|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-B-021-R1 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | New York City, 15-16 November 2018 | |
| **DOCUMENT** | | | | |
| **Source:** | | Ada Health GmbH | | |
| **Title:** | | Proposal: Standardized benchmarking of diagnostic self-assessment apps | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Henry Hoffmann Ada Health GmbH Germany | | Tel: +49 177 6612889 Email: [henry.hoffmann@ada.com](mailto:henry.hoffmann@ada.com) |
| **Contact:** | | Andreas Kühn Ada Health GmbH Germany | | Tel: +49 30 60031987 Email: [andreas.kuehn@ada.com](mailto:andreas.kuehn@ada.com) |

|  |  |
| --- | --- |
| **Abstract:** | This document proposes an AI for Health Focus Group project for the implementation of standardized benchmarking for diagnostic self-assessment applications according to the FG-AI4H-A-102 call on “use cases, benchmarking, and data”. |

Overview

This document proposes the Focus Group on AI for Health to set-up a project for standardized benchmarking of diagnostic self-assessment applications (DSAAs). This class of apps - sometimes referred to as “symptom checkers” - allows app users to describe their health issues in order to receive a pre-diagnosis and advice on reasonable next steps. The project would include the evaluation of different approaches to collect labelled datasets and the definition of adequate measures for assessing the performance and the implementation of the benchmarking infrastructure. This includes the definition of API interfaces, protocols and data formats for testing the systems, as well as all processes needed to eliminate the risk of both AI providers getting access to the undisclosed test data and anyone getting access to the AI providers intellectual property and technology.

(Note that a more detailed discussion on the DSAAs use case can be found in [FGAI4H-A-020](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-A-020.docx).)

Impact

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 This shortage is driven by several factors including growing population, increasing life expectancy and higher health demands. This could lead to several billion people with limited or no access to proper health care, less and less doctor time for individual patients and long patient journeys to a correct diagnosis and proper treatment.2

At the same time, patient-centered DSAAs have become widely available. However, the lack of consistent standardization makes it difficult for organizations like the WHO, governments, and other key players to adopt self-assessment-apps as part of their solutions for addressing global health challenges.

The implementation of a standardized benchmarking for DSAAs by WHO/ITU’s AI for Health Focus Group could therefore be an important step towards closing this gap. Paving the way for the safe and transparent application of AI technology would help to improve access to healthcare for more than a billion people. It would enable earlier diagnosis of conditions, more efficient care-navigation through the health system and ultimately better health as it is currently pursued by WHO’s sustainable development goal number 3 (SDG 3).

Current Standardization and Benchmarking

To-date, there is no advanced standardization of symptom checkers available. Recent standardization approaches are often nationally driven (e.g. by Federal Drug Administration, USA or National Health System, UK). A project outline in the *British Medical Journal Open Journal* addresses this topic for medical applications in general. Further, in 2015 a study published in the BMJ systematically compared 23 symptom checkers via 45 standardized patient vignettes which were created manually.3 Very recently, a publication in *The Lancet* outlined the necessity for a standardized test framework for symptom checkers, while referring to the study in BMJ of 2015.4 To our knowledge, besides these limited works, a general maintained test-set of sufficient quality is not available for neither symptom checkers or diagnostic decision support systems in general.

Previous Work / State-of-the-Art

The ‘state-of-the-art’ of DSAAs is primarily using AI approaches based on statistics (e.g. Bayesian methods) and explicit knowledge. They prompt users for presenting complaints, then collect further symptoms -usually by engaging in chat-like dialog- to finally present the most likely underlying diseases, a general triage classification and relevant next steps. The applications often consider epidemiologic data, risk factors, geolocation and in parts, laboratory results, genetic information or additional AI components e.g. for image classification. Examples for current globally or regionally available symptom checkers include, in alphabetical order and without providing a complete list: Ada, Babylon, Buoy, Infermedica (Symptomate), K Health, Sense.Ly, WebMD, Your.MD.

Data Availability

The main reason explaining why there is no standardized benchmarking for DSAAs and diagnostic decisions support systems is the lack of datasets of necessary quality.

While many clinics and health systems have large repositories with case data, they often contain only semi-structured free text, findings are not mapped to ontologies and the conditions are potentially biased for billing reasons. In many contexts privacy and ethical aspects also play an important role. While the datasets collected by symptom checker providers themselves are often biased by their strategy to collect symptoms from users, they usually have high quality data suitable to test their own technology.

Another source of test data are case vignettes that are produced for scientific publications on the quality of symptom checkers, such as the aforementioned BMJ study2. While they are usually limited to a few dozen cases, the quality of their cases descriptions and condition labels are quite high.

Taking these points into consideration, we propose three different approaches to collect data for benchmarking:

1. **Test cases provided by symptom checker providers**  
   There are currently not many high-quality sources for labelled test cases available. Therefore, the providers of symptom checkers could contribute a subset of their internal test data to the undisclosed data set for benchmarking. To limit the bias of the overall benchmarking result and guarantee objectivity, the contribution of each provider should be limited to a maximum of 10% and be further reduced as the number of cases and contributors increases. The test set as a whole is undisclosed and providers get no access to the data provided by other contributors.
2. **Manually curated test cases**   
   Manually curated case vignettes, such as those used for the BMJ publication, could be created with the help of the scientific / medical community. It would potentially lead to a high quality, mostly unbiased set of case data. However, the costs of creating these cases could be high. It might be possible to get funding to create such datasets on a regular basis.
3. **Other digitally available case sources**  
   There are platforms in development that collect case vignettes to foster a professional exchange of patient cases (e.g. Human Diagnosis Project). The cases of these providers have high quality and could potentially be used for benchmarking if such platforms join the Focus Group.

To initiate the benchmarking process, it could be necessary to combine approaches.

Benchmarking

To benchmark DSAAs independently from their technology, providers need to provide a dedicated API endpoint. The definition of the structure and the protocol of the API should be defined by the Focus Group as part of this project. This includes agreeing on the ontologies for the description of all inputs such as factors (e.g. age, sex, smoker, pregnancy etc.), findings (e.g. fever, nausea, dizziness) and their attributes (e.g. intensity, direction, duration) and all the outputs such as conditions (e.g. myasthenia gravis, malaria, tonsillitis) and triage levels (e.g. self-care, primary care same day, emergency).

The actual benchmarking is performed by sending all test cases to the API endpoint and recording the results of each case. Based on the individual case results, the overall benchmarking metrics are calculated. The measure definitions are crucial to standardization because they define what a "good" DSAA is. It will likely be necessary to define a whole panel of measures for benchmarking different aspects of DSAAs and their applications in different scenarios. The definition of relevant measures in cooperation with AI providers and stakeholders that want to apply such technologies will be an important part of this project. The measures might include, for instance, the probability to see the correct diagnosis among the top 1, 3, or 10 results (adjusted to possible biases in the test set) which is used in some publications on DSAA performance. The measures should also be differentiated by dimensions such as condition (group), incidence/prevalence classes (common vs. rare), speciality, if the providers claims to support the disease and by the source of the test data.

For the successful implementation of a first benchmarking for DSAAs by the Focus Group, the project should be restricted to the common inputs and outputs of most DSAA systems. While the details need to be defined and discussed by the Focus Group, a possible viable starting point could be the benchmarking of the DSAAs pre-diagnoses and triage suggestions as output for given user cases containing patient reportable risk factors, the presenting complaints, as well as all other known symptoms. For simplicity, the entire case evidence should be given to a DSAA at once. Simulating the dialog between users and the app, benchmarking the explainability, including next step advice or extending the inputs to the realm of clinical decision support including professional findings e.g. from physical examination, lab and imaging could be implemented in later phases.

The project should also define how the results of an official benchmarking are presented and published so that they are available for decision makers. This might also include the creation of official quality-label-brands.

An important point is that for companies developing DSAA technology, protecting their IP is as crucial as the protection of their test cases is for data providers. Setting up the benchmarking for DSAAs should therefore include the definition of protocols and procedures that ensure the technology can be tested without any risk for the AI and data providers.

**Organizer Details**

Ada Health GmbH is a digital health company based in Berlin, Germany, developing diagnostic decision support systems since 2011. In 2016 Ada launched the Ada-App, a DSAA for smartphone users, that since then has been used by more than 5 million users for over 8 million health assessments. The app is currently available in 5 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 low and middle-income countries where it partners with governments and NGOs to improve people's health.

By working in different countries it became clear that the lack of a unified benchmarking made it challenging to bring AI-based diagnostic decision support technology into practical use since decisions makers could only rely on our retrospective studies rather than independent benchmarking results. In helping to close this gap by creating an independent, transparent official benchmarking system, Ada sees an important step to bring AI-based diagnostic decision support systems in more wide spread use, which will help to address the increasing shortage of access to healthcare in the world.

For building and validating its diagnostic decision support technology, Ada develops a range of internal benchmarking and measurement tools. This includes the collection and modelling of labelled medical cases, systems for automated benchmarking of multiple competing AI engine technologies and web based tools for visualizing and analysing the benchmarking results. We also have experience in designing and conducting retrospective research studies in clinical context. The knowledge and experience in this field will contribute to implement a successful benchmarking framework for DSAAs.

**References**

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; WHO, The World Bank <http://pubdocs.worldbank.org/en/193371513169798347/2017-global-monitoring-report.pdf>
3. Semigran et al; "Evaluation of symptom checkers for self diagnosis and triage:  
   audit study"; BMJ 2015; 351:h3480
4. Fraser H, Coiera E, Wong D; “Safety of patient-facing digital symptom checkers”; The Lancet 2018; 10.1016/S0140-6736(18)32819-8

Annex: Proposal submission questionnaire

1. **Relevance** - How relevant is the health problem to be addressed?

* more than half of the world population can not obtain essential health services (according to WHO and World Bank in 2017)
* 7.2 million missing healthcare professionals in 2013 - raising to 12.9 million by 2035
* symptom checkers can help users with health problems to get assessment reports with: a general triage, a pre-diagnosis, advice on next steps
* sharing reports with doctors safes time, reduces risk of misdiagnosis and improves outcomes
* health system efficiency can be improved if combined with diagnostic self assessments

1. **Impact** - What level of impact will a benchmark in the context of the proposed project have?

* with official benchmarking reports available, policy makers can decide to apply AI in parts of their health systems
* it can guarantee a certain minimum level of quality and avoid potential harm to users
* standardization will objectivize the reliability of the apps
* it can also help app providers to improve their technology by identifying potential for improvement
* BMJ and Lancet just recently (2018) pointed out the importance of a standardization benchmark to test these apps with the goal of providing a high quality and to avoid harm done by potential insufficient applications

1. **Existing work** - Does the project start from scratch, or are there preliminary experiences?

* there is a wide range of symptom checkers available and their coverage is increasing constantly as there are over 4 billion people online to date
* existing studies on the accuracy of symptom checkers journals (e.g. BMJ 2015) designed viable benchmarking methodologies
* most providers of AI systems in the field have designed benchmarking frameworks for internal use

1. **Feasibility** - Is the project feasible, based on the current state of the art?

* technically it is feasible since every provider has created the necessary technology and datasets already for internal use

1. **Data Availability** - Is there sufficient data available? How much of it can be openly available? How much of it as part of the non-disclosed data set?

* there is a lot of electronic health record (EHR) data available in hospitals and health systems, even though there are legal constraints and quality issues that require careful data set review and selection
* AI providers have high quality datasets for internal testing that could be partially shared as long as every provider contributes only a small fraction of the undisclosed test set and if the reports also shows results by source
* there are scientific platforms collecting cases that could potentially be used for the undisclosed data sets
* following the methodology used by scientific publications on symptom checker performance, new cases for benchmarking could be created (requires funding)
* identification of potential data providers and discussion under which circumstances the data can be used is needed

1. **Data Quality** - Is the available data of high quality?

* the providers internal testing data have high quality but may be specific to the respective application
* the quality of cases used for journal publications on symptom checkers is sufficiently high
* EHR data is often biased and not always optimally structured for benchmarking
* assessing the quality of data sets will therefore be a key part of the project

1. **Annotation / Label Quality** - Are the annotations / labels of the data of high quality?

* the above statements hold also true for the quality of the labels / annotations
* for EHR data the labels most likely require peer review and additional annotation work
* the labels for manually curated cases for journal publication and internal use by the AI providers have in general high quality

1. **Data Provenance** - Has the data been obtained in a professional and ethically correct way?

* usually EHR are highly regulated in respect of data security and professionality
* the internal testing data is not necessarily based on real case data but often created by doctors and therefore ethically unproblematic
* external platforms collecting cases have different policies that require review

1. **Benchmarking** - Do the applicants have a clear proposal about what exactly should be evaluated / measured?

* the general benchmarking process follows the scheme suggested by EPFL
  + individual training/creation of the AI with internal (and public) data
  + submission of the algorithms as e.g. docker container exposing an API endpoint
  + automated collection of results by a benchmarking system
  + computation of metrics/measures based on the collected data
  + detailed report compilation and publication
* the specification of the API as well as the ontologies and selection for **input** and **output** variables are to be defined as part of the benchmarking project
* the **metrics** need to be agreed on to meet providers and AI user needs
* for simplicity and to foster the standardization process, in the first phase of testing following aspects should be excluded but can be considered in the ongoing improvement of the benchmark process:
  + automated AI dialogs, NLP etc.
  + explainability
  + next step advices
  + support for professional findings labs, sensors etc.

1. **Organizers** - Can the Focus Group work with the applicants, and do they have the time / resources to work with the Focus Group on the problem?

* Ada Health has a strong interest on an objective, standardized, transparent benchmarking of diagnostic self-assessment applications (“symptom checkers”) to increase the impact by better integration into health systems
* as a provider of diagnostic decision support technology we have substantial experience in implementing benchmarking frameworks
* Ada Health is committed to drive this topic which is underlined by our participation since the very first ITU meeting in Ljubljana

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