|  |  |  |  |
| --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-G-017 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | | Plenary | New Delhi, 13-15 November 2019 |
| **DOCUMENT** | | | |
| **Source:** | | TG-Symptom Topic Driver | |
| **Title:** | | TDD update: TG-Symptom (Standardized Benchmarking for AI-based symptom assessment) | |
| **Purpose:** | | Discussion | |
| **Contact:** | | Henry Hoffmann Topic Driver 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) |
| **Contact:** | | Jason Maude Isabel Healthcare UK | Tel: +44 1428 644886 Email: [jason.maude@isabelhealthcare.com](mailto:jason.maude@isabelhealthcare.com) |
| **Contact:** | | Shubs Upadhyay Ada Health GmbH Germany | Tel: +44 7737 826528 Email: [shubs.upadhyay@ada.com](mailto:shubs.upadhyay@ada.com) |
| **Contact:** | | Yanwu XU Artificial Intelligence Innovation Business,  Baidu China | Tel: +86 13918541815 Fax: +86 10 59922186 Email: [xuyanwu@baidu.com](mailto:xuyanwu@baidu.com) |
| **Contact:** | | Ria Vaidya Ada Health GmbH Germany | Tel: +49 173 7934642 Email: [ria.vaidya@ada.com](mailto:ria.vaidya@ada.com) |
| **Contact:** | | Isabel Glusman Ada Health GmbH Germany | Email: [isabel.glusman@ada.com](mailto:isabel.glusman@ada.com) |
| **Contact:** | | Saurabh Johri  Babylon Health UK | Tel: +44 (0) 7790 601 032  Email: [saurabh.johri@babylonhealth.com](mailto:saurabh.johri@babylonhealth.com) |
| **Contact** | | Nathalie Bradley-Schmieg Babylon Health UK | Email: [nathalie.bradley1@babylonhealth.com](mailto:nathalie.bradley1@babylonhealth.com) |
| **Contact** | | Piotr Orzechowski Infermedica Poland | Tel: +48 693 861 163  Email: [piotr.orzechowski@infermedica.com](mailto:piotr.orzechowski@infermedica.com) |
| **Contact** | | Irv Loh, MD Infermedica USA | Tel: +1 (805) 559-6107  Email: [irv.loh@infermedica.com](mailto:irv.loh@infermedica.com) |
| **Contact** | | Jakub Winter Infermedica Poland | Tel: +48 509 546 836  Email: [jakub.winter@infermedica.com](mailto:jakub.winter@infermedica.com) |
| **Contact** | | Ally Salim Jr Inspired Ideas Tanzania | Tel: +255 (0) 766439764  Email: [ally@inspiredideas.io](mailto:ally@inspiredideas.io) |
| **Contact** | | Megan Allen Inspired Ideas Tanzania | Tel: +255 (0) 626608190  Email: [megan@inspiredideas.io](mailto:megan@inspiredideas.io) |
| **Contact** | | Anastacia Simonchik Visiba Group AB Sweden | Tel: +46 735885399 Email: [anastacia.simonchik@visibacare.com](mailto:anastacia.simonchik@visibacare.com) |
| **Contact** | | Sarika Jain Ada Health GmbH Germany | Email:[sarika.jain@ada.com](mailto:sarika.jain@ada.com) |
| **Contact** | | Yura Perov Babylon Health UK | Email: [yura.perov@babylonhealth.com](mailto:yura.perov@babylonhealth.com) |
| **Contact** | | Tom Neumark University of Oslo Norway | Email: [thomas.neumark@sum.uio.no](mailto:thomas.neumark@sum.uio.no) |
| **Contact** | | Rex Cooper Your.MD UK | Email: [rex@your.md](mailto:rex@your.md) |
| **Contact** | | Martin Cansdale Your.MD UK | Email: [martin@your.md](mailto:martin@your.md) |

|  |  |
| --- | --- |
| **Abstract:** | This document specifies a standardized benchmarking for AI-based symptom assessments. It follows the structure defined in FGAI4H-C-105 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 4.0 (to be 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 Appendix E with a complete list of all TG meetings and related documents * Added Martin Cansdale, Rex Cooper, Tom Neumark, Yura Perov, Sarika Jain, Anastacia Simonchik and Jakub Winter to author list and/or conflict of interest declaration and/or contributors. * Merged meeting F editing by ITU/TSB (Simão Campos)   **Version 3.0 (submitted as** [**FGAI4H-F-017**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-F-017.docx?d=wf55bdbee8809480ca0635f2aaf39624d&Source=https%3A%2F%2Fextranet%2Eitu%2Eint%2Fsites%2Fitu%2Dt%2Ffocusgroups%2Fai4h%2Fdocs%2FForms%2F190903%2Easpx) **for meeting F in Tanzania)**   * Added new TG members Infermedica, Deepcare and Symptify * Added 5.2 section on the MMVB work * Added 2.5.3 Status Update for Meeting F Submission * Updated 2.6 Next Meetings * Refined 3.5 Robustness details * Removed validation outside science   **Version 2.0 (submitted as** [**FGAI4H-E-017**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-017.docx?d=w2ddecda3a8ad48b6870e6c56a7689b1b) **for meeting E in Geneva)**   * Added new TG members Baidu, Isabel and Babylon to header and appendix A. * Added the list of systems that could not be considered in chapter 3 for transparency reasons as Appendix D. * Started a section on scores & metrics. * Refined triage section. * Started the separation into subtopics "Self Assessment" and "Clinical Symptom Assessment". * Refined introduction for better readability. * Added section on benchmarking platforms including AICrowd. * Refined existing benchmarking in science section. * Started section on robustness.   **Version 1.0 (submitted as** [**FGAI4H-D-016**](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-016.docx?d=w63899e533e3f41f9a72b0df6ead6a507) **for meeting D in Shanghai)**  This is the initial draft version of the TDD. As a starting point it merges the input documents FGAI4H-A-020, FGAI4H-B-021, FGAI4H-C-019, and FGAI4H-C-025 and fits them to the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6). The focus was especially on the following aspects:   * Introduction to topic and ethical considerations * Workflow proposal for Topic Group * Overview of currently available AI-based symptom assessment applications started * Prior works on benchmarking and scientific approaches including first contributions by experts joining the topic. * Brief overview of different ontologies to describe medical terms and diseases |

**Table of Contents**

| Page |
| --- |
| [1 Introduction 8](#_Toc23871141)  [1.1 AI-based Symptom Assessment 8](#_Toc23871142)  [1.1.1 Relevance 8](#_Toc23871143)  [1.1.2 Current Solutions 9](#_Toc23871144)  [1.1.3 Impact of AI-based Symptom Assessment 9](#_Toc23871145)  [1.1.4 Impact of Introducing a Benchmarking Framework for AI-based Symptom Assessment 10](#_Toc23871146)  [1.2 Ethical and cultural considerations 11](#_Toc23871147)  [1.2.1 Technical robustness, safety and accuracy 12](#_Toc23871148)  [1.2.2 Data governance, privacy and quality 12](#_Toc23871149)  [1.2.3 Explicability 12](#_Toc23871150)  [1.2.4 Fairness 13](#_Toc23871151)  [1.2.5 Individual, societal and environmental wellbeing 13](#_Toc23871152)  [1.2.6 Accountability 13](#_Toc23871153)  [2 AI4H Topic Group on "AI-based Symptom Assessment" 14](#_Toc23871154)  [2.1 General Mandate of the Topic Group 14](#_Toc23871155)  [2.2 Topic Description Document 15](#_Toc23871156)  [2.3 Sub-topics 15](#_Toc23871157)  [2.4 Topic Group Participation 15](#_Toc23871158)  [2.5 Status of this Topic Group 15](#_Toc23871159)  [2.5.1 Status Update for Meeting D (Shanghai) Submission 15](#_Toc23871160)  [2.5.2 Status Update for Meeting E (Geneva) Submission 15](#_Toc23871161)  [2.5.3 Status Update for Meeting F (Zanzibar) Submission 16](#_Toc23871162)  [2.5.4 Status Update for Meeting G (Delhi) Submission 17](#_Toc23871163)  [3 Existing AI Solutions 19](#_Toc23871164)  [3.1.1 Topic Group member Systems for AI-based Symptom Assessment 19](#_Toc23871165)  [3.1.2 Other Systems for AI-based Symptom Assessment 21](#_Toc23871166)  [3.2 Input Data 22](#_Toc23871167)  [3.2.1 Input Types 22](#_Toc23871168)  [3.2.2 Ontologies for encoding input data 23](#_Toc23871169)  [3.3 Output Types 24](#_Toc23871170)  [3.3.1 Output Types 24](#_Toc23871171)  [3.4 Scope Dimensions 28](#_Toc23871172)  [3.4 Additional Relevant Dimensions 29](#_Toc23871173)  [3.5 Robustness of systems for AI based Symptom Assessment 29](#_Toc23871174)  [4 Existing work on benchmarking 30](#_Toc23871175)  [4.1 Scientific Publications on Benchmarking AI-based Symptom Assessment Applications 30](#_Toc23871176)  [4.1.1 "Evaluation of symptom checkers for self diagnosis and triage" 31](#_Toc23871177)  [4.1.2 "ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation" 31](#_Toc23871178)  [4.1.3 "Safety of patient-facing digital symptom checkers." 31](#_Toc23871179)  [4.1.4 "Comparison of physician and computer diagnostic accuracy." 31](#_Toc23871180)  [4.1.5 "A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom checkers." 31](#_Toc23871181)  [4.2 Clinical Evaluation of AI-based Symptom Assessment 32](#_Toc23871182)  [4.2.1 "A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. Emergencias" 32](#_Toc23871183)  [4.2.2 "Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department" 32](#_Toc23871184)  [4.2.3 "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence." 32](#_Toc23871185)  [4.2.4 "Evaluating the potential impact of Ada DX in a retrospective study." 32](#_Toc23871186)  [4.2.5 "Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain." 33](#_Toc23871187)  [4.2.6 "How Accurate Are Patients at Diagnosing the Cause of Their Knee Pain With the Help of a Web-based Symptom Checker?" 33](#_Toc23871188)  [4.2.7 "Are online symptoms checkers useful for patients with inflammatory arthritis?" 33](#_Toc23871189)  [4.2.8 "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis" 34](#_Toc23871190)  [4.3 Benchmarking Publications outside Science 34](#_Toc23871191)  [4.4 Existing Regulations 34](#_Toc23871192)  [4.5 Internal Benchmarking by Companies 34](#_Toc23871193)  [4.6 Existing AI Benchmarking Frameworks 35](#_Toc23871194)  [4.6.1 General Requirements 35](#_Toc23871195)  [4.6.2 AICrowd 37](#_Toc23871196)  [4.6.3 Other Platforms 37](#_Toc23871197)  [4.7 Scores & Metrics 37](#_Toc23871198)  [5 Benchmarking 37](#_Toc23871199)  [5.1 Benchmarking Iterations 37](#_Toc23871200)  [5.2 Minimal Minimal Viable Benchmarking - MMVB 38](#_Toc23871201)  [5.2.1 Architecture and Methodology Overview 39](#_Toc23871202)  [5.2.2 AI Input Data 39](#_Toc23871203)  [5.2.3 Expected AI Output Data encoding 40](#_Toc23871204)  [5.2.4 Symptom Assessment AI Benchmarking Interface 40](#_Toc23871205)  [5.2.5 API Output Data Format 41](#_Toc23871206)  [5.2.6 Benchmarking Dataset Collection 41](#_Toc23871207)  [5.2.7 Scores & Metrics 43](#_Toc23871208)  [5.2.8 Reporting 43](#_Toc23871209)  [5.2.9 Status and Next Steps 44](#_Toc23871210)  [5.3 Minimal Minimal Viable Benchmarking - MMVB Version 2.0 44](#_Toc23871211)  [5.3.1 Adding symptom attributes 44](#_Toc23871212)  [5.3.2 Refining factors 45](#_Toc23871213)  [5.3.3 Explicit id handling 46](#_Toc23871214)  [5.3.4 Triage similarity score 46](#_Toc23871215)  [5.3.5 More MMVB 1.0 toy AIs 46](#_Toc23871216)  [5.3.6 Case by case benchmarking and case marking 46](#_Toc23871217)  [5.3.7 Updated benchmarking UI 47](#_Toc23871218)  [5.3.8 MMVB 2.0 Case Creation Guidelines 48](#_Toc23871219)  [5.4 Minimal Viable Benchmarking - MVB 48](#_Toc23871220)  [5.4.1 Architecture and Methodology Overview 48](#_Toc23871221)  [5.4.2 AI Input Data to use for the MVB 49](#_Toc23871222)  [5.4.3 AI Output Data to use for the MVB 49](#_Toc23871223)  [5.4.4 Symptom Assessment AI Benchmarking Interface 50](#_Toc23871224)  [5.4.5 API Input Data Format 50](#_Toc23871225)  [5.4.6 API Output Data Format 50](#_Toc23871226)  [5.4.7 Benchmarking Dataset Collection 50](#_Toc23871227)  [5.4.8 Benchmarking Dataset Format 50](#_Toc23871228)  [5.4.9 Scores & Metrics 50](#_Toc23871229)  [5.4.10 Reporting Methodology 51](#_Toc23871230)  [5.4.11 Technical Architecture of the Official Benchmarking System 51](#_Toc23871231)  [5.4.12 Technical Architecture of the Continuous Benchmarking System 51](#_Toc23871232)  [5.4.13 Benchmarking Operation Procedure 51](#_Toc23871233)  [6 Results from benchmarking 51](#_Toc23871234)  [7 Discussion on insights from MVB 52](#_Toc23871235)  [Appendix A: Declaration of conflict of interest 53](#_Toc23871236)  [Appendix B: Glossary 56](#_Toc23871237)  [Appendix C: References 58](#_Toc23871238)  [Appendix D: Systems not considered in chapter 3 59](#_Toc23871239)  [Appendix E: List of all (e-)meetings and corresponding minutes 61](#_Toc23871240) |

**List of Tables**

| Page |
| --- |
| [Table 1 – Table caption? 10](#_Toc23871261)  [Table 2 – Table caption? 19](#_Toc23871262)  [Table 3 – Table caption? 21](#_Toc23871263)  [Table 4 – Table caption? 22](#_Toc23871264)  [Table 5 – Table caption? 25](#_Toc23871265)  [Table 6 – Table caption? 25](#_Toc23871266)  [Table 7 – Table caption? 38](#_Toc23871267)  [Table 8 – Table caption? 39](#_Toc23871268)  [Table 9 – Table caption? 40](#_Toc23871269)  [Table 10 – Table caption? 40](#_Toc23871270)  [Table 11 – Table caption? 41](#_Toc23871271)  [Table 12 – Table caption? 42](#_Toc23871272) |

**List of Figures**

| Page |
| --- |
| [Figure 1 – Figure caption? 43](#_Toc23871273)  [Figure 2 – Figure caption? 45](#_Toc23871274)  [Figure 3 – Figure caption? 46](#_Toc23871275)  [Figure 4 – Figure caption? 47](#_Toc23871276)  [Figure 5 – Figure caption? 47](#_Toc23871277)  [Figure 6 – General framework proposed by Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL during workshop associated to FG Meeting A 48](#_Toc23871278) |

# 1 Introduction

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a suite of standardized benchmarking 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 for both subtopics 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.

The paper has been developed by interested subject matter experts and leading companies in the field during a series of virtual and face-to-face group sessions. Its objective is to provide clinical professionals, consumers and innovators in this area with internationally recognised baseline standards for the fields of diagnosis, next-steps advice and triage.

## 1.1 AI-based Symptom Assessment

Why is this needed?

There is a huge potential for AISA applications, and the opportunities laid out below are examples of how people could benefit from the successful implementation of this technology. With the speed of development and adoption globally, it is also critical to ensure that there are clear and rigorous methods to test for safety and quality. The WHO/ITU is committed to working with organisations to develop this.

### 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 prompt, adequate treatment if necessary, 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 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[[4]](#footnote-4). Meanwhile, other studies show that internet self-searches are more likely to incorrectly suggest conditions that may cause inappropriate worry (eg cancers)

### 1.1.3 Impact of AI-based Symptom Assessment

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.

Definition

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).

Sub-types

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), possible 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. 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.

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

The case for Benchmarking

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.[[5]](#footnote-5)

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 described in Table 1.

Table 1 – FG-AI4H thematic classification scheme

| Level | | Thematic classification |
| --- | --- | --- |
| 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 and cultural considerations

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.

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.

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.

The aim is to consider how their quality can be assessed in multi-faceted ways. It draws from a range of sources, including ethical guidelines such as the recently published European Union's Ethics Guidelines for Trustworthy AI, reports, and the academic literature.[[6]](#footnote-6)[[7]](#footnote-7)

### 1.2.1 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.

AI solutions must strive for a high degree of accuracy, and this will include considerations of the wider social and cultural context. 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]](#footnote-8)

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.

### 1.2.2 Data governance, privacy and quality

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.

### 1.2.3 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.

### 1.2.4 Fairness

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.

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.

The design of the AISA should also consider fairness. Issues such as access to, and ability to use, the AISA are important - including access to appropriate smartphone devices, language, and digital literacy. The group should also consider how wider infrastructures, such as electricity and internet, interact with a particular AISA to shape fairness.

### 1.2.5 Individual, societal and environmental wellbeing

AISAs will shape how individuals seek care within a healthcare system. 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.

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.

Finally, AISAs will 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 along the digital infrastructure supply chain, including the disposal of outdated or non-functioning hardware.

### 1.2.6 Accountability

AISAs raise serious 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.

# 2 AI4H Topic Group on "AI-based Symptom Assessment"

The first chapter highlighted the potential of AISA to help solve 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. The current list of Topic Groups can be found at the [AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

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.

During meeting

## 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 Sub-topics

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 Sub-topic Groups can be established to pursue different topic-specific specializations. The AISA Topic Group originally started without separate sub topic groups. With Baidu joining in meeting D in Shanghai, the Topic Group was split into the sub topics "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. In chapter 5 where the benchmarking method will be described, the specific requirements for each sub-topic will be described following [FGAI4H-D-022](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-022.docx?d=w8980136edaed424a845f6e3a74a2f3a0).

## 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/Documents/tg/CfP-TG-Symptom.pdf).

Every Topic Group also has its corresponding subpage at the website of the focus group. The page of the AISA Topic Group can be found [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/symptom.aspx).

## 2.5 Status of this Topic Group

During meeting D it was discussed that the TDD should contain an explicit section describing the progress since the last meeting for the upcoming meeting. The following subsections serve this purpose:

### 2.5.1 Status Update for Meeting D (Shanghai) Submission

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.

### 2.5.2 Status Update for Meeting E (Geneva) Submission

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 are also listed in the "change notes" at the beginning of the document.

### 2.5.3 Status Update for Meeting F (Zanzibar) Submission

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

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

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

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

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

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

* 30.5.2019 - Meeting #2 - Meeting E Breakout [Minutes](https://docs.google.com/document/d/1vT6NscHRsDp3Ajbt19Eydp26iPkKHjGQ2-2CpTcVG84/edit)
* 20.06.2019 - Meeting #3 - Telco [Minutes](https://docs.google.com/document/d/1XNLMXchIzTdKkQJd9jdacb1N1ndoRICK9JXEIkR4EQ0/edit#heading=h.srcrgcpa7gyu)
* 11-12.7.2019 - Meeting #4 - London Workshop [Minutes](https://docs.google.com/document/d/1sFitXcM3fystGfXNFHnzHoFRb3EnIzufVngDY1Q2HEA/edit#heading=h.e14xlew0l9i7)
* 15.8.2019 - Meeting #5 - Telco [Minutes](https://docs.google.com/document/d/161WaTaivPcWN8CctiQhYJFA1hz8FkYX-lZAVqgg_z-s/edit#heading=h.15wuq6sdvo0v)
* 23.08.2019 - Meeting #6 - Telco [Minutes](https://docs.google.com/document/d/1SQPLHs8PwdCEZyG-qKGj-aOCr4Z1T6e8SV7zN9djwpI/edit)

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

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

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

### 2.5.4 Status Update for Meeting G (Delhi) Submission

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

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

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

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

* 03.08.2019 - Meeting #7 - Meeting F Breakout [Minutes](https://docs.google.com/document/d/1Tw8DkMbVnk-RUyz9dRK4hqN8NCJXcf7Ev-mZC6KGcAs/edit)
* 27.09.2019 - Meeting #8 - Telco [Minutes](https://drive.google.com/open?id=1mr2r_jVdHJagTDuRHI_pldVFSK0JfkOtFohCLotV_-M)
* 10-11.10.2019 - Meeting #9 - Berlin Workshop [Minutes](https://drive.google.com/open?id=10JgDh8TBDJaRrzPVcfB6bVBPvj_qkToPdsdfYLVEd7k)
* 17.10.2019 - Meeting #10 - Telco [Minutes](https://drive.google.com/open?id=1n7ixrDQRGSvREZ35Ga-1qtQp3BDVNCXwlLr8Qfar7Pc)
* 20.10.2019 - Meeting #11 - Telco [Minutes](https://drive.google.com/open?id=1-G7ZWsfdYWIdX2-NNYV9IfaYBic9nNobmJNsWW15LGE)
* 25.10.2019 - Meeting #12 - Telco [Minutes](https://drive.google.com/open?id=1ya3BV_v4wMFo_4Ye1Nuc5Riqi6iu5bp1manLM5qM544)
* 30.10.2019 - Meeting #13 - Telco [Minutes](https://drive.google.com/open?id=1PULLCrCWXRkhTun1LTPAeCzf1mNTrCzD4QsO_tINo2g)

At the time of submission 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 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)
* 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)
* Symptify (Dr Jalil Thurber)
* Thomas Neumark (Independent Contributor)
* Visiba Care (Anastacia Simonchik)
* Your.MD (Jonathon Carr-Brown, Rex Cooper, Martin Cansdale)

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

**2.6 Next Meetings**

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

* H: Brasilia, Brazil; January 2020
* I: March 2020
* J: Geneva, 4-8 May 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

Table 2 – Table caption?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Provider** | **System** | **Input** | **Output** | **Scope/Comments** |
| [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, Portuguese, French * Top 1300 conditions * For smartphone users * Android/iOS |
| [Babylon Health](https://www.babylonhealth.com/) | Babylon App | * Age, sex, risk factors, country * Chatbot free text input and free text search (multiple inputs are allowed) * Answers to dialog questions for additional symptoms and risk factors including duration of symptoms, intensity | * Pre-clinical triage * Possible causes ("differentials") * Condition information * Recommendation of appropriate local services and products * Text information about treatments or next steps * Shortcuts in case of immediate danger | * Worldwide * English * 80% of medical conditions * For smartphone/web users * Android/iOS/Web |
| Baidu |  |  |  |  |
| Deepcare | Deepcare Symptom Checker |  |  | * Users: Doctor and Patient * Platform: * iOS * Android * Language: * Vietnamese |
| Infermedica | Infermedica API, Symptomate | * Age, sex * Risk factors * Free text input of multiple symptoms * Region/Travel history * Answers to discrete dialog questions * Lab test results | * Differentials for PC * Pre-clinical triage * Shortcuts in case of immediate danger * Explanation of differentials * Recommended further lab testing | * Worldwide * Top 1000 conditions * 15 language versions * Web, mobile, chatbot, voice |
| Inspired Ideas | Dr. Elsa | * Age, gender * Risk factors * Region/ time of year * Multiple symptoms * Travel history * Answers to discrete dialog questions * Lab test results * Clinicians hypothesis | * List of possible differentials * Condition explanations * Referral & lab test recommendations * Recommended next steps * Clinical triage | * Tanzania, East Africa * Languages: English and Swahili * Android/iOS/Web/API * Users: healthcare workers/ clinicians |
| [Isabel Healthcare](https://www.isabelhealthcare.com/) | Isabel Symptom Checker | * Age * Gender * Pregnancy Status * Region/Travel History * Free text input of multiple symptoms all at once | * List of possible diagnoses * Diagnoses can be sorted by 'common' or 'Red flag' * Each diagnosis linked to multiple reference resources * If triage function selected, patient answers 7 questions to obtain advice on appropriate venue of care | * 6,000 medical conditions covered * Unlimited number of symptoms * Responsive design means website adjusts to all devices * APIs available allowing integration into other systems * Currently English only but professional site available in Spanish and Chinese and model developed to make available in tmost languages |
| [Symptify](https://symptify.com/) | [Symptom Checker](https://symptify.com/) |  |  |  |
| Visiba Group AB | Visiba Care app | * Age * Gender * Chatbot free text input * Region/ time of year * Discrete answers * Lab results, inputs from devices enabled | * List of possible diagnoses * pre-clinical triage including format of meeting (digital or physical) * Next-step advice * condition information | * Language: Swedish * Android/iOS/Web * Users: Doctor and Patient |
| [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.

Table 3 – Table caption?

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

## 3.2 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.

### 3.2.1 Input Types

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

Table 4 – Table caption?

|  |  |  |
| --- | --- | --- |
| **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. |  |
| ... |  |  |

### 3.2.2 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

#### 3.2.2.1 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).

#### 3.2.2.2 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[[9]](#footnote-9).

#### 3.2.2.3 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. (TODO refine)

#### 3.2.2.4 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. (TODO refine)

## 3.3 Output Types

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.

### 3.3.1 Output Types

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

Table 5 – Table caption?

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

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

#### 3.3.1.1 Clinical Triage

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

Table 6 – Table caption?

|  |  |  |  |
| --- | --- | --- | --- |
| **Level** | **Status** | **Color** | **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.

#### 3.3.1.2 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.

**Ada Health Pre-Clinical Triage Levels**

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

**Babylon Pre-Clinical Triage Levels**

Generally:

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

With additional information provided per condition.

**Deepcare Triage Levels**

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

**Infermedica Triage Levels**

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

On top of that the system provides information on whether remote care is feasible (e.g. teleconsultation).

Additional information provided per condition (e.g. doctor's specialty in case of medical appointments).

**Inspired Ideas Triage Levels**

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

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

**Isabel Pre-Clinical Triage Levels**

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

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

**Symptify Pre-Clinical Triage Levels**

**Visiba Care Pre-Clinical Triage Levels**

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

Depending on the condition additional adjustments possible.

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

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

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

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

#### 3.3.1.3 Differential Diagnosis

* to be written

#### 3.3.1.4 Next Step Advice

* to be written

#### 3.3.1.5 Treatment 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. Since the granularity varies, starting at continent-level but also going down to the neighbourhood-level. The reporting most likely needs to support a hierarchical or multi-hierarchical structure.

***Condition Set***

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

***Age Range***

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

***Languages***

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

## 3.4 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 spring time 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.

## 3.5 Robustness of systems for AI based Symptom Assessment

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

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

**Memory Stability & Reproducibility**

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

**Empty case response**

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

**Negative evidence only response**

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

**All symptoms response**

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

**Duplicate symptom response**

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

**Wrong symptom response**

Systems should respond properly to unknown symptoms.

**Symptom with wrong attributes response**

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

**Symptom without mandatory attribute response**

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

# 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

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:

### 4.1.1 "Evaluation of symptom checkers for self diagnosis and triage"[[10]](#footnote-10)

TODO

### 4.1.2 "ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation"[[11]](#footnote-11)

TODO

### 4.1.3 "Safety of patient-facing digital symptom checkers."[[12]](#footnote-12)

TODO

### 4.1.4 "Comparison of physician and computer diagnostic accuracy."[[13]](#footnote-13)

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.

### 4.1.5 "A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom checkers."[[14]](#footnote-14)

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.

## 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.

### 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.[[15]](#footnote-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.

### 4.2.2 "Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department"[[16]](#footnote-16)

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.

### 4.2.3 "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.[[17]](#footnote-17)

### 4.2.4 "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%.[[18]](#footnote-18)

### 4.2.5 "Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain."[[19]](#footnote-19)

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%.

### 4.2.6 "How Accurate Are Patients at Diagnosing the Cause of Their Knee Pain With the Help of a Web-based Symptom Checker?"[[20]](#footnote-20)

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 upto 9 self-diagnoses.

### 4.2.7 "Are online symptoms checkers useful for patients with inflammatory arthritis?"[[21]](#footnote-21)

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.

### 4.2.8 "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis"[[22]](#footnote-22)

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.

## 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. Since these articles have not been peer reviewed and are not always follow following scientific standards they will not be discussed in this TDD.

## 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.

### 4.6.1 General Requirements

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 sand-box. It must not be possible to access the file-system 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.

### 4.6.2 AICrowd

In response to the call for benchmarking platforms ([FGAI4H-C-106](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-106.docx?d=w19fd773712344ed4bbdb6797fd4fa4e2)), in meeting D in Shanghai [FGAI4H-D-011](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-011.docx?d=w83459290fe0b423abc4f6dabbdc93c50) 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.

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 sand-boxing and the design of the platform itself outweighs the work of changing an existing platform to the Topic Groups needs, compared to implementing a new specialised solution.

### 4.6.3 Other Platforms

TODO

* analyse kaggle, …

## 4.7 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.

TODO

TODO

TODO

# 5 Benchmarking

Chapter 5 specifies the methodology, technology and protocols necessary for the benchmarking of AI-based symptom assessment systems. The focus of chapter 5 is to specify a concrete benchmarking. Theoretical background and the different options for the numerous decisions to be taken are supposed to be discussed in chapter 4. Since meeting D the Topic Group has the two sub topics "self assessment" and "clinical symptom assessment". Since V2.0 of this document we follow the approach of [FGAI4H-D-022](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-022.docx?d=w8980136edaed424a845f6e3a74a2f3a0) to specify the benchmarking for both subtopics together and elaborate on the specific details at the end of each subtopic.

## 5.1 Benchmarking Iterations

Due to the complexity of a holistic standardized benchmarking framework for AI-based symptom assessment, the benchmarking is developed and refined over several iterations adding more and more features and details. The following tables gives an overview of the different versions and their purpose

Table 7 – Table caption?

|  |  |  |
| --- | --- | --- |
| **Short Name** | **Name** | **Focus/Goals** |
| MMVB | Minimal Minimal Viable Benchmarking | * show a complete benchmarking pipeline including case generation, AI, metrics, reports * with all parts visible to everyone so that we can all understand how to proceed with relevant details for MVB * learn about the needed data structures and scores * write/test some first case annotations guidelines * learn about the cooperation on both software and annotation guidelines * have a foundation for further discussions on if an own benchmarking software is needed or crowdAI could be used * Target: meeting F Zanzibar |
| MMVB#2 | Minimal Minimal Viable Benchmarking Version 2 | * extend the MMVB model to attributes * refine the MMVB factor model * switch to cloud based toy AI hosting * test one-case-at-a-time testing * improve AI error handling * add an informative metric to the scoring system * add a robustes metric to the scoring system * refine the web interface dimension handling * extend the annotation guidelines to attributes and the new factors * Target: meeting G Delhi |
| MVB | Minimal Viable Benchmarking | * first benchmarking with real AI and real data * Target: meeting H |
| Vx.0 | TG Symptom Benchmarking Vx.0 | * the regular e.g. quarterly benchmarkings for this topic group * continuous integration of new features |

## 5.2 Minimal Minimal Viable Benchmarking - MMVB

During the Topic Group meeting #2 it was agreed that in preparation of building a minimal viable benchmarking (MVB) that benchmarks real company AIs and uses real data that none of the participants has seen before, 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. For discussing the requirements and technicalities of such an MMVB the Topic Group met from 11.-12.7.2019 in London. In the weeks that followed, a first MMVB was then implemented based on the outcomes of this meeting.

### 5.2.1 Architecture and Methodology Overview

The main goal of the MMVB was to see a first working benchmarking pipeline for symptom assessment systems. 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". The model can be found at <https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=1175944267>.

The group further agreed on an approach where the AIs are evaluated via REST API endpoints they expose for this purpose. This allows every participant to implement their AI in the technology of their choice. It also allows participants to host their own systems in their data centers and only submit their AI via access to theses endpoints rather than an e.g. docker container containing all the company's IP which is for some companies rated worth > 1 billion USD, even if this implies the need to create a new benchmarking dataset for each benchmarking.

Since for the MMVB there is no need for realistic data the group decided to generate synthetic case data by sampling the agreed upon London-model. This case data is then used by an evaluator to test the AI systems and record the responses in the file systems. For showcasing the pipeline a simplistic web-application was implemented that allows to generate test sets, run the evaluator against all AIs and then present the results as a simple table.

The MMVB was designed to test both pre-clinical triage and pre-diagnosis. The group decided to start with a self-assessment model, assuming that at this stage the learnings also apply to the clinical symptom assessment.

The benchmarking pipeline, the toy-AIs and the web-application have been implemented using Python3. For the meeting F in Zanzibar it is also planned to have AIs integrated running on non-Python technology.

### 5.2.2 AI Input Data

The MMVB uses as input for the AIs a simple use profile, explicit presenting complaints (PC/CC), and additional complaints. The additional complaints might also contain risk factors. The concrete fields used for cases are:

Table 8 – Table caption?

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

This JSON format is used for both, providing the AIs with the inputs, as well as storing cases.

### 5.2.3 Expected AI Output Data encoding

In addition the "public" fields given to the AIs for inference, the generated case data also encodes the expected outputs of triage and the diagnoses.

Table 9 – Table caption?

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

The group also discussed the following fields, but they are not part of the MMVB data yet:

Table 10 – Table caption?

|  |  |
| --- | --- |
| **field name** | **description** |
| otherRelevantDifferentials | * Conditions that would be an important/relevant/niceToHave part of the differentials |
| impossibleConditions | * Conditions that can be ruled out with the given case evidence without any doubts (e.g. ectopic pregnancy in men) |
| correctConditions | * The diseases that actually caused the symptoms - no matter if it can be seen in the case from the symptoms e.g. "brain cancer" even if "headache" is the only symptom. |

### 5.2.4 Symptom Assessment AI Benchmarking Interface

For the MMVB the AIs all shared the same simple interface that accepts a POST request with the caseData object as described in the AI input section. It also supports an aiImplementation parameter with a key of the AI to use. This is mainly motivated by the fact that the initial implementation contains several AIs in one python server. It is also already supported to easily add any aiImplementation that points to any possible server host and port, hence any Python or non-Python AI implementation is supported.

### 5.2.5 API Output Data Format

The AI systems are supposed to respond to the POST requests with an output format similar to the expected values encoded in the case data. In contrast to only one condition they are all allowed to return an ordered list of conditions. The group decided to not include an explicit score yet since the semantics the scores of the group members is different and not comparable.

Table 11 – Table caption?

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

For triage the AI might response with "UNCERTAIN" to declare that with the given evidence no conclusive triage result was possible. The list of conditions might be empty, and if so, it means that with the given evidence no conclusive differential result was possible.

### 5.2.6 Benchmarking Dataset Collection

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

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

One way to approach this 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 the MMVB work 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 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 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 appendix, as well as the [MMVB guideline](https://docs.google.com/document/d/1SLc8yrNr5s1RQQVy5rD4BdG4jvmFOg3rIwbwKminU68/edit) itself, but broadly followed the following process:

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

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

Below is an example of a case that illustrates the structure

Table 12 – Table caption?

|  |  |
| --- | --- |
| **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) - eg endometriosis in a male | ectopic pregnancy |
| **Correct conditions** (from condition template) | appendicitis |

The instructions (with an example) were shared with clinicians in the TG companies and some cases were created for use by the MMVB. Feedback was collected on the quality of guidelines and process.

### 5.2.7 Scores & Metrics

For the MMVB we started with the simple top-n match scores stating in how many percent the expected condition was contained in the first n conditions of the AI result. The current implementation uses n=1, n=3 and n=10.

For triage we only implemented the n=1, in fact presenting in how many cases the AI got the triage exactly as stated in the expectedTriageLevel. In addition to that, as it was suggested during the London meeting, a simple distance measure taking into account the "self care" is worse than "primary care" if "emergency care" was implemented.

### 5.2.8 Reporting

For the reporting the MMVB uses a simplistic web interface rendering an interactive table with all scores for all AI systems:

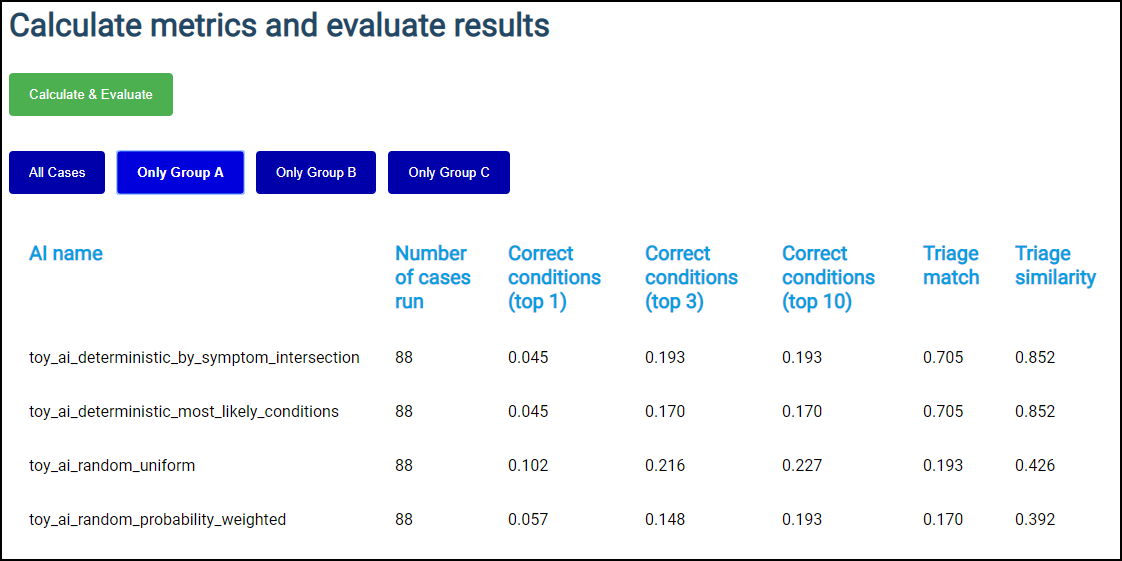


Figure 1 – Figure caption?

From the discussions it was clear that a single table or leaderboard is not sufficient for the benchmarking in this Topic Group. As outlined in section 3.4 Scope Dimensions there are 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 AISA tool. For the MMVB, a simple interactive table is implemented to show it is possible to filter results by different groups. For the illustrative purposes of the MMVB, three simple groups are introduced that filter the results by the age of case patients. More sophisticated filtering, grouping and table generation/interaction is required post the MMVB.

### 5.2.9 Status and Next Steps

As intended, the MMVB reached a point where first promising results can be seen. While it provides a good starting point for further work, a few more details need to be implemented and evaluated until the work on the MVB can start. Among theses things are:

* Adding symptom attributes
* Adding more factors
* Adding scope dimensions and using them in a more interactive reporting
* Implementation of some robustness scores e.g. for determinism of the results
* Better scores dealing with AI's responding with "unsure".
* Scores dealing with AI errors.
* Dedicated handling/evaluation of error e.g. if the evaluator uses invalid symptoms.
* Integrating human created test sets (and marking them with a scope dimension for dedicated reporting).
* Dynamic AI registration through the web-interface.
* Running the benchmarking by case rather than by AI, including some timeout handling.
* The Topic Group member should provide more AI implementations.
* Finding an approach to represent input and output in a standardised way such that all participants can consume input and return results appropriately.
* Finding a way to account for the fact that real patients use vague language to report their input.
* Account for the fact that different AI systems deal with inputs in different ways (dialogue; full input at once; etc).

## 5.3 Minimal Minimal Viable Benchmarking - MMVB Version 2.0

Building on top of the MMVB version 1.0 described in section 5.2 the work after meeting F focused on a next iteration addressing the next steps described in 5.2.9. The improvements that have been made in the model and/or the MMVB implementation will be summarized in the following sections:

### 5.3.1 Adding symptom attributes

The most relevant limitation of the MMVB 1.0 model was the missing support of explicit attributes for describing the details like intensity, time since onset or laterality of symptoms like headache. So far the model contained only so-called compound symptoms grouping a symptom with a specific attribute expression pattern like "abdominal pain cramping central 2 days" or "sharp lower quadrant pain". As next step the attributes have now been added as shown in the following picture:

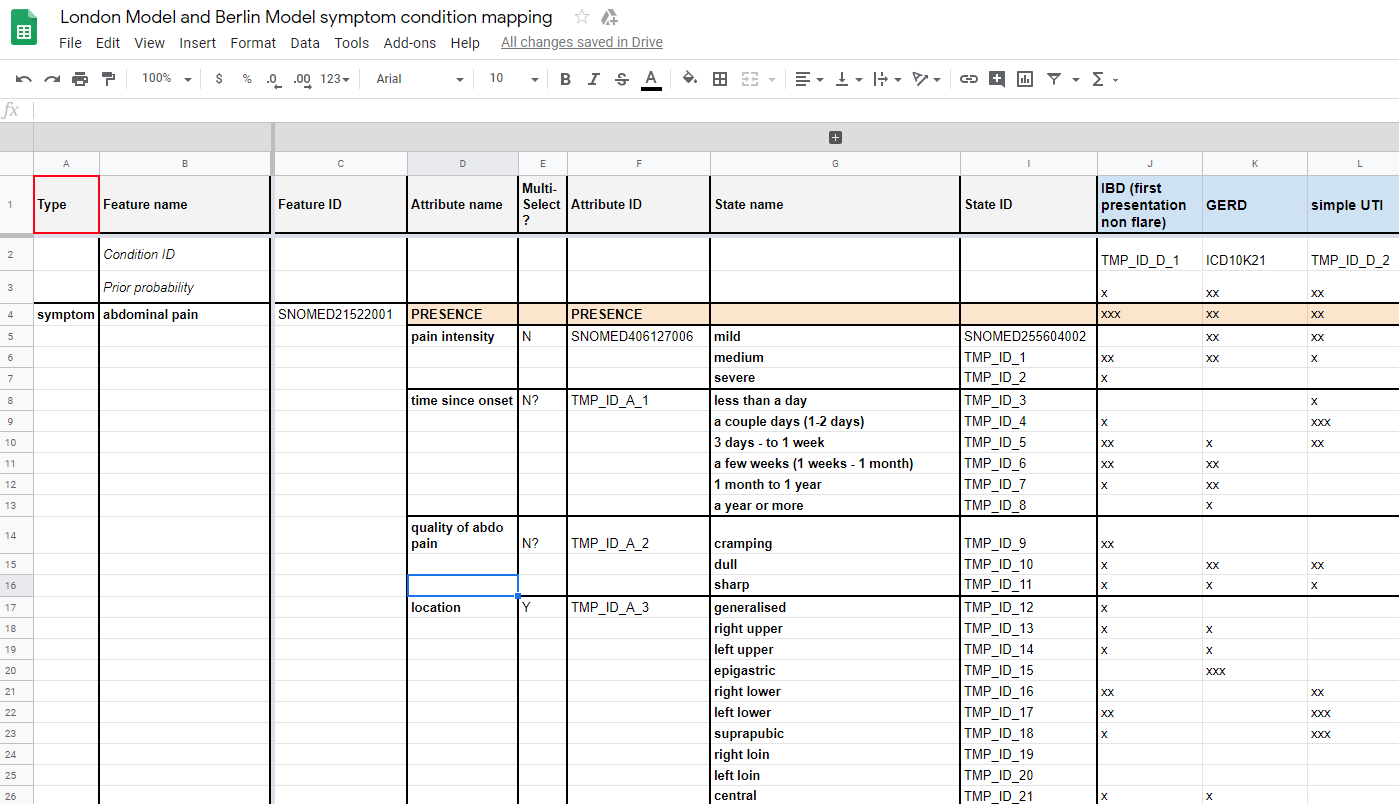


Figure 2 – Figure caption?

The above mentioned compound symptoms have been replaced with a single symptom "abdominal pain" as it is often reported by users of self-assessment applications. For expressing the details the symptom now contains sub structures for each attribute stating the probability distribution of attribute for all the conditions where this is known. As it can happen that no evidence for the attributes is available the "presence" of the symptom has to be expressed explicitly. All symptoms with attributes therefore have an explicit "PRESENCE" attribute, which is responsible for information on whether a symptom is "present", "absent" or a patient is "unsure" (or does not know) about it. The cell on the intersection between symptom's "PRESENCE" 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.

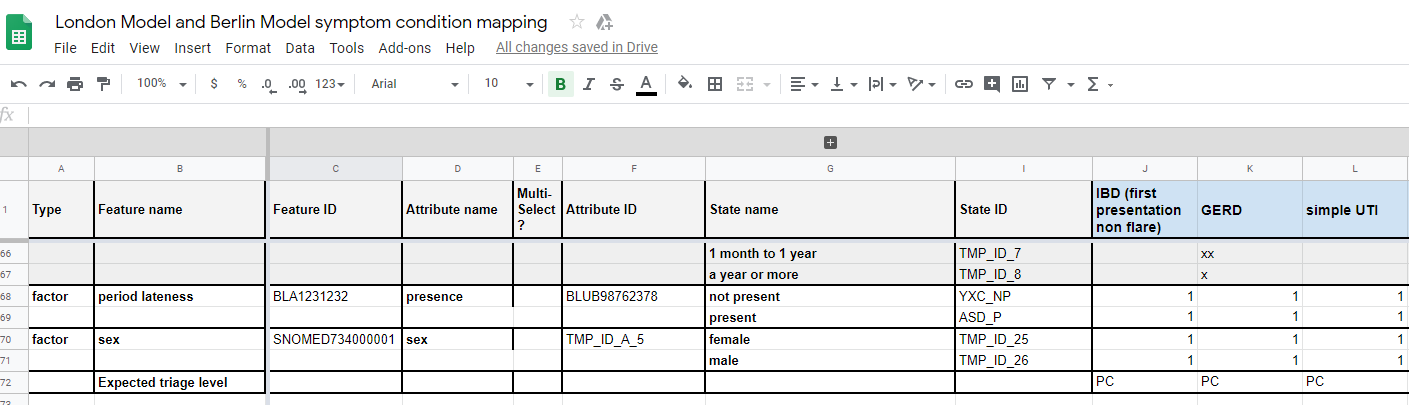
Some symptom attribute states are exclusive (i.e. not multiselect; see column E), meaning that only one attribute state can be "present". Other symptom attribute states are not exclusive (i.e. multiselect), meaning several states might be present at the same time.

If symptom is "absent" or "unsure", then no attributes or attribute states are expected to be provided.

Note that it is acceptable if only some or none of attributes with their states are provided (i.e. only information the presence of the symptom is provided).

### 5.3.2 Refining factors

As second aspect that was improved for the version 2 of the MMVB as the modelling of risk factors. In the initial model it was only informally noted in a comment field that "ectopic pregnancy" is "only females". For later in the MMVB supporting more factors that also influence the AIs in a non-binary way we intruduced explicit probability distributions modulating the prior distributions of the diffent conditions. Factors don't have the same "PRESENCE" as symptoms. Instead, factors are quantified by their state such that that affects the probability of diseases by a multiplier coefficient. The factors are not provided by "present", "absent", "unsure" presence state, but instead they only rely on their attributes and attribute states. Depending on the values of attribute states and corresponding scalar multipliers, the probabilities of diseases are adjusted accordingly:



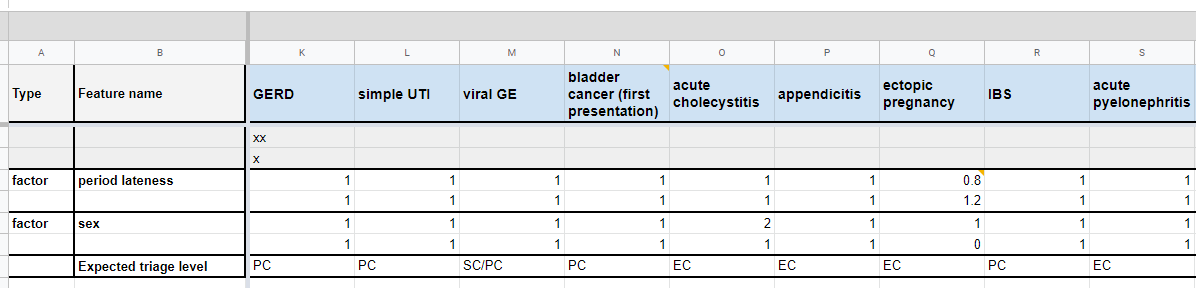


Figure 3 – Figure caption?

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

### 5.3.3 Explicit id handling

All features, attributes, attribute states and diseases now have unique identifiers that are predefined in the spreadsheets rather than being automatically generated in the MMVB code in an opaque way. For now, while we are still deciding on the ontology (or ontologies) to use, we have come up with temporary IDs for most of these objects. In the future, we aim to replace all of them by some ontologies' codes. The definition of IDs for each e.g. symptom is important since it is the base for the communication between the benchmarking systemen and the different AIs.

### 5.3.4 Triage similarity score

We decided to implement a new score "Triage similarity (soft)" score (in addition to the existing "Triage similarity") such that if an AI says that it is "unsure" about a triage, the AI is given a triage similarity score higher than 0 (as it currently happens for "Triage similarity" score). The reason to introduce this illustrative 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".

### 5.3.5 More MMVB 1.0 toy AIs

Topic group participants have agreed to implement their own versions of toy AIs. The initial plan, as discussed in Berlin in October 2019, was to implement the new Berlin model, but there has not been enough time to do it. We aim that by the time of Meeting in Delhi, there will be at least one or several participants, whose cloud hosted toy AIs (at least with London model) will be integrated and tested as part of the MMVB.

### 5.3.6 Case by case benchmarking and case marking

In the MMVB 2.0, the AIs are not tested independently with all cases in a batch, but instead cases are sent to AIs case by case to ensure that each case is exposed to the participants at the same time. The main point is here to step by step harden the benchmarking to make it more robust against cheating e.g. by inter-AI-communication.

Closely related is also the functionality to mark cases as "burnt" such that every case that is sent to AIs is marked as "used" to track the fact that it was exposed to the public.

To reduce the risk that cases are inefficiently "burned" e.g. if there are network issues, a health check end-point has to be implemented for each AI. This is to ensure that before a next case is sent to each AI, we check that all AIs are ready to process that new case. If some AIs are malfunctioning or don't respond, the MMVB can wait for some number of iterations (which is a configurable parameter) before sending the next case.

### 5.3.7 Updated benchmarking UI

For reflecting the changes in the benchmarking processes the web based user interface for the benchmarking systems as extended accordingly. The new UI of the MMVB is now more interactive and allow to see the progress of health checks and sent cases to AIs.

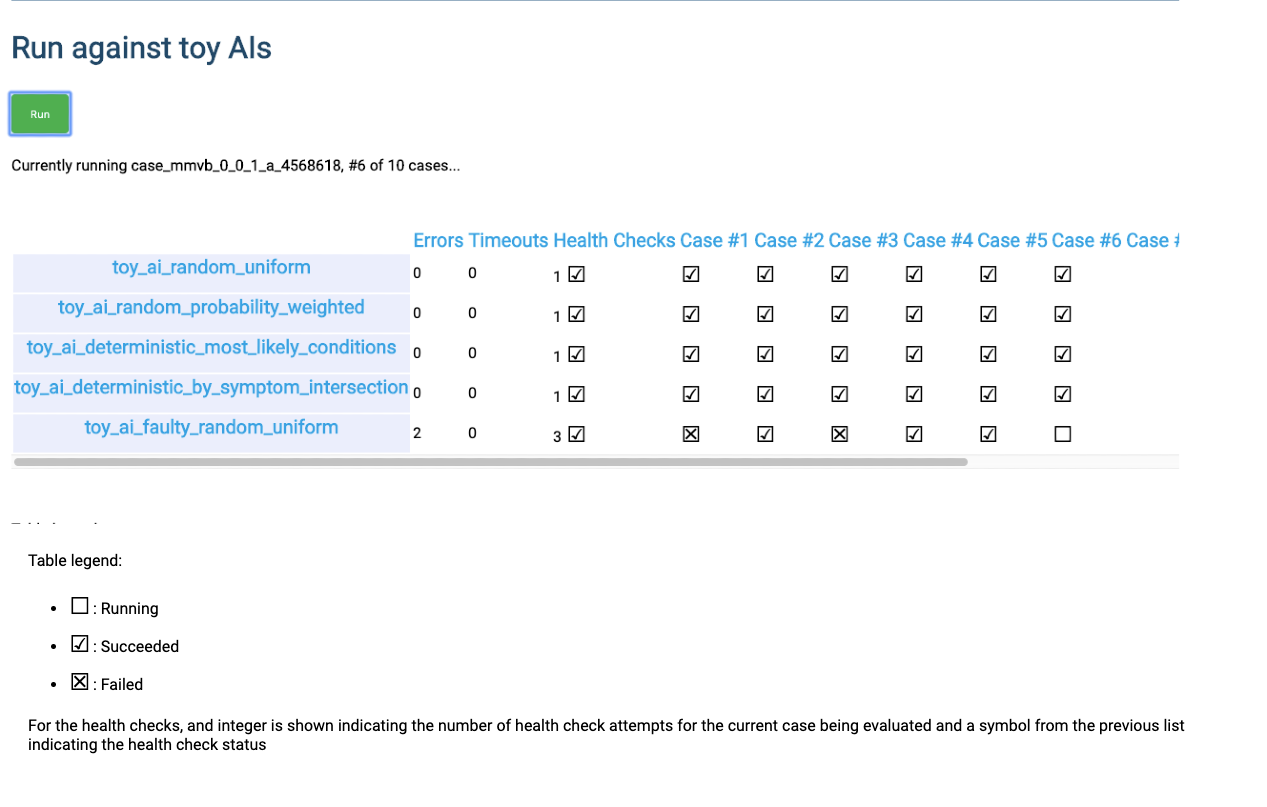


Figure 4 – Figure caption?

The new MMVB 2 interface also output logs in real time:

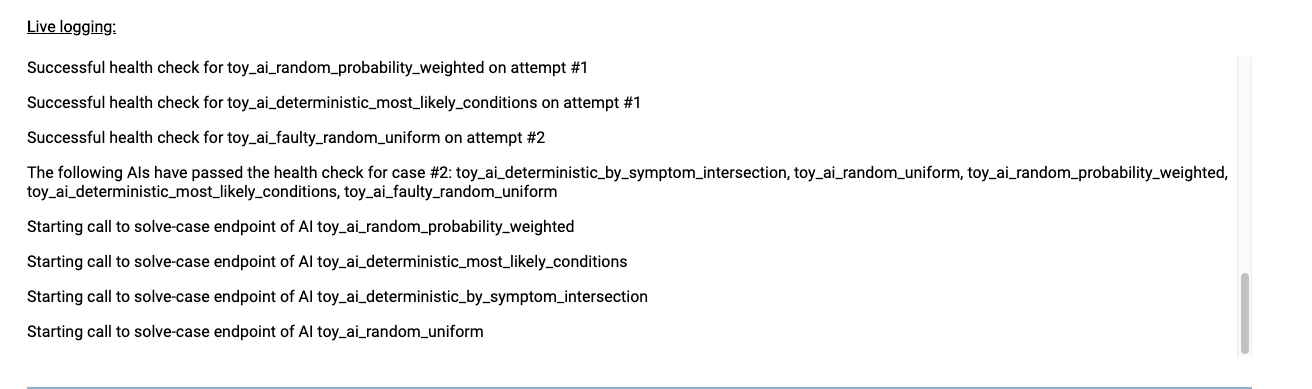


Figure 5 – Figure caption?

The group also discussed a more dynamic multidimensional report filtering to learn how e.g. to get the benchmarking results for e.g. the best systems suited for CVD diagnosis in 60+ patients in Scandinavia. However, given the limited development resources the next steps will only available for meeting H.

### 5.3.8 MMVB 2.0 Case Creation Guidelines

In parallel to the work on the benchmarking software we also updated the cases annotation guidelines to reflect the new attribute and factor structures introduced for MMVB 2.0. Here is the link to these - [MMVB 2.0 Guidelines](https://docs.google.com/document/d/1sKcEHF-6BN8_a8OHpg8tXhAO6E_B1IBuE5JwwKSCiAI/edit). Clinicians in the participating organisations then use these new guidelines to create a new set of cases with a new level of complexity for the MMVB 2.0 benchmarking round. As part of this work it became clear the spread-sheets have reached the limitations as a meaningful tool for creating benchmarking cases and the group needs to consider alternatives like dedicated case creation applications.

## 5.4 Minimal Viable Benchmarking - MVB

### 5.4.1 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):

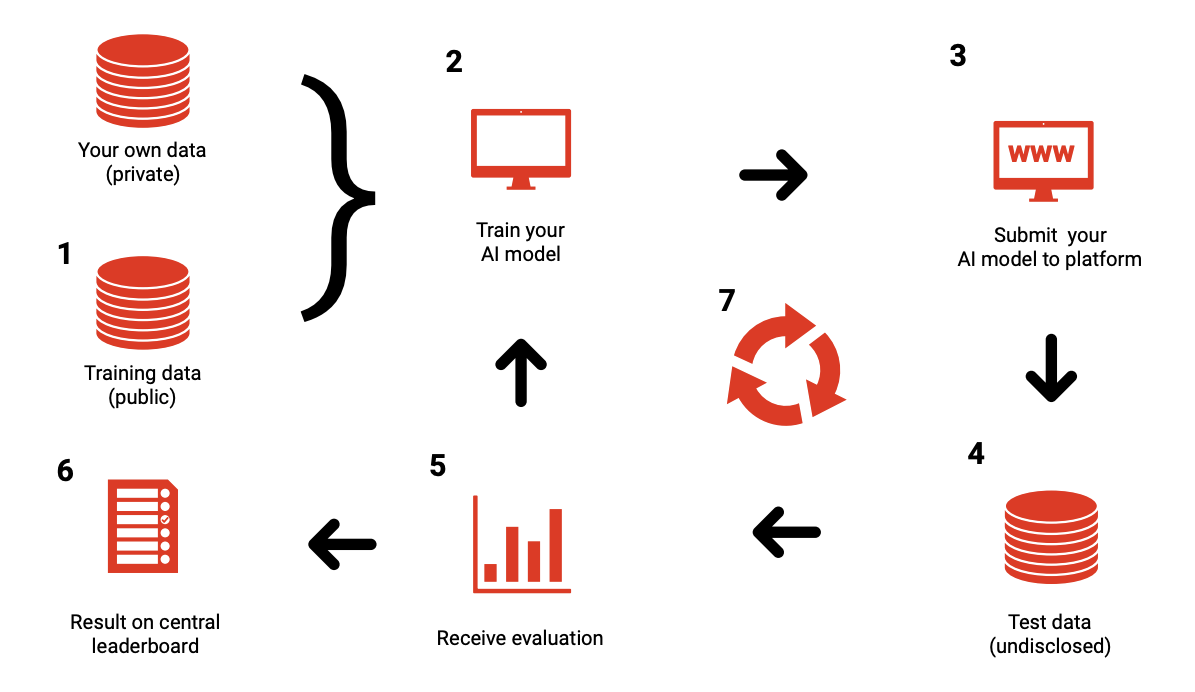


Figure 6 – General framework proposed by Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL during workshop associated to FG Meeting A[[23]](#footnote-23)

**(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.

The general architecture for both subtopics is the same.

### 5.4.2 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):

### 5.4.3 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 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

### 5.4.4 Symptom Assessment AI Benchmarking Interface

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

### 5.4.5 API Input Data Format

* TODO: JSON Format

### 5.4.6 API Output Data Format

* TODO: JSON Format

### 5.4.7 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)

### 5.4.8 Benchmarking Dataset Format

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

### 5.4.9 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

### 5.4.10 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?

### 5.4.11 Technical Architecture of the Official Benchmarking System

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

### 5.4.12 Technical Architecture of the Continuous Benchmarking System

* TODO
* in comparison to the official system
* possibility to test benchmarking before an official benchmarking ("test submission")

### 5.4.13 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 from benchmarking

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 on insights from MVB

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.

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

Further contributions to this document: Andreas Kühn ([andreas.kuehn@ada.com](mailto:andreas.kuehn@ada.com)), Johannes Schröder [(johannes.schroeder@ada.com](mailto:johannes.schroeder@ada.com)), Sarika Jain ([sarika.jain@ada.com](mailto:sarika.jain@ada.com)), Isabel Glusman (TODO), Ria Vaidya ([ria.vaidya@ada.com](mailto:ria.vaidya@ada.com))

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.

Involved people: Saurabh Johri ([saurabh.johri@babylonhealth.com](mailto:saurabh.johri@babylonhealth.com)), Nathalie Bradley-Schmieg ([nathalie.bradley1@babylonhealth.com](mailto:nathalie.bradley1@babylonhealth.com))

Baidu

TODO

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.

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

Infermedica

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

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

Inspired Ideas

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

Involved people: Ally Salim Jr ([ally@inspiredideas.io](mailto:ally@inspiredideas.io)), Megan Allen (megan@inspiredideas.io)

Isabel Healthcare

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

Involved people: Jason Maude ([jason.maude@isabelhealthcare.com](mailto:jason.maude@isabelhealthcare.com))

Symptify

TODO

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.

Involved people: Anastacia Simonchik ([anastacia.simonchik@visibacare.com](mailto:anastacia.simonchik@visibacare.com))

Your.MD Ltd

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

Involved people: Jonathon Carr-Brown ([jcb@your.md](mailto:jcb@your.md)), Matteo Berlucchi([matteo@your.md](mailto:matteo@your.md)), Rex Cooper ([rex@your.md](mailto:rex@your.md)), Martin Cansdale ([martin@your.md](mailto:martin@your.md))

Appendix B:  
Glossary

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

| Acronym | Expansion | Comment |
| --- | --- | --- |
| AI | [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) | While the exact definition is highly controversial, in context of this document it refers to a field of computer science working on machine learning and knowledge based technology that allows to *understand* complex (health related) problems and situations at or above human (doctor) level performance and providing corresponding insights (differential diagnosis) or solutions (next step advice, triage). |
| 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. |
| CP | Chief Complaint | See "Presenting Complaint". |
| 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 |  |
| MMVB | Minimal minimal viable benchmarking | A simple benchmarking sandbox for understanding and testing the requirement for implementing the MVB. See chapter 5.2 for details. |
| MRCGP | [Membership of the Royal College of General Practitioners](https://en.wikipedia.org/wiki/Membership_of_the_Royal_College_of_General_Practitioners) | A postgraduate medical qualification in the United Kingdom run by the Royal College of General Practitioners. |
| 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-101](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-101.docx?d=w8026fe8822d346409b79a04ab605b1c0) 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>

Appendix D:  
Systems not considered in chapter 3

Chapter 3 lists currently existing symptom-based AI decision support systems. Systems that for some reason could not be added are listed in the following table. The table is supposed to be checked on a regular basis. If there is evidence that systems still exist and offer a non-trivial AI bases symptom-based decision support, they might be moved to the chapter 3 tables.

| Provider/System | Last Check | Exclusion Reason |
| --- | --- | --- |
| Amino | 01.2019 | could not find any more |
| BetterMedicine (www.bettermedicine.com/symptom-checker/) | 01.2019 | do they still exist? |
| Doctor on Demand | 01.2019 | no AI symptom checker? |
| Everyday Health Symptom Checker | 01.2019 | Infermedica white label |
| First Opinion | 01.2019 | no AI symptom checker, just chatting online with a doctor |
| FreeMD | 01.2019 | IP address could not be found |
| GoodRX | 01.2019 | no AI symptom checker |
| Harvard Medical School Family Health Guide (USA) | 01.2019 | could not find |
| Heal | 01.2019 | no AI symptom checker, only booking doctor home visits |
| Healthwise | 01.2019 | no AI symptom checker, seems like it's a "patient education" platform |
| Healthy Children | 01.2019 | there is a symptom checker, but it's not AI-based |
| iTriage | 01.2019 | got bought by Aetna, and then the iTriage app was taken off the iOS and Android app stores. ([source](https://www.mobihealthnews.com/content/itriage-early-mobile-health-success-shuts-down-aetna-preps-new-flagship-app)) |
| NHS Symptom Checkers | 01.2019 | looks like these are not available anymore. Only the NHS Conditions list exists, where you can look up conditions. |
| Onmeda.de | 01.2019 | only symptom lookup, no AI output |
| Oscar | 01.2019 | only a telemedicine/tech-focused health insurance provider, no symptom checking |
| Practo | 01.2019 | only a telemedicine app, no symptom checking |
| PushDoctor | 01.2019 | only a telemedicine app, no symptom checking |
| Sherpaa | 01.2019 | looks like tele only |
| Steps2Care | 01.2019 | maybe not available anymore |
| Teladoc | 01.2019 | looks like tele only |
| Zava (Dr Ed) | 01.2019 | no AI symptom checker? |

Appendix E:  
List of all (e-)meetings and corresponding minutes

|  |  |  |
| --- | --- | --- |
| Date | Meeting | Relevant Documents |
| 26-27.9.2018 | Meeting A - Geneva | * [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" |
| 15-16.11.2018 | Meeting B - New York | * [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 |
| 22-25.01.2019 | Meeting C - Lausanne | * [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 * [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 |
| 2-5.4.2019 | Meeting D - Shanghai | * [D-016](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-016.docx?d=w63899e533e3f41f9a72b0df6ead6a507): Standardized Benchmarking for AI-based symptom assessment * [D-041](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-041.pptx?d=we546b5b2934b48b7bdc3022c1ccdc49e&Source=https%3A%2F%2Fextranet%2Eitu%2Eint%2Fsites%2Fitu%2Dt%2Ffocusgroups%2Fai4h%2Fdocs%2FForms%2F190402%2Easpx): TG Symptom Update (Presentation) |
| 29.5.-1.6.2019 | Meeting E - Geneva | * [E-017](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-017.docx?d=w2ddecda3a8ad48b6870e6c56a7689b1b): TDD update: TG-Symptom (Symptom assessment) * [E-017-A01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-017-A01.pptx?d=w23aa109469464e549b813e95b378fd68): TDD update: TG-Symptom (Symptom Assessment) - Att.1 - Presentation |
| 30.5.2019 | Meeting #2 - Meeting E Breakout | * [Minutes](https://docs.google.com/document/d/1vT6NscHRsDp3Ajbt19Eydp26iPkKHjGQ2-2CpTcVG84/edit) |
| 20.06.2019 | Meeting #3 - Telco | * [Minutes](https://docs.google.com/document/d/1XNLMXchIzTdKkQJd9jdacb1N1ndoRICK9JXEIkR4EQ0/edit#heading=h.srcrgcpa7gyu) |
| 11-12.7.2019 | Meeting #4 - London Workshop | * [Minutes](https://docs.google.com/document/d/1sFitXcM3fystGfXNFHnzHoFRb3EnIzufVngDY1Q2HEA/edit#heading=h.e14xlew0l9i7) * [London Model](https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=575520860) * [GitHub](https://github.com/babylonhealth/itu_who_2019_symptom_assessment_mmv_benchmark) |
| 15.8.2019 | Meeting #5 - Telco | * [Minutes](https://docs.google.com/document/d/161WaTaivPcWN8CctiQhYJFA1hz8FkYX-lZAVqgg_z-s/edit#heading=h.15wuq6sdvo0v) |
| 23.08.2019 | Meeting #6 - Telco | * [Minutes](https://docs.google.com/document/d/1SQPLHs8PwdCEZyG-qKGj-aOCr4Z1T6e8SV7zN9djwpI/edit) |
| 2-5.9.2019 | Meeting F- Zanzibar | * [F-017](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-F-017.docx?d=wf55bdbee8809480ca0635f2aaf39624d): TDD update: TG-Symptom (Standardized Benchmarking for AI-based symptom assessment) * [F-017-A01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-F-017-A01.pptx?d=w32dd907fc5f14b8d8b343faee8ce7f2a): TDD update: TG-Symptom - Att.1 - Presentation |
| 3.09.2019 | Meeting #7 - Meeting F Breakout | * [Minutes](https://docs.google.com/document/d/1Tw8DkMbVnk-RUyz9dRK4hqN8NCJXcf7Ev-mZC6KGcAs/edit) |
| 27.09.2019 | Meeting #8 - Telco | * [Minutes](https://drive.google.com/open?id=1mr2r_jVdHJagTDuRHI_pldVFSK0JfkOtFohCLotV_-M) |
| 10-11.10.2019 | Meeting #9 - Berlin Workshop | * [Minutes](https://drive.google.com/open?id=10JgDh8TBDJaRrzPVcfB6bVBPvj_qkToPdsdfYLVEd7k) * [Berlin Model](https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit#gid=980125545) |
| 17.10.2019 | Meeting #10 - Telco | * [Minutes](https://drive.google.com/open?id=1n7ixrDQRGSvREZ35Ga-1qtQp3BDVNCXwlLr8Qfar7Pc) |
| 20.10.2019 | Meeting #11 - Telco | * [Minutes](https://drive.google.com/open?id=1-G7ZWsfdYWIdX2-NNYV9IfaYBic9nNobmJNsWW15LGE) |
| 25.10.2019 | Meeting #12 - Telco | * [Minutes](https://drive.google.com/open?id=1ya3BV_v4wMFo_4Ye1Nuc5Riqi6iu5bp1manLM5qM544) |
| 30.10.2019 | Meeting #13 - Telco | * [Minutes](https://drive.google.com/open?id=1PULLCrCWXRkhTun1LTPAeCzf1mNTrCzD4QsO_tINo2g) |

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

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. PushDoctor. UK Digital Health Report 2015. <https://www.pushdoctor.co.uk/digital-health-report> [↑](#footnote-ref-4)
5. United Nations. Goal 3: Ensure healthy lives and promote well-being for all ages. https://www.un.org/sustainabledevelopment/health/ [↑](#footnote-ref-5)
6. European Union. (2019) *Ethics Guidelines for Trustworthy AI* [↑](#footnote-ref-6)
7. Academy of Medical Royal Colleges. (2019) *Artificial Intelligence in Healthcare* [↑](#footnote-ref-7)
8. Erikson, S. L. (2018), *Cell Phones ≠ Self and Other Problems with Big Data Detection and Containment during Epidemics: Problems with Big Data Detection and Containment*. Medical. Anthropological Quarterly. **32**, 315–339 [↑](#footnote-ref-8)
9. [https://hpo.jax.org/app/](https://www.human-phenotype-ontology.org) [↑](#footnote-ref-9)
10. 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 [↑](#footnote-ref-10)
11. 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 [↑](#footnote-ref-11)
12. 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 [↑](#footnote-ref-12)
13. 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. [↑](#footnote-ref-13)
14. 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. [↑](#footnote-ref-14)
15. 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-15)
16. 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). [↑](#footnote-ref-16)
17. 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-17)
18. 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-18)
19. 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. [↑](#footnote-ref-19)
20. 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. [↑](#footnote-ref-20)
21. 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. [↑](#footnote-ref-21)
22. 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.

    [↑](#footnote-ref-22)
23. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/M\_Salath%C3%A9\_Presentation.pdf [↑](#footnote-ref-23)