|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-C-019 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | N/A | Lausanne, 23-25 January 2019 |
| **DOCUMENT** |
| **Source:** | Ada Health GmbH |
| **Title:** | Status report on the “Evaluating the accuracy of ‘symptom checker’ applications” use case |
| **Purpose:** | Discussion |
| **Contact:** | Henry HoffmannAda Health GmbHGermany | Tel: +49 177 6612889Email: henry.hoffmann@ada.com |
| **Contact:** | Andreas KühnAda Health GmbHGermany | Tel: +49 30 60031987Email: andreas.kuehn@ada.com |
| **Contact:** | Johannes SchröderAda Health GmbHGermany | Tel: +49 30 60031987Email: johannes.schroeder@ada.com |

# Abstract

This is the first progress report of the feasibility study for the standardized benchmarking of diagnostic self assessment applications as proposed in [B-021](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-B-021.docx) and admitted by the focus group in [B‑101](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-B-101.docx). The report focuses on the points requested by the WG-O chair in the mail addressed on 11 January 2019.

# 1. Overview

In document [B-021](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-B-021.docx), we proposed the AI4H use case to benchmark diagnostic self assessment applications (DSAAs, short “symptom-checker”). During focus group meeting B in New York in November 2018, the use case was accepted and Ada was named as the corresponding topic driver. In the time since this meeting, we reached out to several field experts to refine the approach for the proposed benchmarking. This document summarizes our insights from this period by answering the questions issued by the WG-O chair Mr. Markus Wenzel in the email “Report of work done between meetings" circulated on the 11th of January, 2019.

## General Expectations for the First Benchmarking

Our exchanges with experts and other focus group members demonstrated that DSAA benchmarking approaches require the consideration of many dimensions. Based on these discussions we generated the following hypotheses and assumptions:

* We must start with a **minimal viable benchmarking procedure** using the most common inputs and outputs shared by most DSAA providers.
* We have to focus benchmarking on the use cases that we can get data for.
* We must establish metrics reflecting the needs of possible AI users (WHO, NGO’s, health systems, governments) and the use cases that the solutions have been designed for.
* We have to design an architecture that protects benchmarking data, provider AI, the integrity of the benchmarking process, and benchmarking results.
* We have to define the benchmarking framework and all the details of inputs and outputs cooperatively with interested stakeholders to have a viable working hypothesis before reaching out to all other possible DSAA providers.

## Overall Architecture

Fig. 1 shows the general generic benchmarking architecture discussed by the focus group during the previous meetings. Although the architecture can be applied to the DSAA use case, we see the need for some specific adjustments that we will discuss in the following paragraphs.



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

**(2)** Every provider creates its own DSAA 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 use cases such as image recognition, 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 for DSAA benchmarking.

As part of the benchmarking process, the DSAA providers have to provide an API endpoint that accepts test cases and computes the corresponding output **(3)**. Agreeing on the formats for input and output will be part of the focus group’s work. The ontology/terminologies used to refer to factors, symptoms or conditions will play a key role in defining the formats.

In contrast to some use cases of the focus group where sharing might be possible (e.g. a trained neural network in some standard format), DSAA systems are usually too complex. Sharing key parts of the companies’ IP may expose them to the risk of being leaked and reverse-engineered. At the same time, the solution must ensure that no DSAA provider can acquire a copy of the undisclosed test data that would allow them to optimize system performance.

The actual benchmarking will then be performed by the ITU benchmarking infrastructure by sending test cases **(4)** without the labels to the DSAA and recording the corresponding AI results. To generate a report for each DSAA **(5)** the benchmarking system will then compute the previously agreed-upon metrics and scores based on the output datasets. Defining the metrics will be a key part of the work in the focus group. The results of the different DSAAs can finally be presented in a leaderboard **(6)**.

Besides the official benchmarking on undisclosed datasets and submission of AI for testing, we will also consider a continuous benchmarking process **(7)** that uses an open test dataset and API endpoints hosted by the DSAA providers on their systems. This would facilitate the testing of API endpoints and required data format transformations, while also providing a rough estimate of performance before the official benchmarking. It is noteworthy that for most providers, an automatic mapping of cases to their internal symptom ontology is not trivial and requires careful testing before reliable results can be obtained.

# 2. Addressing the Questions

This section addresses the questions directed to all topic drivers in the email “Report of work done between meetings” addressed on 11 January 2019:

a) In which specific form will you provide the data (which file format, how is the database structured)?

The input data for DSAA benchmarking consists of user-reported cases of a specific health issue that users seek advice for. To prepare a minimal viable benchmarking, we expect the cases to contain the following details:

* **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.
* **Additional symptoms** - Additional symptoms and factors that are collected by the DSAA usually by engaging with the user in some kind of dialogue. This also includes absent symptoms and additional symptom details like “side: left” or “intensity: moderate”.
* **Additional metadata** - To document the test cases and to implement and test the DSAA-API, the test data will also contain additional metadata. This will include a case identifier, a description, as well as some versioning details to track changes in the test sets. For the official benchmarking this information will be stripped to avoid providing indirect clues on the correct result.

Additionally, the inputs need to be expressed in a structured and semantically unambiguous way using some **ontology/terminology**.

To keep things simple for a first minimal viable benchmarking, we explicitly suggest to **exclude** from the first benchmarking version:

* **Free text input** - Any unstructured free text e.g. searching the presenting complaint or the conversation between the DSAA and the user.
* **Professional findings** - Symptoms and findings that only a doctors could answer.
* **Additional user data** - Further data e.g. from lab results, sensor and tracking devices or genetics profiles.

While for the first minimal viable benchmarking some batch-processing would be feasible, it is already clear that in later stages more complex and interactive benchmarking systems will have to be implemented e.g. simulating the dialogue between the user and the DSAA. One widely used standard for communication between systems are REST API microservices using JSON to encode the data to be exchanged. Assuming that the DSAA benchmarking system will use this approach, an example case as it would be sent to the DSAA-API endpoint from the ITU could take the following form:

|  |
| --- |
| { "key": "rc31415926", "metaData": { "description": "A basic acute appendicitis test case", "caseSource": "Ada Health", "labelingSource": "Dr. John Smith", "labelingDate": "Fri, 06 Jan 2019 08:04:05 GMT", "creationDate": "Fri, 06 Jan 2019 07:34:05 GMT", "creationUser": "Dr. John Smith", "modificationDate": "Fri, 06 Jan 2019 11:64:55 GMT", "modificationUser": "Dr. Lisa Doe", "version": "3" }, "profileInformation": { "sex": "male", "age": "25 years" }, "presentingComplaints": [ { "key": "HP:0002027", "name": "Abdominal pain", "present": true, "attributes": [ { "key": "timeSinceOnset", "value": "lessThanOneDay" }, { "key": "position", "value": "lowerRightQuadrant" } ] }, { "key": "HP:0004396", "name": "Poor appetite", "present": true, "attributes": [ { "key": "timeSinceOnset", "value": "lessThanOneDay" } ] }, { "key": "HP:0002013", "name": "Vomiting", "present": true, "attributes": [ { "key": "forceFull", "value": "absent" }, { "key": "unableToTolerateFluids", "value": "present" } ] }, { "key": "HP:0001945", "name": "Fever", "present": true, "attributes": [ { "key": "timeSinceOnset", "value": "lessThanOneDay" } ] }  ], "additionalSymptoms": [ { "key": "HP:0002018", "name": "Nausea", "present": true, "attributes": [ { "key": "intensity", "value": "severe" } ] }, { "key": "HP:0001649", "name": "Rapid pulse", "present": true }, { "key": "SCTID: 43478001", "name": "Abdominal Tenderness", "present": true, "attributes": [ { "key": "position", "value": "lowerRightQuadrant" } ] }, { "key": "HP:0012378", "name": "Malaise", "present": true }, { "key": "HP:0002014", "name": "Diarrhea", "present": false }, { "key": "AI4H:factor:00141", "name": "Appendectomy", "present": false } ]} |

As mentioned, the metadata will be stripped for the actual benchmarking from the undisclosed dataset. As for other details, we expect to refine the format together with other DSAA providers as part of the work of the focus group.

b) How will you provide the labels/annotations of the single samples in your data set? What output variables are possible?

The most important decision to initiate DSAA benchmarking is to decide which outputs should be benchmarked. With the focus on minimal viable benchmarking and to avoid overcomplicating the acquisition of an undisclosed test dataset, we will initially focus on the most appropriate explanations of the presenting complaints. It is important to emphasize that, based on user reportable symptoms, a final correct diagnosis is often neither possible nor appropriate. Even if at a later point in time a case turned out to be caused by a certain disease it is usually not correct to rank it at top position when the first symptoms occur and all differential diseases have not yet been ruled out by a doctor. More appropriate than labeling the case with a confirmed diagnosis is a list of acceptable differentials considering only the evidence at this point in time. In addition to these acceptable conditions, in some cases other conditions can already be excluded based on the available evidence. Basic outputs for the first benchmarking iteration can focus on the following:

* **Acceptable Top Conditions** - The conditions that are acceptable as top matching conditions.
* **Ruled Out Conditions** - The conditions that can by ruled out by the given evidence and should not appear in the result set (e.g. female-only diseases in males).

For the first iteration of the DSAA benchmarking we suggest to not include:

* **Pre-clinical triage** - A classification if the case is e.g. an emergency, requires a doctor visit or if self-care is acceptable. For this more work needs to be done to align on a common pre-clinical triage classification as classifications besides clear “emergency” situations are often more a matter of opinion and vary widely.
* **Next Steps and treatment suggestions** - Even if treatment suggestions could help (especially in contexts where doctors are not available), it is currently not clear in which format and to which extent this could be tested across different providers.
* **AI explanations** - While the explainability of AI will play a role in future AI benchmarking, currently most providers do not support this feature. There is also still some research necessary to score the quality of an explanation across different providers.

As for the input data, the DSAA providers must agree on an **ontology/terminology** to encode the acceptable output data (the conditions). An example using ICD-10 to encode the acceptable considered and ruled-out conditions can be seen in the figure below:

|  |
| --- |
|  "acceptableTopConditionsKeys": [ { "key": "ICD10CM:K35", "name": "Acute appendicitis" } ], "ruledOutConditions": [ { "key": "ICD10CM:N70-N77", "name": "Inflammatory diseases of female pelvic organs" } ] |

The labels will be embedded directly in the test datasets and then omitted when sent to the DSAA-API endpoint for the official benchmarking.

The response generated by a DSAA will contain a similar list of conditions as the test cases. In addition to the key and an optional name it contains the the probability that the condition caused the presenting complaints of the given case. An example output could take the following form:

|  |
| --- |
|  "matchingConditions": [ { "key": "ICD10CM:A08", "name": "Viral and other specified intestinal infections", "probability": 0.5527 }, { "key": "ICD10CM:A04.5", "name": "Campylobacter enteritis", "probability": 0.3527 }, { "key": "ICD10CM:N10", "name": "Acute tubulo-interstitial nephritis", "probability": 0.0256 }, { "key": "ICD10CM:A02.0", "name": "Salmonella enteritis", "probability": 0.0099 } ] |

c) Are you ready to show a few labelled samples, as actual files on your computer, in Lausanne?

While the above examples have been written by hand, we will work on automatically transforming some of the Ada test cases into the above JSON format before the focus group meeting C, to show more explicit real-world examples.

d) How many labelled samples can you actually provide? (This is something the focus group needs to know.)

Ada Health has a variety of test cases for all its conditions. All theses cases use the Ada symptom and condition ontologies so that they can only be used after careful mapping to the ontologies that the focus group has agreed on. After agreeing on the ontologies, Ada currently plans to map parts of its reference case database and provide it as part of the undisclosed dataset if this does not affect our rights to use it for our internal purposes.

# 3. Communication & Outreach

Since the last focus group meeting B in New York in November 2018, we reached out to several experts in benchmarking “symptom checkers”, clinical decision support systems and conducting clinical studies. While there was no resonance before Christmas and in the first weeks of the year, in the week before meeting C we had several informal teleconferences and email exchanges.

Theses meetings have provided valuable insights that will allow the refinement of the outlined approach. This also includes contributions to the lists of further experts, DSAA providers and test data providers that may be contacted in the upcoming weeks. We also expect some of them to participate in one of the next focus group meetings.

One of the most noteworthy insights from experts with medical background is that they see the dialog between DSAAs and users as a key part that should not be omitted, even for simplicity in a first minimal viable benchmarking. This new insight should be discussed at meeting C.

# 4. Next Steps / Roadmap

The general approach of the focus group has been outlined as follows:

1. *Invite and select proposals for use cases and data.*
2. *Collect undisclosed test data.*
3. *Publish description of test data and benchmarking metrics.*
4. *Invite submissions of AI technology.*
5. *Benchmark AI technology with undisclosed test data*

With the admission during focus group meeting B, the first step has already been completed for the DSAA use case.

As outlined in section 3, an important part of the work of this focus group is agreeing with DSAA providers on a **joint ontology** to formalize input symptoms and output conditions. Therefore, for this use case, we must consider the additional points:

* 1a) Collect DSAA provider candidates and reach out to them based on to-be-defined criteria
* 1b) Agreeing with the participating DSAA providers on the input and output ontologies.

We also consider issuing “calls for proposals” to collect descriptions of the structure of the benchmarking datasets and general characteristics of DSAA systems to be tested. If the focus group agrees on this, such calls could already be issued as output documents of meeting C.

Based on the lists of possible DSAA dataset and AI providers, we will then reach out to the candidates and send them these calls. The first input documents specific to the DSAA use case can then be discussed during meeting D in Shanghai.

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

1. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/M\_Salath%C3%A9\_Presentation.pdf [↑](#footnote-ref-1)