|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-J-021-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 30 September – 2 October 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Symptom Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Symptom) | | |
| **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) |

|  |  |
| --- | --- |
| **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. This document was submitted to FG-AI4H meeting J (e-meeting), 30 September - 2 October 2020. |

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

**Contributors**

|  |  |  |
| --- | --- | --- |
| **Contact:** | Andreas Kühn Germany |  |
| **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 EQL UK | Email: [yura@eql.ai](mailto:yura@eql.ai) |
| **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) |
| **Contact** | Martina Fischer Germany |  |
| **Contact** | Lina Elizabeth Porras Santana 1DOC3 Colombia | Email: [linaporras@1doc3.com](mailto:linaporras@1doc3.com) |
| **Contact** | Juan Sebastián Beleño 1DOC3 Colombia | Email: [jbeleno@1doc3.com](mailto:jbeleno@1doc3.com) |
| **Contact** | María Fernanda González Alvarez 1DOC3 Mexico | Email: [mgonzalez@1doc3.com](mailto:mgonzalez@1doc3.com) |
| **Contact** | Adam Baker  Babylon Health UK | Email: [adam.baker@babylonhealth.com](mailto:adam.baker@babylonhealth.com) |
| **Contact** | Xingxing Cao  Baidu China | Email: [caoxingxing@baidu.com](mailto:caoxingxing@baidu.com) |
| **Contact** | Clemens Schöll Ada Health GmbH Germany | Email: [clemens.schoell@ada.com](mailto:clemens.schoell@ada.com) |
| **Contact** | Audrey Menezes Your.MD UK | Email: [audrey@your.md](mailto:audrey@your.md) |

**Table of Contents**

| Page |
| --- |
| [1 Introduction 13](#_Toc52122570)  [1.1 AI-based Symptom Assessment 13](#_Toc52122571)  [1.1.1 Relevance 13](#_Toc52122572)  [1.1.2 Current Solutions 14](#_Toc52122573)  [1.1.3 Impact of AI-based Symptom Assessment 14](#_Toc52122574)  [1.1.4 Impact of Introducing a Benchmarking Framework for AI-based Symptom Assessment 15](#_Toc52122575)  [1.2 Ethical and cultural considerations 16](#_Toc52122576)  [1.2.1 Technical robustness, safety and accuracy 16](#_Toc52122577)  [1.2.2 Data governance, privacy and quality 17](#_Toc52122578)  [1.2.3 Explicability 17](#_Toc52122579)  [1.2.4 Fairness 17](#_Toc52122580)  [1.2.5 Individual, societal and environmental wellbeing 17](#_Toc52122581)  [1.2.6 Accountability 18](#_Toc52122582)  [2 AI4H Topic Group on "AI-based symptom assessment" 18](#_Toc52122583)  [2.1 General mandate of the Topic Group 19](#_Toc52122584)  [2.2 Topic description document 19](#_Toc52122585)  [2.3 Sub-topics 20](#_Toc52122586)  [2.4 Topic Group participation 20](#_Toc52122587)  [2.5 Status of this Topic Group 20](#_Toc52122588)  [2.5.1 Status Update for Meeting D (Shanghai) submission 20](#_Toc52122589)  [2.5.2 Status Update for Meeting E (Geneva) submission 20](#_Toc52122590)  [2.5.3 Status Update for Meeting F (Zanzibar) submission 21](#_Toc52122591)  [2.5.4 Status Update for Meeting G (Delhi) submission 22](#_Toc52122592)  [2.5.5 Status Update for Meeting H (Brasilia) submission 23](#_Toc52122593)  [2.5.6 Status Update for Meeting I (Online) submission 25](#_Toc52122594)  [2.5.7 Status Update for Meeting J (Online) submission 27](#_Toc52122595)  [3 Existing AI solutions 29](#_Toc52122596)  [3.1 Existing Systems for AI-based symptom assessment 29](#_Toc52122597)  [3.1.1 Topic Group member systems for AI-based symptom assessment 29](#_Toc52122598)  [3.1.2 Other systems for AI-based symptom assessment 31](#_Toc52122599)  [3.2 Input Data 33](#_Toc52122600)  [3.2.1 Input types 33](#_Toc52122601)  [3.2.2 Ontologies for encoding input data 34](#_Toc52122602)  [3.3 Output types 35](#_Toc52122603)  [3.3.1 Output types 35](#_Toc52122604)  [3.4 Scope dimensions 38](#_Toc52122605)  [3.5 Additional relevant dimensions 39](#_Toc52122606)  [3.6 Robustness of systems for AI based symptom assessment 40](#_Toc52122607)  [4 Existing work on benchmarking 41](#_Toc52122608)  [4.1 Scientific publications on benchmarking AI-based symptom assessment applications 41](#_Toc52122609)  [4.1.1 "Evaluation of symptom checkers for self-diagnosis and triage" 42](#_Toc52122610)  [4.1.2 "ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation" 42](#_Toc52122611)  [4.1.3 "ISABEL: Accuracy of a Machine Learning Based Ddx Generator 42](#_Toc52122612)  [4.1.4 ISABEL: “Asking ISABEL” for diagnostic dilemmas in pediatrics: How does a web based diagnostic checklist perform? 42](#_Toc52122613)  [4.1.5 ISABEL: Performance of a Web-Based Clinical Diagnosis Support System for Internists 42](#_Toc52122614)  [4.1.6 "Safety of patient-facing digital symptom checkers." 43](#_Toc52122615)  [4.1.7 "Comparison of physician and computer diagnostic accuracy." 43](#_Toc52122616)  [4.1.8 "A novel insight into the challenges of diagnosing degenerative cervical myelopathy using web-based symptom checkers." 43](#_Toc52122617)  [4.2 Clinical evaluation of AI-based symptom assessment 44](#_Toc52122618)  [4.2.1 "A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application. Emergencias" 44](#_Toc52122619)  [4.2.2 "Evaluation of a diagnostic decision support system for the triage of patients in a hospital emergency department" 44](#_Toc52122620)  [4.2.3 "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence." 44](#_Toc52122621)  [4.2.4 "Evaluating the potential impact of Ada DX in a retrospective study." 45](#_Toc52122622)  [4.2.5 "Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain." 45](#_Toc52122623)  [4.2.6 "How Accurate Are Patients at Diagnosing the Cause of Their Knee Pain With the Help of a Web-based Symptom Checker?" 45](#_Toc52122624)  [4.2.7 "Are online symptoms checkers useful for patients with inflammatory arthritis?" 45](#_Toc52122625)  [4.2.8 "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis" 46](#_Toc52122626)  [4.3 Benchmarking publications outside science 46](#_Toc52122627)  [4.4 Existing regulations 46](#_Toc52122628)  [4.5 Internal benchmarking by companies 47](#_Toc52122629)  [4.6 Existing AI benchmarking frameworks 48](#_Toc52122630)  [4.6.1 General requirements 48](#_Toc52122631)  [4.6.2 AICrowd 49](#_Toc52122632)  [4.6.3 Other platforms 49](#_Toc52122633)  [4.7 Scores & Metrics 49](#_Toc52122634)  [4.7.1 Outline of clinical considerations for scores and metrics 49](#_Toc52122635)  [4.7.2 What do we mean by clinical metrics? 50](#_Toc52122636)  [4.7.3 More complex than image classification 51](#_Toc52122637)  [4.7.4 Important considerations 52](#_Toc52122638)  [4.7.5 Metrics for symptom assessment 56](#_Toc52122639)  [4.7.6 Putting it all together for clinicians 64](#_Toc52122640)  [4.7.7 Additional clinical considerations and limitations 67](#_Toc52122641)  [4.7.8 Conclusion 68](#_Toc52122642)  [5 Benchmarking 68](#_Toc52122643)  [5.1 Benchmarking Iterations 68](#_Toc52122644)  [5.2 Minimal Minimal Viable Benchmarking - MMVB 69](#_Toc52122645)  [5.2.1 Architecture and methodology overview 69](#_Toc52122646)  [5.2.2 AI input data 70](#_Toc52122647)  [5.2.3 Expected AI output data encoding 71](#_Toc52122648)  [5.2.4 Symptom assessment AI benchmarking interface 71](#_Toc52122649)  [5.2.5 API output data format 72](#_Toc52122650)  [5.2.6 Benchmarking dataset collection 72](#_Toc52122651)  [5.2.7 Scores & metrics 74](#_Toc52122652)  [5.2.8 Reporting 74](#_Toc52122653)  [5.2.9 Status and next steps 75](#_Toc52122654)  [5.3 Minimal Minimal Viable Benchmarking - MMVB Version 2.0 76](#_Toc52122655)  [5.3.1 Adding symptom attributes 76](#_Toc52122656)  [5.3.2 Refining factors 77](#_Toc52122657)  [5.3.3 Explicit id handling 77](#_Toc52122658)  [5.3.4 Triage similarity score 77](#_Toc52122659)  [5.3.5 More MMVB 1.0 toy AIs 77](#_Toc52122660)  [5.3.6 Case by case benchmarking and case marking 78](#_Toc52122661)  [5.3.7 Updated benchmarking UI 78](#_Toc52122662)  [5.3.8 MMVB 2.0 case creation guidelines 79](#_Toc52122663)  [5.4 Minimal Minimal Viable Benchmarking - MMVB Version 2.1 79](#_Toc52122664)  [5.4.1 Input and Output ontology encoding and model schema 80](#_Toc52122665)  [5.4.2 New benchmarking frontend 82](#_Toc52122666)  [5.4.3 New benchmarking backend 86](#_Toc52122667)  [5.4.4 Annotation tool 86](#_Toc52122668)  [5.4.5 Status and next steps 88](#_Toc52122669)  [5.5 Minimal Minimal Viable Benchmarking - MMVB Version 2.2 88](#_Toc52122670)  [5.5.1 Benchmarking frontend 88](#_Toc52122671)  [5.5.2 New benchmarking backend 91](#_Toc52122672)  [5.5.3 Annotation tool 91](#_Toc52122673)  [5.5.4 Status and next steps 91](#_Toc52122674)  [5.6 Minimal Viable Benchmarking - MVB 92](#_Toc52122675)  [5.6.1 Architecture and methodology overview 92](#_Toc52122676)  [5.6.2 AI input data to use for the MVB 93](#_Toc52122677)  [5.6.3 AI output data to use for the MVB 93](#_Toc52122678)  [5.6.4 Symptom assessment AI benchmarking interface 94](#_Toc52122679)  [5.6.5 API input data format 94](#_Toc52122680)  [5.6.6 API output data format 94](#_Toc52122681)  [5.6.7 Benchmarking dataset collection 94](#_Toc52122682)  [5.6.8 Benchmarking dataset format 94](#_Toc52122683)  [5.6.9 Scores & metrics 94](#_Toc52122684)  [5.6.10 Reporting methodology 94](#_Toc52122685)  [5.6.11 Technical architecture of the official benchmarking system 95](#_Toc52122686)  [5.6.12 Technical architecture of the continuous benchmarking system 95](#_Toc52122687)  [5.6.13 Benchmarking operation procedure 95](#_Toc52122688)  [5.7 Case creation funding considerations 95](#_Toc52122689)  [5.7.1 Funding agencies 95](#_Toc52122690)  [5.7.2 Funding by benchmarking participants 95](#_Toc52122691)  [5.7.3 Funding by governments 96](#_Toc52122692)  [6 Results from benchmarking 96](#_Toc52122693)  [7 Discussion on insights from MVB 96](#_Toc52122694)  [Appendix A: Declaration of conflict of interest 97](#_Toc52122695)  [Appendix B: Glossary 101](#_Toc52122696)  [Appendix C: References 103](#_Toc52122697)  [Appendix D: Systems not considered in chapter 3 104](#_Toc52122698)  [Appendix E: List of all (e-)meetings and corresponding minutes 106](#_Toc52122699) |
|  |

**List of Tables**

| Page  [Table 1 – FG-AI4H thematic classification scheme 15](#_Toc52122700)  [Table 2 – List of regular TG Symptom input documents 19](#_Toc52122701)  [Table 3 – Symptom assessment systems inside the Topic Group 29](#_Toc52122702)  [Table 4 – Symptom assessment systems outside the Topic Group 32](#_Toc52122703)  [Table 5 – Overview symptom assessment system inputs 33](#_Toc52122704)  [Table 6 – Overview symptom assessment system outputs 35](#_Toc52122705)  [Table 7 – Manchester Triage System levels 36](#_Toc52122706)  [Table 8 – Benchmarking iterations 68](#_Toc52122707)  [Table 9 – MMVB input data format 70](#_Toc52122708)  [Table 10 – MMVB AI output expectation encoding 71](#_Toc52122709)  [Table 11 – Additional fields not included in the MMVB 71](#_Toc52122710)  [Table 12 – MMVB API output encoding 72](#_Toc52122711)  [Table 13 – Case example for the London Model 73](#_Toc52122712) |
| --- |

**List of Figures**

| Page |
| --- |
| [Figure 1 – Example “Model Facts” label for sepsis machine learning model from Sendak et al, 2020. (Nature) 66](#_Toc52122713)  [Figure 2 – MMVB interactive result table 75](#_Toc52122714)  [Figure 3 – Abdominal Pain symptom with attributes inside the Berlin Model 76](#_Toc52122715)  [Figure 4 – Factors with attribute details inside the Berlin Model 77](#_Toc52122716)  [Figure 5 – Refined factor distributions for ectopic pregnancy inside the Berlin Model 77](#_Toc52122717)  [Figure 6 – Benchmarking UI showing the realtime updated benchmarking progress 78](#_Toc52122718)  [Figure 7 – Live logging section of the Benchmarking UI 79](#_Toc52122719)  [Figure 8 – Excerpt of the schema: the definition of “Abdominal pain”, a symptom 81](#_Toc52122720)  [Figure 9 – Excerpt of the schema: the definition of “Finding site”, a symptom-scoped attribute 82](#_Toc52122721)  [Figure 10 – Landing page for the MMVB frontend (as seen in dark-mode) 83](#_Toc52122722)  [Figure 11 – List of registered AI implementations (including system-provided toy AIs) with health status 84](file:///D:/Usr/Campos/outlook/Attachments/Inbox/_Discard/FGAI4H-J-021-A01.docx#_Toc52122723)  [Figure 12 – Statistical analysis of a synthesized case set 84](#_Toc52122724)  [Figure 13 – Parameter selection for starting a new benchmark (currently only case set and participators) 85](#_Toc52122725)  [Figure 14 – Progress view while running a benchmark showing errors, timeouts and in-progress actions. 85](#_Toc52122726)  [Figure 15 – Evaluation page showing different metrics 85](#_Toc52122727)  [Figure 16 – Excerpt of a case vignette showing the presenting complaint with attributes and values 87](#_Toc52122728)  [Figure 17 – Example of a drop-down deep in the data-structure offering the related options 87](#_Toc52122729)  [Figure 18 – 2.2 Version of the Benchmarking start page 89](#_Toc52122730)  [Figure 19 – The AI implementations list now featuring the ability of adding of new AIs and editing existing ones. 90](#_Toc52122731)  [Figure 20 – 2.2 Version of the benchmarking result screen 90](#_Toc52122732)  [Figure 21 – General framework proposed by Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL during workshop associated to FG Meeting A 92](#_Toc52122733) |

# 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.

## 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.

### 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.

### 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 (e.g. cancers)

### 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.

### 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 |

## 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)

### 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.

### 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.

### 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.

### 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.

### 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.

### 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.

# 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

## 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 labelled 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

## Topic description document

The primary output deliverable 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. While in the beginning, the primary way of suggesting changes to the TDD was submitting input documents and discussing them during the official meetings of the Focus Group, the current mode of operation is to join the Topic Group and to discuss changes online between the meetings.

During meeting G the deliverable structure of the Focus Group as a whole was revised. The final output of our Topic Group will be deliverable “DEL10.14 Symptom assessment (TG-Symptom)“. For every meeting the Topic Group is supposed to submit input documents reflecting the updates on the work on this deliverable (Table 2).

Table 2 – List of regular TG Symptom input documents

| Number | Title |
| --- | --- |
| FGAI4H-x-021-A01 | Latest update of the Topic Description Document of TG Symptom |
| FGAI4H-x-021-A02 | Latest update of the Call for Topic Group Participation (CfTGP) |
| FGAI4H-x-021-A03 | The presentation summarizing the latest update of the Topic Description Document of TG Symptom |

## 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 subtopic groups. With Baidu joining in meeting D in Shanghai, the Topic Group was split into the subtopics "self-assessment" and "clinical symptom assessment". The first group addresses the symptom-checker apps used by non-doctors while the second group focuses on symptom-based diagnostic decision support systems for doctors. This document will discuss both sub-topics together. 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).

## 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).

## 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:

### 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.

### 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.

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

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

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

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

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

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

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

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

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

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

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

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

### Status Update for Meeting H (Brasilia) submission

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

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

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

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

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

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

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

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

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

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

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

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

### Status Update for Meeting I (Online) submission

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### Status Update for Meeting J (Online) submission

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

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

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

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

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

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

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

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

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

We also published a topic group internal summary of meeting I that can be found here:

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

Since Meeting I, we have been joined by:

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

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

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

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

**2.6 Next Meetings**

The Focus Groups meets about every two months at changing locations. Due to the ongoing SARS-CoV-2 pandemic the schedule for the upcoming meetings is not clear. 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).

# 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, … )

## Existing Systems for AI-based symptom assessment

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

### Topic Group member systems for AI-based symptom assessment

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

Table 3 – Symptom assessment systems inside the Topic Group

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

### Other systems for AI-based symptom assessment

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

Table 4 – Symptom assessment systems outside the Topic Group

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

## Input Data

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

### Input types

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

Table 5 – Overview symptom assessment system inputs

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

### Ontologies for encoding input data

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

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

#### SNOMED clinical terms

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

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

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

#### Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) (www.human-phenotype-ontology.org) is an ontology focused on phenotyping patients especially in context of hereditary diseases, containing more than 13,000 terms. In context of rare disease it is the most commonly used ontology and was adopted by OrphanNet for encoding the conditions in their rare disease database. Other examples are the 100K Genomes UK, NIH UDP, Genetic and Rare Diseases Information Center (GARD). The HPO is part of the Monarch Initiative, an NIH-supported international consortium dedicated to semantic integration of biomedical and model organism data with the ultimate goal of improving biomedical research[[9]](#footnote-9).

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

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

#### Unified Medical Language System (UMLS)

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

## Output 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.

### Output types

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

Table 6 – Overview symptom assessment system outputs

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

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

#### Clinical triage

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

Table 7 – Manchester Triage System levels

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

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

#### Pre-clinical triage

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

**1DOC3**

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

**Ada Health Pre-Clinical Triage Levels**

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

**Babylon Pre-Clinical Triage Levels**

Generally:

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

With additional information provided per condition.

**Deepcare Triage Levels**

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

**Infermedica Triage Levels**

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

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

**Inspired Ideas Triage Levels**

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

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

**Isabel Pre-Clinical Triage Levels**

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

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

**Symptify Pre-Clinical Triage Levels**

**Visiba Care Pre-Clinical Triage Levels**

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

Depending on the condition additional adjustments possible.

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

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

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

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

#### Differential diagnosis

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

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

#### Next step advice

* to be written

#### Treatment advice

* to be written

## Scope dimensions

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

***Regional Scope***

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

***Condition Set***

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

***Age Range***

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

***Languages***

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

## Additional relevant dimensions

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

***Dealing with "No-Answers" / missing information***

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

***Dialog Engines***

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

***Number of Presenting Complaints***

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

***Multimorbidity***

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

***Symptom Search***

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

***Natural Language Processing***

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

***Seasonality***

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

## Robustness of systems for AI based symptom assessment

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

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

***Memory Stability & Reproducibility***

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

***Empty case response***

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

***Negative evidence only response***

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

***All symptoms response***

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

***Duplicate symptom response***

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

***Wrong symptom response***

Systems should respond properly to unknown symptoms.

***Symptom with wrong attributes response***

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

***Symptom without mandatory attribute response***

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

# 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.

## 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:

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

TODO

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

TODO

### "ISABEL: Accuracy of a Machine Learning Based Ddx Generator

<https://www.isabelhealthcare.com/pdf/DEM_2017_Isabel_Accuracy.pdf>: TODO reference and commentary on relevance to benchmarking

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

### ISABEL: “Asking ISABEL” for diagnostic dilemmas in pediatrics: How does a web based diagnostic checklist perform?

<https://www.isabelhealthcare.com/pdf/ISABEL_PAS_POSTER_2013.pdf> TODO reference and commentary on relevance to benchmarking

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

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

### ISABEL: Performance of a Web-Based Clinical Diagnosis Support System for Internists

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2150633/> TODO reference and commentary on relevance to benchmarking

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

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

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

TODO

### "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.

### "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.

## 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.

### "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.

### "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.

### "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)

### "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)

### "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%.

### "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.

### "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.

### "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.

**4.2.9 "CONSORT-AI and SPIRIT-AI reporting guidelines**

[TOD](http://www.consort-statement.org/checklists/view/32--consort-2010/66-title)O

## 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.

## 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
* US Food and Drug Administration (FDA)
  + *Clinical Decision Support Software Draft Guidance for Industry and Food and Drug Administration Staff, Draft guidance, issued on September 27, 2019*
* The Center of Medical Device Evolution in China (CMDE)
  + *Verification Points for Decision Supporting Medical Device Software based on Deep learning(深度学习辅助决策医疗器械软件审评要点及相关说明),* [*https://www.cmde.org.cn/CL0030/19342.html*](https://www.cmde.org.cn/CL0030/19342.html)*, June 28, 2019*
* ISO13485
* (CE)
* clinical trials, evidence levels (RCT's etc.)
* scores & metrics used

## 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 positions

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

The role of the secondary matches

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

discuss scores & metrics used

## 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.

### 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 sandbox. It must not be possible to access the filesystem of the benchmarking machine, databases, network etc. The AI must not be able to leak the test data to the outside world, nor see the correct labels, nor manipulate the recorded benchmarking results, nor access other AIs or their results. The experience with protecting all kinds of manipulation attempts is the biggest advantage that using a ready-made benchmarking platform could provide.

Online-Mode

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

### 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 sandboxing and the design of the platform itself outweighs the work of changing an existing platform to the Topic Group’s needs, compared to implementing a new specialized solution.

### Other platforms

TODO

* analyse kaggle, …

## 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.

### Outline of clinical considerations for scores and metrics

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

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

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

Examples of stakeholders that might require clinically relevant metrics:

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

It must be noted that clinical evaluation of AI tools in healthcare is already performed through clinical trials. The special, nuanced considerations over and above clinical studies for medical hardware, medications or surgical interventions have been carefully addressed in the [CONSORT-AI reporting guidelines](https://www.nature.com/articles/s41591-020-1034-x/tables/2) released in September 2020. In addition, as part of Software as a Medical Device Regulations (e.g. the FDA, EU MDR), it is a requirement that companies building these tools carry out Post Market Clinical Follow Up (PMCF) in order to demonstrate claimed clinical benefits really do occur in the real world.

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

*Oversight*

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

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

### What do we mean by clinical metrics?

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

* Performance and Accuracy measures
* Safety measures
* Clinical outcome measures

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

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

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

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

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

### More complex than image classification

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

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

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

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

### Important considerations

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

1. *The Ground truth problem*

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

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

|  |  |  |
| --- | --- | --- |
| Gold Standard Case Vignettes | Creation of vignette cases by clinicians. In these, the gold standard conditions and triage level can be defined by the author    Examples in literature include Semigran et al paper | These might be based on clinician experience or real cases they have seen, and as such are subject to the clinicians own heuristics and biases.    Require quality and peer review. Consensus on gold standards and disagreement/variability in opinion between clinicians is common.    Clinician opinion is of lower certainty compared to a definitive imaging/lab/histopathology report (we have learned from our colleagues in these topic groups that even for histopath/imaging reports, there can be issues with gold standards owing to intra and inter-observer variability) |
| EHR records (retrospective) | Converting anonymised/pseudonymised electronic health record episodes into cases, with the coded condition as Gold standard. | There are many issues with using EHRs as ‘gold standard’.  Examples:   * Coding issues (usually optimised for billing) * Depend on whether the clinician actually documented info (for example important negative information) * May not the same point of information (outlined below) * The final diagnosis may occur later down the line (after the evolution of symptoms and time, as well as added data points e.g. labs imaging) * Geographic variability * Lack of standard structure |

*2. Defining metrics in the context of point of information*

*Scope of data*

To create a safe and effective symptom checker, the datasets used in each symptom checker need to have defined and standardised limits with respect to information that can be asked of the user, and to conditions that are provided to the user as possible outcomes. Symptoms are the most logical inclusion in a symptom checker, as the name would suggest; as features that can be described by the user in a typical medical history-taking exercise, they reasonably easily translate to a symptom-checking environment (notwithstanding the nuances of symptom descriptions). But while there has historically been great emphasis on the power of a well-taken medical history in determining diagnoses[[23]](#footnote-23), other elements of a medical assessment add critical data that enhance the diagnostic process. To this end, the criteria for inclusion becomes more complex.

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

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

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

*Defining the scope*

When considering which information is considered appropriate for a symptom checker, it is also important to consider who defines which information is considered appropriate. As there is no established gold standard, the most appropriate method of defining the boundaries of these metrics is to use a panel of clinicians to reach consensus. This is a method also used in clinical medicine for identifying diagnostic reference standards in the absence of gold standard tests[[24]](#footnote-24). The use of ‘expert panels’ for diagnosis is not without issues, such as intra- and inter-observer variability. It is therefore important to define a clear methodology for how consensus will be reached to minimise this variability and increase reproducibility as much as possible. Furthermore, the make-up of members on the panel is dependent on the focus of the symptom checker, so it needs to be considered whether the panel is limited to general practitioners, or a mix of general practitioners and specialists, or whether there are multiple panels appropriate for different conditions or presentations. Numbers of panel members also need to be decided (an odd number is most practical), as does the criteria for inclusion on a panel (e.g. area of expertise, number of years of experience). There is also the consideration of whether a panel of patients is necessary, particularly for defining the validity of user symptoms and the finer details relating to language and presentation of information.

*Getting the comparison right*

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

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

*3. Performance does not equal Utility*

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

*4. Different metrics for different use cases/contexts*

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

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

Some examples of contextual variations include:

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

### Metrics for symptom assessment

*(note, through this section, the terms tool and app are used interchangeably, and refer to patient facing symptom assessment tools)*

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

Performance and Safety measures (How accurate is this device? How safe is this device?)

Clinical outcome and income measures (How does it impact clinical practice - for patients, clinician workflow, or health systems)

*Some basic considerations*

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

* **Condition suggestions.** Some tools provide an ordered list of possible conditions (with varying nomenclature, the most common being differential diagnosis), others have an unordered list of possibilities. Another variable between tools is the indication of probabilities for each suggestion. Sometimes informally any of this is called a “[differential] diagnosis” but it is only an informal name due to the fact that symptom assessment apps generally don’t have enough accuracy (yet); also due to liability considerations and due to important safety & regulatory reasons.
* **Pre clinical triage.** This gives advice about what level of care the user should seek. There is a large variance between tools in the exact levels, which can range from ‘self care’ to “call ambulance”. The triage advice might be given as direct advice or just as “information” (the choice depends on the accuracy of a tool and the estimation of that accuracy by an app provider; by the amount of the liability that an app provider would like to take; on regulations requirements in a particular country; etc.).

There are other tasks that require their own metrics

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

**Performance and Accuracy**

*Condition Suggestions*

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

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

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

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

* The top match condition metric assumes that it is always possible to get to the “final diagnosis” indicated in the gold standard from the inputs available to a symptom assessment tool. In other words, it assumes that taking the information that is available from a medical history will be enough to accurately predict what the patient has in future. This has been discussed further in the Point of Information section. Much depends on the stated intended use of the tool, but in general, at the time of writing, it is not a goal (and it simply can’t be a goal) to give a diagnosis. Aside from the most clear cut cases, the process of diagnosis requires so much more information: observing the evolution over time, clinical examination, bedside tests, lab and imaging tests. (in future, we will be able to come back to this when benchmarking clinician facing decision support tools)
* The top 3 or top 5 condition matching metric is useful as it starts looking at the list of conditions suggested by the AI tool. However, there is no measure of how good the other suggestions on the list are.

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

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

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

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

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

**Similar presentations, varying outcomes**

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

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

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

**Basic Metrics for Differential Diagnosis**

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

Let’s say a patient has a specific presentation (including: symptoms, if any; their medical history; etc.) and let’s say that there is the “ground truth” that he/she has e.g. new presentations of conditions {X\_1, …, X\_n} (where n is likely to be equal to or less than 1), and does not have any other conditions.

Let’s say that a symptom assessment tool, after collecting all information it could collect from a patient, has identified that a patient might have some conditions {Z\_1, …, Z\_m} (for simplicity, let’s assume that the app just says whether each condition is present or not; generally, it might return some likelihood/probability of it, or some other degree of certainty).

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

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

|  |  |
| --- | --- |
| Recall (also called: true positive rate, sensitivity) | It is the ratio of conditions “of interest” that the patient has and which were identified by the app (i.e. presumed by the app to be happening to the patient) to the number of the conditions that the patient has.    Note that in the case of a “simple” “one-condition diagnostic test” for a specific condition, the recall is much “simpler” and is usually calculated in an aggregated way across all patient cases: it is the ratio of sick people (i.e. people with that specific condition) correctly identified as sick (i.e. presumed to have that specific condition) to the number of sick people (i.e. having that specific condition). |
| Precision  (also called: positive predictive value) | It is the ratio of conditions that the patient “of interest” has and which were identified by the app to the number of the conditions that the app has identified. |
| F1-score and Fn-score | It is the harmonic mean of precision and recall. In Fn-score, recall is considered n times as important as precision (n > 0). |
| Specificity  (also called: selectivity; true negative rate) | it is the ratio of the conditions “of interest” that the patient does not have and which were not flagged (i.e. were not highlighted as present/”likely”) by the app to the conditions which the patient does not have.    Note:   * False positive rate (also called: fall-out or false alarm ratio) can be calculated as follows: 1.0 - specificity. * Since there are quite a lot of conditions that a patient might have in general, and since symptom assessment apps generally can rule out most of conditions that a patient does not have, specificity might often be close to 1.0. * Because of that, a receiver operating characteristic curve (ROC curve) (calculated over many cases, not just for one case) that is created by plotting the recall (sensitivity) against the false positive rate (i.e. 1.0 - specificity) might look almost “trivial” in many cases since the false positive rate might often be close to 0.0. |
| Accuracy | Different things might be meant by “accuracy” and there are multiple ways to define “accuracy”. Informally, different metrics can be called “accuracy”. Because of this, it is recommended to always clarify what is meant by “accuracy” if this term is used. |
| Top-N  (Top matching condition) | One of the simple ways to calculate it: for each case top-N is equal to 1.0 if an app’s output (i.e. a “differential diagnosis”) contains the condition of “interest” in its top N conditions in the output (assuming that the app returns ordered conditions). If there are M conditions of “interest” for the case, and K of them are in the top-N conditions from an app’s output, then top-N for the case for the app is equal to K/M. Note that an app might return less then N conditions in its output for a case: in this scenario, if an app returns J < N conditions, it might be assumed that top-N is equal to top-J for this particular case for this app (note that J might be equal to 0).    Note that if an app provides an unordered list of conditions, then it is not possible/trivial to calculate Top-N. |

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

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

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

It should also be noted that there are always special cases: some patients might be healthy or they might have a condition that is not known by an app, and so for the purpose of the metric calculation there might be zero conditions of “interest” that a patient has. Such situations might cause recall to be ill-defined. Also, an app might return an outcome with no conditions at all (i.e. it might assume that a patient does not have any conditions at all, at least of those that it is aware of), and in this case the precision might be ill-defined. Special rules must be applied for such cases as appropriate.

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

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

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

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

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

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

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

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

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

*Basic metrics for triage*

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

One way to measure triage performance is to have one “perfect”, “ground truth” triage option for each patient case (provided by an expert clinician/panel of experts), and to match it with a tool’s triage, such that there is a match or no match. This can be averaged (and weighted if appropriate similarly to the differential diagnosis as described above) across multiple patient cases. This way we capture the “accuracy” of triage.

However, this approach has limitations because:

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

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

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

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

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

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

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

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

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

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

About statistical significance of metrics

[To be added]

“Ground truth” for differential diagnosis/triage

[To be added:]

How to estimate it?

Should it be perfect?

How to average if we have multiple approximations of ground truth from different experts. How to use an expert panel’s opinion?

On balancing/biasing/weighting/stratification

[To be added]

**Other performance measures**

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

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

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

* Quality of question flow: How much of the relevant positive and negative evidence was actually elicited? Were there any key serious symptoms (red flags) not elicited by the tool?
* Do users actually understand/comprehend the questions? (this may be difficult to assess within benchmarking. It might be captured within usability testing)
* Measuring the task of converting the lived experience of the user, into the tool. Different tools use different methods for this - some use Natural Language Processing, for example. It is important to have metrics on the performance/accuracy of this task.
* Metrics to consider that there are certain conditions that may be explicitly ruled out due to the age or the biological sex of the user. An example of this would be pregnancy related conditions (e.g. pre-eclampsia) should usually not be included in the list of differentials for someone whose biological sex is male[[25]](#footnote-25))

**Clinical Outcomes/Impact**

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

Examples of clinical outcomes and impact measurements could be:

**Patient Journey:**

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

**Clinical Workflow**

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

**Health System**

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

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

### Putting it all together for clinicians

A big challenge is communicating this range of metrics (relevant to context) to clinicians and other healthcare stakeholders, in a way that aids understanding. A paper published in Nature in March 2020 by Sendak et al explores this in great detail, and provides a fantastic example inspired by the nutrition information on a cereal box[[26]](#footnote-26). Benchmarking metrics summarised in a such a digestible way that takes the most relevant parts is worth seriously considering.

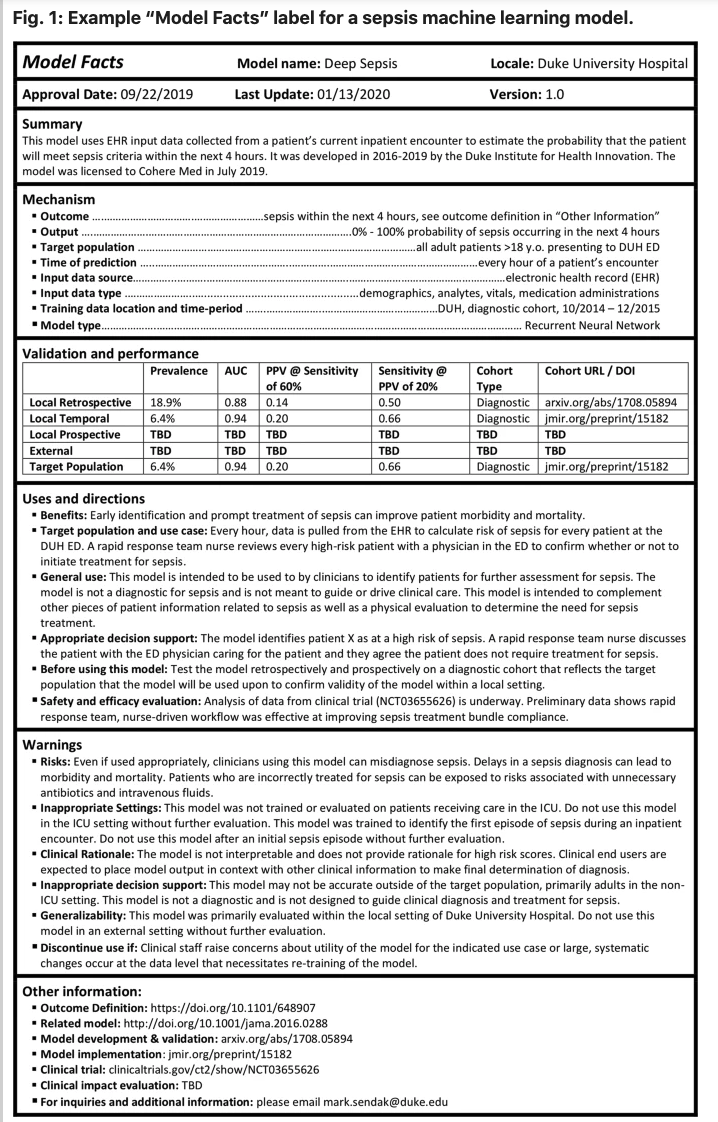


Figure 1 – Example “Model Facts” label for sepsis machine learning model from Sendak et al, 2020. (Nature)

### Additional clinical considerations and limitations

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

*Mapping Ontologies*

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

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

*Explainability*

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

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

*Addressing clinician variability and the ground truth problem going forward*

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

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

### Conclusion

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

* TODO Mean average TopN accuracy for recommendation task (such as diagnosis or treatment recommendation) for given input: N can be set to 1, 3, 5 or Dynamic truncation
* TODO Specificity and Sensitivity for given diagnosis
* TODO Performance: QPS in specific hardware and software environment
* TODO Reliability: Continuous stable running time rate in given time period
* TODO Security
* TODO Robustness
* TODO Performance

# 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.

## 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. Table 7 gives an overview of the different versions and their purpose.

Table 8 – Benchmarking iterations

| 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 |
| MMVB#2.1 | Minimal Minimal Viable Benchmarking Version 2.1 | * a new dedicated benchmarking frontend * a new backend infrastructure * a first simple case annotation tool |
| *MMVB#2.2* | *Minimal Minimal Viable Benchmarking Version 2.2* | * full implementation of the Berlin model in frontend, backend and annotation tool * improve AI error handling / health check * improved usability of the frontend |
| MVB | Minimal Viable Benchmarking | * first benchmarking with real AI and real data * Target: end of 2020 |
| Vx.0 | TG Symptom Benchmarking Vx.0 | * the regular e.g. quarterly benchmarking for this topic group * continuous integration of new features |

## 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.

### 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](https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit).

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 these 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.

### AI input data

The MMVB uses as input for the AIs a simple user profile, explicit presenting complaints (PC/CC), and additional complaints. The additional complaints might also contain risk factors. Table 8 shows the concrete fields with corresponding examples.

Table 9 – MMVB input data format

| Field name | Example | Description |
| --- | --- | --- |
| profileInformation | "profileInformation": {  "age": 38,  "biologicalSex": "male"  } | * General information about the patient * Age is unrestricted, however for the case creation it was agreed to focus on 18-99years. * 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.

### Expected AI output data encoding

In addition to the "public" fields given to the AIs for inference, the generated case data also encodes the expected triage and the diagnosis outputs (Table 9).

Table 10 – MMVB AI output expectation encoding

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

The group also discussed the fields shown in Table 10, but they are not part of the MMVB data yet.

Table 11 – Additional fields not included in the MMVB

| 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. |

### 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.

### 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 of the scores of the group members is different and not comparable.

Table 12 – MMVB API output encoding

| 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.

### 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. Table 12 shows a case example illustrating the structure.

Table 13 – Case example for the London Model

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

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

As part of the work for meeting H, the MMVB was extended by supporting benchmarking based on the cases manually created by our doctors.

### 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.

### Reporting

For the reporting the MMVB uses a simplistic web interface rendering an interactive table with all scores for all AI systems Figure 1).

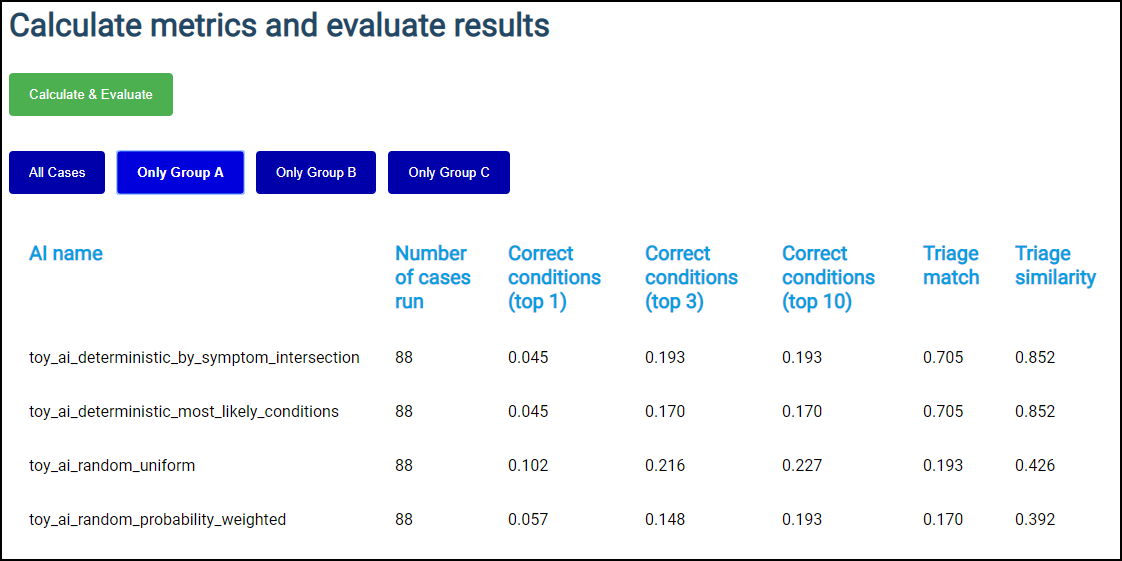


Figure 2 – MMVB interactive result table

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.

### 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.
* 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).

## 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:

### 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 Figure 2.

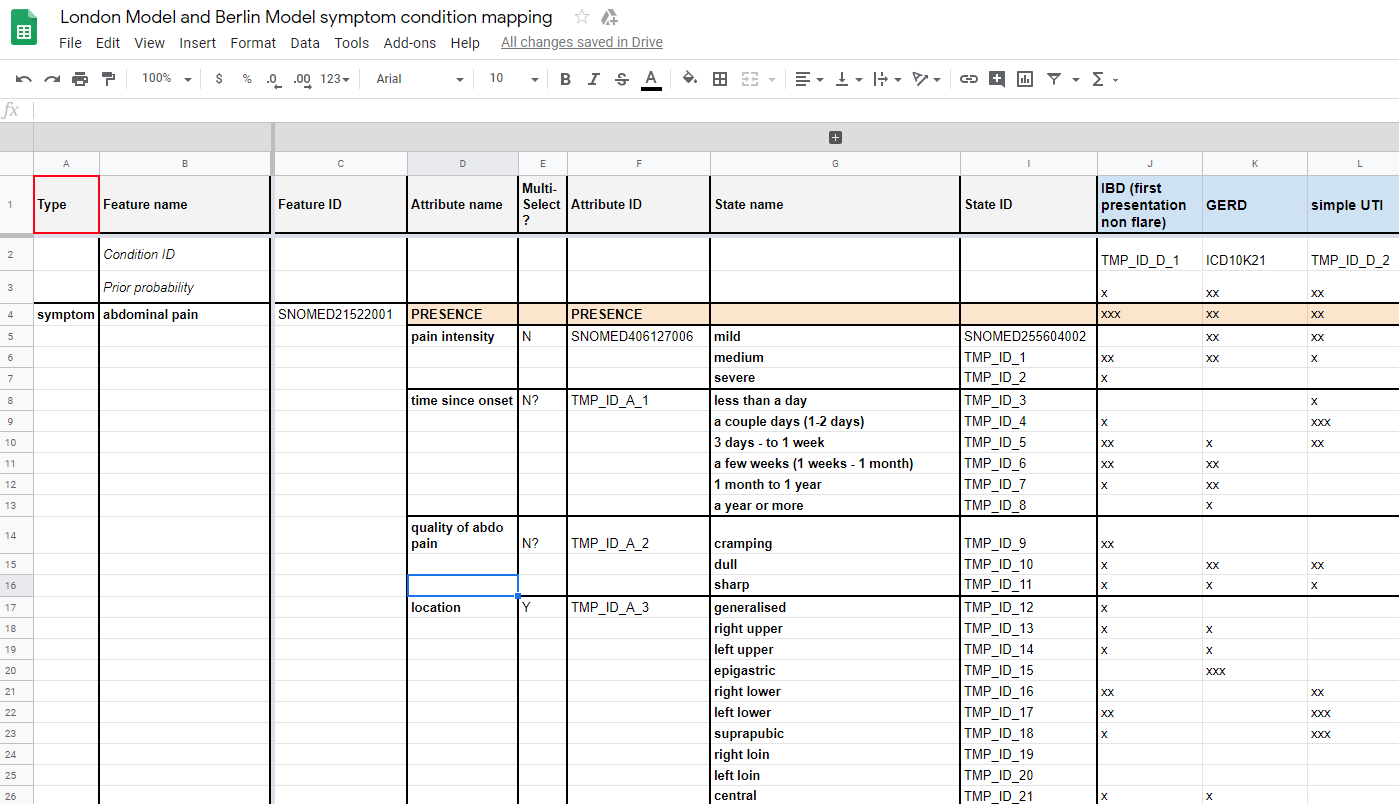


Figure 3 – Abdominal Pain symptom with attributes inside the Berlin Model

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).

### 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 as shown Figure 3 in and Figure 4.

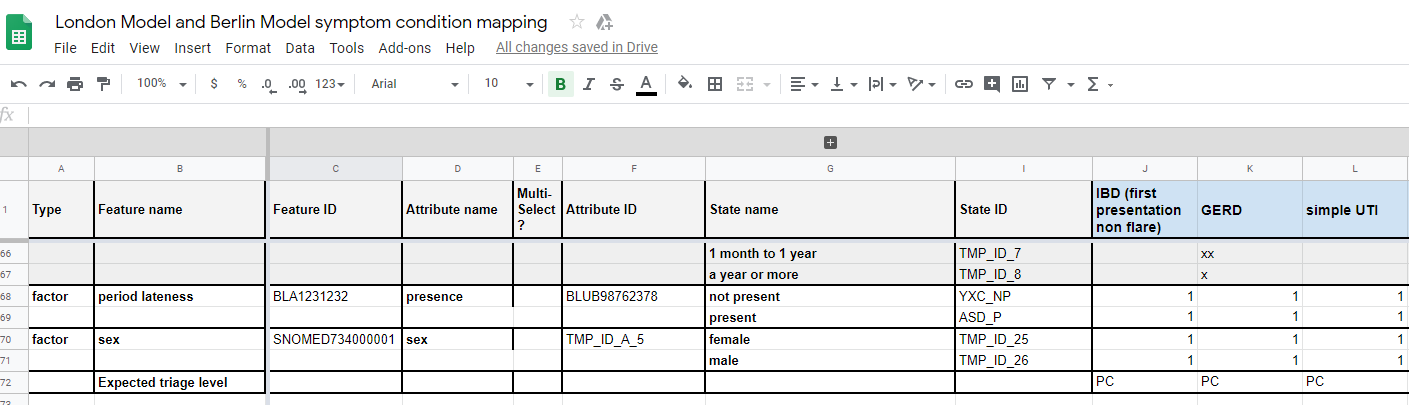


Figure 4 – Factors with attribute details inside the Berlin Model

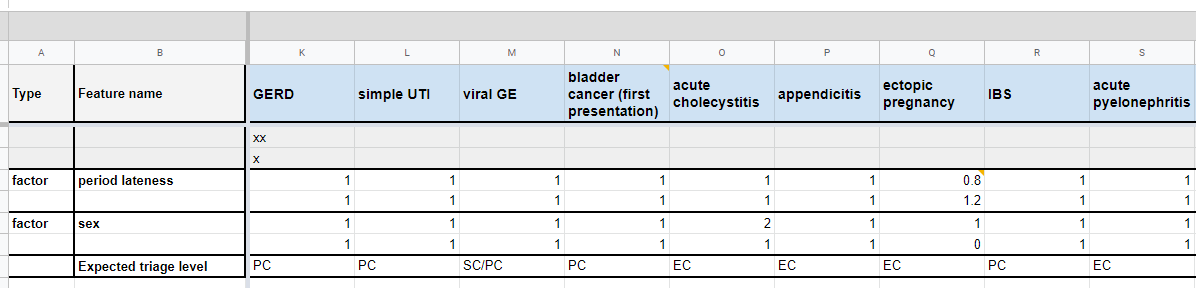


Figure 5 – Refined factor distributions for ectopic pregnancy inside the Berlin Model

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

### 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.

### 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".

### 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.

### 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.

### Updated benchmarking UI

For reflecting the changes in the benchmarking processes the web based user interface for the benchmarking systems was 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 5).

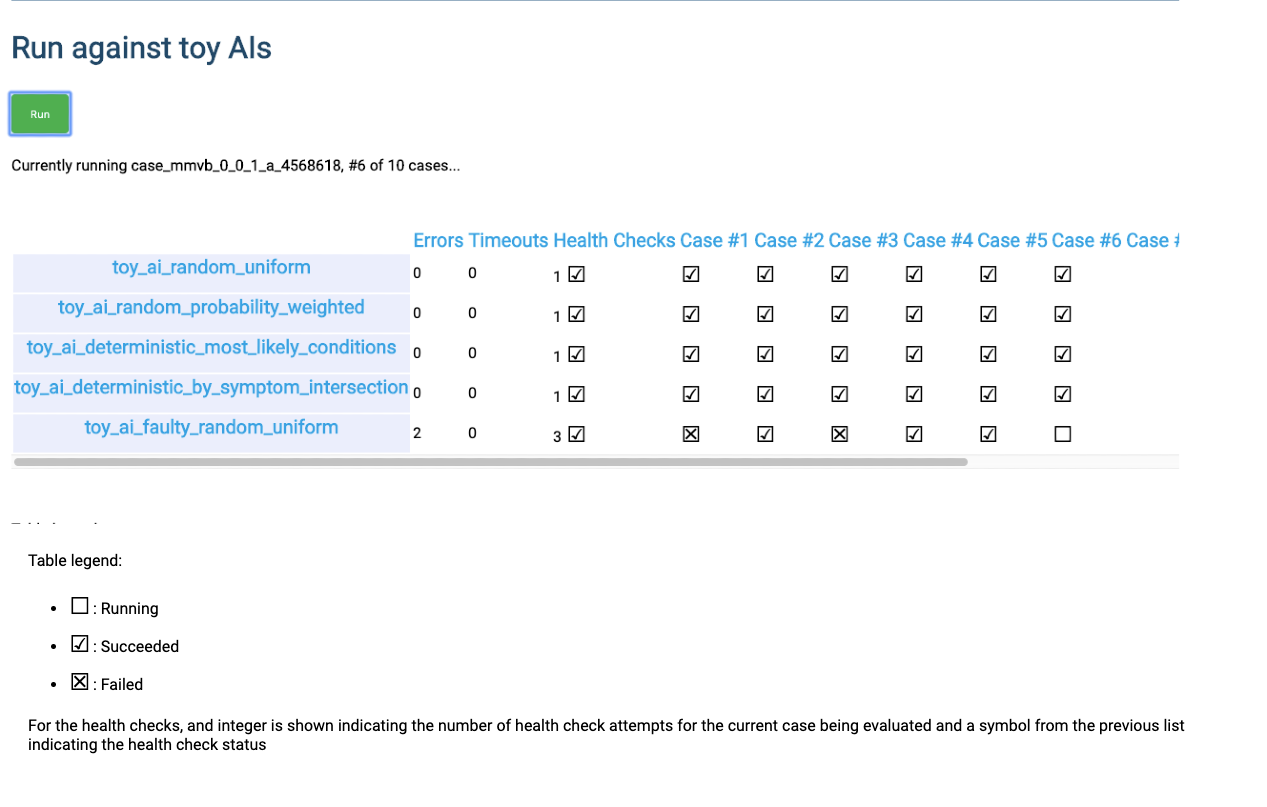


Figure 6 – Benchmarking UI showing the realtime updated benchmarking progress

The new MMVB 2 interface also output logs in real time (Figure 6).

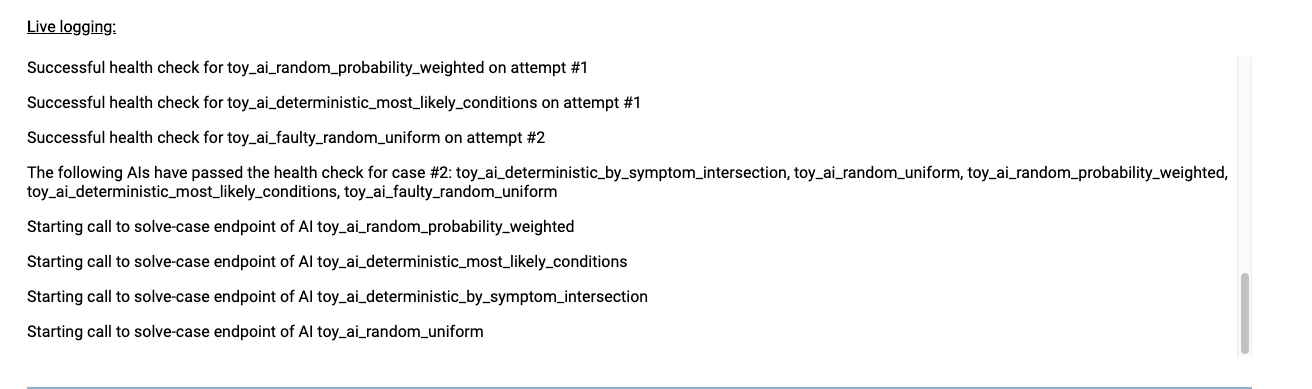


Figure 7 – Live logging section of the Benchmarking UI

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.

### 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 organizations 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.

## Minimal Minimal Viable Benchmarking - MMVB Version 2.1

Continuing the work on the MMVB version 2.0 described in section 5.3 the focus after meeting H was preparing the benchmarking platform for the implementation of the Berlin model.

In a joint document the requirements for the MVB benchmarking system as well as the intermediary MMVB were outlined again. Based on this first steps were identified to be taken to both lay a solid foundation for future development as well as approaching the MVB.

Among these requirements are:

* user accounts with fine-grained access-levels
* a formalized case annotation/creation interface
* a review process for cases/case sets
* running benchmarks without user interaction/supervision
* scheduled benchmarks
* semi-interactive benchmarks
* domain-specific metrics
* interactive drilldowns of results

These changes are substantial and while the first iteration of the MMVB was fit for its purposes it was deemed necessary to undergo a complete re-write for both frontend and backend. To accommodate increasing complexity needs in the frontend, it was decided to split the development effort off into a separate code base / repository. The main aim was to provide a future-proof, extensible, and scalable foundation for future development. The two systems communicate via a semantic REST API using JSON for transport serialization. Documentation for the API is auto-generated from the backend implementation and is available in the OpenAPI format as well as a human friendly web version.

To facilitate collaboration all source code as well as tickets for technical planning are hosted in a joint GitHub organization with access right distributed among contributors.

### Input and Output ontology encoding and model schema

As the London model, the Berlin model is also based on the simple domain with only 12 diseases chosen by the doctors in the topic group. It did not use any ontology or terminology and all the keys have been chosen more or less arbitrary. After Alejandro Osornio joining the topic group in meeting H we started to work on an approach for mapping this simple model to the SNOMED CT ontology. The goal was to use as much of the SNOMED CT knowledge representation capabilities as possible, complemented with an ad-hoc information model that adds attributes and values for the unsupported requirements. Additional attributes and values are defined using SNOMED CT concepts as well, and in the future, it could be represented into a possible SNOMED concept model extension. The initial experiments worked well for all concepts in the model – including the attributes specifying the details of symptoms. Based on these results the topic group decided to switch to an explicit SNOMED CT ontology mapping for the implementation of the next MMVB iteration.

Based on the above decisions the Berlin model was encoded as a JSON Schema. JSON Schemas are an emerging standard for platform-independent specification of data structures, usually used for data expressed in JSON. JSON Schema is foundational to the OpenAPI specification, thus a complete schema for the Berlin model is an important step towards an AI implementation endpoint API specification.

For now, the Berlin model exists as a stand-alone ontology that links to SNOMED in several ways. Wherever possible the concepts link to related SNOMED concepts by including a URI to SNOMED in a list of standard ontology URIs. Other ontologies could be added in the future to extend these lists. Further the

* Validation / encompasses every possible “legal” case
* Possibility to generate input forms

A screenshot of a social media post

Description automatically generated

Figure 8 – Excerpt of the schema: the definition of “Abdominal pain”, a symptom

A screenshot of a social media post

Description automatically generated

Figure 9 – Excerpt of the schema: the definition of “Finding site”, a symptom-scoped attribute

### New benchmarking frontend

Previously the frontend was a single file containing all necessary code to communicate with the backend and run basic benchmarks. Crucial data was stored in-memory and it was impossible to e.g. return to a running benchmark or viewing the results of a benchmark that was run in another browser. This was deemed unscalable and it was thus decided to build a new frontend from scratch which should meet the following criteria: state-less, proper API communication, interactive, user-friendly, extensible (to in the future include features like interactive drilldowns).

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

A screenshot of a cell phone

Description automatically generated

Figure 10 – Landing page for the MMVB frontend (as seen in dark-mode)

A screenshot of a cell phone

Description automatically generated

Figure 11 – List of registered AI implementations (including system-provided toy AIs) with health status

A screenshot of a cell phone

Description automatically generated

Figure 12 – Statistical analysis of a synthesized case set

A screenshot of a cell phone

Description automatically generated

Figure 13 – Parameter selection for starting a new benchmark (currently only case set and participators)

A screenshot of a computer

Description automatically generated

Figure 14 – Progress view while running a benchmark showing errors, timeouts and in-progress actions.

A screenshot of a video game

Description automatically generated

Figure 15 – Evaluation page showing different metrics

### New benchmarking backend

With a view towards adding future functionality required for the MVB such as improved user authentication and authorisation, it was decided to reimplement the original backend implemented using Flask using the Django framework. Django is a well stablished and well documented Python framework that provides multiple relevant features out of the box. It plays well with other frameworks that will make things easier to extend in future, for example Django Rest Framework for the development of REST API. Django has most of the basic foundation work already implemented which allows developers to focus more on the development of features. A further advantage is that it provides an out of the box solution for user accounts which may be customised at some level for different users and permission levels. This will make it easier for developing the MVB where we need to develop features related to different types of users or task (for example submitting a case set or running a benchmark). Django also provides an out of the box customisable solution for admin management which would allow a simple UI to be implemented for admin management on the backend if required.

The previous backend separated each aspect of the benchmarking process (case generation, toy AIs, case evaluation and metrics calculation) into independent microservices. For the new backend it was decided that this microservices architecture was unnecessary, and is instead implemented as separate Django applications within the same project.

Previously all data (such as cases) were contained within the GitHub repository as spreadsheets or static json files. As part of implementing the new backend it was decided to implement a database to contain cases and other data relevant to the benchmarking application. MySQL was chosen as the database, as Django works in the relational realm and MySQL is a stable solution. It was decided that there weren’t any specific specifications that would require us to be concerned about any performance specific details that would motivate the use of an alternative database. Django provides an out of the box Object Relational Mapping (ORM) layer to interact with MySQL.

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

### Annotation tool

The benchmarking for the MMVB iterations uses mainly synthetic data sampled from the London model / Berlin model defined by the doctors in the topic group. However, for learning how to create representative real-world cases for the later MVB version of the benchmarking the topic group also uses cases created by doctors. So far, we used spread sheets for creating the cases, however with the introduction of the Berlin model this reached its limits and it became clear that already for the Berlin model a dedicated annotation tool would be needed.

A first draft of such an annotation tool has been implemented. It is a semi-generic React web solution, based on the above described JSON Schema. It is generic in that it deduces the available options for any particular field from the schema and thus if a symptom with the associated attributes/values were to be added to the schema they could be used in the case creator without further changes. It is not generic in that is hand-crafted to the structure of case vignettes. This trade-off can be generalized as one between reusability (such as for other focus groups, as discussed in [FG-AI4H-H-038-R01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-H-038-R01.docx?d=wbd600bafe71d48038ecab7784b6a3171&csf=1&e=NZwGkA)) and user experience. The latter will be of high importance once a large number of cases will need to be created by human medical doctors, which will happen through the web tool.

By being almost exclusively composed of (multi-)selection fields all cases created are automatically valid according to the schema. The few free inputs (such as age) are validated automatically based on their schema definition.

For the time being it is exclusively aimed at the creation of synthetic cases. It does not incorporate ways of extracting information from a medical record in a documented fashion. Further there is no review mechanism implemented, but the (backend) infrastructure has been designed in a way to accommodate this. No user experience plans exist for the review.



Figure 16 – Excerpt of a case vignette showing the presenting complaint with attributes and values

A screenshot of a cell phone

Description automatically generated

Figure 17 – Example of a drop-down deep in the data-structure offering the related options

### Status and next steps

This iteration of the MMVB was mainly intended as rewrite and refactoring to enable the implementation of the Berlin model benchmarking. After finishing this work until mid-Maya 2020, the next step is then to implement the Berlin model and to perform the benchmarking for it. The steps for this include:

* Finalization of the backend rewrite
* Finalization of the new frontend implementation
* Finalization of the new case annotation tool
* Implementation of authentication
* Implementation of a Berlin model case synthesizer
* Implementation of the new scores and metrics discussed in Berlin workshop
* Publication of the new Toy-AI API specifications
* Implementation of a Berlin model Toy-AI by each TG member
* Test of case creation with the annotation tool by the TG doctors
* Conduction of the benchmarking on the Berlin model

Once the benchmarking of the Berlin model is running, the next big step towards MVB is then to switch to a much ontology supporting the all the factors, symptoms and attributes that would be needed for benchmarking the real AI systems.

## Minimal Minimal Viable Benchmarking - MMVB Version 2.2

Version 2.2 of the minimal minimal viable benchmarking concludes the transition from the London Model to the Berlin Model started after the workshop defining the requirements in October 2019. Most of the necessary changes and extensions in the software have already been described in section 5.4 Minimal Minimal Viable Benchmarking - MMVB Version 2.1, and so this section focuses on the learnings from the finalization of this work.

### Benchmarking frontend

With version 2.1 we separated into a frontend and a backend application. The main motivation was enabling the upcoming work on more complex features for the minimal viable benchmarking. The first version has been implemented in a way that is compatible with the London model. With the current version the data structures, the user interface (UI) and also most of the server communication have been adapted to the new attribute and factors structures. We also introduced some features for increasing the usability and visual appearance. The latest version of the start page can be seen in Figure 17. It now features short cards as shortcuts to the AI-Implementations, datasets and benchmarking sessions, all with indications of the number of entities currently available there. We also added an extra navigation bar on the left side to have all relevant options permanently available.

A screenshot of a social media post

Description automatically generated

Figure 18 – 2.2 Version of the Benchmarking start page

Another added feature is that AI developers can now themselves register new AIs and also change the end points if necessary. For the current phase this has proven to be more practicable and is also closer to the MVB where AI developers would register and submit AIs for official benchmarking too. The current version, however, has not any protection mechanisms implemented so that e.g. all AI developers could change the endpoints of all the other AIs. Adding authentication, rights and roles is therefore one of the important next steps towards an MVB. Figure 18 show the list of AI implementations with the new features.

A screenshot of a cell phone

Description automatically generated

Figure 19 – The AI implementations list now featuring the ability of adding of new AIs and editing existing ones.

The latest version of the benchmarking result screen can be seen in Figure 19. In this latest version we introduced more human readable names for the AIs instead of keys and also added details like the name of the dataset that was used for generating the benchmarking results

A screenshot of a cell phone

Description automatically generated

Figure 20 – 2.2 Version of the benchmarking result screen

### New benchmarking backend

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

Based on the new data structures we also changed the API calls to the toy AIs to include new attribute and factor details in the case data. This implies also that all topic group members had to update their toy AIs to support the new model. For the build-in toy AIs “Weighted Random Conditions Solver” and “Uniform Random Condition Solver” this was already implemented by the team working on the backend. They choose a condition at random or weighted by condition prior and now consider the new factor model within the calculation. At the time of submission of the meeting J version the partners have started to work on updating their toy AIs, with one of them finished and the others to follow in the upcoming weeks.

### Annotation tool

As with the frontend and the backend, we also adjusted the case annotation and creation tool to the Berlin model. Since it already supported attributes and the new factors in the last iteration, the changes here focused adding and refining the encoding of the expected case outcomes. In the week before meeting J it reached the point where doctors could now start to use the tool for creating case sets for the benchmarking.

### Status and next steps

With the completion of the Berlin model implementation the next steps are using the Berlin model and then planning and implementing the next iteration. For using the Berlin model the next steps are:

* Implementing the transfer between the case annotation tool and the benchmarking
* Using the annotation tool to create a dataset created by doctors
* Implementing more Berlin model toy AIs

In parallel to this the topic groups plans to have a 2-day online workshop for specifying the next iteration of the model and the benchmarking system. The most relevant points to address there are:

* Implementation of authentication, rights, roles
* Design and implementation of the dataset and AI meta data annotation system (sometimes called “dimensions” system)
* Agreeing on a SNOMED sub-space, extended with a binding of symptoms to attributes and attributes to states that can be used for describing the full AI diagnosis relevant symptom space.

The last point in particular is the largest and most important open task for implementing the first real minimal viable benchmarking.

## Minimal Viable Benchmarking - MVB

### Architecture and methodology overview

Figure 7 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):

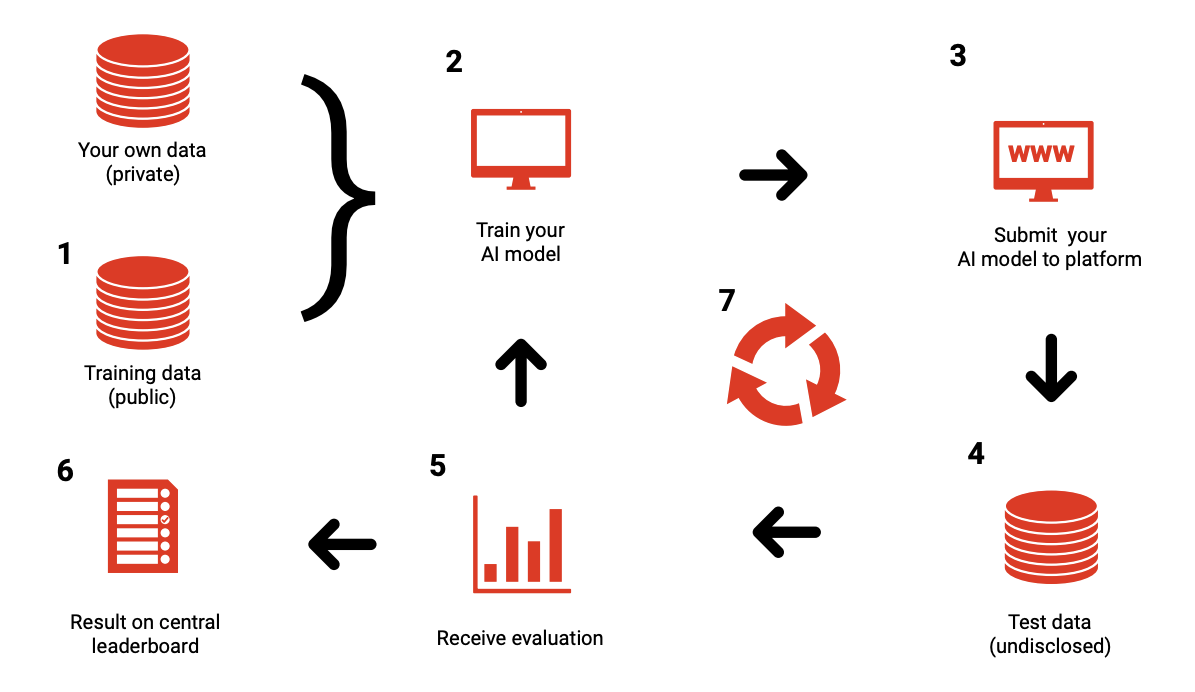


Figure 21 – General framework proposed by Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL during workshop associated to FG Meeting A[[27]](#footnote-27)

**(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.

### 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):

### 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

### Symptom assessment AI benchmarking interface

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

### API input data format

* TODO: JSON Format

### API output data format

* TODO: JSON Format

### 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)

### Benchmarking dataset format

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

### Scores & metrics

* which metrics & scores to use for benchmarking
  + 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 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

### 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?

### Technical architecture of the official benchmarking system

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

### Technical architecture of the continuous benchmarking system

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

### 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

## **Case creation funding considerations**

The benchmarking of this topic group relies on representative high-quality test data. It was identified that the few sources of sufficient quality have been used by the companies of our topic group to build their systems and therefore cannot be used for benchmarking. It was agreed by the topic group that the most viable solution is to design a reproducible process that creates the benchmarking data without any topic group member or participant in the benchmarking ever getting access to the data, to guarantee independent and fair results. However, this process requires a significant amount of funding for clinicians to create case cards which can be used instead.

### Funding agencies

As one of the options the group identified different organisations for funding. It was therefore agreed to approach various of these organisations to compile the requirements and conditions for getting financial support for the creation of case cards. So far, the topic group has reached out to the Wellcome Trust and Botnar Foundation, as both funding bodies have been involved in the ITU/WHO AI4H focus group from the very beginning. The results from the ongoing discussions will be summarized in this chapter in upcoming versions of the document. This might also include approaching additional funding agencies to secure sufficient funding for the generation of case cards on a regular base.

### Funding by benchmarking participants

Beside funding agencies, an alternative option could consider funding by the companies participating in the benchmarking. Given that the benchmarking results help companies to prove that their solutions are viable for certain contexts, this potentially generates revenue that could be used to contribute to the case creation. However, this approach would effectively exclude small or new companies that might provide dedicated innovative new solutions for e.g. a local context. Thus, it is currently not considered the preferred approach.

### Funding by governments

During meeting G in Delhi, it was discussed for the first time to investigate the option to have a topic group overarching system for collecting and annotating cases in a distributed way. Such a platform might be a potential future key part for the world’s digital health infrastructure. And the regular creation of case data for relevant modalities might be recommended by the WHO to guarantee that every country is included in a representative way. In the following meetings the focus group will investigate the potential of such a platform and related models for case creation.

# 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 mid 2020.

# 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 transparency rules this section lists the conflict of interest declarations for everyone who contributed to this document.

1DOC3

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

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

Ada Health GmbH

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

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

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

Babylon Health

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

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

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

Baidu

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

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

***Barkibu***

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.

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

***EQL***

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

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

Infermedica

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

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

Inspired Ideas

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

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

Isabel Healthcare

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

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

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.

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

Your.MD Ltd

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

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

# Appendix B: Glossary

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

| Acronym/Term | Expansion | Comment |
| --- | --- | --- |
| AI | [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) | While the exact definition is highly controversial, in context of this document it refers to a field of computer science working on machine learning and knowledge based technology that allows to *understand* complex (health related) problems and situations at or above human (doctor) level performance and providing corresponding insights (differential diagnosis) or solutions (next step advice, triage). |
| 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. |
| CC | Chief Complaint | See "Presenting Complaint". |
| CONSORT-AI | Consolidated Standards of Reporting Trials | TODO SHUBS |
| 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 |  |
| MTS | Manchester Triage System | A commonly used systems for the initial assessment of patients e.g. in emergency departments. |
| 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) |
| PMCF | Post Market Clinical Follow Up | A requirement by regulators for Software as a medical device. This refers to clinical studies of the product in the real world that serve to show evidence of the claimed benefits of a medical device. |
| PROMs | Patient Reported Outcome Measures | This are outcomes reported by patients (usually through questionnaires) about their quality of life |
| 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 the 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 entry 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://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2030-05-2019%20meeting%202%20minutes.docx?d=w6418637cd3e6475f8a5318789527721b) |
| 20.06.2019 | Meeting #3 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2020-06-2019%20%20meeting%203%20minutes.docx?d=w215896dfe442471cb07f634fbaebe5a6) |
| 11-12.7.2019 | Meeting #4 – London Workshop | * [London Model](https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit) * [GitHub](https://github.com/babylonhealth/itu_who_2019_symptom_assessment_mmv_benchmark) * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-12-2019%20meeting%2014%20minutes.docx?d=w587cebcda07343f0a38184465a1d0f19) |
| 15.8.2019 | Meeting #5 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2015-08-2019%20meeting%205%20minutes.docx?d=wbe4764613f8f46df905a8efc1a6757fa) |
| 23.08.2019 | Meeting #6 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2023-08-2019%20meeting%206%20minutes.docx?d=wf9c51f181a9f49258fd28861494de696) |
| 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://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2003-09-2019%20meeting%207%20minutes.docx?d=w32c9a3a90f0645d9bbe335fe88af79de) |
| 27.09.2019 | Meeting #8 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2027-09-2019%20meeting%208%20minutes.docx?d=w416a633039c545afa0bc485cba1ffabb) |
| 10-11.10.2019 | Meeting #9 – Berlin Workshop | * [Berlin Model](https://docs.google.com/spreadsheets/d/111D40yoJqvvHZEYI8RNSnemGf0abC9hQjQ7crFzNrdk/edit) * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2010-10-2019-11-10-2019%20meeting%209%20minutes.docx?d=w28f04cfffbe047998a45bc006ac1bd15) |
| 17.10.2019 | Meeting #10 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2017-10-2019%20meeting%2010%20minutes.docx?d=w11b253bdb08f4477b39300c857e6ffdb) |
| 20.10.2019 | Meeting #11 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2020-10-2019%20meeting%2011%20minutes.docx?d=w43d3b5f99d224d89811c7b48410c8e52) |
| 25.10.2019 | Meeting #12 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2025-10-2019%20meeting%2012%20minutes.docx?d=wee119da8d7a64fb796b9a4c182a662e8) |
| 30.10.2019 | Meeting #13 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2030-10-2019%20meeting%2013%20minutes.docx?d=w0bfe95524aab406699210287322793a7) |
| 6.12.2019 | Meeting #14 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-12-2019%20meeting%2014%20minutes.docx?d=w587cebcda07343f0a38184465a1d0f19) |
| 6.1.2020 | Meeting #15 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2006-01-2019%20meeting%2015%20minutes.docx?d=wa35cec942d7b4fcda60b0832e3cf9613&Source=https%3A%2F%2Fextranet%2Eitu%2Eint%2Fsites%2Fitu%2Dt%2Ffocusgroups%2Fai4h%2Ftg%2FSitePages%2FTG%2DSymptom%2Easpx) |
| 30.1.2020 | Meeting #16 – Tech Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BBA8BF40A-D2D0-4FE3-8C3F-276DB30C588A%7D&file=FGAI4H%20TG%20Symptom%2030-01-2019%20meeting%2016%20minutes.docx&action=default) |
| 28.03.2020 | Meeting #17 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B24C1610B-A03B-40B8-A47C-9D539E5827CC%7D&file=FGAI4H%20TG%20Symptom%2028-02-2020%20meeting%2017%20minutes.docx&action=default) |
| 12.03.2020 | Meeting #18 – Tech Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BDD2C5742-4422-465B-863C-A737B58BAA6D%7D&file=FGAI4H%20TG%20Symptom%2012-03-2020%20meeting%2018%20minutes.docx&action=default) |
| 13.03.2020 | Meeting #19 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B18E9043E-AB29-4428-BA7D-251F662C06A7%7D&file=FGAI4H%20TG%20Symptom%2013-03-2020%20meeting%2019%20minutes.docx&action=default) |
| 20.03.2020 | Meeting #20 – Tech Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B8236F039-4128-484B-8318-2A1A0CA98D5F%7D&file=FGAI4H%20TG%20Symptom%2020-03-2020%20meeting%2020%20minutes.docx&action=default) |
| 27.03.2020 | Meeting #21 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0877EB0F-D11F-4D5A-95BE-EFD332F5014A%7D&file=FGAI4H%20TG%20Symptom%2027-03-2020%20meeting%2021%20minutes.docx&action=default) |
| 15.04.2020 | Meeting #22 – Tech Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/symptom/FGAI4H%20TG%20Symptom%2015-04-2020%20meeting%2022%20minutes.pdf) |
| 22.04.2020 | Meeting #23 – Tech Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BFEC194B5-F4CA-4AF4-AF1C-F9FAEA48D339%7D&file=FGAI4H%20TG%20Symptom%2022-04-2020%20meeting%2023%20minutes.docx&action=default) |
| 21.04.2020 | Meeting #24 – Clinical Telco | * (no minutes) |
| 24.04.2020 | Meeting #25 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B9072195C-7A17-4DAC-B5CA-ED8768885ECD%7D&file=FGAI4H%20TG%20Symptom%2024-04-2020%20meeting%2025%20minutes.docx&action=default) |
| 29.05.2020 | Meeting #26 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B597F0813-4A2A-479D-AE85-08FEB7304699%7D&file=FGAI4H%20TG%20Symptom%2029-05-2019%20meeting%2026%20minutes.docx&action=default) |
| 11.06.2020 | Meeting #27 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B5FCF70C5-DDEE-473F-9A14-83DCD71E7307%7D&file=FGAI4H%20TG%20Symptom%2011-06-2019%20meeting%2027%20minutes.docx&action=default) |
| 26.06.2020 | Meeting #28 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BFAB2B733-18B9-4EDC-B7A4-9E4B6E305B8D%7D&file=FGAI4H%20TG%20Symptom%2026-06-2019%20meeting%2028%20minutes.docx&action=default) |
| 10.07.2020 | Meeting #29 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BAEF6C2DF-2F39-4D04-911C-C8829AADC88E%7D&file=FGAI4H%20TG%20Symptom%2010-07-2019%20meeting%2029%20minutes.docx&action=default) |
| 07.08.2020 | Meeting #30 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BD8E9FC2C-0C57-42A7-B206-2531808A130A%7D&file=FGAI4H%20TG%20Symptom%2007-08-2019%20meeting%2030%20minutes.docx&action=default) |
| 21.08.2020 | Meeting #31 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7BC5F08531-5CC4-4DEA-BD39-44948F2A27D7%7D&file=FGAI4H%20TG%20Symptom%2021-08-2019%20meeting%2031%20minutes.docx&action=default) |
| 04.09.2020 | Meeting #32 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B81433DBE-97AE-405F-8F8A-84E840E65489%7D&file=FGAI4H%20TG%20Symptom%2004-09-2019%20meeting%2032%20minutes.docx&action=default) |
| 18.09.2020 | Meeting #33 – Telco | * [Minutes](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/_layouts/15/WopiFrame.aspx?sourcedoc=%7B7D2AF17D-BC4B-41DB-809F-E3507CC5B893%7D&file=FGAI4H%20TG%20Symptom%2018-09-2019%20meeting%2033%20minutes.docx&action=default) |

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

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. Summerton N. The medical history as a diagnostic technology. Br J Gen Pract. 2008; 58(549) [↑](#footnote-ref-23)
24. Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10(10):e1001531. [↑](#footnote-ref-24)
25. *it is also important to consider this within the context of people who have congenital sex differentiation disorders, where this notion may not be so clear cut for some users* [↑](#footnote-ref-25)
26. Sendak, M.P., Gao, M., Brajer, N. *et al.* Presenting machine learning model information to clinical end users with model facts labels. *npj Digit. Med.* **3,**41 (2020). https://doi.org/10.1038/s41746-020-0253-3 [↑](#footnote-ref-26)
27. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/M\_Salath%C3%A9\_Presentation.pdf [↑](#footnote-ref-27)