services
- Effect on demand for primary care appointments
- Effects and impact of undertriage (e.g. advising people with a serious condition to self-care)
- Effects and impact of overtriage (e.g. inappropriately advising someone with a non-urgent condition to attend emergency department)

Going forward, it may be of interest to work with health economists in the topic group to explore the utility of health economic metrics within benchmarking that might be a function of cost, QALYs, etc. These may be important for health system level stakeholders (e.g. health ministries, providers, payers).

5.1.7 Putting it all together for Clinicians

A big challenge is communicating this range of metrics (relevant to context) to clinicians and other healthcare stakeholders, in a way that aids understanding. A paper published in Nature in March 2020 by Sendak et al explores this in great detail, and provides a fantastic example inspired by the nutrition information on a cereal box²⁶. Benchmarking metrics summarised in a such a digestible way that takes the most relevant parts is worth seriously considering.

Model Facts		Model name: Deep Sepsis	Locale: Duke University Hospital			
Approval Date: 09/22/2019		Last Update: 01/13/2020	Version: 1.0			
Summary						
This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.						
Mechanism						
<ul style="list-style-type: none"> ▪ Outcomesepsis within the next 4 hours, see outcome definition in "Other Information" ▪ Output0% - 100% probability of sepsis occurring in the next 4 hours ▪ Target populationall adult patients >18 y.o. presenting to DUH ED ▪ Time of predictionevery hour of a patient's encounter ▪ Input data source.....electronic health record (EHR) ▪ Input data typedemographics, analytes, vitals, medication administrations ▪ Training data location and time-periodDUH, diagnostic cohort, 10/2014 – 12/2015 ▪ Model type..... Recurrent Neural Network 						
Validation and performance						
	Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%	Cohort Type	Cohort URL / DOI
Local Retrospective	18.9%	0.88	0.14	0.50	Diagnostic	arxiv.org/abs/1708.05894
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
Local Prospective	TBD	TBD	TBD	TBD	TBD	TBD
External	TBD	TBD	TBD	TBD	TBD	TBD
Target Population	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
Uses and directions						
<ul style="list-style-type: none"> ▪ Benefits: Early identification and prompt treatment of sepsis can improve patient morbidity and mortality. ▪ Target population and use case: Every hour, data is pulled from the EHR to calculate risk of sepsis for every patient at the DUH ED. A rapid response team nurse reviews every high-risk patient with a physician in the ED to confirm whether or not to initiate treatment for sepsis. ▪ General use: This model is intended to be used to by clinicians to identify patients for further assessment for sepsis. The model is not a diagnostic for sepsis and is not meant to guide or drive clinical care. This model is intended to complement other pieces of patient information related to sepsis as well as a physical evaluation to determine the need for sepsis treatment. ▪ Appropriate decision support: The model identifies patient X as at a high risk of sepsis. A rapid response team nurse discusses the patient with the ED physician caring for the patient and they agree the patient does not require treatment for sepsis. ▪ Before using this model: Test the model retrospectively and prospectively on a diagnostic cohort that reflects the target population that the model will be used upon to confirm validity of the model within a local setting. ▪ Safety and efficacy evaluation: Analysis of data from clinical trial (NCT03655626) is underway. Preliminary data shows rapid response team, nurse-driven workflow was effective at improving sepsis treatment bundle compliance. 						
Warnings						
<ul style="list-style-type: none"> ▪ Risks: Even if used appropriately, clinicians using this model can misdiagnose sepsis. Delays in a sepsis diagnosis can lead to morbidity and mortality. Patients who are incorrectly treated for sepsis can be exposed to risks associated with unnecessary antibiotics and intravenous fluids. ▪ Inappropriate Settings: This model was not trained or evaluated on patients receiving care in the ICU. Do not use this model in the ICU setting without further evaluation. This model was trained to identify the first episode of sepsis during an inpatient encounter. Do not use this model after an initial sepsis episode without further evaluation. ▪ Clinical Rationale: The model is not interpretable and does not provide rationale for high risk scores. Clinical end users are expected to place model output in context with other clinical information to make final determination of diagnosis. ▪ Inappropriate decision support: This model may not be accurate outside of the target population, primarily adults in the non-ICU setting. This model is not a diagnostic and is not designed to guide clinical diagnosis and treatment for sepsis. ▪ Generalizability: This model was primarily evaluated within the local setting of Duke University Hospital. Do not use this model in an external setting without further evaluation. ▪ Discontinue use if: Clinical staff raise concerns about utility of the model for the indicated use case or large, systematic changes occur at the data level that necessitates re-training of the model. 						
Other information:						
<ul style="list-style-type: none"> ▪ Outcome Definition: https://doi.org/10.1101/648907 ▪ Related model: http://doi.org/10.1001/jama.2016.0288 ▪ Model development & validation: arxiv.org/abs/1708.05894 ▪ Model implementation: jmir.org/preprint/15182 ▪ Clinical trial: clinicaltrials.gov/ct2/show/NCT03655626 ▪ Clinical impact evaluation: TBD ▪ For inquiries and additional information: please email mark.sendak@duke.edu 						

Figure 11 – Example “Model Facts” label for sepsis machine learning model from Sendak et al, 2020. (Nature)

5.1.8 Additional clinical considerations and limitations

Whilst some important considerations have already been outlined, there are some additional discussions that relate indirectly to clinical metrics.

Mapping Ontologies

Another problem to address is one of ‘mapping’. This refers to variability in nomenclature and ontologies of symptoms and conditions. Each symptom assessment tool might have slightly different names for certain conditions. An example might be “heart attack”. Common synonyms could be “myocardial infarction”, or “acute coronary syndrome”. A gold standard case may have defined the true condition to be either one of those. In addition, there is not one standard ontology.

A robust, reliable and trusted approach is required to map these to a common, agreed ontology. This is discussed elsewhere, but is, from a clinical perspective, very important.

Explainability

Discussions about AI in healthcare naturally arrive at this aspect. It is important to clinicians and decision makers that, in many cases, that the reasons for outputs from AI tools are understood. There is concern that purely black box AI tools that give outputs that are ‘blindly followed’ in a clinical context could have detrimental effects. There is an example (still in preprint as of Sept 2020) of an AI system being trained to detect Covid-19 related changes on chest x-rays in emergency departments. Whilst initial results appeared positive (even on external test images) it was discovered that the system was using other artefacts within the image to determine Covid-19 presence.

Coming back to symptom assessment tools, explainability might relate to the presence of communicated messages to users to justify certain outputs. For example - if a triage outcome is ‘Seek emergency care’, there may be an explanation to the user as to what led to that outcome. In terms of metrics, this could be measured as a) the presence of a justification as well as b) its quality and accuracy but a concrete, widely accepted, quantifiable measure of explainability does not currently exist.

Addressing clinician variability and the ground truth problem going forward

Within the topic group, there has been fervent debate about the issues with EHRs and clinician variability in defining gold standards or ground truth. Another approach that has been discussed as an enhancement to the current benchmarking framework is as follows.

Benchmarking could involve each AI tool being paired with clinical sites across the globe that are relevant to its intended use and users. Connecting anonymised data of the patients that come through over a period of time, the AI systems are given the symptom information of the patients using the service. The outputs are then compared to the outputs at the ‘end of the episode’ – i.e. after the diagnostic process is completed. The “final” (for that episode) triage and/or differential diagnosis in the real world is then compared to the AI tool’s outcome. This has the advantages of being ‘real world’ test data, and reduces the problems of clinician variability. However, examples of

some challenges to overcome are: data security, reaching a critical mass of participating sites, standardisation of processes to achieve fair, comparable benchmarking tests across all sites.

5.1.9 Conclusion

This section has highlighted the myriad complexities, challenges and considerations that need to be addressed and applied for the successful adoption and implementation of symptom assessment tools. Benchmarking can capture and answer some of the questions relevant for clinical stakeholders, but it is important to acknowledge the limitations as discussed. What this should lead to is a pragmatic approach that brings alignment about what clinically questions can be answered for stakeholders within benchmarking, and which questions may need to be answered in other ways (e.g. robust prospective clinical studies).

- 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?

5.2 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.)

6 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](#) “Data specification” (introduction to deliverables 5.1-5.6), [DEL05_1](#) “Data requirements” (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), [DEL05_2](#) “Data acquisition”, [DEL05_3](#) “Data annotation specification”, [DEL05_4](#) “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](#) “Data handling” (which outlines how data will be handled once they are accepted), [DEL05_6](#) “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](#) “AI training best practices specification” (which reviews best practices for proper AI model training and guidelines for model reporting), [DEL07](#) “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](#) “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](#) “AI technical test specification” (which specifies how an AI can and should be tested *in silico*), [DEL07_3](#) “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](#) “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](#) “FG-AI4H assessment platform” (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), [DEL09](#) “AI for health applications and platforms” (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), [DEL09_1](#) “Mobile based AI applications,” and [DEL09_2](#) “Cloud-

based AI applications” (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 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 9 – Benchmarking iterations

Short Name	Name	Focus/Goals
MMVB 1.0	Minimal Minimal Viable Benchmarking	<ul style="list-style-type: none"> ● 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	<ul style="list-style-type: none"> ● 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	<ul style="list-style-type: none"> ● 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	<ul style="list-style-type: none"> ● 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	<i>Minimal Minimal Viable Benchmarking Version 3.3</i>	<ul style="list-style-type: none"> ● a first benchmarking using a SNOMED CT-based ontology for encoding case symptoms and their attributes ● Target: Q4 of 2021
MVB	<i>Minimal Viable Benchmarking</i>	<ul style="list-style-type: none"> ● <i>first benchmarking with real AI and real data</i> ● Target: end of 2021
Vx.0	<i>TG Symptom Benchmarking Vx.0</i>	<ul style="list-style-type: none"> ● <i>the regular e.g. quarterly benchmarking for this topic group</i> ● <i>continuous integration of new features</i>

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.

6.1.1 Benchmarking version MMVB 1.0

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

6.1.1.1 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 **Figure 12 – "London Model" used for sampling cases for MMVB 1.0** 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.

	IBD (first presentation non flare)	GERD	simple UTI	viral GE	bladder cancer (first presentation)	acute cholecystitis	appendicitis	ectopic pregnancy	IBS	acute pyelonephritis	Abdo Pain NOS (Idiopathic)
<i>How Common Condition?</i>	x	xx	xx	xx	x	x	x	x, only females	xxx	x	xx
Abdo Pain Cramping Central 2 days	xx			xx		x	x	x	xx	x	x
sharp lower quadrant pain	x		x				xx	xx		xx	
Diarrhoea	xx			xx			x		xx		
Vomiting	x	x	x	xx		xx	x	xx	x	xx	
Dysuria			xx							xx	
Increased Urination Freq.			xx		x		x			xx	
Haematuria			x		xx		x	x		xx	
Weight Loss	x				xx						
Fever	x		x	x		xx	xx	x		xx	
heartburn		xx									
Factor: missed period								x			
Expected triage level	PC	PC	PC	SC/PC	PC	EC	EC	EC	PC	EC	PC

Figure 12 – "London Model" used for sampling cases for MMVB 1.0

6.1.1.2 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).

6.1.1.2.1 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 trival toy-AIs directly implemented in the benchmarking backend, as well the actual toy-AIs hosted by the benchmarking participants in their own datacenters. 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 13**. 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 datacenter using a technology of their choice.

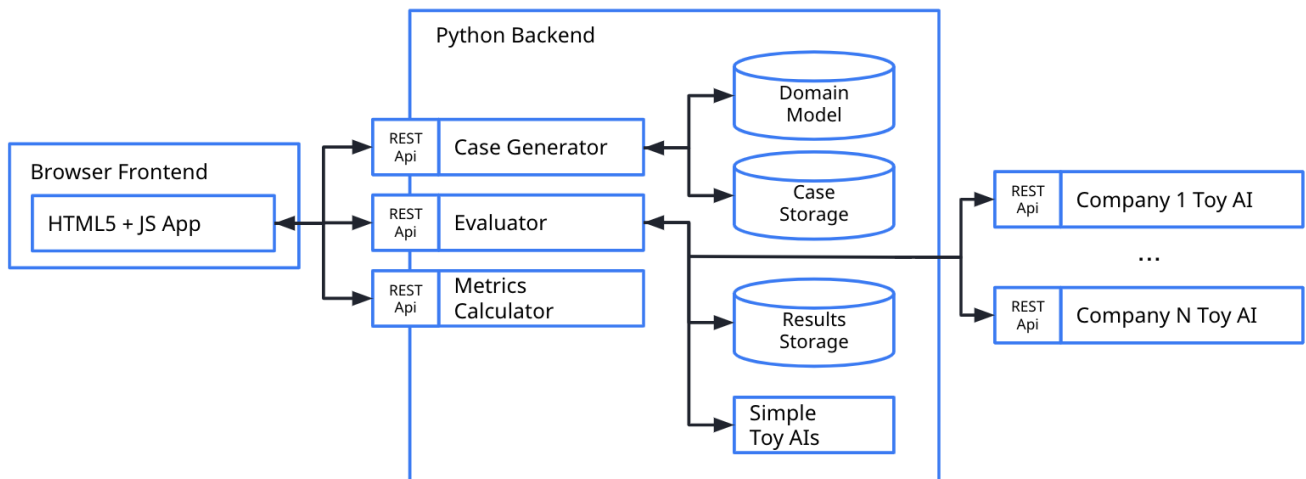


Figure 13 – MMVB 1.0 High-level architecture

6.1.1.2.2 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

6.1.1.2.3 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.

6.1.1.2.4 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 14** . 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.

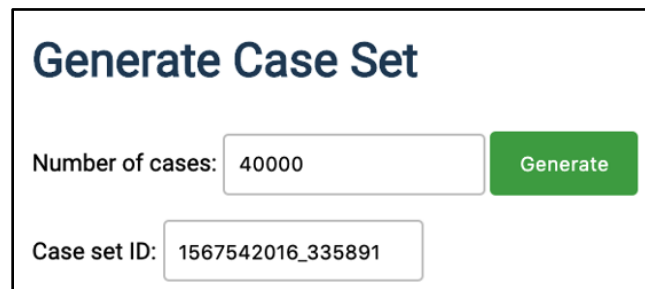


Figure 14 – MMVB 1.0 case generation UI

The user was then able to trigger the execution of the benchmarking in the screen shown **Figure 15** . 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.

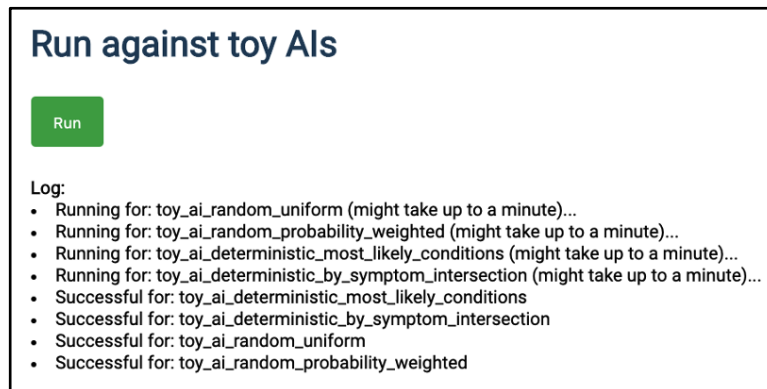


Figure 15 – MMVB 1.0 screen for running a benchmarking session

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 16** . As result the report table show in this figure was generated.

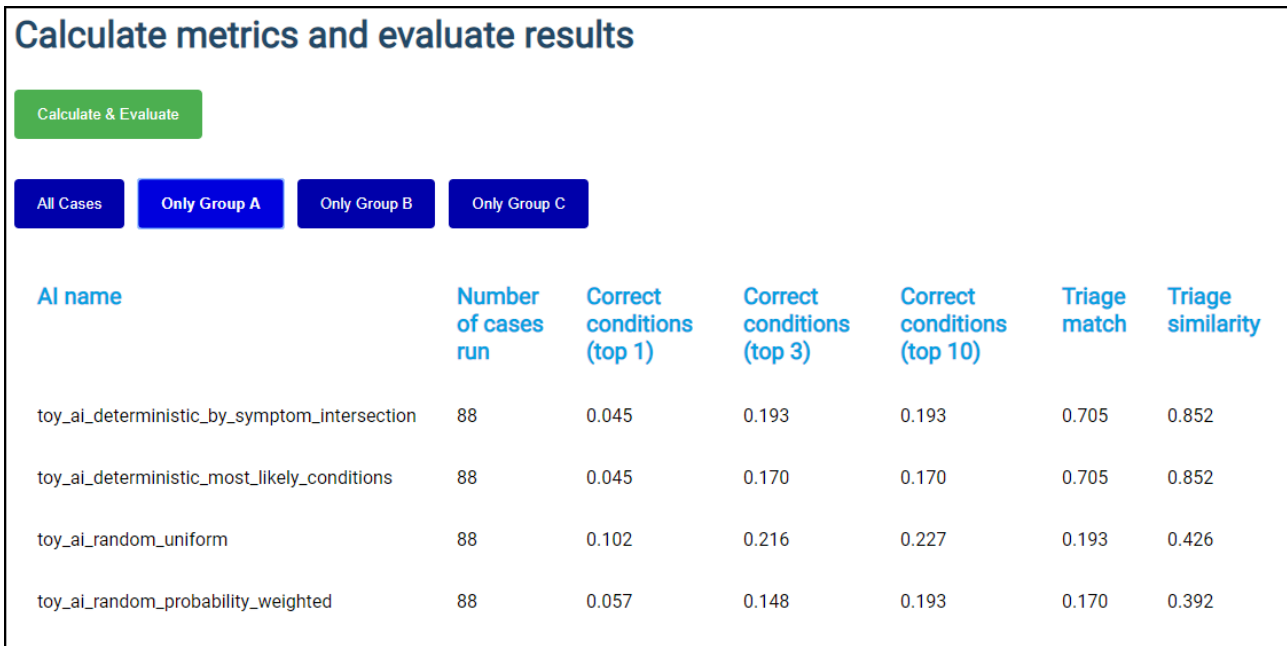


Figure 16 – MMVB 1.0 result screen

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.

6.1.1.3 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 10** shows the concrete fields with corresponding examples.

Table 10 – MMVB 1.0 input data format

Field name	Example	Description
profileInformation	<pre>"profileInformation": { "age": 38, "biologicalSex": "male" }</pre>	<ul style="list-style-type: none"> • General information about the patient • Age is unrestricted, however <i>for the case creation it was</i>

	<pre> }</pre>	<p>agreed to focus on 18-99 years.</p> <ul style="list-style-type: none"> ● As sex we started with the biological sex "male" and "female" only
presentingComplaints	<pre> "presentingComplaints": [{ "id": "c643bff833aaa9a47e3421a", "name": "Vomiting", "state": "present" }]</pre>	<ul style="list-style-type: none"> ● The complaints the user seeks and explanation/advice for ● Always present ● A list, but for the MMVB always with exactly one entry
otherFeatures	<pre> "otherFeatures": [{ "id": "e5bcdaa4cf15318b6f021da", "name": "Increased Urination Freq.", "state": "absent" }, { "id": "c643bff833aaa9a47e3421a", "name": "Vomiting", "state": "unsure" }],</pre>	<ul style="list-style-type: none"> ● 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.

6.1.1.4 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 to benchmarking API calls 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.

Table 11 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 11 – MMVB 1.0 API output encoding example

Field name	Example	Description
conditions	<pre>"conditions": [{ "id": "ed9e333b5cf04cb91068bbcde643", "name": "GERD" }]</pre>	<ul style="list-style-type: none"> • The conditions the AI considers best explaining the presenting complaints. • Ordered by relevance descending
triage	<pre>"triage": "EC"</pre>	<ul style="list-style-type: none"> • The triage level the AI considers adequate for the given evidence • Uses the same abbreviations defined by the London-model EC, PC, SC, UNCERTAIN

6.1.1.5 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. **Table 12** shows an example of how the expected output is encoded.

Table 12 – MMVB 1.0 AI output label encoding

Field name	Example	Description
condition	<pre>"condition": [{ "id": "85473ef69bd60889a208bcla6", "name": "simple UTI" }]</pre>	<ul style="list-style-type: none"> • The conditions expected/accepted as top result for explaining the presenting complaints based on the given evidence.

]	<ul style="list-style-type: none"> • A list, but only one entry for mono-morbid cases as it is the case for MMVB
expectedTriageLevel	"expectedTriageLevel": "PC"	<ul style="list-style-type: none"> • 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. **Table 13** shows an example of a complete case set structure combining profile information, presenting complaints, other features and the expected values to predict.

Table 13 – 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"  
  }  
}  
]  
}
```

6.1.1.6 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 within the first n suggested conditions. The score is used internally by most topic group members and can also be found in 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.

6.1.1.7 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 from an equal distribution [18; 80] years. Sex is sampled 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 from 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](#) 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 14 – 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.

6.1.1.8 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](#) on *data handling* and [DEL05_6](#) 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/e dit#gid=1175944267>

6.1.1.9 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.

6.1.1.10 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 9**.

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.

6.1.1.11 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.

6.1.1.12 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.

6.1.1.13 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>

6.1.2 Benchmarking version MMVB 2.0 - 2.2

This section includes all technological and operational details of the benchmarking process for the benchmarking versions 2.0-2.2. As the sub-versions have been introduced primarily for structuring the reporting for different focus group meetings, we summarize them here as one entity. As the version 2.2 is an extension of version 1.0 we focus here on the incremental differences.

6.1.2.1.1 Overview

After finishing the MMVB 1.0 version which was centred around the “London model” that was described in the previous section 6.1.1, the work then focused on addressing the next steps pointed out in 6.1.1.12. How to approach the different challenges mentioned there was discussed by the topic group during the second workshop held 10.10.2019 - 11.10.2019 in Berlin. The corresponding extension of the London model was then accordingly called the “Berlin model”.

6.1.2.1.2 Adding symptom attributes

The most relevant limitation of the MMVB 1.0 model was the missing support of explicit *attributes* for describing details like intensity, time since onset or laterality of symptoms like headache. So far, the model contained only so-called “pre-coordinated” combining a symptom with a specific attribute expression pattern like "abdominal pain cramping central 2 days" or "sharp lower quadrant pain". For MMVB 2.2 the attributes have been explicitly added as shown in **Figure 17**.

A	B	C	D	E	F	G	I	J	K	L
Type	Feature name	Feature ID	Attribute name	Multi-Select ?	Attribute ID	State name	State ID	IBD (first presentation non flare)	GERD	simple UTI
	Condition ID							TMP_ID_D_1	ICD10K21	TMP_ID_D_2
	Prior probability							x	xx	xx
symptom	abdominal pain	SNOMED21522001	PRESENCE		PRESENCE			xxx	xx	xx
			pain intensity	N	SNOMED406127006	mild	SNOMED255604002		xx	xx
						medium	TMP_ID_1	xx	xx	x
						severe	TMP_ID_2	x		
			time since onset	N?	TMP_ID_A_1	less than a day	TMP_ID_3			x
						a couple days (1-2 days)	TMP_ID_4	x		xxx
						3 days - to 1 week	TMP_ID_5	xx	x	xx
						a few weeks (1 weeks - 1 month)	TMP_ID_6	xx	xx	
						1 month to 1 year	TMP_ID_7	x	xx	
						a year or more	TMP_ID_8		x	
			quality of abdo pain	N?	TMP_ID_A_2	cramping	TMP_ID_9	xx		
						dull	TMP_ID_10	x	xx	xx
						sharp	TMP_ID_11	x	x	x
			location	Y	TMP_ID_A_3	generalised	TMP_ID_12	x		
						right upper	TMP_ID_13	x	x	
						left upper	TMP_ID_14	x	x	
						epigastric	TMP_ID_15		xxx	
						right lower	TMP_ID_16	xx		xx
						left lower	TMP_ID_17	xx		xxx
						suprapubic	TMP_ID_18	x		xxx
						right loin	TMP_ID_19			
						left loin	TMP_ID_20			
						central	TMP_ID_21	x	x	

Figure 17 – Abdominal Pain symptom with attributes inside the Berlin Model

The above-mentioned pre-coordinated symptoms have been replaced with a single symptom like "abdominal pain" as it is often reported by users of self-assessment applications. For expressing the details, the symptoms now contain sub structures for each attribute stating the probability distribution of the attribute for all the conditions causing this symptom.

6.1.2.1.3 Factor distributions

The second aspect that was improved for the MMVB version 2.2 was the more detailed modelling of risk factors. In the initial model it was only informally noted in a comment field that e.g. "ectopic pregnancy" allows "only females". For later supporting more factors that also influence the AIs in a non-binary way, we introduced explicit probability distributions modulating the prior distributions of the different conditions. Figure 18 and Figure 19 show the refined probability distributions for ectopic pregnancy.

A	B	C	D	E	F	G	I	J	K	L
Type	Feature name	Feature ID	Attribute name	Multi-Select ?	Attribute ID	State name	State ID	IBD (first presentation non flare)	GERD	simple UTI
						1 month to 1 year	TMP_ID_7		xx	
						a year or more	TMP_ID_8		x	
factor	period lateness	BLA1231232	presence		BLUB98762378	not present	YXC_NP	1		1
						present	ASD_P	1		1
factor	sex	SNOMED734000001	sex		TMP_ID_A_5	female	TMP_ID_25	1		1
						male	TMP_ID_26	1		1
	Expected triage level							PC	PC	PC

Figure 18 – Factors with attribute details inside the Berlin Model

A	B	K	L	M	N	O	P	Q	R	S
Type	Feature name	GERD	simple UTI	viral GE	bladder cancer (first presentation)	acute cholecystitis	appendicitis	ectopic pregnancy	IBS	acute pyelonephritis
		xx								
		x								
factor	period lateness	1	1	1	1	1	1	0.8	1	1
		1	1	1	1	1	1	1.2	1	1
factor	sex	1	1	1	1	2	1	1	1	1
		1	1	1	1	1	1	0	1	1
	Expected triage level	PC	PC	SC/PC	PC	EC	EC	EC	PC	EC

Figure 19 – Refined factor distributions for ectopic pregnancy inside the Berlin Model

For example, a chosen attribute value "male" for factor "sex" implies that the probability of "ectopic pregnancy" is zero.

6.1.2.1.4 New benchmarking backend application

With a view towards adding future functionality required for the MVB, the topic group agreed to reimplement the original backend implemented using Flask using the Django framework providing all benchmarking logic as REST API endpoints to the new frontend application. As part of implementing the new backend it was also decided to switch to a dedicated database to store cases and other data relevant to the benchmarking application.

6.1.2.1.5 New benchmarking frontend application

For MMVB 2.2 we also implemented a new web-based frontend application. Previously the frontend was a single file containing all necessary code to communicate with the backend and run basic benchmarks. Crucial data was stored in-memory and it was not supported to e.g. return to a running benchmark or viewing the results of a benchmark that was run in another browser. To address those issues we implemented a new frontend based meeting the following criteria: stateless, proper API communication, interactive, user-friendly, extensible (to in the future include features like interactive drilldowns). The new user interface was also designed to support the new Berlin model attributes and factors. We also improved the usability and visual appearance of the frontend.

6.1.2.1.6 Case Annotation Tool

The benchmarking of the MMVB 2.2 version uses mainly synthetic data sampled from the Berlin model defined by the doctors in the topic group. However, for learning how to create representative real-world cases for the later MVB version of the benchmarking, the doctors in the topic group also created cases. In MMVB 1.0, we used spreadsheets for describing the cases, however with the introduction of the Berlin model this reached its limit, and it became necessary to develop a dedicated annotation tool. As part of MMVB 2.2 we therefore implemented a first simple annotation tool using the same frontend technology stack as the new benchmarking frontend. The annotation tool was then used by the doctors in the topic group to create benchmarking cases on top of the synthetic ones to collect evidence on future annotation tools for the MVB.

6.1.2.2 Benchmarking methods

This section provides details about the methods of the MMVB 2.2 benchmarking. 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).

6.1.2.2.1 Benchmarking system architecture

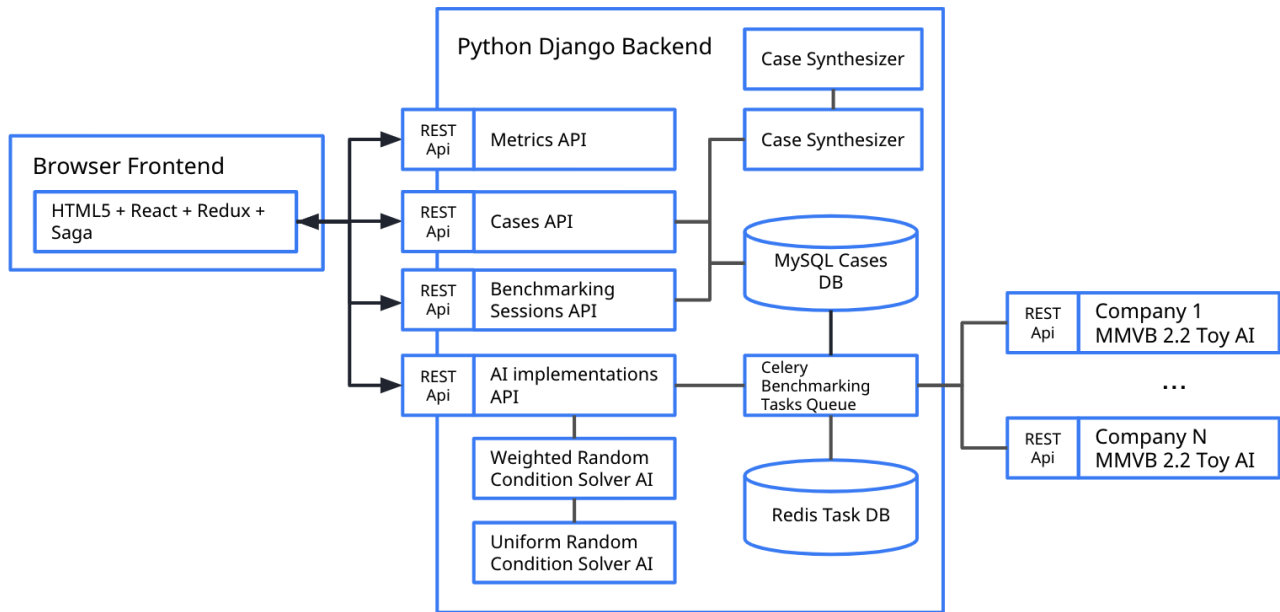


Figure 20 – MMVB 2.2 High-level architecture

6.1.2.2.1.1 Backend architecture changes

The new backend for the MMVB 2.2 benchmarking was based on Django, which is a well-established and well documented Python framework. It plays well with other frameworks which will make the benchmarking easier to extend in future. Django has most of the basic foundation work already implemented which allows developers to focus more on the development of features. A further advantage is that it provides an out of the box solution for user accounts which may be customised at some level for different users and permission levels. This will make it easier to develop the MVB where we need to implement features related to different types of users or tasks (for example submitting a case set or running a benchmark). Django also provides an out of the box customisable solution for admin management which would allow a simple UI to be implemented for admin management on the backend if required. Early discussions towards a focus group wide benchmarking software are also considering Django, which will help migration if necessary.

MySQL was chosen as the database for the new backend, as Django works in the relational realm and MySQL is a stable solution. It was decided that there weren't any specific specifications that would require us to be concerned about any performance specific details that would motivate the use of an alternative database. Django provides an out of the box Object Relational Mapping (ORM) layer to interact with MySQL.

In order to support executing the benchmark on multiple AIs it was decided to use Celery and Redis to manage the task queue. Celery is a Python based task queuing package that enables execution of asynchronous tasks. It is often used in combination with Redis which is a performant in-memory key-value data store which is used as a message broker to store messages between the task queue and the application code.

The latest MMVB 2.2 backend was designed to support the new Berlin model data structures. The most important part was implementation of a Berlin model case synthesizer. As with the previous London model, synthetic cases are generated based on the simple medical domain model for abdominal conditions, findings, factors and profile information. First a condition is sampled at random according to its prior probability, taking into account the weight of factors associated with the synthetic patient sampled from the patient factor distribution. Then clinical findings are then sampled for that condition according to their strength of association with the condition. For the Berlin model, attributes are sampled for each finding where possible.

Based on the new data structures we also changed the API calls to the toy AIs to include new attribute and factor details in the case data. This implies also that all topic group members had to update their toy AIs to support the new model. For the build-in toy AIs “Weighted Random Conditions Solver” and “Uniform Random Condition Solver” this was already implemented by the team working on the backend. They choose a condition at random or weighted by condition prior and now consider the new factor model within the calculation.

The previous backend separated each aspect of the benchmarking process (case generation, toy AIs, case evaluation and metrics calculation) into independent microservices. For the new backend we implemented the different aspects as separate Django applications within the same project. The MMVB 2.2 design contained the following Django applications:

Common

This application aims to act as an aggregator of all the common/shared functionalities for the other applications.

Case Synthesizer

This application is responsible for implementing the structure for synthesizing Cases and CaseSets from the Berlin model.

Toy AIs

This application is responsible for the structure needed for implementing Toy AIs and registering them as available AI implementations.

AI Implementations (API)

This application is responsible for implementing the data models and API for AI Implementations. It allows users to register a new AI Implementation, list, delete and update the registered AI Implementations or ask for their health status.

GET	/api/v1/ai-implementations
POST	/api/v1/ai-implementations
GET	/api/v1/ai-implementations/{id}
PUT	/api/v1/ai-implementations/{id}
PATCH	/api/v1/ai-implementations/{id}
DELETE	/api/v1/ai-implementations/{id}
GET	/api/v1/ai-implementations/{id}/health-check

Figure 21 – MMVB 2.2 ai-implementations API

Cases (API)

This application is responsible for implementing the data models and API for Cases and CaseSets. it offers to list, retrieve, create, and update Cases and CaseSets.

GET	/api/v1/cases
POST	/api/v1/cases
GET	/api/v1/cases/{id}
PUT	/api/v1/cases/{id}
PATCH	/api/v1/cases/{id}
DELETE	/api/v1/cases/{id}
GET	/api/v1/case-sets
POST	/api/v1/case-sets
GET	/api/v1/case-sets/{id}
PUT	/api/v1/case-sets/{id}
PATCH	/api/v1/case-sets/{id}
DELETE	/api/v1/case-sets/{id}
POST	/api/v1/cases/synthesize
POST	/api/v1/case-sets/synthesize

Figure 22 – MMVB 2.2 cases and case-sets API

Benchmarking Sessions (API)

This application is responsible for implementing the structure for creating, retrieving and handling benchmarking sessions.

GET	/api/v1/benchmarking-sessions
POST	/api/v1/benchmarking-sessions
GET	/api/v1/benchmarking-sessions/{id}
PUT	/api/v1/benchmarking-sessions/{id}
PATCH	/api/v1/benchmarking-sessions/{id}
DELETE	/api/v1/benchmarking-sessions/{id}
GET	/api/v1/benchmarking-sessions/{id}/results
GET	/api/v1/benchmarking-sessions/{id}/status
POST	/api/v1/benchmarking-sessions/{id}/run

Figure 23 – MMVB 2.2 benchmarking-sessions API

Metrics

This application is responsible for the structure needed for implementing metrics as well as calculating the metrics for a given benchmarking session result. In contrast to single leaderboard based AI benchmarking systems sufficient for other topic groups for symptom assessment we need an interactive drill down to a given context which requires the dynamic recomputation of sub-set metrics via this API endpoint.

GET	/api/v1/metrics
-----	-----------------

Figure 24 – MMVB 2.2 metrics API

6.1.2.2.1.2 Frontend architecture changes

To ensure both a high code quality as well as an easy onboarding for new developers it was decided to use the React library with a TypeScript (version 3.8, thus a superset of JavaScript ES7). React was chosen for being the presumably most common frontend technology at the moment of decision. TypeScript helps to ensure and enforce both maintainable and understandable code, albeit at the cost of having a learning curve for developers who previously only used JavaScript and sometimes being more verbose. In the background Redux is used in conjunction with Saga to handle state-management and asynchronous communication with the backend. For basic design and user-friendly building blocks a React-specific community implementation of Google's Material Design guidelines called Material UI (material-ui.com) was chosen. Most of the current design-needs are covered by the provided components. For charts the Baidu-backed ECharts library was chosen, currently a candidate project for the Apache Foundation. It was deemed the most versatile option, especially concerning interaction, but is complex in its usage.

6.1.2.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture. In the MMVB 2.2 version the overall flow is similar to the one for MMVB 1.0 and has still the following components:

Model Generation

- The more complex Berlin model was defined by the doctors directly in a google spreadsheet.
- Derived from this was then a technically cleaner spreadsheet version.
- From there it was exported into several CSV files for the findings, conditions, attribute value sets and the condition-finding relations.
- These CSV files are then read by the backend directly with no JSON intermediate format as in MMVB 1.0.

Synthetic Case Generation

- The user triggers the creation of a new case-set in the web-interface.
- The cases are then sampled from the Berlin model.
- The case-sets are then stored in the MySQL database.

Manual Case Generation

- The doctors created manually curated benchmarking cases using the newly developed case annotation tool.
- The corresponding case-sets are then stored in the same database as the synthetic case-sets.

Benchmarking

- The evaluator reads a selected case-set from the case storage and sends it to the AIs. In contrast to earlier versions the cases are sent case by case.
- The AIs respond with their results which are then stored by the evaluator in the database.
- The web-app then uses the metrics API to compute the metrics based on the results stored in the results storage and the corresponding cases stored in the case storage.

The computed results are then finally displayed by the web-application. All case data and results from the benchmarking runs are stored in the database of the backend. As this is still not the MVB yet there was no need for any backend strategy etc..

6.1.2.2.3 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 for MMVB 2.2 was hosted by Ada Health in a GCP VM instance. All the toy-AIs that participated in the benchmarking have been hosted by the individual companies following their own standards for safe and secure operation.

While for the first benchmarking iterations only Babylon, hosting the first benchmarking system, had access to it and all the data stored including the REST API endpoints of all toy-AIs, for the MMVB 2.2 we allowed self-registration of API endpoints with no special protection.

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 in a MySQL database.

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. As the new implementation was much more stable, this was rarely needed.

6.1.2.2.4 Benchmarking process

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

The MMVB 2.2 is still benchmarking toy AIs with synthetic data. The process is therefore similarly simple as the one for MMVB 1.0. The following sections will describe the most important steps.

Landing Page

Starting point for benchmarking is the landing page shown in Figure 25. It features cards as shortcuts to the AI-Implementations, datasets and benchmarking sessions, all with indications of the number of entities currently available there. We also added an extra navigation bar on the left side to have all relevant options permanently available.

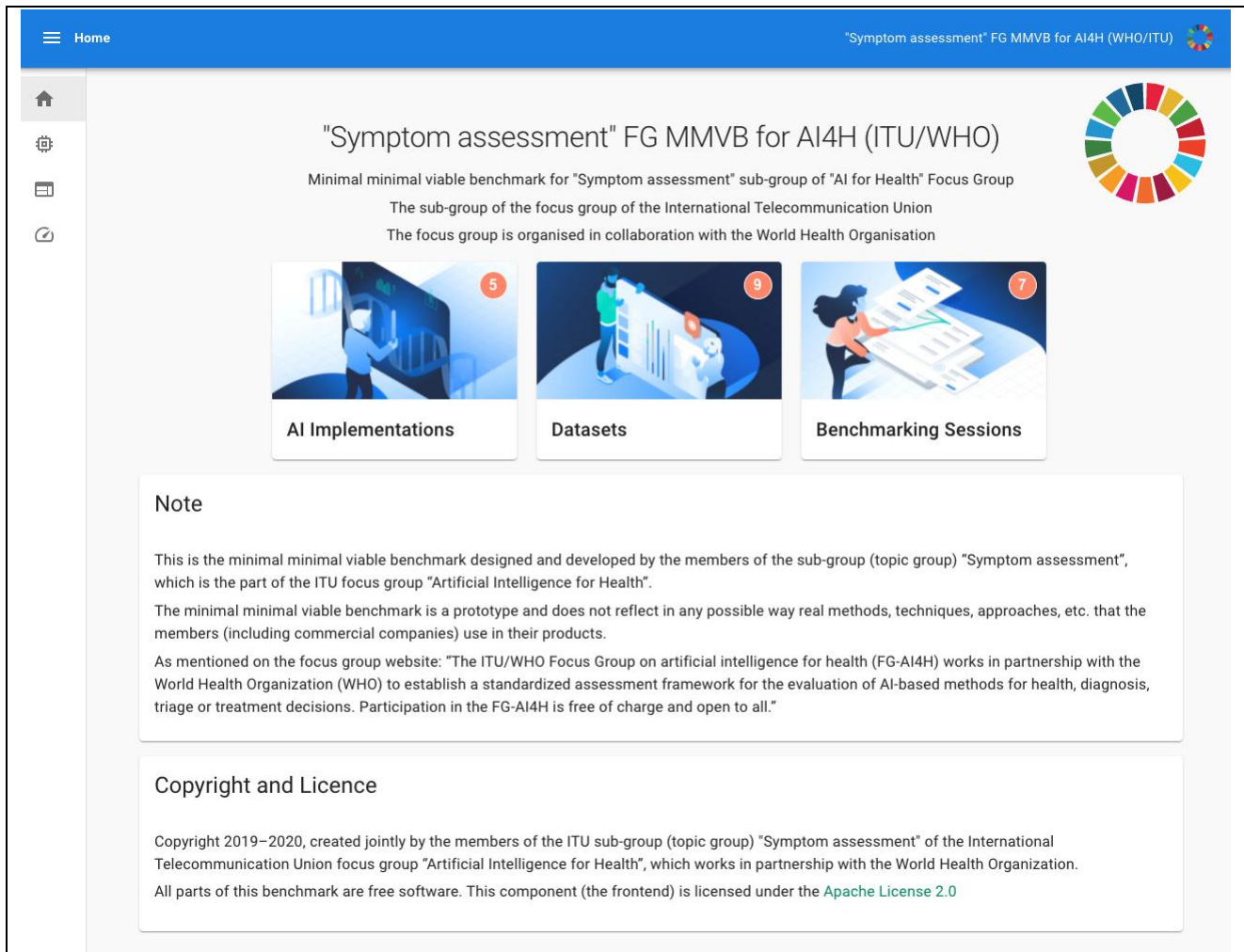


Figure 25 – 2.2 Version of the Benchmarking start page

AI Registration

As a second step AI developers can register their AIs with name and API endpoint. For the current phase this has proven to be more practicable and is also closer to the MVB where AI developers would register and submit AIs for official benchmarking too. The current version, however, has no protection mechanisms implemented so that e.g. all AI developers could change the endpoints of all the other AIs. Adding authentication, rights and roles is therefore one of the important steps towards the MVB.

Figure 26 shows the list of AI implementations with the button for registering a new AI. As AI developers have to register their AI in the current system once, this step can in most cases be skipped.

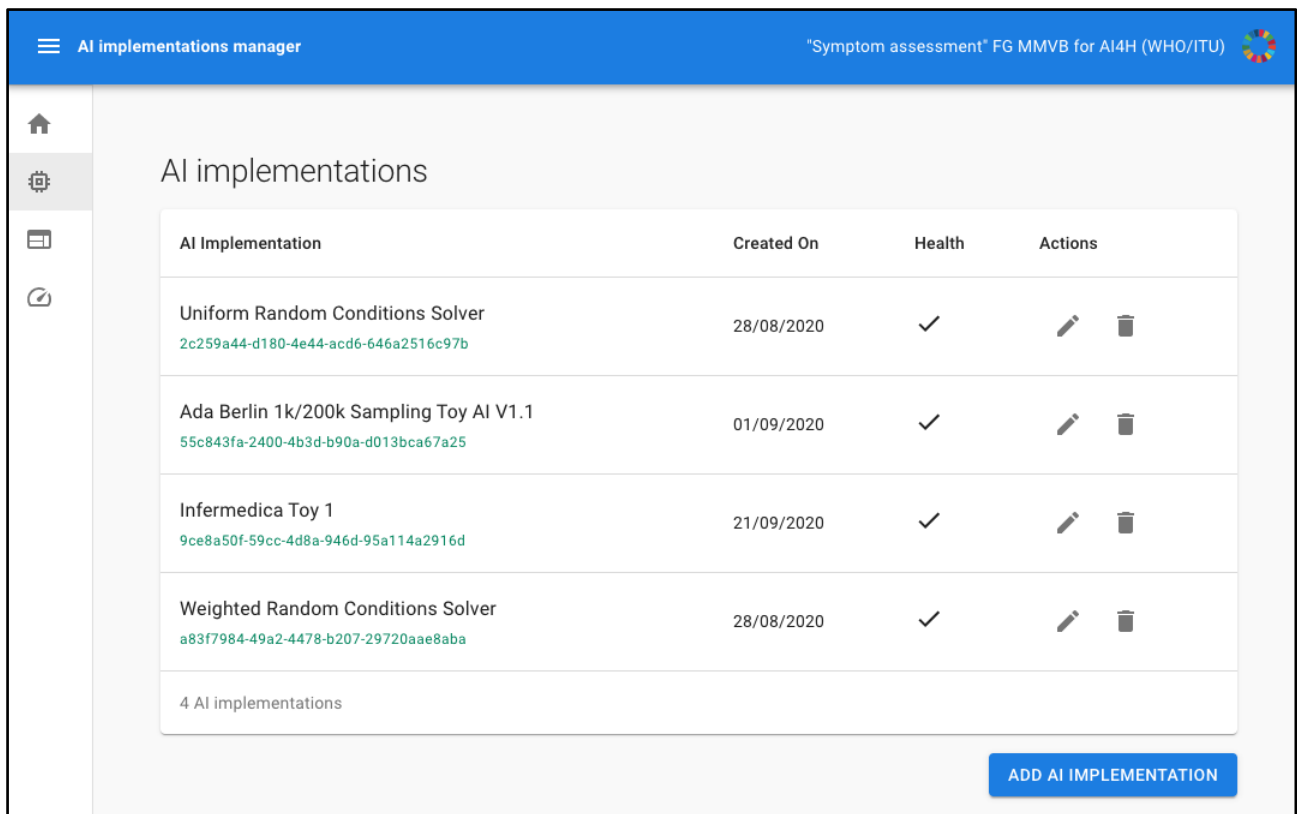


Figure 26 – The AI implementations list now featuring the ability of adding new AIs and editing existing ones.

Creating a benchmarking data set

The next step is creating a case set for the benchmarking. Figure 20 shows the page with all the casesets created so far. Users can create new casesets by clicking the “Add AI Implementation” button in the screen shown in Figure 21. Beside the name the user can define the number of synthetic cases to sample from the Berlin model. Manual cases have to be directly registered with the backend and can currently not be added through the UI.

The screenshot displays the 'Case sets manager' interface. The top navigation bar includes a hamburger menu, the text 'Case sets manager', and a title 'Symptom assessment* FG MMVB for AI4H (ITU/WHO)' with a logo. A left sidebar contains navigation icons for home, settings, list, and refresh. The main content area is titled 'Case sets' and features a table with the following data:

Name	Created On	Size	Labels	Actions
funny 1 edf7bc47-88e3-43f5-a483-891cf494c62a	11/11/2020 16:37	100		
Official FG AI4H Meeting I Benchmarking test data set 71e9c086-2d8f-4cf9-bbe4-106117c95c5c	23/09/2020 13:54	1000		
showEditor 572d8988-68de-4841-8f74-7d1a4a253a75	12/11/2020 10:38	2		
some100 a12ed29d-fef6-478e-aaca-4a76bd70dc9c	13/11/2020 16:22	100		
test10 f266e564-135f-4959-8a60-2f6d3f5cec9d	30/10/2020 11:22	10		
test1K 3a9b95e3-3c2e-4631-ac8c-16024e2e9c7b	06/11/2020 13:25	1000		
Testing Shubs 6cd94abc-02e1-4170-a11a-411ca3b44274	06/10/2020 14:03	11		

7 case sets

[GENERATE CASE SET](#)

Figure 27 – MMVB 2.2 case sets overview page

Generate new case set

Parameters

Name
Official FG AI4H Meeting L Benchmarking test data set

Number of cases
10000

GENERATE CASE SET →










Figure 28 – MMVB 2.2 case set creation page

Creating a benchmarking session

Once there are AIs and a case set the user can create a new benchmarking session. **Figure 29** shows the overview page with the list of the existing benchmarking sessions. By clicking on the corresponding button the user can create a new session using the screen shown in **Figure 30** by selecting the caseset and the AIs that should participate in the benchmarking.

Benchmarking sessions manager "Symptom assessment" FG MMVB for AI4H (ITU/WHO)

Benchmarking sessions

Benchmarking session ID	Created On	Als	Dataset	Status	Actions
01566449-4fa5-40fc-887d-16b52b27e798	13/11/2020 16:22	<ul style="list-style-type: none">Uniform Random Conditions SolverAda Berlin 1k/200k Sampling Toy AI V1.1Your.MD Berlin Model toy AIInfermedica Toy 1Weighted Random Conditions Solver	some100 100 cases	✓	  
f76c1477-16e1-443a-9e9b-423c5c73938a	12/11/2020 10:39	<ul style="list-style-type: none">Uniform Random Conditions SolverAda Berlin 1k/200k Sampling Toy AI V1.1Your.MD Berlin Model toy AIInfermedica Toy 1Weighted Random Conditions Solver	test10 10 cases	✓	  
d58c19b6-b8b9-403a-ada4-06d831b7c6d5	06/11/2020 13:24	<ul style="list-style-type: none">Uniform Random Conditions SolverAda Berlin 1k/200k Sampling Toy AI V1.1Your.MD Berlin Model toy AIinfermedica_toy_1Infermedica Toy 1Weighted Random Conditions Solver	test10 10 cases	✓	  

3 benchmarking sessions

[CREATE BENCHMARKING SESSION](#)

Figure 29 – MMVB 2.2 benchmarking sessions overview page

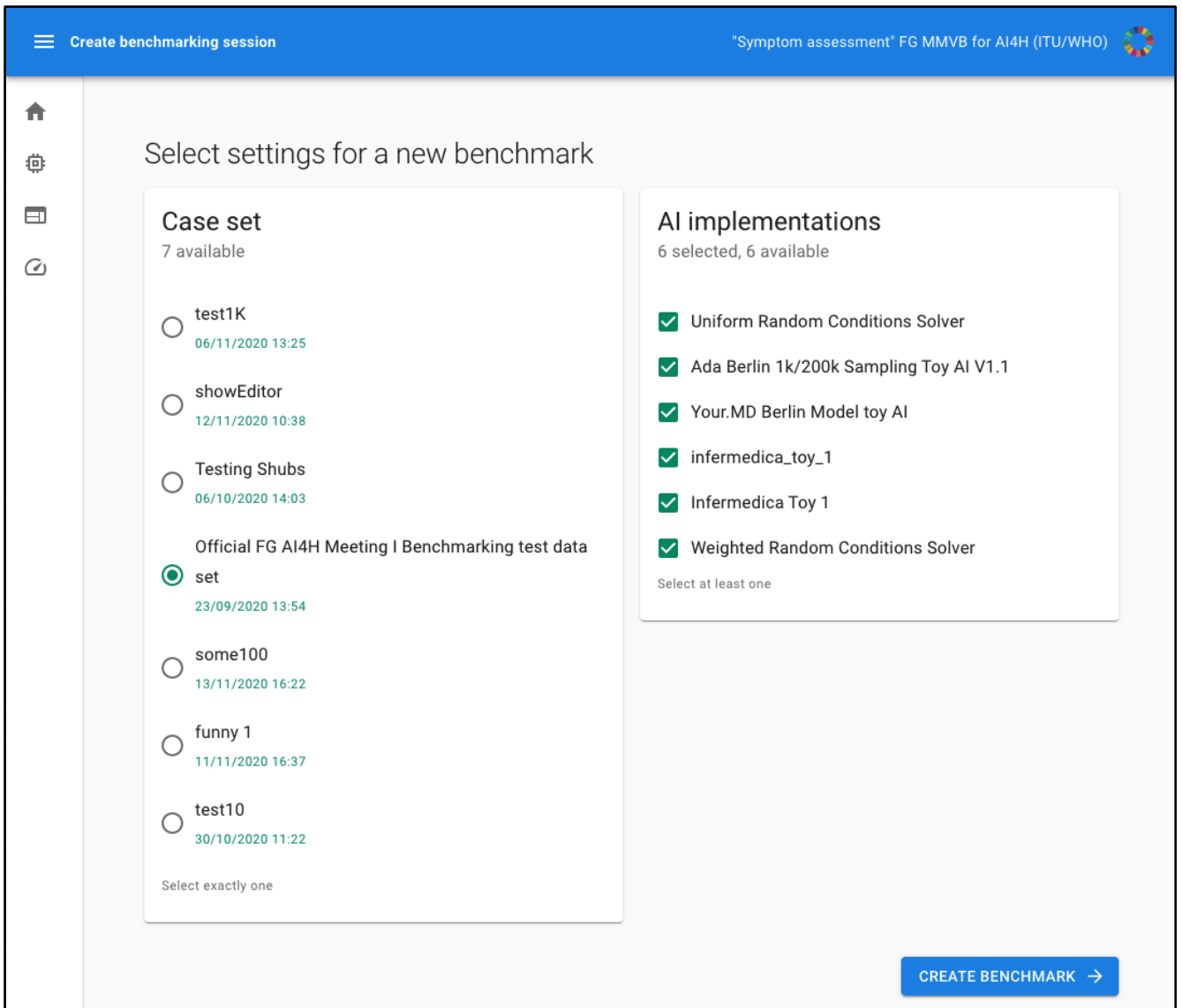
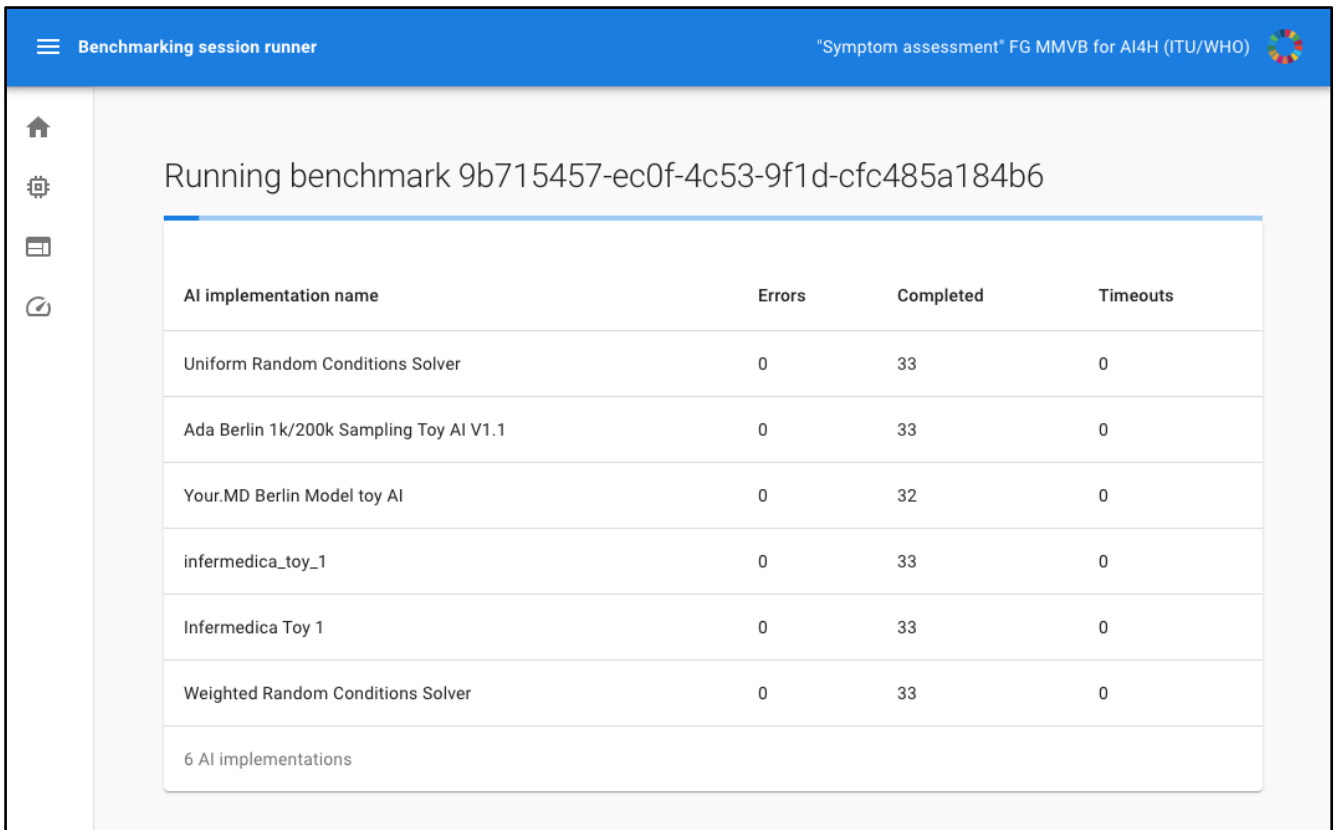


Figure 30 – MMVB 2.2 benchmarking session creation page

Running a benchmarking session

Once the benchmarking session was created it can be run by clicking the play button. The benchmarking runs asynchronously on the server. While the benchmarking is running the benchmarking session shows the progress screen display in **Figure 31** with a progress bar, number of completed cases, errors and timeouts.



The screenshot shows a web interface for a benchmarking session runner. The header is blue and contains the text "Benchmarking session runner" on the left and "Symptom assessment" FG MMVB for AI4H (ITU/WHO) on the right. Below the header, there is a sidebar with navigation icons (home, gear, list, refresh). The main content area displays the title "Running benchmark 9b715457-ec0f-4c53-9f1d-cfc485a184b6" and a table with the following data:

AI implementation name	Errors	Completed	Timeouts
Uniform Random Conditions Solver	0	33	0
Ada Berlin 1k/200k Sampling Toy AI V1.1	0	33	0
Your.MD Berlin Model toy AI	0	32	0
infermedica_toy_1	0	33	0
Infermedica Toy 1	0	33	0
Weighted Random Conditions Solver	0	33	0
6 AI implementations			

Figure 31 – MMVB 2.2 benchmarking session runner

Reviewing the benchmarking results

Once the benchmarking calculation was completed, the session shows a result screen displayed in **Figure 32**. The metrics are calculated dynamically based on the stored responses of the AIs. While this might be unusual for many classical AI competition platforms, for this topic group we need the flexibility to drill down into a given context relevant to a stakeholder and recompute the metrics based on this filtered subset.

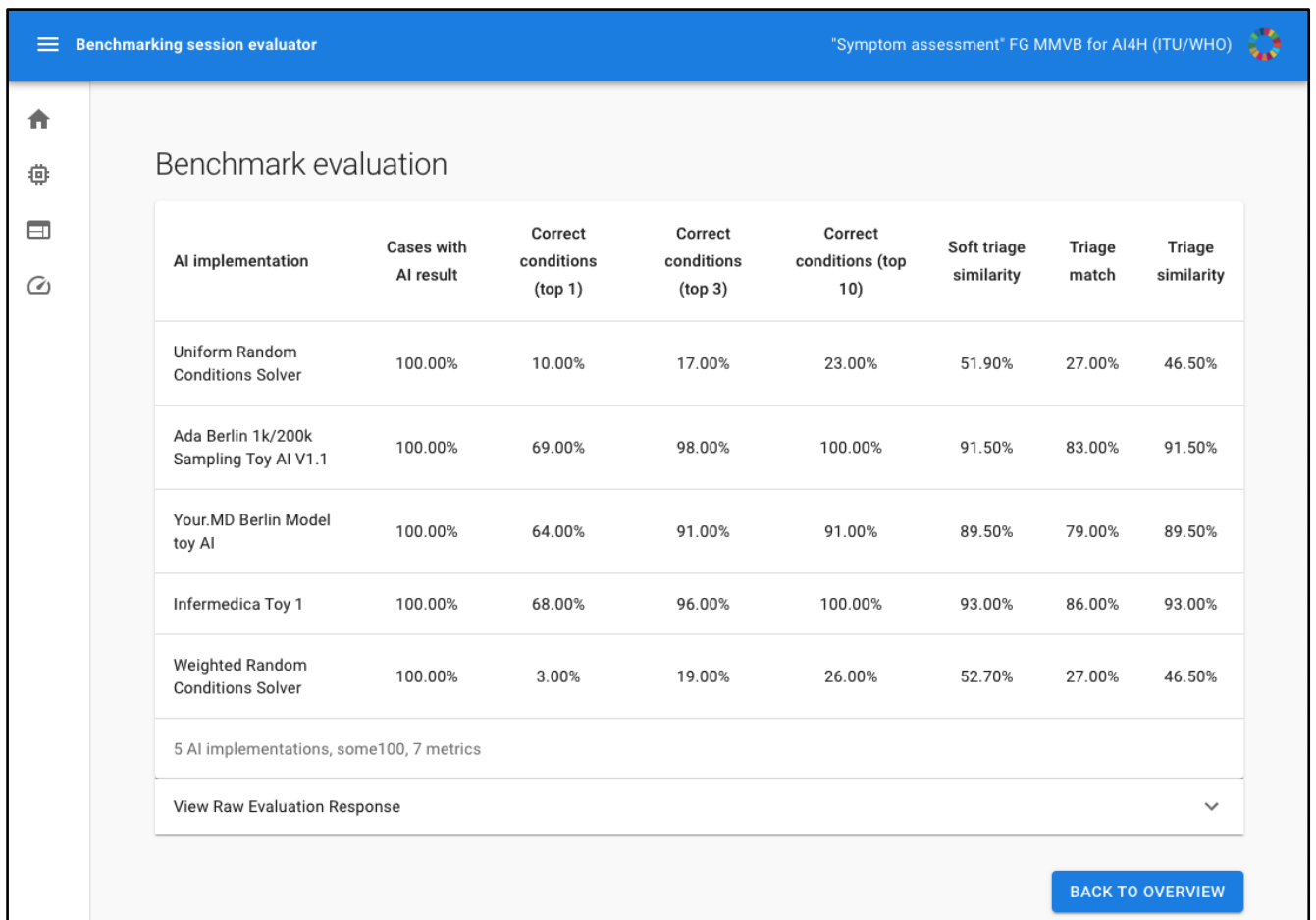


Figure 32 – MMVB 2.2 benchmarking result page

As for MMVB 1.0 for the MMVB 2.2 version there was also no scheduled benchmarking. Every developer could run a benchmarking session at any time for building their own toy AI which helped both the development of the pipeline and the toy AIs. In contrast to MMVB 1.0 the step of submitting the toy AI to Yura Perov was also not necessary anymore as we added the AI self-registration feature.

6.1.2.3 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 2.2 uses as input for the AIs a simplistic user profile, explicit presenting/chief complaints (PC/CC), and additional symptoms, findings and factors. The general case structure that is sent to a toy AI can be seen in **Figure 33**. Since MMVB 1.0 we added an explicit field for the name of the AI implementation so that AI developers can host all AIs at the same systems and route the requests to the correct AIs by name. **Table 15** shows the details of the different fields in the case structure.

```
{
  "caseData": {
    "otherFeatures": [
      ...
    ]
  }
}
```

```

    ],
    "profileInformation": {
        ...
    },
    "presentingComplaints": [
        ...
    ]
  },
  "aiImplementation": "Company X Berlin Model toy AI"
}

```

Figure 33 – MMVB 2.2 General input case structure

Table 15 – MMVB 2.2 input data format

Field name	Example	Description
profileInformation	<pre> "profileInformation": { "age": 38, "biologicalSex": "male" } </pre>	<ul style="list-style-type: none"> ● 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	<pre> "presentingComplaints": [{ "id": "0ef0...2d87", "name": "Diarrhea (finding)", "state": "present", "attributes": [{ "id": "7a0c...5d3c", "name": "Time since onset", "value": { "id": "4395...90a23", "name": "3 days - to 1 week", "standardOntologyUris": ["CUSTOM:102"] } }, { "id": "7a0c...5d3c", "name": "Time since onset", "value": { "id": "4395...90a23", "name": "3 days - to 1 week", "standardOntologyUris": ["CUSTOM:102"] } }], "standardOntologyUris": ["CUSTOM:1"] }], "standardOntologyUris": [</pre>	<ul style="list-style-type: none"> ● The complaints the user seeks and explanation/advice for ● Always present ● A list, but for the MMVB 2.2 always with exactly one entry ● For MMVB 2.2 the PC now also contain attributes ● In addition to MMVB 1.0 also standardOntologyUris (even if in this example the it is only an ID not an URI)

	<pre>"62315008"] }]</pre>	
otherFeatures	<pre>"otherFeatures": [{ "id": "356ae...a492", "name": "Vomiting (disorder)", "state": "unsure", "attributes": [], "standardOntologyUris": ["422400008"] }, { "id": "b200...d5e8", "name": "Dysuria (finding)", "state": "present", "attributes": [{ "id": "7a0c...5d3c", "name": "Time since onset", "value": { "id": "b505...ae52", "name": "a year or more", "standardOntologyUris": ["CUSTOM:105"] } }, { "id": "7a0c...5d3c", "name": "Time since onset", "value": { "id": "b505...ae52", "name": "a year or more", "standardOntologyUris": ["CUSTOM:105"] } }], "standardOntologyUris": ["CUSTOM:1"] }], "standardOntologyUris": ["49650001"] }],</pre>	<ul style="list-style-type: none">● Similar to the presenting complaints now with attributes and standard ontology identifiers

The MMVB 2.2 case data explicitly encoded the presence of each symptom. All symptoms have an explicit "state" attribute, which is responsible for information on whether a symptom is "present", "absent" or a patient is "unsure" (or does not know) about it.

With the introduction of the Berlin model and Alejandro Lopez Osornio joining the topic group as an ontology expert, we started the first steps towards mapping the model to the SNOMED CT ontology. The goal was to use as much of the SNOMED CT knowledge representation capabilities as possible, complemented with an ad-hoc information model that adds attributes and values.

6.1.2.4 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.

The case object described in the previous section is sent to the “/solve-case” context of the API endpoints specified by the toy AIs as a JSON post request payload. The expected response is a JSON object with the fields described in **Table 16**. It has not changed compared to MMVB 1.0.

Table 16 – MMVB 2.2 AI output structure

Field name	Example	Description
conditions	<pre>"conditions": [{ "id": "ed9e333b5cf04cb91068bbcbde643", "name": "GERD" }]</pre>	<ul style="list-style-type: none"> • The conditions the AI considers best explaining the presenting complaints. • Ordered by relevance descending
triage	<pre>"triage": "EC"</pre>	<ul style="list-style-type: none"> • The triage level the AI considers adequate for the given evidence • Uses the same abbreviations defined by the London-model EC, PC, SC, UNCERTAIN

In addition to the “solve-case” endpoints the toy AIs are also supposed to listen to the “health-check” endpoint. It is used during the benchmarking to make sure that the AIs are ready to process cases. In the later MVB the benchmarking would pause if an AI is not responding anymore. The expected response for the health check is an JSON object like { "data": "OK" }. Every answer other than “OK” would be considered an error.

6.1.2.5 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.

MMVB 2.2 relies on synthetic data sampled from the Berlin model. All cases are stored as case sets in the MySQL database. Internally the case sets are encoded similar to the cases sent to the AI but with the additional fields listed in **Table 17**.

Table 17 – MMVB 2.2 case with labels included

Field name	Example	Description
<i>correctCondition</i>	<pre>"correctCondition": { "id": "2333...13c8", "name": "GERD", "standardOntologyUris": ["http://snomed.info/id/23559 5009"] }</pre>	<ul style="list-style-type: none"> • The correct condition i.e. in MMVB 2.2 the condition this case was sampled from. Note: the correct condition might not be the correct expected one given the evidence! For instance, in the case of only headache a common cold is expected even if the case was a brain cancer case.
<i>expectedCondition</i>	Same as for correctCondition	<ul style="list-style-type: none"> • 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 2.2
<i>impossibleConditions</i>	Same as for correctCondition	<ul style="list-style-type: none"> • An optional list of diseases that must not be contained in the results e.g. because have contradict sex e.g. male diseases for females in vice versa
<i>otherRelevantDifferentials</i>	Same as for correctCondition	<ul style="list-style-type: none"> • Other diseases that should be present in the result as relevant differentials to rule out.
<i>expectedTriageLevel</i>	<pre>"expectedTriageLevel": "PC"</pre>	<ul style="list-style-type: none"> • The expected triage level (EC, PC, SC, UNCERTAIN)
<i>name</i>	<pre>"name": "BPPV test case #1",</pre>	<ul style="list-style-type: none"> • The name of the case. This is especially helpful for cases created by doctors.
<i>caseSets</i>	<pre>"caseSets": ["4354...3a75"]</pre>	<ul style="list-style-type: none"> • The list of case-sets containing this case. • Only used for case-set management.

The overall case-set structure can be seen in **Table 18**.

Table 18 - MMVB 2.2 overall case-set structure

<pre>{ "id": "f266e564-135f-4959-8a60-2f6d3f5cec9d", "name": "test10", "cases": [</pre>

```
{
  "id": "29959f5c-c4e6-4341-9a1a-4802ce451629",
  "data": {
    "caseData": { ... },
    "metaData": {
      "name": "Synthesized case a5e425a2",
      "caseCreator": "MMVB Berlin model case synthesizer"
    },
    "valuesToPredict": {
      "correctCondition": {
        "id": "9d27455f69d907a7dad7bb471f3717a4",
        "name": "Appendicitis (disorder)",
        "standardOntologyUris": [
          "74400008"
        ]
      },
      "expectedCondition": {
        "id": "9d27455f69d907a7dad7bb471f3717a4",
        "name": "Appendicitis (disorder)",
        "standardOntologyUris": [
          "74400008"
        ]
      },
      "expectedTriageLevel": "EC",
      "impossibleConditions": [],
      "otherRelevantDifferentials": []
    }
  },
  "caseSets": [
    "f266e564-135f-4959-8a60-2f6d3f5cec9d"
  ]
  ...
}
```

The frontend application also offers a feature to preview and download all the case set data (see Figure 34).

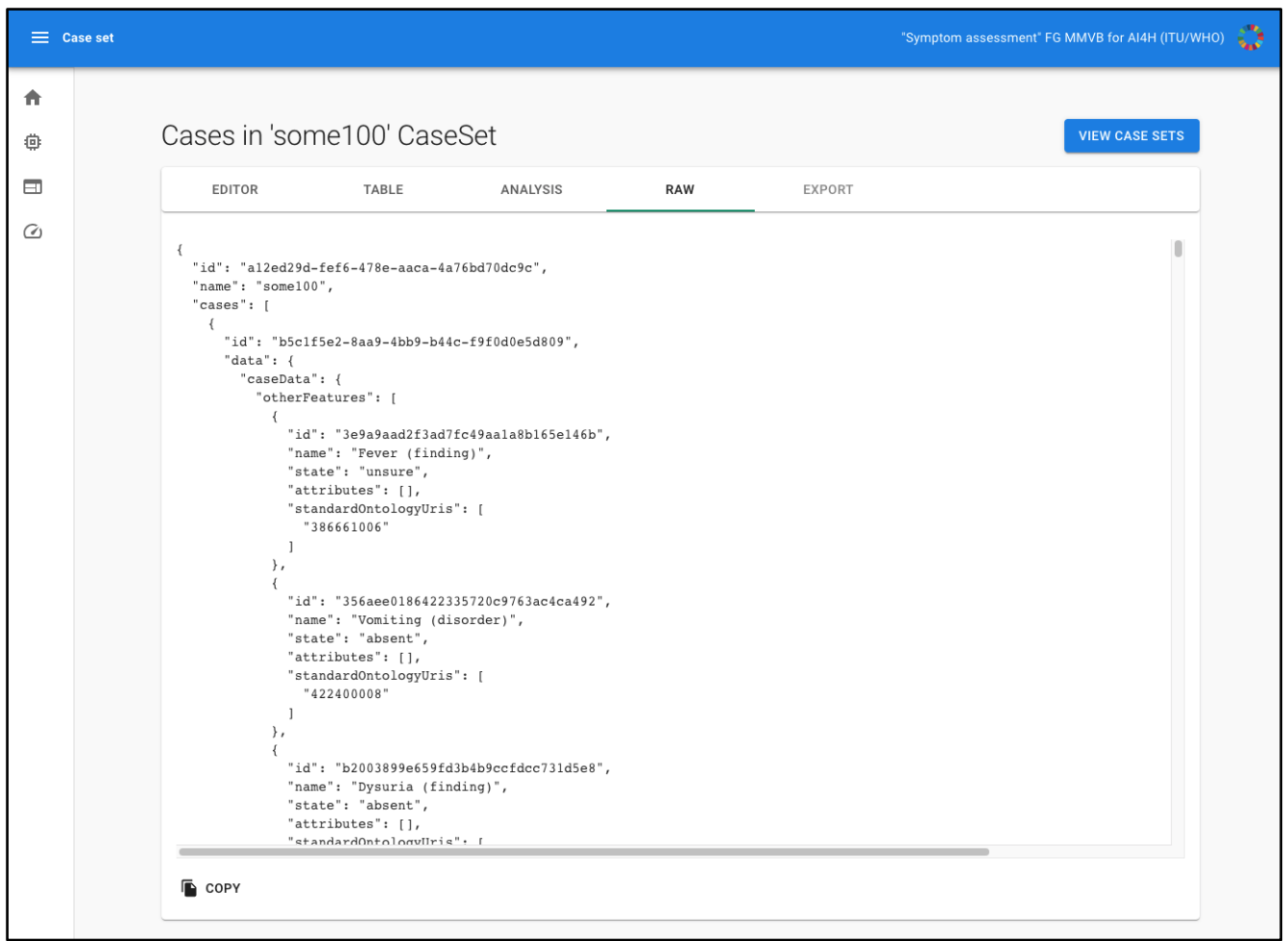


Figure 34 – MMVB 2.2 case-set raw viewer

6.1.2.6 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 1.0 benchmarking, MMVB 2.2 also only used toy AIs and toy data and so the focus was still to have metrics for implementing the benchmarking in the first place. For this purpose we implemented the same metrics:

- Cases with AI result (success rate)
- Correct conditions (top 1)
- Correct conditions (top 3)
- Correct conditions (top 10)
- Triage match
- Triage similarity
- Soft triage similarity

We decided to implement a new "Triage similarity (soft)" score such that if an AI says that it is "unsure" about a triage, the AI is given a triage similarity score higher than 0. The reason to introduce this score is to learn how to integrate "unsure" into the scoring calculations. In future

iterations, looking toward MVB we might want to treat "unsure" answers for triage and/or condition list differently to the "worst answer".

6.1.2.7 Test dataset acquisition

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

For the MMVB 2.2 benchmarking iteration we used both synthetic data and case vignettes created by doctors.

6.1.2.7.1 Synthetic test data acquisition

The sampling is performed by the case synthesizer in the backend based on the CSV exported spreadsheet of the Berlin model:

<https://docs.google.com/spreadsheets/d/1dxzHFA8Rz2erN16dKKf8Sq9MffHiEWsHEsyDEtYW7dg/edit#gid=2083361879>

The cell on the intersection between symptom's first row and a disease is a rough estimate of a link strength (captured by "x", "xx" or "xxx" labels where "xxx" stands for the strongest link) between a disease and a symptom. Each attribute state also might have a link with a disease, however it is already conditioned on the presence of the symptom. Some symptom attribute states are exclusive (i.e. not multiselect), meaning that only one attribute state can be "present". Other symptom attribute states are not exclusive (i.e. multiselect), meaning several states might be present at the same time. The encoding for this can be found in the "attributes-value_sets" sheet. If symptom is "absent" or "unsure", then no attributes or attribute states are sampled.

In general the case synthesizer first samples a patient from a uniform sex distribution and uniform age distribution between 18 and 80 years. It also samples then the remaining factors obeying their sex dependencies. Based on this a condition that is not contradicting the factors is sampled based on their prior probability which is assumed linear to the number of "x" in the model. Based on the condition then the symptoms are samples based on their conditional probability which is defined as 30% times the number of "x" in the model. It is also sampled if the symptoms should be part of the case, marked as "unsure" or omitted. In the last state the attributes are samples based on their conditional probabilities but also considering their multi-selectable / single-selectable type.

For checking the sampling we also implemented a first statistical tool showing the sex and age distributions (see **Figure 35**).

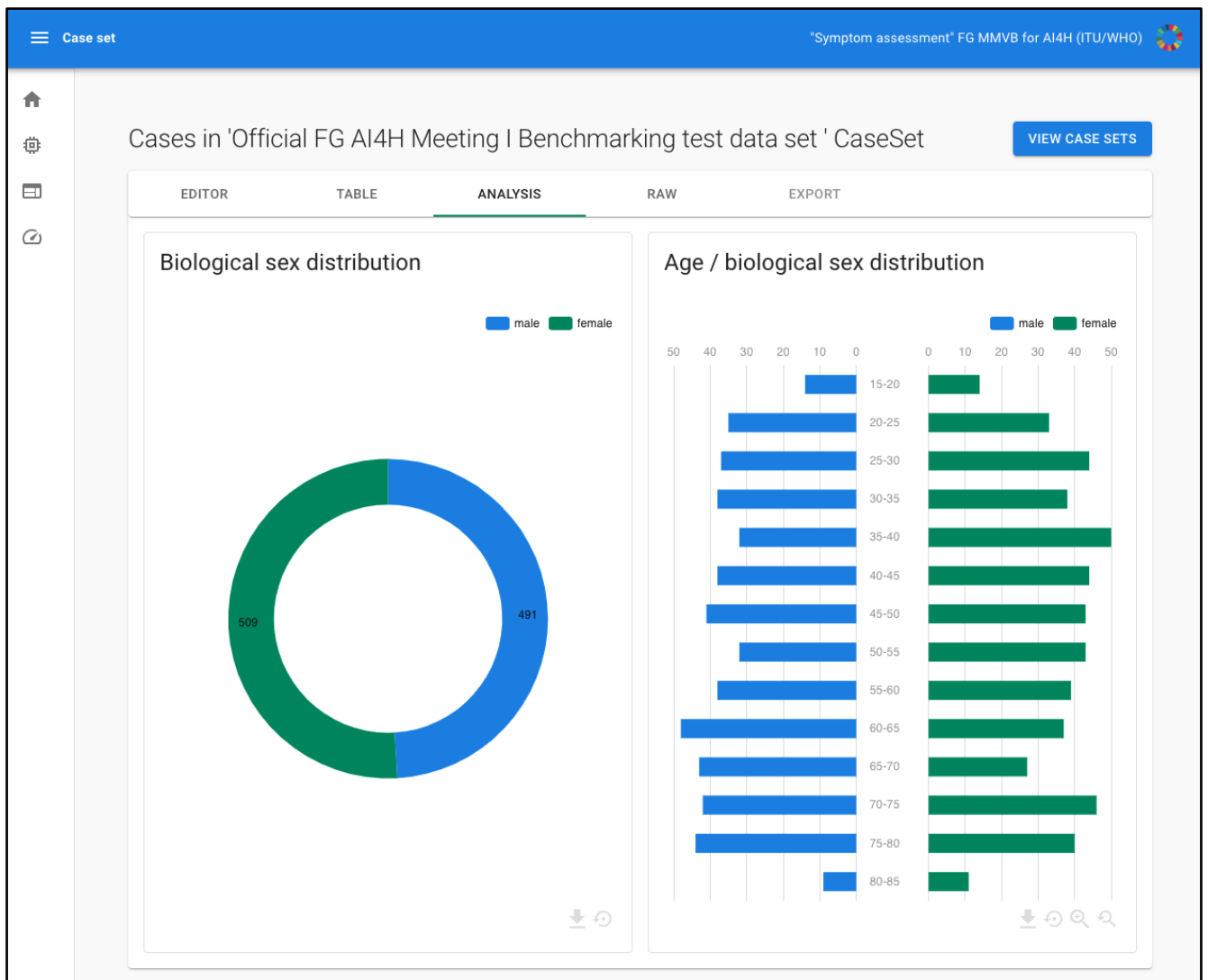


Figure 35 – MMVB 2.2 case-set statistics view

6.1.2.7.2 Manual test data acquisition

The manual case creation was based on the annotation tool created for MMVB 2.2. It is a generic tool largely automatically generated from the Berlin model. This way it makes sure that only valid cases can be created by offering e.g. only the available attribute if adding a symptom. This generic approach implies a trade-off between reusability (such as for other focus groups, as discussed in [FG-AI4H-H-038-R01](#)) and user experience. The latter will be of high importance once a large number of cases will need to be created by human medical doctors, which will happen through the web tool.

By being almost exclusively composed of (multi-)selection fields all cases created are automatically valid according to the model.

For the time being it is exclusively aimed at the creation of synthetic cases. It does not incorporate ways of extracting information from a medical record in a documented fashion. Currently there is no review mechanism implemented, but the (backend) infrastructure has been designed in a way to accommodate this. **Figure 36** shows examples of the case annotation tool.

Case set "Symptom assessment" FG MMVB for AI4H (ITU/WHO)

Home, Settings, Case set, Refresh icons

Edit case 'cbd31a68-c544-4306-9750-dc88c82a7a2f'

in case set 'Testing Shubs'

Correct Condition acute pyelonephritis **Case Name** acute pyelonephritis **Age** 31 **Biological sex** female

Case creator Shubs Upadhyay **Triage Level** PC **Expected condition** acute pyelonephritis

PRESENTING COMPLAINT

Clinical Finding Name: Abdominal pain, State: present

ATTRIBUTES 4/4

Multi-Select Attribute: Finding site (Left loin)

Attribute: Characteristic of pain (Sharp)

Attribute: Pain intensity (Moderate)

Attribute: Time since onset (3 days - to 1 week)

OTHER FEATURES

Clinical Finding Name: Fever, State: present

Clinical Finding Name: Vomiting, State: absent

ATTRIBUTES 0/1

ADD ATTRIBUTE

Clinical Finding Name: Diarrhea, State: absent

ATTRIBUTES 0/2

ADD ATTRIBUTE

Clinical Finding Name: Dysuria, State: present

ATTRIBUTES 1/1

Attribute: Time since onset (3 days - to 1 week)

Clinical Finding Name: Blood in urine, State: present

ATTRIBUTES 1/1

Attribute: Time since onset (less than a day)

Clinical Finding Name: Increased frequency of uri..., State: present

ADD CLINICAL FINDING

RELEVANT DIFFERENTIALS 2/11

Condition: ectopic pregnancy

Condition: simple UTI

ADD

IMPOSSIBLE CONDITIONS 2/11

ADD

SUBMIT

Figure 36 – MMVB 2.2 example of a case defined by a doctors using the case annotation tool

For the manual case creation we also created cases annotation guidelines helping the doctors with what to consider when creating cases. The link to these can be found here - [MMVB 2.2 Guidelines](#). Clinicians in the participating organizations then use these new guidelines to create a new set of cases for the MMVB 2.2 benchmarking.

The latest set of 11 cases has then been manually imported into the benchmarking system where they are available as a case set for then benchmarking.

6.1.2.8 Data sharing policies

For the MMVB 2.2 iteration only synthetic cases and some cases manually created by the doctors in the topic group have been used. The cases are highly specific to 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>

As the benchmarking system for MMVB 2.2 is freely available under <http://35.228.161.168:3000/> all the synthetic data and all the benchmarking results are also available without any limitation despite the fact that there is no protection against someone deleting old data sets or benchmarking results as there is currently no legal or regulatory need for retaining them.

6.1.2.9 Baseline acquisition

For the MMVB 2.2 assessing any baseline was out of scope.

6.1.2.10 Reporting methodology

As the MMVB 2.2 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 contrast to the MMVB 1.0 version, the current frontend does not contain any interactive drill-down feature yet. However, it is still clear that this will be an important feature in the MVB

6.1.2.11 Result

As this benchmarking iteration was only an intermediate development step, no final result was recorded, nor would it be of any practical relevance. As expected, the best performing toy AI uses a brute force sampling approach which is asymptotically optimal and cannot be outperformed. **Figure 32** shows the results for a 100 cases test set.

6.1.2.12 Discussion of the benchmarking

The release of the MMVB 2.2 version concluded the implementation of the Berlin model with the more complex domain model considering also attributes and more detailed factors. The work included a complete rewrite of the frontend and backend application and also the implementation of a dedicated case annotation/creation tool.

The work on the benchmarking pipeline also underlined that in general the benchmarking with AI systems hosted by the participants is technically feasible.

During the work on the MMVB the Focus Group started to pursue the idea of a Focus Group wide open source initiative for a benchmarking platform. If feasible TG-Symptoms would adopt this platform for the benchmarking of AI-based symptom assessment systems. This would however shift the focus of our work on the parts that are specific for our topic group, namely:

- A joint ontology for encoding case vignettes for benchmarking.
- The case annotation and creation tool or a component for such a tool in a general annotation tool.
- The scores and metrics specific for AI-based symptom assessment.
- The interactive result drill-down and filtering system provides the benchmarking results for a specific context relevant to a stakeholder.
- The approach of distributed AIs hosted by the participants.

The first point in particular is the largest and most important open task for implementing the first real minimal viable benchmarking so that we will focus all our work on that and will not implement another benchmarking iteration until this work is completed.

6.1.2.13 Retirement

The MMVB 2.2 is still an intermediate development system. Frontend, backend and the annotation tool will be hosted as a demo and topic group internal reference system until a new benchmarking iteration is available under:

<http://35.228.161.168:3000/>

The code of the benchmarking system is available in GitHub at:

<https://github.com/FG-AI4H-TG-Symptom/fgai4h-tg-symptom-benchmarking-frontend>

and

<https://github.com/FG-AI4H-TG-Symptom/fgai4h-tg-symptom-assessment-mmvb-backend>

The Berlin Model used for generating the synthetic test cases are available at:

<https://docs.google.com/spreadsheets/d/111D40yoJqvvhZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=575520860>

6.2 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.

7 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 6.1.2.11).

- 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)?

8 Regulatory considerations

*Topic Driver: This section reflects the requirements of the working group on **Regulatory considerations on AI for health (WG-RC)** 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) compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL02 “*AI4H regulatory considerations*” (which provides an educational overview of some key regulatory considerations), DEL02_1 “*Mapping of IMDRF essential principles to AI for health software*”, and DEL02_2 “*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 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.

8.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as 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)?
 - MDR 2017/745
 - 21 CFR Volume 8 (US FDA)
 - FDA Clinical Decision Support Software (Draft Guidance for Industry and Food and Drug Administration Staff) - updated May 2020
 - FDA Clinical Decision Support Software (Draft Guidance for Industry and Food and Drug Administration Staff) - updated May 2020
 - GDPR, HIPAA, CCPA
 - ISO13485, ISO14971, IEC62304, ISO27001
 - UK Medicines and Medical Devices Act 2021

- Are there any aspects to this AI system that require additional specific regulatory considerations?
 - The EU Commission’s proposed AI Regulatory Framework will be relevant (but not finalised)

8.2 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)?

8.3 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?
 - ISO13485 standard has sections on ‘Measurement’, as well as Software Verification/Validation

- Are viable solutions with the necessary certifications already available?
 - At the time of meeting M in 2021 there are no such systems known to the group.

- Could the TG implement such a solution?
 - Based on the topic group's current development capacities it is unlikely that the group alone can implement and own an ISO13485 compliant benchmarking platform. However, the focus group started an Opensource initiative to implement a benchmarking platform that could be used by most topic groups. Implementing the components specific to the topic group for an otherwise ISO13485 compliant platform is more realistic.

8.4 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 "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?

9 References

- [rx10] Semigran Hannah L, Linder Jeffrey A, Gidengil Courtney, Mehrotra Ateev. Evaluation of symptom checkers for self diagnosis and triage: audit study *BMJ* 2015; 351 :h3480
<https://www.bmj.com/content/351/bmj.h3480>
- [rx11] P Ramnarayan, A Tomlinson, A Rao, M Coren, A Winrow, J Brit. ISABEL:a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation *ADC BMJ* 2002 <https://adc.bmj.com/content/88/5/408.long>
- [rx12] Hamish Fraser, Enrico Coiera, David Wong. Safety of patient-facing digital symptom checkers. *Lancet* Vol 392 Issu 10161 Correspondence
[https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(18\)32819-8/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32819-8/fulltext)
- [rx13] J. Maude, Accuracy of a Machine Learning Based Ddx Generator, DEM, 10th International Conference, Boston, MA, Oct 8–10, 2017
https://www.isabelhealthcare.com/pdf/DEM_2017_Isabel_Accuracy.pdf
- [rx14] Paul Manicone MD, Claire Stewart MD, Jeremy Kern MD, Mary Ottolini, MD MPH Children's National Medical Center, Washington, DC, ,, 'ASKING ISABEL' FOR DIAGNOSTIC DILEMMAS IN PEDIATRICS: HOW DOES A WEB BASED DIAGNOSTIC CHECKLIST PERFORM?“, PAS Poster 2013,
https://www.isabelhealthcare.com/pdf/ISABEL_PAS_POSTER_2013.pdf
- [rx15] Mark L. Graber, MD, Ashlei Mathew, Performance of a Web-Based Clinical Diagnosis Support System for Internists, *J Gen Intern Med.* 2008 Jan; 23(Suppl 1): 37–40.,
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2150633/>
- [rx16] Semigran, H. L., Levine, D. M., Nundy, S., & Mehrotra, A. (2016). Comparison of physician and computer diagnostic accuracy. *JAMA Internal Medicine*, 176(12), 1860-1861.
- [rx17] Davies, B. M., Munro, C. F., & Kotter, M. R. (2019). A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom checkers. *Journal of Medical Internet Research*, 21(1), e10868.
- [rx18] Moreno BE, Pueyo FI, Sánchez SM, Martín BM, Masip UJ. A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. *Emergencias* 2017; 29:391-396. <https://www.ncbi.nlm.nih.gov/pubmed/29188913>
- [rx19] Nazario Arancibia, J. C., Martín Sanchez, F. J., Del rey Mejías, A. L., del Castillo, J. G., Chafer Vilaplana, J., Briñon, G., ... & Seara Aguilar, G. (2019). Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department. *International Journal of Interactive Multimedia & Artificial Intelligence*, 5(4).
- [rx20] Liang H, Tsui BY, Ni H et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med.* 2019; 25(3):433-438. doi: 10.1038/s41591-018-0335-9.
<https://www.nature.com/articles/s41591-018-0335-9>
- [rx21] Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet Journal of Rare Diseases*, forthcoming.
- [rx22] Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. *Am J Sports Med.* 2014;42:2371–2376.
- [rx23] Bisson, L. J., Komm, J. T., Bernas, G. A., Fineberg, M. S., Marzo, J. M., Rauh, M. A., ... & Wind, W. M. (2016). How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker?. *Orthopaedic Journal of Sports Medicine*, 4(2), 2325967116630286.

[rx24] Powley, L., McIlroy, G., Simons, G., & Raza, K. (2016). Are online symptoms checkers useful for patients with inflammatory arthritis?. *BMC musculoskeletal disorders*, 17(1), 362.

[rx25] Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, Arnold DoRosario, Megan Mahoney, Saurabh Johri: A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis <https://arxiv.org/abs/1806.10698>.

[rx26] Summerton N. The medical history as a diagnostic technology. *Br J Gen Pract.* 2008; 58(549)

[rx27] Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med.* 2013;10(10):e1001531.

[rx28] Liu, X., Cruz Rivera, S., Moher, D. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 26, 1364–1374 (2020). <https://doi.org/10.1038/s41591-020-1034-x>

[HPO] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglou, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, Peter N Robinson, The Human Phenotype Ontology in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D1207–D1217, <https://doi.org/10.1093/nar/gkaa1043>, <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 3] 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	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 <i>understand</i> 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	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	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	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	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	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	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	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 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](#) 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), Juan Beleño (jbeleno@1doc3.com) and María Fernanda González (mgonzalez@1doc3.com)

Ada Health GmbH

[Ada Health GmbH](#) 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), Shubhanan Upadhyay (shubs.upadhyay@ada.com),

Involved before: Andreas Kühn, Clemens Schöll, Johannes Schröder, Sarika Jain, Isabel Glusman, Ria Vaidya, 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), , Adam Baker (adam.baker@babylonhealth.com)

Nathalie Bradley-Schmieg (nathalie.bradley1@babylonhealth.com)

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](#) 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), Ernesto Hernández Cura (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), Hoan Dinh (hoan.dinh@deepcare.io), Anh Phan (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).

Flo Health

Flo Health is international company with offices London, Villnius, Minsk, Cyprus and USA. We are focused purely on women’s health and our aim is to help women and girls prioritise their health by giving access to expert information, knowledge and support. We encourage our users to better understand how physiology affects their wellbeing. We are mainly a B2C company, available in 22

languages, we offer products including a menstrual cycle tracker, symptoms tracking and predictions, and a dialog service that provides potential differentials for symptoms they are experiencing. We also have a very large content library for users to use and educate themselves on conditions and symptoms.

People actively involved: Dr Anna Klepchukova (CMO, a_klepchukova@flo.health) and Dr Saddif Ahmed (Medical Director, s_ahmed@flo.health)

Infermedica

[Infermedica](#), 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), Piotr Orzechowski (piotr.orzechowski@infermedica.com), Jakub Winter (jakub.winter@infermedica.com), Michał Kurtys (michal.kurtys@infermedica.com)

Inspired Ideas

[Inspired Ideas](#) 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](#), 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), Megan Allen (megan@inspiredideas.io)

Isabel Healthcare

[Isabel Healthcare](#) 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)

Kahun

Kahun is an Israeli based med-tech venture, founded in 2018, developed an AI virtual clinical intake technology. Kahun has built an evidence-based medical knowledge graph (20M+ relations) and an AI engine that utilizes the graph to generate real-time insights. It enables Kahun to perform a patient interview, and supply a patient decision support dashboard to the provider.

People actively involved: Michal Tzuchman Katz (michal@kahun.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)

Your.MD Ltd

[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), Matteo Berlucchi (matteo@your.md), Martin Cansdale (martin@your.md), Audrey Menezes (audrey@your.md)

Involved before: Rex Cooper
