|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-I-028 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 7-8 May 2020 |
| **DOCUMENT** |
| **Source:** | Editors |
| **Title:** | Updated DEL07: AI for Health ~~Specification~~ Evaluation Considerations |
| **Purpose:** | Discussion |
| **Contact:** | Markus WenzelFraunhofer HHI, Germany | Email: markus.wenzel@hhi.fraunhofer.de  |
| **Contact:** | David Neumann Fraunhofer HHI, Germany | Email: david.neumann@hhi.fraunhofer.de  |
| **Contact:** | Sandeep Reddy Deakin University, Australia | Email: sandeep.reddy@deakin.edu.au |
| **Contact:** | Annika Reinke DKFZ, Germany | Email: a.reinke@dkfz-heidelberg.de |
| **Contact:** | Manuel Wiesenfarth DKFZ, Germany | Email: m.wiesenfarth@dkfz-heidelberg.de  |
| **Contact:** | Alberto MerolaAICURA medical GmbH, Germany | Email: alberto.merola@aicura-medical.com  |
| **Contact:** | Steffen VoglerBayer AG, Germany | Email: steffen.vogler@bayer.com  |

|  |  |
| --- | --- |
| **Abstract:** | In this introductory document, characteristics of health AI evaluation that are novel or otherwise essential are identified, best practices for the health AI model assessment are collected from selected sources, and requirements for a benchmarking platform are considered. This sets the scene for the four related documents DEL07.1-4 that dive into the details of health AI evaluation including a process description, technical tests, metrics and clinical validation. |

**Call for participation**

If you would like to contribute to the *AI for Health Evaluation Considerations*, please contact Markus Wenzel (markus.wenzel@hhi.fraunhofer.de) and the Focus Group Secretariat (tsbfgai4h@itu.int) with a brief introduction of yourself and your relevant expertise.

Keywords

AI/ML model evaluation, software testing, quality assessment, statistics, evidence-based medicine, clinical trials, medical informatics, fair/accountable/transparent machine learning

Change Log

* This deliverable draft “*AI for Health Evaluation Considerations*” will be submitted on 30 April 2020 to the Secretariat of the ITU/WHO Focus Group on “AI for Health” (tsbfgai4h@itu.int), uploaded by the Secretariat as document FG-AI4H-I-207 to <http://itu.int/go/fgai4h/collab>, and presented by the authors in the virtual meeting, 7-8 May 2020.
* The document is based on two earlier versions:
	+ Working draft sent as update on 24 March 2020 to tsbfgai4h@itu.int.
	+ FG-AI4H-G-207, titled “DEL07: AI4H evaluation specification (Outline)”.

**CONTENTS**

|  Page |
| --- |
| [1 Scope 5](#_Toc39825192)[2 References 5](#_Toc39825193)[3 Terms and definitions 6](#_Toc39825194)[3.1 Terms defined elsewhere 6](#_Toc39825195)[3.2 Terms defined here 6](#_Toc39825196)[4 Abbreviations 6](#_Toc39825197)[5 Introduction 6](#_Toc39825198)[6 Background 6](#_Toc39825199)[7 Novelty 7](#_Toc39825200)[8 Independent standardized model benchmarking 9](#_Toc39825201)[8.1 Benchmarking platform with evaluation in a closed environment 10](#_Toc39825202)[8.2 Benchmarking platform with evaluation via interface 10](#_Toc39825203)[8.3 Federated benchmarking platform 11](#_Toc39825204)[9 Preliminary considerations on the evaluation process 12](#_Toc39825205)[10 Overview of the evaluation deliverables 13](#_Toc39825206)[10.1 DEL07.1: AI for health evaluation process description 13](#_Toc39825207)[10.2 DEL07.2 AI technical test specification 13](#_Toc39825208)[10.3 DEL07.3 AI technical test metric specification 14](#_Toc39825209)[10.4 DEL07.4 Clinical Validation of AI for Health 14](#_Toc39825210)[11 Test data, ethical and regulatory aspects 14](#_Toc39825211)[Appendix I Requirements of a benchmarking platform with evaluation in a closed environment 15](#_Toc39825212)[I.1 System overview 15](#_Toc39825213)[I.1.1 Administrative backend 15](#_Toc39825214)[I.1.2 Public frontend 17](#_Toc39825215)[I.1.3 Execution environment 18](#_Toc39825216)[I.2 General considerations 19](#_Toc39825217)[I.2.1 Security 19](#_Toc39825218)[I.2.2 Hosting 19](#_Toc39825219)[I.2.3 Computing resources 20](#_Toc39825220)[I.2.4 Availability 20](#_Toc39825221)[Appendix II Best practices from the scientific literature and other documents 21](#_Toc39825222) |
|  |
| **List of Tables**

|  Page |
| --- |
| [Table 1: Aspects of health AI model evaluation that are novel, unique or otherwise essential 7](#_Toc39169802)[Table 2: Overview of the evaluation deliverables. 13](#_Toc39169803) |

**List of Figures**

|  Page |
| --- |
| [Figure 1: Exemplary data sources and tasks/use cases for health AI/ML models 7](#_Toc39169804)[Figure 2: Independent model evaluation with standard procedures on separate test data 9](#_Toc39169805)[Figure 3: Cyclical evaluation process 12](#_Toc39169806)[Figure I.1: Overview of the architecture of the system. 16](#_Toc39169807) |

 |

Artificial Intelligence for Health
Evaluation Considerations

Summary

In this document, we aim to identify characteristics of health AI assessment that are novel (have not yet been dealt with before when assessing other digital health technologies), unique or otherwise essential. Furthermore, we collect best practices for the AI model evaluation/testing from selected sources and consider requirements for a benchmarking platform. In this way, we set the scene for four related documents that dive into the details of health AI evaluation: “AI for Health Evaluation Process Description” (DEL07.1), “AI Technical Test Specification” (DEL07.2), “AI Technical Test Metric Specification” (DEL07.3) and “Clinical Validation of AI for Health” (DEL07.4).

# Scope

Introduction with considerations on novel aspects of/best practices for health AI evaluation. It is a living document, iteratively refined on each focus group meeting. This document belongs to a series of deliverables listed in FGAI4H-H-200:

<https://itu.int/en/ITU-T/focusgroups/ai4h/Documents/listdeliverables.pdf>

Therefore, we do not address e.g. data, ethics, regulation and other topics addressed in separate deliverables.

# References

[Nilsson 1998] Nilsson, N. J., & Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. Morgan Kaufmann.

[WHO Constitution] Constitution of the World Health Organization (1946, 2006). Basic Documents, Forty-fifth edition. <http://www.who.int/governance/eb/who_constitution_en.pdf>

[DEL01] FG-AI4H Deliverable DEL01, *AI4H ethics considerations*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-201.docx>

[DEL02] FG-AI4H Deliverable DEL02, *AI4H regulatory considerations*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-202.docx>

[DEL05] FG-AI4H Deliverable DEL05, *Data specification*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-205.docx>

[DEL07.1] FG-AI4H Deliverable DEL07.1, *AI for health evaluation process description*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-207-A01.docx>

[DEL07.2] FG-AI4H Deliverable DEL07.2, *AI technical test specification*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/Deliverables/DEL07_2.docx>

[DEL07.3] FG-AI4H Deliverable DEL07.3, *AI technical test metric specification*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-207-A03.docx>

[DEL07.4] FG-AI4H Deliverable DEL07.4, *Clinical validation of AI for health*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/Deliverables/DEL07_4.docx>

[DEL10] FG-AI4H Deliverable DEL10, *AI4H use cases: topic description documents.*. Last found: <https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/Deliverables/DEL10_0.docx>

NOTE: Additional literature references are listed in the bibliography.

# Terms and definitions

## Terms defined elsewhere

This document uses the following terms defined elsewhere:

**3.1.1 Artificial Intelligence [Nilsson 1998]:** “Artificial intelligence, broadly (and somewhat circularly) defined, is concerned with intelligent behavior in artefacts. Intelligent behavior, in turn, involves perception, reasoning, learning, communicating, and acting in complex environments.”

**3.1.2 Health [WHO Constitution]:** “Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.”

## Terms defined here

This document does not define any new terms.

# Abbreviations

|  |  |
| --- | --- |
| AI4H | Artificial Intelligence for Health |
| FG-AI4H | ITU/WHO Focus Group on Artificial Intelligence for Health |
| ITU | International Telecommunication Union |
| ML | Machine learning |
| WHO | World Health Organization |

# Introduction

*Evidence-based trust* is essential for any technology to be adopted by the healthcare community including solutions based on artificial intelligence (AI) and machine learning (ML). Are the AI/ML models effective, safe, accurate, robust, transparent, usable, fair, free from bias, plausible, and therefore trustworthy?

Best practices and recommendations for creating this evidence are collected and discussed in the present introductory document (DEL07) and in the associated deliverables (DEL07.1-4). The considerations will cover the entire evaluation process (DEL07.1) starting with technical tests (DEL07.2) using meaningful test metrics (DEL07.3). Technical tests are an important step in the evaluation procedure, but need to be informed by clinical/public health expertise, and should be followed up by clinical trials or field/scientific tests and post market surveillance procedures to evaluate the efficacy, safety, usability, and the outcome of the technology in practice (DEL07.4).

# Background

Health AI/ML models can analyse various modalities of data. These input data can be microscope or radiology images, electronic health records, laboratory test results, epidemiological maps, or input to mobile phone apps, to list a few examples (cf. Figure 1). The models — whether they are data-driven ML models (e.g. artificial neural networks) or knowledge-based AI models (e.g. expert systems) — can be applied in a range of use cases such as diagnostics, forecasting, triage, image segmentation, and many others. Typically, every model is trained to perform *one* (or few) specific task(s), such as classification or regression of specific input data that are mapped to associated output labels. The associated output labels are task specific and can be for example pixel labels in a segmentation solution, International Classification of Diseases (ICD) codes in the case of a diagnostics application, or a risk score in a triage or forecasting task.



Figure 1: Exemplary data sources and tasks/use cases for health AI/ML models

AI/ML models and the circumstances where they are applied can be highly complex, and customised to the context, so it is non-trivial to assess their merits and limitations. Fortunately, we can build on a large corpus of previous work driven by the objective to create evidence about the quality of health technologies and interventions, and of ML models. We can also profit from experience related to other digital health technologies such as Computer-Aided Detection (CAD) in radiology and Software as a Medical Device (SaMD). A non-comprehensive collection of best practices, recommendations, insights and perspectives from the scientific literature and other documents concerned with the evaluation and testing of health technology and of AI/ML models is presented and discussed in Appendix II.

# Novelty

Considering this prior knowledge, we should start with carefully thinking about characteristics of health AI model evaluation that are novel and unique. Which characteristics have not been dealt with before, when assessing other digital health technologies? What cannot be captured by established standard evaluation and assessment methods, such as in-house software verification and validation? What cannot be captured by randomized controlled clinical trials? How can these gaps be addressed best? Table 1 contains an open, non-comprehensive list with some major points.

Table 1: Aspects of health AI model evaluation that are novel, unique or otherwise essential

| N° | Aspect |
| --- | --- |
| 1 | The significance of technical test results depends on appropriate test data sets. However, separate high-quality standard test data sets from different sources (geographically, measurement devices, patient cohorts of different ages or with comorbidities etc.) are scarce. Usually, only a very small subset of all conceivable test cases can be covered. It is known that algorithms do not generalize well across centers, presumably due the domain gap between medical centers and devices. Hence, we need more data sets with data from different locations. Yet, more data sets do not always help. Careful attention must be paid to define a population of interest and systematically collect samples (test cases) which cover this population. It is very much a question of design of experiments and careful choice of test cases. A proper sampling paradigm/scheme (that says we need exactly more of e.g. “male; 10-15 y”, “female; 70-80y; smoker”) would help do a data-informed and targeted data search. Otherwise, even with more data, there is the risk that it is still not the right data.Possible solution: Community efforts to gather standardized test data sets from around the world. This test data set collection could either be organized on a central evaluation platform or in a federated fashion (see section 8 below). |
| 2 | Limits of in-house technical tests: lack of transparency, results not comparable. Possible solution: external validation (through independent benchmarking by trusted third parties), using standard technical test procedures designed by a multidisciplinary expert team. |
| 3 | Technical test criteria for AI models are potentially clinically irrelevant. Possible solution: Setting clinical objectives for the technical tests (and involving health domain experts in the test design). Subsequent verification with patient outcomes. |
| 4 | Clinical trials take time, put test subjects at risks, cost much, and may result in a limited number of sample points. Nevertheless, clinical trials have the advantage of being controlled experiments, and are designed such that the study population ideally is representative of the population of interest. This is currently lacking in most benchmarking exercises (where not even a population of interest is properly stated). Accordingly, every effort should be made to the chance to *properly* evaluate the models in silico first, and check them for different quality criteria, and then follow up with complex clinical trials. |
| 5 | Concerns that the unprecedented model complexity applied in complex settings makes it difficult to assess the models. However, black-box tests can be conducted irrespective of the model complexity. Appropriate testing procedures and metrics carefully designed in a community effort by multidisciplinary experts can help here. |
| 6 | Frequent model/software updates require frequent tests. The same applies to so called “self-learning” or “adaptive” algorithms that are automatically being re-trained based on new incoming data. (Note: AI models are often “locked”/“frozen” and not necessarily “self-learning”.) Assuming that a self-learning model might also perform worse over time, gets tested and then loses permission to operate in the clinic (from one day to another). What would happen? Hardly any software provider would take the risk of delivering a model that self-learns. From a business risk perspective one would prefer frozen models. But then in turn, we don’t realize the potential for increased accuracy of self-learning systems. So the patient and healthcare system are not leveraging AI to full potential (cf. [Gerke et al, 2020]). Automated pre-assessment via a platform could be a solution. The benchmarking platform can frequently assess the updated model versions and assure that there are no drastic changes that deteriorated the performance, at least on the test data. This check could support post market surveillance. |
| 7 | Aiming at becoming close to a technical equivalent to clinical trials, benchmarking challenges/competitions are applied to assess the technical performance of AI algorithms. They have a very high impact on the research field but there is almost no quality control. Solution: Integrate standardized guidelines and peer reviews for the benchmarking design, publish benchmarking designs in order to ensure transparency and reproducibility. (Note: Challenges can also be seen as collaborative challenges in which researchers work together on the best solution of a specific problem and not only as competitions.) |
| 8 | The human factor needs to be considered in a systemic view (cf. [Gerke et al, 2020]): In a clinical setting, the models are not operated autonomously but are embedded in the workflow of professional healthcare providers (HCP). This implies that actually the mode of AI usage by the HCP is an equally relevant part. Professionals with different grades of seniority will surely use the AI differently (i.e. more experienced, may be technology critical radiologist more often overrule the AI output; that might be correct or wrong). Models that are tuned for high sensitivity might have too many false positives. Hence, they get ignored by HCPs after a while (considering as not trustworthy).  |
| 9 | Similar to (8) if the reimbursement or legal frameworks either prefer or discourage use of AI, the HCPs could subconsciously be biased to use an accurate tool in the wrong way (training is needed?). As an example from [Gerke et al, 2020]: if payers only reimburse if recommendation is according to AI system, one gets a very strong emphasize on the AI although the system was designed as a “human-in-the-loop” setup.  |
| 10 | There are AI systems (sys1) that identify patients and design clinical trials. If these trials are meant to assess AI systems (sys2), then AI is assessing AI. If the sys1 is built on false data, then sys2 is basically also erroneous, right? While many would feel very uncomfortable if AI assesses AI, it is unclear whether this concern is justified. Theoretically it could be better than “humans assessing AI”. |
| … | … |

*These are initial ideas only. Please send additional novel/unique aspects which come to your mind that are not contained in this open, non-comprehensive list to the editor [markus.wenzel@hhi.fraunhofer.de]!*

# Independent standardized model benchmarking

*Independent benchmarking* *by a trusted third party using agreed-upon, standardized test procedures and metrics on separate high-quality test data from different sources* is the core idea for the technical testing step, pursued by the ITU/WHO Focus Group on “AI for Health” (cf. Figure 2). This approach could be a valuable complement to in-house or local technical tests and subsequent clinical trials. It doesn’t put test subjects at risk, can be repeated e.g. in the case of model/software updates, can be based on large amounts of high-quality test data from different sources, and is fast. Moreover, the approach can lead to comparable and transparent results using standardized testing procedures with meaningful test objectives, test tasks, quality criteria and test metrics, defined by a community of experts. In addition, releasing a new benchmarking data set drives the community and can serve as incentive to shift R&D resources into desired directions.



Figure 2: Independent model evaluation with standard procedures on separate test data

Standard benchmarking tasks with corresponding standard evaluation metrics and standard test data need to be agreed upon and clearly defined for every health subject area.

For this purpose, the ITU/WHO Focus Group on “AI for Health” established *Topic Groups* that work on *Topic Description Documents* ([DEL10] and applicable sub-documents) where these standards are defined for selected AI tasks from a range of health subject areas.

The *Topic Description Documents* and the *Data Specification* DEL05 with DEL05.1-6 will also address questions such as: How is the test data distributed? Are the samples balanced to represent the real world? How many cases will there be per benchmarking task and is this number sufficient to obtain the necessary statistical power? Will there be test data to check generalizability capabilities of submitted algorithms (e.g. data from another procedure/organ/hospital)?

On the long-term, a clear qualification process is needed for the recruitment of experts that construct the test data and metric, accompanied by a thorough review process and appropriate checks and balances. Moreover, it must be defined who calls the experts for a given health subject area, considering that there are always conflicting “schools of thought”.

## Benchmarking platform with evaluation in a closed environment

A benchmarking platform of health AI models should be able to produce meaningful test results

* by preventing overfitting of the models to the test data set, and
* by prohibiting attempts of cheating as good as possible.

Hence, the test data set must be unpublished and be withheld from the AI developer, and the evaluation should happen in a completely closed environment without connection to the internet.

A model that has been overfitted to the test data, can achieve an excellent test result without actually being able to perform well in practice when “fresh”, unknown data points are coming in and need to be processed. For instance, the model could simply memorize the association between test data points and corresponding labels (if they were known) and then correctly reproduce the labels during the test/benchmark, but be “helpless” in the “real world” when the model has to infer the label from “fresh” data points without knowing the label in advance.

The benchmarking platform with evaluation in a closed environment works – in brief - as follows:

The developer submits the to-be tested and already trained AI/ML model to the platform. In a closed environment, the model is provided with the test data points, processes these data, creates the corresponding output, which is then compared by evaluation functionality of the platform with the “ground truth”, using standardized quality criteria and metrics (as defined by the *Topic Groups* in their *Topic Description Documents* [DEL10] and applicable sub-documents). The evaluation results are returned to the AI/ML developer and the benchmark organizer.

The concept is explained in much **more detail in** **Appendix I**.

A benchmarking platform of this type might be considered as a ‘data safe haven’ open to developers and expert evaluators, where (a) the model performance can be assessed based on pre-set, standardized criteria, and where (b) models/code and test data are only open to restricted pass holders.

The approach can be compared with popular “challenge” platforms from the machine learning community to some extent (e.g. [aicrowd.com](https://www.aicrowd.com/)/[crowdai.org](http://crowdai.org/), [kaggle.com](http://kaggle.com/), [evalai.cloudcv.org](https://evalai.cloudcv.org/), [grand-challenge.org](https://grand-challenge.org/), [ichallenge.baidu.com](http://ichallenge.baidu.com/), [compression.cc/challenge](https://www.compression.cc/challenge), [codalab.org](http://www.codalab.org/), [ramp.studio](https://ramp.studio/), [carpl.ai/](https://carpl.ai/)). Learnings from conducting the aforementioned ML challenges are of paramount importance for the conceived health AI benchmarking platform, because we want to guarantee meaningful test results and fair procedures. Therefore, we should be aware of leader board probing, about how “weaknesses in biomedical challenge design and organization” can be exploited [Reinke et al., 2018], and “why rankings of biomedical image analysis competitions should be interpreted with care” [Maier-Hein et al., 2018].

Two alternatives to this centralized approach are presented in the next two sections 8.2 and 8.3.

## Benchmarking platform with evaluation via interface

The concept for a benchmarking platform with evaluation *via interface* is an alternative to the concept described above where the assessment happens in a *closed environment*.

The to-be-tested (and already trained) AI model connects to the benchmarking platform via an interface, e.g. over the internet. The platform sends test data points to the AI model, which processes the data, computes the corresponding output (e.g. labels) and returns this output to the benchmarking platform. The platform again compares the received output with the ground truth and computes the benchmarking result. The AI models are in this case not uploaded to the benchmarking platform but remain on the computers of the AI developers.

Advantage of this concept is that it meets concerns of developers who are hesitant to provide their AI models with business relevant trade secrets to the closed environment of the trusted third party described earlier.

This concept is followed by the *Topic Group on Symptom Assessment* and described in more detail in their Topic Description Document [**DEL10.14]**.

However, for obtaining meaningful results in a fair evaluation procedure, where cheating is prohibited, this concept requires the creation of new test data every time a benchmark is conducted. Otherwise, the test data could be stored and included in the model, which would greatly improve the chances for better results at the next benchmarking run. This inclusion in the model could happen e.g. by letting humans label the data by hand and then re-training (=overfitting) the model on these “test” data. Also unsupervised approaches (without label information) are conceivable, since information about the distribution of the test data is already valuable.

Obviously, creating new test data requires much effort, which is a disadvantage in comparison to the concept with the closed environment described earlier. (A possible solution: Generative adversarial networks or other algorithms might be able to synthesise new - but hopefully still realistic - data points in some cases.)

Moreover, all (potentially competing) AI models must be benchmarked at the very same moment, in the case of evaluation via interface. Otherwise, test data received by “model A” could be stored and included in a separate “model B” by the same developer, which could take part in the benchmark at a later time point. Model B would have greater chances for better, but meaningless test results.

(If the benchmarking procedure is sequential and data point after data point is sent to the model in random order, it must be guaranteed that each competitor takes part in the benchmark with only one model at the same time, for the same reason. Otherwise, information obtained from one model could be cross fed from model A to a different model B, similar to the description above.)

## Federated benchmarking platform

Evaluation in a *closed environment* can potentially also be conducted in a federated fashion (unlike the centralized concept described in section 8.1 and without transferring samples like in section 8.2).

The approach: the test data sets remain where they had been acquired, e.g. in different hospitals. The to-be-tested (and already trained) AI/ML model is sent to the locations where the data are stored (to the hospitals in this example). Here, the model is benchmarked against the local test data with standardized test procedures and metrics. The results are returned to the party that is organizing the benchmark.

While retaining the advantages of closed environment benchmarking, the federated benchmarking platform approach improves on the evaluation *via interface* by:

1. Lowering overhead related to curation for transfer, in that the test data is not transferred
2. Lowering communication load, in that less data is transferred overall
3. Lowering security risks related to data transfer, in that sensitive data remains where the acquisition took place

Nevertheless, appropriate security measures must be put in place to assure that the test data cannot be leaked.

The to-be tested AI/ML model must also be protected from undesired access aimed for example at getting hold of the source code itself (intellectual property) or of the training data via the model (e.g. through model inversion or adversarial attacks). Access to the models would also make it possible for data-providers to unnoticeably manipulate the test data (to make it harder for competitors to achieve good results; e.g. by adding adversarial noise).

Finally, from the benchmarking perspective, it must also be assured that neither model-provider nor data-provider can manipulate or interfere with the evaluation procedure.

Both these last points could be addressed by preventing access to the original source, for example via encapsulation into containerization software (e.g. Docker or Mesos) or pre-compilation.

Still, security mechanisms need to be put in place that keep track of when/how often the to-be-tested containerized model had been “touched” to avoid copying or reverse engineering, e.g. by creating a substitute model through feeding many data points in while observing the output [cf. Juuti et al 2019]. Moreover, the security mechanisms need to assure that all models have processed the very same data. Only then, the procedure is trustworthy and the results are comparable. Of course, federation does not solve all potential privacy and security issues and, similar to the other benchmarking approaches, will always carry some risks. As for minimising these, several strategies can be adopted, some of which are not specific to AI or federated benchmarking. Encapsulation helps by offering by default a series of tools such as:

- efficiently creating and managing controlled containers. For example containers offering only a small set of tools for loading the model/running the analysis and with all ports for communication blocked; sandbox environments for testing against adversarial attacks; buffer environments, etc.

- implementing tracking and versioning of the containers, so that only containers from a safe registry and from a certain version (ID, hash, etc.) can be run.

# Preliminary considerations on the evaluation process

While the overall *AI for health evaluation process* is described in detail in [DEL07.1], some preliminary considerations are discussed here.

Figure 3: Cyclical evaluation process

The cyclical process depicted above, with its distinct components, might present a robust path to evaluate the application of AI in the healthcare context. The evaluation commences with a clear definition of the respective task that the AI/ML models under assessment are expected to perform (cf. [DEL10]). For the technical tests, i.e. benchmarking procedure, (cf. [DEL07.2] and [DEL10]), quality criteria with corresponding metrics (cf. [DEL07.3]) are specified and the required test data properties are defined (cf. [DEL05] and [DEL10]). The test data set is collected according to these requirements and audited for quality, realism, representativeness and fidelity to the target population. Then, the models are automatically pre-assessed on a benchmarking platform using the previously defined metrics and test data sets. If the results of the benchmarking procedure indicate a sufficient level of quality, the next step can be initiated. Then, the safety, efficacy and usability in the clinic or field is validated (cf. [DEL07.4]). Pre- and post-deployment of the model, safety and quality mechanisms need to be put in place to ensure there is no harm to the patients because of the use of the model. This assessment continues throughout the life cycle of the model. Monitoring procedures allow for measuring the impact and for detecting anomalies and reporting them back to both developers and user-institution and if necessary relevant authorities. The evaluation process is meant to be connected and cyclical to ensure continuous quality improvement is enshrined.

# Overview of the evaluation deliverables

The present “AI for health evaluation considerations” serve as introduction to four documents that address important aspects for the evaluation of AI/ML models for health.

Table 2 and the next subsections give an overview of these deliverables.

Links to these documents can be found above in section 2.

Table 2: Overview of the evaluation deliverables.

| N° | Deliverable title | Editor |
| --- | --- | --- |
| DEL07.1 | AI for health evaluation process description | Sheng Wu (World Health Organization) |
| DEL07.2 | AI technical test specification | Auss Abbood (Robert-Koch Institut) |
| DEL07.3 | AI technical test metric specification | Luis Oala (Fraunhofer HHI) |
| DEL07.4 | Clinical validation of AI for health | Naomi Lee and Rupa Sarkar (The Lancet) |

## DEL07.1: AI for health evaluation process description

The deliverable “AI for health evaluation process description” [DEL07.1] summarizes the state-of-the-art of the evaluation process in a general overview, identifies evaluation gaps and research needs, and suggests potential solutions. It also provides an outlook to future developments of the AI technology that could lead to novel evaluation requirements.

## DEL07.2 AI technical test specification

The deliverable “AI technical test specification” [DEL07.2] discusses best practices and particular requirements for conducting the *in silico* technical testing and benchmarking of the AI models, covering insights from software verification and validation as well as from model testing in machine learning research.

## DEL07.3 AI technical test metric specification

The deliverable “AI technical test metric specification” [DEL07.3] discusses relevant quality criteria, e.g. performance, robustness, plausibility (sound explanations given by “explainable AI” methods), fairness, and associated statistical test approaches and test metrics. These metrics are suited to measure the AI model quality and can be employed in the aforementioned technical tests/ benchmarks.

## DEL07.4 Clinical Validation of AI for Health

The deliverable “Clinical Validation of AI for Health” [DEL07.4] considers the clinical/medical perspective, including e.g. clinical trials (or field tests e.g. in the case of public health interventions). The document also discusses how it can be assured that the aforementioned technical tests actually consider relevant, correct and meaningful objectives and clinical endpoints.

# Test data, ethical and regulatory aspects

Test **data** play a crucial part in the evaluation procedure and determine the significance of the test result to a great extent. All aspects related to data are treated in the separate documents **DEL05** and **DEL05.1-6**. Moreover, **ethical** and **regulatory** aspects must be considered for the evaluation of AI for health. These aspects are considered in the separate deliverables **DEL01** and **DEL02** with **DEL02.1-2**.

# Appendix IRequirements of a benchmarking platformwith evaluation in a closed environment

## I.1 System overview

In order to be able to assess the quality of different AI-based solutions to a variety of medical problems, a software system is required, which makes it easy to manage and discover benchmarking tasks, submit solutions, automatically evaluate them, and provide results in an aggregated human-readable format. Such a system should allow presenting medical benchmarking tasks to AI developers. It should provide all the details needed to prepare an AI-based solution for submission to the benchmarking procedure. AI developers may sign up for benchmarking tasks and submit their software solutions and auxiliary documentation. The software solutions must be submitted in a standardized form, so that they can be tested automatically by the system. The results of these tests and the auxiliary documentation should be made available to the administrators to simplify the evaluation.

The functional concept of the envisioned benchmarking platform is explained in the following. This system overview will broadly lay out the needed components and the most important requirements these components have to meet. The proposed system consists of three major components: an *administrative backend*, a *public frontend*, and an *execution environment* (cf. Figure I.1). The *administrative backend* comprises a web application where administrators are able to develop and publish benchmarking tasks and can inspect the benchmarking results. The *public frontend* constitutes the public interface to the proposed benchmarking tasks. AI developers and other interested parties can discover current benchmarking tasks and sign up for them. Participants are provided with all the necessary details about the benchmarking task, including but not limited to descriptions, public data sets, documentation, and examples. They can upload their AI solution for the benchmarking task to the platform in a standardized packaging format. The software projects contained in these solutions must adhere to a standardized interface for benchmarking. Once a solution has been handed in, it will be queued for benchmarking, which will be performed automatically by the *execution environment*. The *execution environment* will report back the results to the *administrative backend*, where they will be aggregated into a central overview (customizable tables, visualisations and/or rankings with potentially multiple ranking schemes). The following sections provide detailed information about each of these components.

### I.1.1 Administrative backend

The *administrative backend* of the platform serves as the management interface for the administrative staff. It is here where new AI health benchmarking tasks will be developed and published. It also provides the data storage for the system: a *database*, which contains the benchmarking tasks, and a *data set store* for managing data sets. Furthermore it provides an *internal interface*, e.g. in the form of a REST API, for the other components.

A new benchmarking task must at least consist of a name, a description, a private test data set, and a deadline. Furthermore, administrators may add further documentation, examples, a public data set and other auxiliary information. The private test data set will remain undisclosed and will be used to evaluate the submitted solutions. It represents the “gold standard” with the included true labels or annotations as “ground truth”. In contrast, the public data set will be made available to all participants.

In addition, it must be possible to specify how often the AI developers can submit their models to a benchmarking task. Even if the test data is hidden, it is possible to overfit to the test data by analysing the achieved metric values, checking which adjustments make them increase if multiple submissions are possible.

As soon as a benchmarking task has been developed, it can be published to the *public frontend*.



Figure I.1: Overview of the architecture of the system.

The *administrative backend* may consist of the following sub-components:

**Internal Interface**–The *internal interface* is a web application, which manages all benchmarking procedures and provides other components accesses to them. Separating the internal data management from other components makes it easy to share this functionality across multiple components: the *administration web client*, the *front-facing web client*, and the *execution manager* will all depend on the *internal interface*. The *internal interface* may, for example, be implemented as a REST API and should provide the following functionality:

* User management, i.e. identity management, identification, authorization, and access control through state-of-the art technologies. The user management will manage both the administrative accounts for the *administration web client*, as well as the user accounts of the benchmarking participants in the *front-facing web client*. Applicable data protection laws must be considered (e.g. in a privacy policy, stating what happens with the user’s personal data; “Right to be Forgotten” etc.).
* API security, e.g. through client IDs and client secrets for backend services or through host verification for frontend websites. This is needed, so that the *administration web client*, the *front-facing web client*, and the *execution manager* are able to authorize against the *internal interface* and to block all other systems from accessing the internal data.
* Add, update, and delete benchmarking tasks.
* Upload, update, and remove data sets.
* Manage benchmarking participants and registrations for benchmarking tasks.
* Maintain benchmarking results, and aggregate them to customizable tables, visualisations and/or rankings (with potentially multiple ranking schemes).
* A search engine, which will be used by the *front-facing web client* to retrieve and display benchmarking tasks.

**Database**–A *database*, which will store all structured information about users (administrators and participants), benchmarking tasks, benchmarking task registrations, submitted solutions, and benchmarking results. The *database* serves as a data storage for the *internal interface* and must only be accessed by it. The *database* may for example be implemented as a relational database. The transactional nature of most relational database systems ensures that, even in the event of failure, the *database* remains in a consistent state and lost data may be recovered. Consistence is vital to the *database* as it stores highly sensitive information. Data need to be stored encrypted.

**Data Set Store**–The *data set store* should be able to efficiently manage large amounts of unstructured data, e.g. by employing a binary large object (BLOB) store. The *internal interface* will store uploaded data sets here. Encryption of the data is mandatory. Ideally, the public data sets and the private test data sets should be “physically” separated from each other. While the public data sets should be downloadable by participants, the private test data sets must never be disclosed to the participants. This is particularly important, because the private test data serve for testing the generalization capabilities of the AI-based solutions in a fair manner. Crucially, and in contrast to some “data science challenges”, not only the true labels or annotations of the test data, but the entire test data sets including the “features” or “raw data” have to remain undisclosed. This has the following reason: Access to the private test data might tempt some participants to take unfair advantage by (asking human experts to label the test data and then) tuning their AI solution to produce good results on these test data (“overfitting”). Yet, this overfitting does not ensure that the solutions can generalize to other, previously unseen data, which is the core idea of the benchmarking framework. Besides, the private test data set may contain sensitive medical data, which must not be accessible to the public.

**Administration Web Client**–The *administration web client* is a website only available to administrators. It provides a management frontend for benchmarking tasks, data sets, submitted solutions, and benchmarking results. The *administration web client* interfaces with the *internal interface* for the management tasks.

### I.1.2 Public frontend

The *public frontend* is comprised of a *public-facing web client*, which is a website that is the portal to all proposed benchmarking tasks. The *public-facing web client* interfaces with the *internal interface* to provide AI developers and other interested parties access to the published benchmarking tasks. Users are able to create a user account. Unauthenticated users and authenticated users, which are not signed up for a benchmarking task, will be presented a list of current benchmarking tasks and are able to search for benchmarking tasks by keywords. For each benchmarking task, a details page exists, which contains the description of the benchmarking task and a deadline for submissions. Authenticated users will be able to sign up for a benchmarking task on the details page. Participants who have signed up for a benchmarking task will be able to access further documentation and examples, as well as the public (training or example) data set of the benchmarking task. Furthermore, the website should provide participants with detailed information about the submission process.

A submission will consist of two parts: the software solution and the documentation. The software solution must be packaged in a standardized format, which contains everything needed to execute the solution. The documentation should be contained in a single document (e.g. PDF or text file). The details page of a benchmarking task provides the means to upload solutions and documentations. The upload must be performed via an encrypted communication channel (e.g. *HTTPS*), because the solutions may contain secret intellectual property/trade secrets of the participant. After evaluation of the submissions, it should be possible to display the results in customizable tables and visualisations to the participants, to benchmark organizers, and to selected expert evaluators. Results might be aggregated in optional and possibly anonymous rankings, with potentially multiple ranking schemes.

### I.1.3 Execution environment

The *execution environment* consists of the *execution manager service* and an *execution server pool* on which *execution clients* can be run. The *execution manager service* orchestrates the benchmarking of the submitted software solutions and interfaces with the *internal interface* to retrieve queued submissions as well as the private test data sets. The *execution server pool* is a set of servers on which the actual execution of the software solutions is performed. Once a solution has been submitted by a participant, the software solution package is queued for benchmarking by the *administrative backend*. The *execution manager service* will go through the queued submissions and schedule them to run on an *execution client* in the *execution server pool*. When the *execution client* has completed the computations of the submitted software solution on the private test data, it will report the results back to the *execution manager service*, which will in turn calculate the benchmarking metrics and communicate them back to the *administrative backend*.

Participants submit software solutions in the form of a standardized package. The packaging format could, for example, be a compressed file (e.g. ZIP or TAR GZIP), which contains all files necessary to execute the solution. The package should contain a top level executable (e.g. an EXE file or a Batch file for a Windows environment, or a Linux executable or a Bash script file for a Linux environment), which is run by the *execution client*. It is important, that the software must adhere to a standardized interface, so that it can be automatically executed. This interface must be defined and documented, and may be implemented, for instance, through inter-process communication (IPC). The *execution client* runs the executable from the solution package and passes it all necessary information to use the IPC interface. The executable can then establish a connection to a process running on the *execution client*, from which it will receive the samples from the test data set and report back the inference result. This IPC interface could be implemented, e.g., as a local REST web service, which is only available via the local loopback address. This would ensure that the IPC protocol is cross-platform and could be used on Windows and Linux *execution clients* alike.

Since different AI developers may write software solutions in an array of different programming languages on different operating systems using a wide variety of AI software packages, the *execution environment* must be very flexible in order to be able to cater to these unique specifications. This can be best implemented by using some sort of containerization software (e.g. Docker, Mesos, or Singularity) or through virtualization (e.g. VMWare, Virtual Box, or HyperV). Each solution package must contain – besides the actual software – information about the desired execution environment (e.g. in the form of an XML or JSON document), which consists of a specific operating system and version (e.g. Windows 10, Ubuntu 18.04, Debian Stretch, etc.), a runtime (e.g. .NET, Python, Java, etc.), and a list of dependencies that have to be installed (e.g. .NET NuGet packages, Python PyPI packages, Debian APT packages, Chocolatey packages etc.). This is needed by the *execution manager service* to set up an *execution client* for the benchmarking. For this purpose, the *execution manager service* should maintain a set of “base images” for all supported operating systems. These base images should contain a stripped-down version of the operating system as well as a service, which is started once the container/virtual machine is launched. This service can receive and execute commands from the *execution manager service*.

When scheduling a submission for benchmarking, the *execution manager service* finds the next server in the *execution server pool*, which has sufficient available resources to run an execution client (e.g. using a scheduling algorithm like round robin). Then, the *execution manager service* reads the environment specification from the package and chooses the correct base image, based on the operating system specified in the environment specification. It creates a new *execution client* by starting a new container/virtual machine using the selected base image. Then, it connects to the service running inside the container/virtual machine and issues commands to install the specified runtime (e.g. .NET, Python, Java etc.) and the specified dependencies. When the *execution client* is ready, the private test data are retrieved from the *administrative backend.* The true labels or annotations of the test data are withhold. Only the unlabeled/ unannotated data points (i.e. “features”) are uploaded to the *execution client*. Subsequently, the software solution is uploaded to the *execution client*, unpacked, and run. Now, the solution generates output variables y = f(x) from the test data x. These results are reported back to the *execution manager service* that calculates the benchmarking metrics by comparing the reported results with the true labels or annotations using the respective statistical metric(s) for benchmarking. These metrics can be specific for each benchmarking task (e.g. accuracy, F1 score, precision, recall, ROC/AuC, Jaccard index etc.).

The *execution clients* must not be able to connect to any public network in order to keep the test data secret. This is of paramount importance for a fair evaluation and for protecting sensitive, personal, medical data. In addition, *execution clients* must not be able to connect to other *execution clients*, in order to protect intellectual property/trade secrets by keeping every software solution secret. Therefore, the *execution manager service* must establish a private network for each *execution client*. Error handling and monitoring must be addressed appropriately, without leaking any information about the test data to the participants (or about the solutions of other participants). Submitted solutions are expected to require graphics processing unit (GPU) access in the execution environment - for model execution, not for the typically more resource intensive model training.

## I.2 General considerations

Besides the specific requirements for each of the described components, there are general considerations, which apply to the entire system and to every component. These considerations subdivide into security, hosting, computing resources, and availability.

### I.2.1 Security

Test data and submitted AI solutions to-be-evaluated must be protected with the highest possible security standards and have to remain undisclosed. Thus, the assessment framework and all of its software components must adhere to the state of the art in computer security and the current universally valid standards in this field, and all involved server infrastructure must be kept up-to-date. Communication between the components as well as all external communication should be done using secure protocols (i.e. *HTTPS*). The safety of the software system as well as the server infrastructure should be automatically tested for vulnerabilities on a regular basis.

### I.2.2 Hosting

Another consideration for the assessment system is the question of hosting. The whole system should be run by a trusted third party in a private data centre, which appears be the most secure option in terms of data protection. Sensitive medical data and intellectual property will remain in the hands of the trusted third party only. This hosting option is the favoured solution, because it provides a maximum of security and trust for the contributors of data and the submitters of the health AI solutions to-be-benchmarked.

(Managing a custom data center takes a lot of effort. The alternative is to run the system in the public cloud, e.g. Microsoft Azure or Amazon AWS. These services already offer professional solutions to many problems stated in this document: professionally managed computing, highly available and secure databases, BLOB storage etc. Furthermore, public clouds operate on a pay-as-you-go basis, which makes them highly cost effective. For example, the load on the *execution server pool* is not constant but highly temporally limited to the period immediately after the deadline of a benchmarking task and low in the meantime. This poses a dilemma: on the one hand, there should be enough resources for times of peak workload, but on the other hand, the machines should not be idle in times of low load. When hosting the system in a private data center, an adequate amount of servers must be managed and kept available at all times, while from a cloud service, virtual machines may be rented by the minute. Nevertheless, security and trust are priorities.)

### I.2.3 Computing resources

The computing resources required for the benchmarking platform depend on the number and difficulty of the offered benchmarking tasks and on the quantity and properties of the future AI submissions. The submitted AI solutions to-be-benchmarked will vary in terms of the required computing resources. Some solutions may need graphics processing unit (GPU) access, others not. Memory requirements will differ between submitted AI solutions. Storage space for test data in the data set store will depend on the benchmarking tasks.

### I.2.4 Availability

It is to be expected that AI developers and other interested parties from all over the world will participate in the offered benchmarking tasks. Therefore, the *frontend web client* should work reliably worldwide, even under high workload. Finally, server infrastructure may fail due to several reasons, e.g. software or hardware failures. Furthermore, single servers need to be taken down regularly for software updates. In order to be able to continue operation even in the event of failure, the whole system should be implemented with fault tolerance in mind, and regular backups need to be considered.

# Appendix IIBest practices from the scientific literature and other documents

A non-comprehensive collection of best practices, recommendations, insights and perspectives from the scientific literature and other documents concerned with the evaluation and testing of AI/ML models and health technology is presented and discussed here.

Boldface type highlights text passages that were considered to be particularly relevant for the evaluation of AI for health (boldface in this document; not the original format).

*Please send additional literature suggestions and comments to the editor [markus.wenzel@hhi.fraunhofer.de]!*

|  |
| --- |
| **Mathews at al. (2019)** |
| Mathews, S. C., McShea, M. J., Hanley, C. L., Ravitz, A., Labrique, A. B., & Cohen, A. B. (2019). Digital health: a path to validation. *NPJ Digital Medicine, 2*(1), 1-9. <https://doi.org/10.1038/s41746-019-0111-3>  |
| *“All healthcare stakeholders would benefit from a more standardized, objective, rigorous, and transparent process for validation. Specifically, the* ***validation domains*** *would be* ***technical*** *validation (e.g., how accurately does the solution measure what it claims?),* ***clinical*** *validation (e.g., does the solution have any support for improving condition-specific outcomes?), and* ***system*** *validation (e.g., does the solution integrate into patients’ lives, provider workflows, and healthcare systems).”* |
| **Considerations:** * Especially interesting is Fig. 2 where the proposed “Independent Evaluator” matches well with the independent benchmarking proposed in the present document [DEL07].
* This article addresses digital health in general and points out important issues. Now, the aspects of AI that are unique and novel in comparison to other digital health solutions need to be carved out - see Table 1 in the present document [DEL07].
 |

|  |
| --- |
| **Consort-AI & Spirit-AI (2019)** |
| Liu, X., Rivera, S.C., Faes, L. et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med 25*, 1467–1468 (2019). <https://doi.org/10.1038/s41591-019-0603-3>(Additional publication: Liu, X., Faes, L., Calvert, M. J., & Denniston, A. K. (2019). Extension of the CONSORT and SPIRIT statements. *The Lancet*, *394*(10205), 1225. [https://doi.org/10.1016/S0140-6736(19)31819-7](https://doi.org/10.1016/S0140-6736%2819%2931819-7) ) |
| *“As artificial intelligence moves into the realm of clinical trials, consideration is needed on whether the current CONSORT and SPIRIT reporting statements are sufficient to ensure transparency. [...]**Most AI interventions thus far, particularly diagnostic algorithms, have been evaluated only in the context of diagnostic accuracy. Although this initial validation stage is important, a demonstration of good diagnostic accuracy does not necessarily translate to improved patient outcomes. Yet if the ultimate goal of introducing AI into healthcare is to bring about patient benefit, then demonstration of improved patient outcome is needed. This should be done in a prospective clinical trial, in which the AI intervention is placed within its intended clinical pathway, with patient outcomes as the primary endpoint, and with an evaluation of demonstrable downstream effects in the broader management strategy. [...]**Although this guidance has substantially improved the completeness of clinical trials reporting, there are* ***challenges in trials involving AI interventions that are not addressed by the current guidance****. For example, elements that require detailed and specific reporting include the* ***study setting and its ability to administer a machine learning intervention in real time, the criteria for inclusion at the input-data level as well as at the participant level, the interactions between the human and the algorithm and its potential knock-on effects downstream, and the effects of adaptive machine learning technologies (which have the potential to continuously improve in performance)****. Without complete and transparent reporting, readers cannot assess the validity and generalizability of the findings, which can result in widespread misconception of overstated efficacy and utility. The risk is that an AI intervention that might not be effective or feasible in the real world could be commissioned and implemented. [...]**To address these challenges, the* ***CONSORT-AI and SPIRIT-AI Steering Group is preparing international, consensus-based, AI-specific extensions to the CONSORT and SPIRIT statements that will focus specifically on clinical trials in which the intervention includes a machine learning or other AI component****, using the EQUATOR (Enhancing Quality and Transparency of Health Research) Network methodological framework for guideline development5. This* ***initiative will be complementary to*** *the efforts of others working on reporting standards such as the* ***TRIPOD-ML*** *(TRIPOD, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) initiative of Collins and Moons, which seeks to improve the reporting of machine-learning-driven predictive model development and validation6. [...]**A final consensus meeting and the* ***publication of final recommendations will be completed in spring 2020****. Once the new guidance becomes available, we recommend that investigators planning to take an AI intervention through to clinical trials consult the new reporting standards as early as possible.”* |
| **Considerations:*** The independent benchmarking proposed in the present document [DEL07] is complementary to CONSORT-AI, SPIRIT-AI and TRIPOD-ML and provides chances of thorough technical tests prior to clinical trials.
 |

|  |
| --- |
| **Consort-AI (2020)** |
| Forthcoming journal paper.Website: The CONSORT-AI Extension: Reporting Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (registered on 8th of May, 2019). <https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials/#AI> |
| *“The group plans to publish the reporting guideline by March 2020, in an open-access journal.”*  |
| * See *Consort-AI & Spirit-AI (2019)* above.
 |

|  |
| --- |
| **Spirit-AI (2020)** |
| Forthcoming journal paper. Website: Protocol Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (SPIRIT-AI Extension) (registered 21 June 2019)<https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials-protocols/#AI>  |
|  *“The group plans to publish the reporting guideline as an open-access document in March 2020.”*  |
| * See *Consort-AI & Spirit-AI (2019)* above.
 |

|  |
| --- |
| **TRIPOD-ML** |
| Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, *393*(10181), 1577-1579. [https://doi.org/10.1016/S0140-6736(19)30037-6](https://doi.org/10.1016/S0140-6736%2819%2930037-6)  |

|  |
| --- |
| **TRIPOD** |
| Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ 2015; 350:g7594. PMID: [25569120](http://www.ncbi.nlm.nih.gov/pubmed/25569120)  |
| *“Reporting of studies developing,* ***validating****, or updating a* ***prediction model****, whether for diagnostic or prognostic purposes.”* (Quote from <https://www.equator-network.org/reporting-guidelines/tripod-statement/> ) |
| **Considerations:*** Basis for the extension towards machine learning “TRIPOD-ML”, cited above.
 |

|  |
| --- |
| **Studer et al. (2020)** |
| Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. R. (2020). Towards CRISP-ML (Q): A Machine Learning Process Model with Quality Assurance Methodology. *arXiv preprint arXiv:2003.05155*. <https://arxiv.org/abs/2003.05155> |
| *“This evaluation phase consists of three tasks:* ***evaluation of performance, robustness and explainability****. […] it is important to assure the* ***correctness of the results but also to study its behaviour on false inputs****. A* ***major risk is caused by the fact that a complete test coverage of all possible inputs is not tractable because of the large input dimensions****. However, extensive testing reduces the risk of failures. When testing, one has to always keep in mind that the stochastic nature of the data resulting in label noises bounds the test accuracy from the top. That means, 100% test accuracy can be rarely achieved.****Validate performance****: A risk occurs during the validation of the performance by using feedback signals from the test set to optimize the model. To avoid this, it is good practice to hold back an additional* ***test set****, which is* ***disjoint from the training (and validation) set and stored only for a final evaluation and never shipped to any partner to be able to measure the performance metrics in a kind of blind-test way****. To not bias the performance of a model, the* ***test set should be assembled and curated with caution and ideally by a team of experts that are capable to analyze the correctness and ability to represent real cases****. In general, the* ***test set should cover the whole input distribution and consider all the invariances in the data****. Invariances are transformations of the input that should not change the label of the data. (Zhou and Sun, 2019; Tian et al., 2018; Pei et al., 2017) have shown that a highly sophisticated model for autonomous driving could not capture those invariances and found extreme cases which led to false predictions by transforming a picture taken on a sunny day to a rainy day picture or by darkening the picture. It is recommended to* ***separate the teams and the procedures collecting the training and the test data to erase dependencies and avoid false methodology propagating from the training set to the test set. On that test set, the prior defined performance metrics should then be evaluated****. Additionally, it is recommended to perform a sliced performance analysis to highlight weak performance on certain classes or time slices. A full test set evaluation may mask flaws on certain slices.****Determine robustness****: The robustness of the model, in terms of the model's ability to generalize to a perturbation of the data set, can be determined with K-fold cross-validation. Hereby, the algorithm is repeatedly validated by holding disjoint subsets of the data out of the training data as validation data. The mean performance and variance of the cross-validation can be analyzed to check the generalization ability of the model on different data sets. […] Moreover, robustness should be checked when adding different kinds of noise to the data or varying the hyper-parameters which characterize the model indirectly (e.g. the number of neurons in a deep neural network). In addition, it is recommended to assure robustness of a model when given wrong inputs e.g. missing values, NaNs or data out of distribution as well as signals which might occur in case of malfunctions of input devices such as sensors. A different challenge is given by adversarial examples (Goodfellow et al., 2014) that perturbs the image by an imperceptible amount and fool classifiers to make wrong predictions. […]* *[…]* ***to avoid spurious correlations*** *(compare clever hans phenomenon in (Lapuschkin et al., 2019)), it is best practice to* ***carefully observe the features which impact the model's prediction the most and check whether they are plausible from a domain experts' point of view****. For example, heat maps highlight the most significant pixels in image classification problem (Lapuschkin et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017; Lapuschkin et al., 2019) or the most significant words in NLP tasks (Arras et al., 2017). [...]**Model evaluation under production condition: As training and test data is gathered to train and evaluate the model, the possible* ***risk*** *persists* ***that the production data does not resemble the training data or didn't cover corner cases****. Previous assumptions on the training data might not hold in production and the hardware that gathered the data might be different. Therefore it is best practice to evaluate the performance of the model under incrementally increasing production conditions by iteratively running the tasks […]**Assure user acceptance and* ***usability****: Even after passing all evaluation steps, there might be the risk that the user acceptance and the usability of the model is underwhelming. The model might be incomprehensible and did not cover corner cases. It is best practice to build a prototype and run an exhaustive field test with end users.”* |
| **Considerations:*** Excellent summary concerning ML model evaluation methods.
* This cross-industry perspective could be adapted to health domain.
* Clinical validation needs to be added/integrated in this technical perspective.
 |

|  |
| --- |
| **Zhang et al. (2020)** |
| Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*. <https://doi.org/10.1109/TSE.2019.2962027> (or: <https://arxiv.org/abs/1906.10742>) |
| *“This paper provides a comprehensive survey of Machine Learning Testing (ML testing) research. It covers 144 papers on testing properties (e.g., correctness, robustness, and fairness), testing components (e.g., the data, learning program, and framework), testing workflow (e.g., test generation and test evaluation), and application scenarios (e.g., autonomous driving, machine translation).”* |
| **Considerations:** * Valuable recent overview of the state of the art of machine learning model testing.
* The document is a general survey. Hence, the specifics for the health/medicine domain could deserve additional attention, e.g. the clinical perspective (Do the test consider a relevant and correct clinical endpoint/objective?; clinical trials; etc.)
 |

|  |
| --- |
| **IMDRF Software as a Medical Device Working Group (2017)** |
| IMDRF Software as a Medical Device Working Group (2017). Software as a Medical Device (SaMD): Clinical Evaluation. <http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf>  |
| *“5.3 Analytical / Technical Validation of a SaMD**Analytical validation* ***measures the ability of a SaMD to accurately, reliably and precisely generate the intended technical output from the input data****. Said differently, analytical validation:** *Confirms and provides objective evidence that the software was correctly constructed – namely, correctly and reliably processes input data and generates output data with the appropriate level of accuracy, and repeatability and reproducibility (i.e., precision); and*
* *Demonstrates that (a) the software meets its specifications and (b) the software specifications* ***conform to user needs and intended uses****. The analytical validation is* ***generally evaluated and determined by the manufacturer*** *during the verification and validation phase of the software development lifecycle using a QMS. [...]*

*A SaMD can best be described as software that utilizes an algorithm (logic, set of rules, or* ***model****)* ***that operates on data input*** *(digitized content)* ***to produce an output that is intended for medical purposes as defined by the*** *SaMD* ***manufacturer*** *(Figure 9). The risks and benefits posed by SaMD outputs are largely related to the risk of inaccurate or incorrect output of the SaMD, which may impact the clinical management of a patient.”* |
| **Considerations:** * In-house validation by the manufacturer is not as trustworthy as external validation by independent third party. Expert knowledge and large amounts of high quality, independent test data from different sources are required for high quality, meaningful validation. Various flaws in the testing pipeline can lead to meaningless testing results.
* In-house test data are often limited and close to training data. Hence, significance of test results might be limited and findings might not translate to actual application in the field.
* Test data should not be known to the developer for meaningful testing.
 |

|  |
| --- |
| **EU Regulation on medical devices (2017)** |
| Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance) ELI: <http://data.europa.eu/eli/reg/2017/745/2017-05-05>  |
| *„(22)  ‘****performance’*** *means the ability of a device to* ***achieve its intended purpose as stated by the manufacturer****; […]“**“(51) ‘clinical evidence’ means clinical data and clinical evaluation results pertaining to a device of a sufficient amount and quality to allow a qualified assessment of whether the device is safe and achieves the intended clinical benefit(s), when used as intended by the manufacturer;**(52) ‘clinical performance’ means the ability of a device, resulting from any direct or indirect medical effects which stem from its technical or functional characteristics, including diagnostic characteristics, to achieve its intended purpose as claimed by the manufacturer, thereby leading to a clinical benefit for patients, when used as intended by the manufacturer;**(53) ‘clinical benefit’ means the positive impact of a device on the health of an individual, expressed in terms of a meaningful, measurable, patient-relevant clinical outcome(s), including outcome(s) related to diagnosis, or a positive impact on patient management or public health; […]”**„15.1. Diagnostic devices and devices with a measuring function, shall be designed and manufactured in such a way as to* ***provide sufficient accuracy, precision and stability for their intended purpose, based on appropriate scientific and technical methods. The limits of accuracy shall be indicated by the manufacturer****. […]“**“17.1. Devices that incorporate electronic programmable systems, including* ***software****, or software that are devices in themselves,* ***shall be designed to ensure repeatability, reliability and performance in line with their intended use****. […]”**“The documentation shall contain the results and critical analyses of all* ***verifications and validation tests and/or studies*** *undertaken to demonstrate conformity of the device with the requirements of this Regulation and in particular the applicable* ***general safety and performance requirements****.[…]”**“(b)* ***detailed information*** *regarding* ***test design, complete test or study protocols, methods of data analysis, in addition to data summaries and test conclusions*** *regarding in particular: [...]**— software verification and validation (describing the software design and development process and evidence of the validation of the software, as used in the finished device. This information shall typically include the summary* ***results of all verification, validation and testing performed both in-house and in a simulated or actual user environment prior to final release.*** *[…]”**“[…]* ***The notified body shall*** *have documented procedures, sufficient expertise and facilities for the type-examination of devices in accordance with Annex X including the capacity to:[…]**—* ***establish a test plan identifying all relevant and critical parameters which need to be tested by the notified body or under its responsibility****; […]**—* ***carry out the appropriate examinations and tests*** *in order to verify that the solutions adopted by the manufacturer meet the general safety and performance requirements set out in Annex I. Such examinations and* ***tests shall include all tests necessary to verify that the manufacturer has in fact applied the relevant standards it has opted to use****; […]**— assume full responsibility for test results.* ***Test reports*** *submitted by the manufacturer shall only be taken into account if they have been issued by conformity assessment bodies which are competent and* ***independent of the manufacturer****. […]”* |
| **Considerations:*** The regulatory perspective is discussed in more detail in the *AI4H regulatory considerations* [DEL02].
 |

|  |
| --- |
| **BIAS Initiative: Transparent reporting of biomedical image analysis challenges** |
| Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., ... & Saez-Rodriguez, J. (2019). BIAS: Transparent reporting of biomedical image analysis challenges. *arXiv preprint*. https://arxiv.org/abs/1910.04071Also registered to the Equator Network: |
| *“The Biomedical Image Analysis ChallengeS (BIAS) initiative was founded by the challenge working group of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society board with the* ***goal of bringing biomedical image analysis challenges to the next level of quality****”**“This paper of the initiative presents a* ***guideline to standardize and facilitate*** *the writing and reviewing process of biomedical image analysis challenges and* ***help readers of challenges interpret and reproduce results by making relevant information explicit****.”**“An increasingly relevant problem is that it typically* ***remains unknown which specific feature of one algorithm actually makes it better than competing algorithms*** *[18]. For example, many researchers are convinced that the method for data augmentation often has a much bigger influence on the performance of a deep learning algorithm than the network architecture itself. For this reason, a* ***structured description*** *(e.g. using ontologies) not only of the challenge but* ***also of the participating algorithms may be desirable****.”* |
| **Considerations:*** Standardized guideline for challenge design
 |

|  |
| --- |
| **Benchmarking Visualization Toolkit** |
| Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A. (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. *arXiv preprint* <https://arxiv.org/abs/1910.05121> |
| *“The presentation of results in publications is commonly* ***limited to tables and simple visualization of the metric values*** *for each algorithm. […]* ***crucial information on the stability of the ranking is not conveyed.****”**“[Example:] The* ***rankings*** *of these challenges are* ***identical****, although the* ***distributions of metric values are radically different****.”**“The purpose of this paper is therefore to propose methodology along with an* ***open-source framework*** *for systematically* ***analyzing and visualizing results*** *of challenges. Our work will help challenge organizers and participants* ***gain further insights into both the algorithms’ performance and the assessment data set*** *itself in an intuitive manner.”**“Whereas the methodology and toolkit proposed were designed specifically for the analysis and visualization of challenge data, they may also be* ***applied to presenting the results of validation studies performed in the scope of classical original papers****. In these papers it has become increasingly common to compare a new methodological contribution with other previously proposed methods.* ***Our methods can be applied to this use case in a straightforward manner****.”* |
| **Considerations:*** Could be valuable for the benchmarking platform as well in order to visualize the results.
 |

|  |
| --- |
| **To Be Continued** |
| Add literature reference here |
| Quote from the reference here and highlight particularly relevant text passages in boldface. |
| **Considerations:*** Add brief discussion about merits and limitations of the quoted literature source here.
 |

It might be worthwhile to add excerpts from e.g. [Dustler et al., 2020], [Kim et al., 2020], McKinney at al., 2020], [Talmon et al, 2009], [The English National Institute for Health and Care Excellence, 2019], [Van Calster et al., 2019], [Gerke et al., 2020] (and from many more publications) to this non-comprehensive collection of best practices.

Bibliography

Anderson‐Cook et al. (2019). How to Host an Effective Data Competition: Statistical Advice for Competition Design and Analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(4), 271-289. <https://doi.org/10.1002/sam.11404>

Collins, G. S., & Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, *393*(10181), 1577-1579. [https://doi.org/10.1016/S0140-6736(19)30037-6](https://doi.org/10.1016/S0140-6736%2819%2930037-6)

Dustler, M. (2020). Evaluating AI in breast cancer screening: a complex task. The Lancet Digital Health. [https://doi.org/10.1016/S2589-7500(20)30019-4](https://doi.org/10.1016/S2589-7500%2820%2930019-4)

Gerke, S., Babic, B., Evgeniou, T. et al. (2020) The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digit. Med. 3*, 53. <https://doi.org/10.1038/s41746-020-0262-2>

Juuti, M., Szyller, S., Marchal, S., & Asokan, N. (2019, June). PRADA: protecting against DNN model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 512-527). IEEE. <https://doi.org/10.1109/EuroSP.2019.00044>

Kim, H. E., Kim, H. H., Han, B. K., Kim, K. H., Han, K., Nam, H., ... & Kim, E. K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health. [https://doi.org/10.1016/S2589-7500(20)30003-0](https://doi.org/10.1016/S2589-7500%2820%2930003-0).

Liu, X., Rivera, S.C., Faes, L. et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med 25*, 1467–1468 (2019). <https://doi.org/10.1038/s41591-019-0603-3>

Liu, X., Faes, L., Calvert, M. J., & Denniston, A. K. (2019). Extension of the CONSORT and SPIRIT statements. *The Lancet*, *394*(10205), 1225. [https://doi.org/10.1016/S0140-6736(19)31819-7](https://doi.org/10.1016/S0140-6736%2819%2931819-7)

Maier-Hein et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1), 5217. <https://doi.org/10.1038/s41467-018-07619-7>

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., ... & Saez-Rodriguez, J. (2019). BIAS: Transparent reporting of biomedical image analysis challenges. *arXiv preprint*. <https://arxiv.org/abs/1910.04071>

Mathews, S. C., McShea, M. J., Hanley, C. L., Ravitz, A., Labrique, A. B., & Cohen, A. B. (2019). Digital health: a path to validation. *NPJ Digital Medicine*, *2*(1), 1-9. <https://doi.org/10.1038/s41746-019-0111-3>

McKinney et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89-94. <https://doi.org/10.1038/s41586-019-1799-6>

Michaelis et al. (2019). Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv preprint*. <https://arxiv.org/abs/1907.07484>

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, *162*(1), W1-W73. <https://doi.org/10.7326/M14-0698>

Protocol Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (SPIRIT-AI Extension) (registered 21 June 2019). <http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials-protocols/#AI>

Reinke et al. (2018). How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 388-395). Springer, Cham. <https://doi.org/10.1007/978-3-030-00937-3_45>

Salathé, M., Wiegand, T., Wenzel, M. and Kishnamurthy, R. (2018). White Paper of the Focus Group on Artificial Intelligence for Health. <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf>

Talmon, J., Ammenwerth, E., Brender, J., De Keizer, N., Nykänen, P., & Rigby, M. (2009). STARE-HI—Statement on reporting of evaluation studies in Health Informatics. *International Journal of Medical Informatics*, *78*(1), 1-9. <https://doi.org/10.1016/j.ijmedinf.2008.09.002>

The CONSORT-AI Extension: Reporting Guidelines for Artificial Intelligence and Machine Learning Interventions in Randomised Trials (registered on 8th of May, 2019). <http://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-clinical-trials/#AI>

The English National Institute for Health and Care Excellence (2019). Evidence Standards Framework for Digital Health Technologies. <https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf>

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association, 26*(12), 1651-1654. <https://doi.org/10.1093/jamia/ocz130>

WHO (2013). WHO Evaluation Practice Handbook. ISBN 978 92 4 154868 7. <https://apps.who.int/iris/bitstream/handle/10665/96311/9789241548687_eng.pdf>

WHO (2016). Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. Geneva: World Health Organization; 2016. <https://apps.who.int/iris/bitstream/handle/10665/252183/9789241511766-eng.pdf>

Wiegand, T., Krishnamurthy, R., Kuglitsch, M., Lee, N., Pujari, S., Salathé, M., Wenzel, M. and Xu, S. (2019). WHO and ITU establish benchmarking process for artificial intelligence in health. *The Lancet, 394*(10192), 9-11. [https://doi.org/10.1016/S0140-6736(19)30762-7](https://doi.org/10.1016/S0140-6736%2819%2930762-7)

Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A. (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. *arXiv preprint* <https://arxiv.org/abs/1910.05121>

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