|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | **FG-AI4H-D-016** | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | N/A | Shanghai, 2-5 April 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Symptom Topic Driver | | |
| **Title:** | | Standardized Benchmarking for AI-based symptom assessment. | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Henry Hoffmann Ada Health GmbH Germany | | Tel: +49 177 6612889 Email: [henry.hoffmann@ada.com](mailto:henry.hoffmann@ada.com) |
| **Contact:** | | Andreas Kühn Ada Health GmbH Germany | | Tel: +49 30 60031987 Email: [andreas.kuehn@ada.com](mailto:andreas.kuehn@ada.com) |
| **Contact:** | | Jonathon Carr-Brown Your.MD UK | | Tel: +44 7900 271580 Email: [jcb@your.md](mailto:jcb@your.md) |
| **Contact:** | | Matteo Berlucchi Your.MD UK | | Tel: +44 7867 788348 Email: [matteo@your.md](mailto:matteo@your.md) |

|  |  |
| --- | --- |
| **Abstract:** | This document specifies a standardized benchmarking for AI-based symptom assessments. It follows the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6) and covers all scientific, technical and administrative aspects relevant for setting up this benchmarking. The creation of this document is an ongoing process until it will be finally approved by the Focus Group. This draft will be a continuous Input- and Output-Document. |
| **Change Notes:** | Version 1.0  This is the initial draft version of the TDD. As a starting point it merges the input documents FGAI4H-A-020, FGAI4H-B-021, FGAI4H-C-019, and FGAI4H-C-025 and fits them to the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6). The focus was especially on the following aspects:   * Introduction to topic and ethical considerations * Workflow proposal for Topic group * Overview of currently available AI-based symptom assessment applications started * Prior works on benchmarking and scientific approaches including first contributions by experts joining the topic. * Brief overview of different ontologies to describe medical terms and diseases |

# 

**Table of Contents**

[**1 Introduction**](#_yf3fhnt5cjli) **5**

[1.1 AI-based Symptom Assessment](#_1ci93xb) 5

[1.1.1 Relevance](#_9udlyca9d4mq) 5

[1.1.2 Current Solutions](#_g5qpj4sthqww) 5

[1.1.3 Impact of AI-based Symptom Assessment](#_dkdk00jmavmb) 6

[1.1.4 Impact of Introducing a Benchmarking Framework for AI-based Symptom Assessment](#_c6slbj8rjz6l) 6

[1.2 Ethical Considerations](#_3whwml4) 7

[**2 AI4H Topic Group on “AI-based Symptom Assessment”**](#_2bbacaul1rd3) **8**

[2.1 General Mandate of the Topic Group](#_3l57wsidlzu5) 9

[2.2 Topic Description Document](#_ytvgh6q9nsgy) 9

[2.3 Subtopics](#_o1aeed3h3bls) 9

[2.4 Topic Group Participation](#_b0oxkxyc9) 10

[2.5 Status of this Topic Group](#_bb4hjmjconbk) 10

[2.6 Next Meetings](#_d7ijh2xxhwst) 10

[**3 Existing AI Solutions**](#_hlihxu1gimgx) **10**

[3.1 Existing Systems for AI-based Symptom Assessment](#_44mohcuivca6) 10

[3.1.1 Topic Group member Systems for AI-based Symptom Assessment](#_51ldpjwipgxe) 10

[3.1.2 Other Systems for AI-based Symptom Assessment](#_7gzgh15fve27) 11

[3.2 Input Types](#_eodei16upk1u) 12

[3.3 Output Types](#_f3cfvy4kvss3) 13

[3.3.1 Pre-Clinical Triage](#_wxvlvzycqado) 13

[3.3.2 Differential Diagnosis](#_8t03jbjte5fn) 14

[3.3.3 Next Step Advice](#_m2cx4lsdsvak) 14

[3.4 Scope Dimensions](#_r4s5xqtuib2a) 14

[3.4 Additional Relevant Dimensions](#_e7nnxl8q3f9w) 14

[**4 Existing work on benchmarking**](#_qsh70q) **15**

[4.1 Scientific Publications on Benchmarking AI-based Symptom Assessment Applications](#_fd33s9nwdsh0) 15

[4.2 Clinical Evaluation of AI-based Symptom Assessment](#_wf0zipvg921f) 16

[4.3 Benchmarking Publications outside Science](#_2snkw7y4nsfj) 16

[4.4 Existing Regulations](#_y4njxlqnsmji) 17

[4.5 Internal Benchmarking by Companies](#_jexmf83oojni) 17

[4.6 Existing AI Benchmarking Frameworks](#_on3s5v2da8pu) 18

[**5 Method of the Minimal Viable Benchmarking**](#_j1s3n2rtp1vr) **18**

[5.1 Architecture and Methodology Overview](#_6n83p9txclog) 18

[5.2 AI Input Data to use for the MVB](#_1pxezwc) 19

[5.3 AI Output Data to use for the MVB](#_otqsae2805pu) 20

[5.4 Symptom Assessment AI Benchmarking Interface](#_9ro9u53xlro7) 21

[5.5 API Input Data Format](#_fab9r0ps7h3l) 21

[5.6 API Output Data Format](#_9scjlo5zunvx) 21

[5.7 Benchmarking Dataset Collection](#_stcxnwe4d1w9) 21

[5.8 Benchmarking Dataset Format](#_9gsdlq3uxzih) 21

[5.9 Scores & Metrics](#_8hvjzsq44b7d) 21

[5.10 Reporting Methodology](#_ihv636) 22

[5.11 Technical Architecture of the Official Benchmarking System](#_wl3r5pqyzjsx) 22

[5.12 Technical Architecture of the Continuous Benchmarking System](#_6tm9kyrs2jt9) 22

[5.13 Benchmarking Operation Procedure](#_al7de9c2zanu) 22

[**6 Results**](#_32hioqz) **22**

[**7 Discussion**](#_1hmsyys) **22**

[**Appendix A: Declaration of Conflict of Interest**](#_pr53dcsqrrb) **23**

[**Appendix B: Glossary**](#_q2zzw2rqkl40) **23**

[**Appendix C: References**](#_1w0kchntb7ko) **24**

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 1 Introduction

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a standardized benchmarking approach for AI-based symptom assessment (AISA) applications.

The document is structured in seven chapters. Chapter 1 introduces the topic and outlines its relevance and the potential impact that a benchmarking will have. Chapter 2 provides an overview of the Topic Group that created this document and will implement the actual benchmarking as part of the AI4H Focus Group. Chapter 3 collects all benchmarking-relevant background knowledge on the state-of-the-art of existing AI-based symptom assessment systems. Chapter 4 describes the approaches for assessing the quality of such systems and details that are likely relevant to set up a new standardized benchmarking. Chapter 5 specifies the actual benchmarking methodology at a level of detail that includes technological and operational implementation. Chapter 6 summarizes the results of the different iterations to perform benchmarking according to the specified methods. Chapter 7 discusses learnings from working on this document, the implementation of the methods and the performed benchmarking. It also discusses insights from using the benchmarking results.

## 1.1 AI-based Symptom Assessment

### 1.1.1 Relevance

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.[[1]](#footnote-1) 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.[[2]](#footnote-2) 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.

While the problem in low and middle income countries (LMIC) 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[[3]](#footnote-3). 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 less resources.

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

### 1.1.2 Current Solutions

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. 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. Meanwhile, other studies show that internet self-searches are more likely to return inappropriate severe diagnoses like cancer.

### 1.1.3 Impact of AI-based Symptom Assessment

In recent years, a 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 or sex or other risk factors.

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 some 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), likely differential diagnoses (DD), and advice on what to do next.

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.

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 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).

### 1.1.4 Impact of Introducing a Benchmarking Framework for AI-based Symptom Assessment

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. Very few papers 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 (see Chapter 4).

The implementation of a standardized benchmarking for AISA applications by the WHO/ITU’s AI4H-FG 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. It will enable 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.[[4]](#footnote-4)

According to the current version of the thematic classification scheme document C-027 of the FG, the categorization of the Topic “AI-based symptom assessment” is applicable as follows:

|  |  |  |
| --- | --- | --- |
| Level 1 | Public Health (Level-1A) | 1.5 Health surveillance  1.6. Health emergencies  1.9. Communicable diseases  1.10. Non-communicable diseases  sub-classes applicable:   * 1 epidemiology * 3 biostatistics * 4 health services delivery * 6 community health * 8 health economics * 9 informatics * 10 public health interventions |
| Clinical Health  (Level-1B) | 1.2. Diagnosis  sub-classes applicable:  1-35 (potentially all specialities) |
| Level 2 (Artificial Intelligence) | | 3. Knowledge representation and reasoning   * 3.1. default reasoning * 3.3. ontological engineering   4. Artificial Intelligence   * 4.1. generative models * 4.2. autonomous systems * 4.3. distributed systems |
| Level 3 (nature of data types) | | 1. Anonymized electronic health record data  3. Non-medical data (socio economic, environmental, etc)  4. lab test results (later)  -. structured medical information (e.g. based on ontology) |
| Level 4 (origin of the data) | | 3. PHR (Personal health record)  4. Medical Device |
| Level 5 (data collectors) | | 1. Service provider |

## 1.2 Ethical Considerations

This section needs review and amendment by a health-AI-ethics field experts

**Ethical implications using AI systems in health**

Currently there is no sufficient evidence yet that AISA applications are safe to use and can in any way compete with the gold standard of medical professional assessment. At the same time such systems are used by millions of smartphone users every day. While the majority of users gets an appropriate advice, these systems also fail on some cases. This may lead to users being influenced to act while there is no need to act or stopped from acting e.g. by not seeing a doctor while they should see one. At the same time large parts of the world population have no easy access do doctors so that AI-based systems are their only option. It is also noteworthy that even doctors have a high variance in their performance. So far there is no reliable data available to compare the ethical implications of using AI-based symptom assessment with the consultation of doctors or not having access to health care at all. Setting up a benchmarking of AI-bases symptom assessment will help in making the quality of AI systems more transparent and allow to better assess the ethical implications.

**Ethical implications of benchmarking and data acquisition**

Even with introducing a benchmarking process that proves a certain overall quality, systems fail on individual cases. While providing a more sound and transparent foundation for decision making the introduction of benchmarking will lead to a broader application of such systems. While this will help many, this will also increase the overall number of people that will experience a sub-optimal performance of such systems.

Benchmarking of AI-based symptom assessment systems requires labeled test data. Depending on the approach for creating the data set this might involve real anonymized patient cases. In this case the high standards for privacy and protection of the personal health data need to be met. Since this is an issue of high importance, the Focus Group actively works on making sure that the used data meets high standards for ethics and protection of the personal health data.

**Implications of using AI systems in inappropriate context**

Compared to medical professional assessment and conventional diagnostics an AI system should in the end lead to both, an increase in specificity and sensitivity in the context of diagnosis. A trade-off of specificity against sensitivity is possible, in certain context. This context has to be made clear before establishing a benchmark. So, for emergency settings it might be advisable to increase sensitivity even if specificity is slightly reduced. The benchmark has to be adapted to these settings. In the end to be judged “better” than conventional diagnostics, an AI system or medical professionals using this system have to prove the superiority to the prior gold standard. The risk of misinterpreting and misusing the benchmarking results for justifying the application of AI in an inappropriate context has to be addressed by the Focus Group.

# 2 AI4H Topic Group on “AI-based Symptom Assessment”

The first chapter highlighted the potential of AISA to help solving important health issues and that the creation of a standardized benchmarking would provide decision makers with the necessary insights to successfully address these challenges. To develop this benchmarking framework, the AI4H Focus Group decided at the January 2019 meeting C in Lausanne to create the Topic Group “AI-based symptom assessment”. It was based on the “symptom checkers” use case, which was accepted at the November 2018 meeting B in New York building on proposals by Ada Health:

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

and on a similar initiative by Your.MD:

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

In addition to the “AI-based symptom assessment” Topic Group, the ITU/WHO Focus Group created nine other Topic Groups for additional standardized benchmarking of AI:

* Cardiovascular disease risk prediction
* Dermatology
* Falls among the elderly
* Histopathology
* Neuro-cognitive diseases
* Ophthalmology (retinal imaging diagnostics)
* Psychiatry
* Snakebite and snake identification
* Tuberculosis

As the work by the Focus Group continues, new Topic Groups will be created. To organize the Topic Groups, for each topic the Focus Group chose a topic driver. The exact responsibilities of the topic driver are still to be defined and are likely to change over time. The preliminary and yet-to-confirm list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document.
* Reviewing the input documents for the topic and moderating the integration in a dedicated session at each Focus Group meeting.
* Organizing regular phone calls to coordinate work on the topic description document between meetings.

During meeting C in Lausanne, Henry Hoffmann from Ada Health was selected topic driver for the “AI-based Symptom Assessment” Topic Group.

## 2.1 General Mandate of the Topic Group

The Topic Group is a concept specific to the AI4H-FG. The preliminary responsibilities of the Topic Groups are:

1. Provide a forum for open communication among various stakeholders
2. Agree upon the benchmarking tasks of this topic and scoring metrics
3. Facilitate the collection of high quality labeled test data from different sources
4. Clarify the input and output format of the test data
5. Define and set-up the technical benchmarking infrastructure
6. Coordinate the benchmarking process in collaboration with the Focus Group management and working groups

## 2.2 Topic Description Document

The primary output of each Topic Group is the topic description document (TDD) specifying all relevant aspects of the benchmarking for the individual topics. **This document is the TDD for the Topic Group on “AI-based symptom assessment” (AISA)**. The document will be developed cooperatively over several FG-AI4H meetings starting from meeting D in Shanghai. Suggested changes to the document will be submitted as input documents for each meeting. The relevant changes will then be discussed and integrated into an official output document until the TDD ready for the first official benchmarking.

## 2.3 Subtopics

Topic groups summarize similar AI benchmarking use cases to limit the number of use case specific meetings at the Focus Group meetings and to share similar parts of the benchmarking. However, in some cases, it is expected that inside a Topic Group different subTopic Groups can be established to pursue different topic-specific specializations. The AISA Topic Group will start without separate subTopic Groups. They will introduce once participants with different benchmarking requirements join the Topic Group. It is expected to introduce subTopic Groups for “patient facing symptom assessment” and “clinical symptom assessment” as soon as partners that are interested in benchmarking systems for clinical symptom assessment join the group.

## 2.4 Topic Group Participation

The participation in both the focus and Topic Group is generally open and free of charge. Anyone who is from a member country of the ITU may participate. On the 14. of March 2019 the ITU published an official “call for participation” document outlining the process for joining the Focus Group and the Topic Group. For this topic, the corresponding call can be found [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/symptom.aspx).

## 2.5 Status of this Topic Group

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 expects at the upcoming meetings and the in between.

## 2.6 Next Meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* E: Geneva, Switzerland; ​29 May-1 June ​2019
* F: Zanzibar, Tanzania; 2-5 September 2019
* G: New Delhi, India; November 2019
* H: Brasilia, Brazi; January 2020

An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

* Tools/process of TG cooperation - to be filled out according to FG regulations
* TG interaction with WG, FG- to be filled out according to FG regulations

# 3 Existing AI Solutions

* Some words on the history of expert systems for diagnostic decision support and how it lead to the new generation of AI systems (INTERNIST, EXPERT, GLAUCOM, CASNET, … )

## 3.1 Existing Systems for AI-based Symptom Assessment

This section presents a table with the providers currently available and known to the Topic Group. The table summarizes 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. The table is split into members of the Topic Group and non-members since the initial benchmarking will most likely start with the providers that joined the Topic Group.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Provider** | **System** | **Input** | **Output** | **Scope** |
| [Ada Health GmbH](https://ada.com/) | Ada app | * Age, sex, risk factors * Free text PC search * Discrete answers to dialog questions for additional symptoms including attribute details like intensity | * Differentials for PC * Pre-clinical triage * Shortcuts in case of immediate danger | * Worldwide * English, German, Spanish, Portugues * Top 1300 conditions * For smartphone users * Android/iOS |
| [Your.MD Ltd](https://www.your.md/) | Your.MD app | * Age, sex,medical risk factors, * Chatbot free text input * User consultation output (report) | * Differentials for PC * Pre-clinical triage * Shortcuts in case of immediate danger * Condition information * Recommendation of appropriate local services and products * Medical factors | * Worldwide * English, * Top 370 conditions (building to 500). * For smartphone users Android /iOS and web and messaging groups Skype etc |

## 3.1.2 Other Systems for AI-based Symptom Assessment

The list of systems of providers that have not joined the Topic Groups is most likely incomplete. Suggestions for systems to add are appreciated. The same applies for suggestions for the missing columns. 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provider** | **System** | **Input** | **Output** | **Scope** | **Comments** |
| [Aetna](https://www.aetna.com/individuals-families.html) | [Symptom checker](https://www.healthwise.net/aetna/Content/CustDocument.aspx?XML=STUB.XML&XSL=CD.FRONTPAGE.XSL) |  |  |  |  |
| AHEAD Research | [Symcat](http://www.symcat.com/) |  |  |  |  |
| [Babylon Health](https://www.babylonhealth.com/) |  |  |  |  |  |
| [Buoy Health, Inc](https://www.buoyhealth.com) |  |  |  |  |  |
| [DocResponse](https://www.docresponse.com/) | [DocResponse](https://www.docresponse.com/) |  |  | * for doctors |  |
|  | Doctor Diagnose |  |  | * Android |  |
| [Drugs.com](https://www.drugs.com) | [Symptom Checker](https://www.drugs.com/symptom-checker/) |  |  | * Triage | Harvard Health decision guide used |
| [EarlyDoc](https://www.earlydoc.com/) |  |  |  | * Web |  |
| [FamilyDoctor.org](https://familydoctor.org/) | [Symptom Checker](https://familydoctor.org/your-health-resources/health-tools/symptom-checker/) |  |  | * Web |  |
| [Healthline](https://www.healthline.com/) | [Symptom Checker](https://www.healthline.com/symptom-checker) |  |  |  |  |
| [Healthtap](https://www.healthtap.com/) | [Symptom Checker](https://www.healthtap.com/member/login) (for members) |  |  |  |  |
| [Infermedica](https://infermedica.com/) | (Symptomate) |  |  | * worldwide * Arabic, English, French, German, Italian, Polish, Portuguese, Portuguese (Brazilian), Simplified Chinese, Slovak, Spanish, Russian, Ukrainian |  |
| [Isabel Healthcare](https://symptomchecker.isabelhealthcare.com/) | [Isabel Symptom Checker](https://symptomchecker.isabelhealthcare.com/suggest_diagnoses_advanced/landing_page) |  |  |  |  |
| [K Health](https://www.khealth.ai/) | K app chatbot |  |  |  |  |
| [Mayo Clinic](https://www.mayoclinic.org/) | [Symptom Checker](https://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075) |  |  |  |  |
| [MDLive](https://www.mdlive.com/) | Symptom checker on MDLive app |  |  |  |  |
| [MEDoctor](https://www.medoctor.io/) | [Symptom Checker](https://www.medoctor.com/Freemium/interview) |  |  |  |  |
| [Mediktor](https://www.mediktor.com/en) | [Web-based symptom checker,](https://www.mediktor.com/en) or Mediktor app |  |  |  |  |
| [NetDoktor](https://www.netdoktor.de/) | [Symptom Checker](https://www.netdoktor.de/symptom-checker/) |  |  |  |  |
| [PingAn](http://www.pingan.cn/en/index.shtml) | Good Doctor app |  |  |  |  |
| [Sharecare, Inc.](https://www.sharecare.com/) | [AskMD](https://www.sharecare.com/askmd/get-started) |  |  |  |  |
| [Symptify](https://symptify.com/) | [Symptom Checker](https://symptify.com/) |  |  |  |  |
| [WebMD](https://www.webmd.com/) |  |  |  |  |  |

## 3.2 Input Types

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. The following table gives an overview of the different input types used by the listed AI systems:

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

## 3.3 Output Types

The corresponding outputs of existing AI systems are presented in the following table:

|  |  |  |
| --- | --- | --- |
| **Output Type** | **Short Description** | **Number of Systems** |
| 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:

## 3.3.1 Pre-Clinical Triage

This section will be discussed in the next Topic Group meeting

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

## 3.3.2 Differential Diagnosis

* to be written

## 3.3.3 Next Step Advice

* to be written

## 3.4 Scope Dimensions

The table of existing solutions also lists the scope of the intended application of these systems. Analyzing 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.

***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. The benchmarking therefore needs to be categorized by different condition sets to account for the different system capabilities.

***Age Range***

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

***Languages***

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

## 3.4 Additional Relevant Dimensions

Beside 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 theses 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.

# 4 Existing work on benchmarking

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 sub categories that will be discussed in their own subsections.

## 4.1 Scientific Publications on Benchmarking AI-based Symptom Assessment Applications

While rare, there exist a few publications 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:

* 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>
* 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>
* 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

## 4.2 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.

Please add more publications. For the existing ones the sections should discuss, scores, metrics, methodology of case creationion etc.

**4.2.1 “A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. Emergencias”**

One report was published in 2017 assessing a single AI-based symptom assessment in a Spanish Emergency Setting.[[5]](#footnote-5) 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.

**4.2.2 “Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence.”**

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.[[6]](#footnote-6)

**4.2.3 “Evaluating the potential impact of Ada DX in a retrospective study.”**

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%.[[7]](#footnote-7)

## 4.3 Benchmarking Publications outside Science

In addition to scientific benchmarking attempts, there are several newspaper articles reporting tests of primarily user-facing symptom assessment applications. While not always following scientific standards, they provide relevant insights on how such systems can be compared. The following section gives an overview of some of these articles (without claim of completeness).

**4.3.1 “Can you really trust the medical apps on your phone?”**

The Wired magazine addressed the issue of “symptom checkers” creating a test set for the applications Ada, Babylon and Your.MD for a small sample of common diseases and based on symptoms described by the NHS Choices website. They not only focussed on the result but also on the advice the application gave after the assessment and the questions generated by the apps and how they were asked.[[8]](#footnote-8)

continue with other articles

## 4.4 Existing Regulations

Complementary to explicit benchmarking attempts, in many countries there is strict regulation of health-related products in place. While the original regulatory focus was more on hardware devices, the regulatory environment has been rapidly adapting to the needs of software. This section reviews the existing regulation to collect criteria that could be part of a standardized automatic benchmarking.

* medical product regulation and the upcoming class II requirement
* FDA(C)
* (CE)
* clinical trials, evidence levels (RCT’s etc.)
* scores & metrics used

## 4.5 Internal Benchmarking by Companies

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 position***

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.

discuss scores & metrics used

## 4.6 Existing AI Benchmarking Frameworks

Triggered by the hype around AI, recent years have seen the development of several benchmarking platforms where AIs can compete for the best performance on a given dataset. Document [C031](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-031.pptx?d=wda212412d98541f5a4fbb4757b620828) provides a list of the available platforms. While not specific for symptom assessment they provide important examples for many aspects of benchmarking ranging from operational details, over scores & metrics, leaderboards, 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.

* analyse kaggle
* analyse crowdAI
* scores & metrics
* architecture
* submission process
* leaderboards
* reports
* data and AI security
  + special focus on protection of the AIs since this seems to be not a concern for these communities
* the role of [FG-AI4H-C-106](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FG-AI4H-C-106.docx) and that we might be able to use an existing technology

# 5 Method of the Minimal Viable Benchmarking

Chapter 5 specifies the methodology, technology and protocols necessary for the benchmarking of AI-based symptom assessment systems. Due to the complexity of a holistic standardized benchmarking framework for AI-based symptom assessment, this specification begins with a **“minimal viable benchmarking”** **(MVB)** that will iteratively extend based on the insights learned from each benchmarking iteration. The focus of chapter 5 is on specifying a concrete benchmarking. Theoretical background and the different options for the numerous decisions to be taken are supposed to be discussed in chapter 4.

## 5.1 Architecture and Methodology Overview

Fig. 1 shows the general generic benchmarking architecture defined by the Focus Group that will serve as the basis for the symptom assessment topic. The different components will be explained in the following sections (figure will be adapted to Topic):

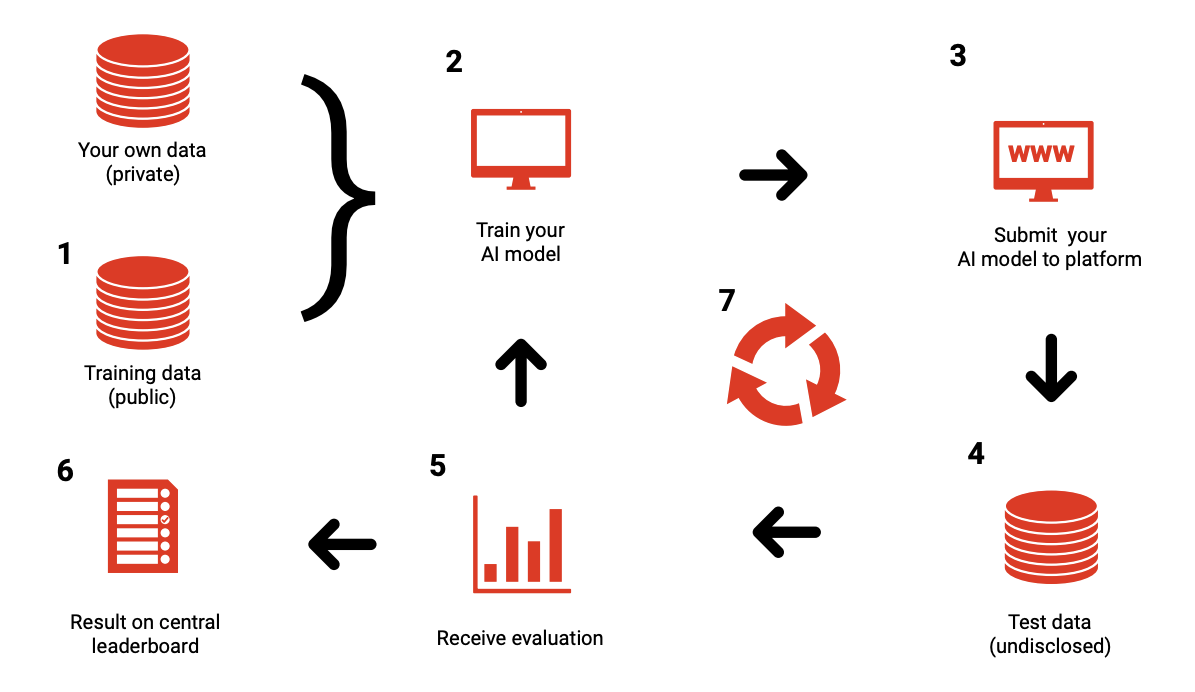


Fig. 1: General framework proposed by Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL during workshop associated to FG Meeting A[[9]](#footnote-9)

**(2)** Every benchmarking participant creates its own system based on its technology of choice. This will likely involve machine learning techniques relying on private and/or public data sets **(1)**. In contrast to other Topic Groups, there are currently no public training datasets available. Given the large number of conditions and that some are very rare, it is unlikely that large public training data sets will soon be available.

As part of the benchmarking process, participants have to implement an application program interface (API) endpoint that accepts test cases and returns the corresponding computed output **(3)**.

The actual benchmarking will then be performed by the United Nations International Computing Centre (ICC) benchmarking infrastructure by sending test cases **(4)** without the labels to the AI and recording the corresponding results. To generate a report for each system, **(5)** the benchmarking system will then compute the previously agreed-upon metrics and scores based on the output datasets. The results of the different DSAAs can finally be presented in a leaderboard **(6)**.

In addition to official benchmarking on undisclosed datasets and submission of AI for testing, there will also be a continuous benchmarking process **(7)** that uses an open test dataset and API endpoints hosted by the AI providers on their systems. This will facilitate the testing of API endpoints and required data format transformations, while also providing a rough estimate of performance before the official benchmarking.

## 5.2 AI Input Data to use for the MVB

Section 3.2 outlined the different input types of the known symptom assessment systems. For the MVB the following input types will be selected:

* **Profile information** - General background information on the user including age and sex (and later additional information e.g. location).
* **Presenting complaints** - The initial health problem(s) that the user seeks an explanation for in form of symptoms with detailing attributes e.g. “side: left” or “intensity: moderate”
* **Additional symptoms** - Additional symptoms and factors including detailing attributes and ability to answer with “no” and “don’t know”

To make this information usable for the MVB the Topic group or the Focus Group, respectively will have to agree on a standardized way to describe these inputs. Currently there are various classification systems for these medical terms available, each with own Pros and Cons.

The following list gives an overview of some of these classification systems and will be extended in more detail (without claim of completeness):

**SNOMED Clinical Terms**

SNOMED CT is a hierarchical classification of medical Terms, which is considered to be computer processable. Maintenance and distribution takes place by 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. The hierarchy and connections between the terms are multidimensional and partly redundant. SNOMED CT is also currently adapted to fit the needs of ICD-11 to link both classification systems (see below).

**Human Phenotype Ontology (HPO)**

HPO has been created in the context of describing phenotype deviation in human due to hereditary diseases. It is an ontology that describes phenotypic abnormalities and includes currently around 13,000 terms that are linked with hereditary diseases.

**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.

**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 and DSM (see below) and links these systems creating an ontology of medical terms.

## 5.3 AI Output Data to use for the MVB

Of the output types listed in 3.3 Output Types the MVB we benchmark the following types:

* **Differential Diagnosis** - The most likely explanations for the initial presenting complaints of the patient.
* **Pre Clinical Triage** - The general classification of what to to next e.g. “see doctor today”

As for the Input data, the output data has to be described in a standardized way for the MVB. The following list presents main established classification systems and describes the main features and usage of these

**International Statistical Classification of Diseases and Related Health Problems (ICD)**

The ICD system is the worldwide mostly used system for coding and describing diagnoses. It dates back until the 19th century and it was under revision from 2007 on from ICD-10 to ICD-11 which was accomplished recently. The coding system is based on agreement of a huge network of experts and working groups. The version ICD-11 holds a complex underlying semantic network of terms and thus connects the different entities in a new way and is referred to as “digital ready”.

**Diagnostic and Statistical Manual of Mental Disorders (DSM)**

This system (currently DSM-5) is widely used in the US and worldwide for the classification of mental disorders. It is maintained by the American Psychiatric Association.

**Triage / Advice - Scales**: to be defined / agreed upon

## 5.4 Symptom Assessment AI Benchmarking Interface

* TODO: specify an JSON REST API endpoint for benchmarking
* version

## 5.5 API Input Data Format

* TODO: JSON Format

## 5.6 API Output Data Format

* TODO: JSON Format

## 5.7 Benchmarking Dataset Collection

* raw data acquisition / acceptance
* test data source(s): availability, reliability,
* labelling process / acceptance
* bias documentation process
* quality control mechanisms
* discussion of the necessary size of the test data set for relevant benchmarking results
* specific data governance derived by general data governance document (currently C-004)

## 5.8 Benchmarking Dataset Format

* TODO: JSON Format
* mainly the JSON format for the API
* additional metadata

## 5.9 Scores & Metrics

* which metrics & scores to use for benchmarking
  + probability to see the right diagnosis among the first N results
  + consider the probability reported by AI
  + consider not only the probability of the correct one but also probability of the other wrong ones
  + consider conditions explicitly ruled out by e.g. sex/age
  + consider how far a diagnosis is off and how dangerous this is
  + consider if all relevant questions have been asked
  + should the response time be measured?
* considering relation to parameters stakeholders need for decision making
  + some stakeholders ask for PPV, FP, TP, F-Score, etc.
* considering scores that providers use
* considering the scope providers designed their solutions for
  + group by all dimensions from 3.4 Scope Dimensions
* considering the state of the art in RCT, statistics, AI benchmarking etc.
* considering bias transparency
  + group results by source of dataset parts in case we use different datasets

## 5.10 Reporting Methodology

* Report publication in papers or as part of ITU documents
* identify journals that could be interest in publication (to be discussed)
* Online reporting
* interactive dashboards (it might be that due to the large number of dimensions to consider an interactive dashboard is the only way to fully understand all details.
* public leaderboards vs. private leaderboards
* collect opinion on this once more AI providers joined the Topic Group
* Credit-Check like on approved sharing with selected stakeholders
* Report structure including an example
* Frequency of benchmarking
* once per quarter?

## 5.11 Technical Architecture of the Official Benchmarking System

* TODO
* servers, systems,
* IIC infrastructure
* implementation
  + publishing the benchmarking software on github would be transparent

## 5.12 Technical Architecture of the Continuous Benchmarking System

* TODO
* in comparison to the official system

## 5.13 Benchmarking Operation Procedure

* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations
* how long is the data stores
* how long are AI systems stored

# 6 Results

Chapter 6 will outline the results from performing the benchmarking based on the methodology specified in this document. Since the benchmarking is still in its specification phase, there are no results available yet. Depending on the progress made on this document, first preliminary test benchmarking results on small public data sets are expected by the end of 2019. The first official results form a MVB are expected in early 2020.

# 7 Discussion

This chapter will discuss the insights from the first MVB results described in chapter 6 as soon as they are available.

# Appendix A: Declaration of Conflict of Interest

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

***Ada Health GmbH***

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

Involved people: Henry Hoffmann (henry.hoffmann@ada.com), Andreas Kühn (andreas.kuehn@ada.com), Johannes Schröder (johannes.schroeder@ada.com)

***Your.MD Ltd***

[Your.MD](https://www.your.md/) is a Norwegian company based in London. We have four years’ experience in the field, a team of 40 people and currently delivers next steps health advice based on symptoms and personal factors to 570,000 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.

Involved people: Jonathon Carr-Brown ([jcb@your.md](mailto:jcb@your.md)), Matteo Berlucchi(matteo@your.md)

# Appendix B: Glossary

This section lists all the relevant abbreviations and acronyms used in the document. If there is an external source

* **AI** - [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) - While the exact definition is highly controversial, in context of this document it refers to a field of computer science working on machine learning and knowledge based technology that allows to *understand* complex (health related) problems and situations at or above human (doctor) level performance and providing corresponding insights (differential diagnosis) or solutions (next step advice, triage).
* **AuI** - Augmented Intelligence
* **AI4H** - AI for health - An [ITU-T SG16 Focus Group](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx) founded in cooperation with the WHO in July 2018.
* **AISA** - AI-based symptom assessment - The abbreviation for the topic of this Topic Group.
* **API** - [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) - the software interface systems communicate through.
* **DD** - Differential Diagnosis -
* **PC -** [Presenting Complaint](https://en.wikipedia.org/wiki/Presenting_problem) - The health problems the user of an symptom assessment systems seeks help for.
* **FG** - [Focus Group](https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx) - An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas.
* **IIC** - International Computing Centre - The United Nations data center that will host the benchmarking infrastructure.
* **ITU** - [International Telecommunication Union](https://www.itu.int) - The United Nations specialized agency for information and communication technologies – ICTs.
* **LMIC** - Low and Middle Income Countries
* **MVB** - minimal viable benchmarking
* **NGO** - [Non Governmental Organization](https://en.wikipedia.org/wiki/Non-governmental_organization) - NGOs are usually non-profit and sometimes international organizations independent of governments and international governmental organizations that are active in humanitarian, educational, health care, public policy, social, human rights, environmental, and other areas to affect changes according to their objectives. (from Wikipedia.en)
* **SDG** - [Sustainable Development Goals](https://www.un.org/sustainabledevelopment/) - The United Nations Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. Currently there are 17 goals defined. SDG 3 is to “Ensure healthy lives and promote well-being for all at all ages” and is therefore the goal that will benefit from the AI4H Focus Groups work the most.
* **TDD** - Topic Description Document - Document specifying the standardized benchmarking for a topic FG AI4H Topic Group works on. This document is the TDD for the Topic Group “AI-based symptom assessment”.
* **Triage** - A [medical term](https://en.wikipedia.org/wiki/Triage) describing a heuristic scheme and process for classifying patients based on the severity of their symptoms. It is primarily used in emergency settings to prioritize patients and to determine the maximum acceptable waiting time until actions need to be taken.
* **TG** - Topic Group - Structures inside AI4H FG summarizing similar use cases and working on a TDD specifying the setup of a standardized benchmarking for the corresponding topic. The Topic Groups have been first introduced by the FG at the Meeting C, January 2019 in Lausanne. See protocol FG-AI4H-C-10x for details.
* **WHO** - [World Health Organization](https://www.who.int) - The United Nations specialized agency for international public health.

A new term should be introduced “... this is a text with a new term (NT) ...” and then added to the glossary list in the format “NT - new term - Description of the new term.”, possibly with a link e.g. Wikipedia.

# Appendix C: References

This section lists all the references to external sources cited in this document. Please use Vancouver style for adding references, if possible.

1. WHO. “Global health workforce shortage to reach 12.9 million in coming decades”.WHO 2013. <http://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en/>
2. 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
3. Pillay N. The economic burden of minor ailments on the national health service in the UK. SelfCare 2010; 1:105-116
4. United Nations. Goal 3: Ensure healthy lives and promote well-being for all ages. https://www.un.org/sustainabledevelopment/health/
5. 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
6. 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. https://www.nature.com/articles/s41591-018-0335-9
7. 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 2019.
8. Burgess M. Can you really trust the medical apps on your phone? Wired Magazine Online 2017-10-01; <https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy>
9. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/M\_Salath%C3%A9\_Presentation.pdf

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

1. WHO. “Global health workforce shortage to reach 12.9 million in coming decades”.WHO 2013. <http://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en/> [↑](#footnote-ref-1)
2. 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 [↑](#footnote-ref-2)
3. Pillay N. The economic burden of minor ailments on the national health service in the UK. SelfCare 2010; 1:105-116 [↑](#footnote-ref-3)
4. United Nations. Goal 3: Ensure healthy lives and promote well-being for all ages. https://www.un.org/sustainabledevelopment/health/ [↑](#footnote-ref-4)
5. 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 [↑](#footnote-ref-5)
6. 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 [↑](#footnote-ref-6)
7. 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. [↑](#footnote-ref-7)
8. Burgess M. Can you really trust the medical apps on your phone? Wired Magazine Online 2017-10-01; <https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy> accessed: 2019-03-20 [↑](#footnote-ref-8)
9. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/M\_Salath%C3%A9\_Presentation.pdf [↑](#footnote-ref-9)