ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Focus Group Deliverable

(03/2023)

Focus Group on Artificial Intelligence for Health (FG-AI4H)

FG-AI4H DEL07

Artificial intelligence for health evaluation considerations

ITU-T FG-AI4H Deliverable DEL07

Artificial intelligence for health evaluation considerations

Summary

In this Technical Report, considerations on the evaluation and benchmarking of health artificial intelligence (AI) are presented, novel characteristics of health AI validation and evaluation are identified, and the concept of standardized model benchmarking is discussed. Moreover, requirements for a benchmarking platform are considered in detail, and best practices for a health AI model assessment are collected from selected sources.

Keywords

Artificial intelligence, assessment, clinical evaluation, health, machine learning, medical informatics, software, technical validation, testing, medicine, trustworthy.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1 of Deliverable DEL07 on "*Artificial intelligence for health evaluation considerations*" approved on 16 March 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H)].

Editor: Markus Wenzel, Fraunhofer HHI, Germany, Email: markus.wenzel@hhi.fraunhofer.de

Contributors: (in alphabetical order)

Alberto Merola, AICURA medical GmbH, Germany, Email: alberto.merola@aicura-medical.com

David Neumann, Fraunhofer HHI, Germany, Email: david.neumann@hhi.fraunhofer.de

Sandeep Reddy, Deakin University, Australia, Email: sandeep.reddy@deakin.edu.au

Annika Reinke, DKFZ, Germany, Email: a.reinke@dkfz-heidelberg.de

Steffen Vogler, Bayer AG, Germany, Email: steffen.vogler@bayer.com

Table of Contents

ITU-T FG-AI4H Deliverable DEL07

Artificial intelligence for health evaluation considerations

1 Scope

Considerations on the testing, validation, benchmarking and evaluation of Artificial Intelligence (AI) and machine learning (ML) models, methods, or systems in the healthcare domain. The Technical Report belongs to a series of International Telecommunication Union ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) deliverables (DEL) listed in the overview [DEL0]. Thus, separate deliverables cover several aspects not addressed here, e.g., data, ethics, regulation, etc.

2 References

[DEL0] FG-AI4H deliverable DEL0, *Overview of the FG-AI4H deliverables*.

[DEL0.1] FG-AI4H deliverable DEL0.1, *Common unified terms in artificial intelligence for health.*

[DEL1] FG-AI4H deliverable DEL01, *AI4H ethics considerations*.

[DEL2] FG-AI4H deliverable DEL02, *Overview of regulatory considerations on artificial intelligence for health*.

[DEL5] FG-AI4H deliverable DEL05, *Data specification*.

[DEL7] FG-AI4H deliverable DEL07, *Artificial intelligence for health evaluation consideration*.

[DEL7.2] FG-AI4H deliverable DEL07.2, *AI technical test specification*.

[DEL7.4] FG-AI4H deliverable DEL07.4, *Clinical evaluation of AI for health*.

[DEL10] FG-AI4H deliverable DEL10, *AI4H use cases: topic description documents.*

[DEL10.14] FG-AI4H deliverable DEL10.14, *Topic group on Symptom assessment*.

NOTE – The Bibliography lists additional literature references.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report adopts the terms defined in [DEL0.1].

3.2 Terms defined in this Technical Report

None.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

AI Artificial Intelligence

AI4H Artificial Intelligence for Health

DEL Deliverable

DL Deep Learning

FG-AI4H ITU/WHO Focus Group on Artificial Intelligence for Health

HCP	Healthcare Providers
ITU	International Telecommunication Union
ML	Machine learning
WHO	World Health Organization
XAI	Explainable AI

5 Background

Evidence-based trust is essential for any health technology. Obviously, this applies to solutions based on AI and ML too. Therefore, it must be ensured that the AI/ML models are accurate, robust, transparent, fair, free from bias, plausible, and therefore trustworthy. Moreover, they must be usable, effective, and safe in practice. The present Technical Report collects and discusses best practices and recommendations for creating this evidence with a particular focus on the benchmarking of AI/ML models for health.

Health AI/ML models and the circumstances where they are applied can be highly complex, and hence it is important to thoroughly assess the merits and limitations of each model (respectively system/method). AI/ML models can analyse various modalities of data, e.g., microscopic 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 learns to perform one or a few specific task(s), such as classification, or regression, or segmentation 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.

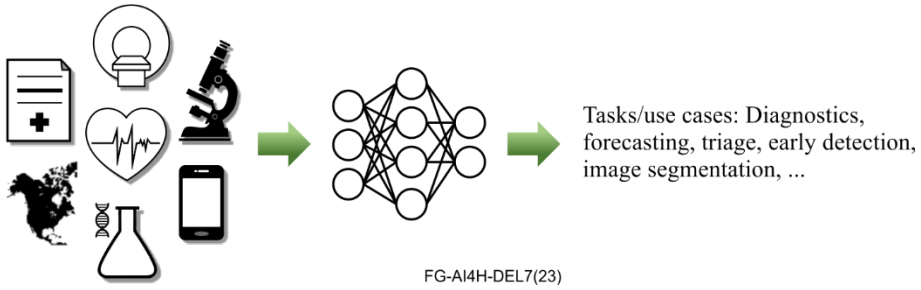

Figure 1 – Illustration of exemplary data sources and tasks/use cases for health AI/ML models

6 Evaluation process

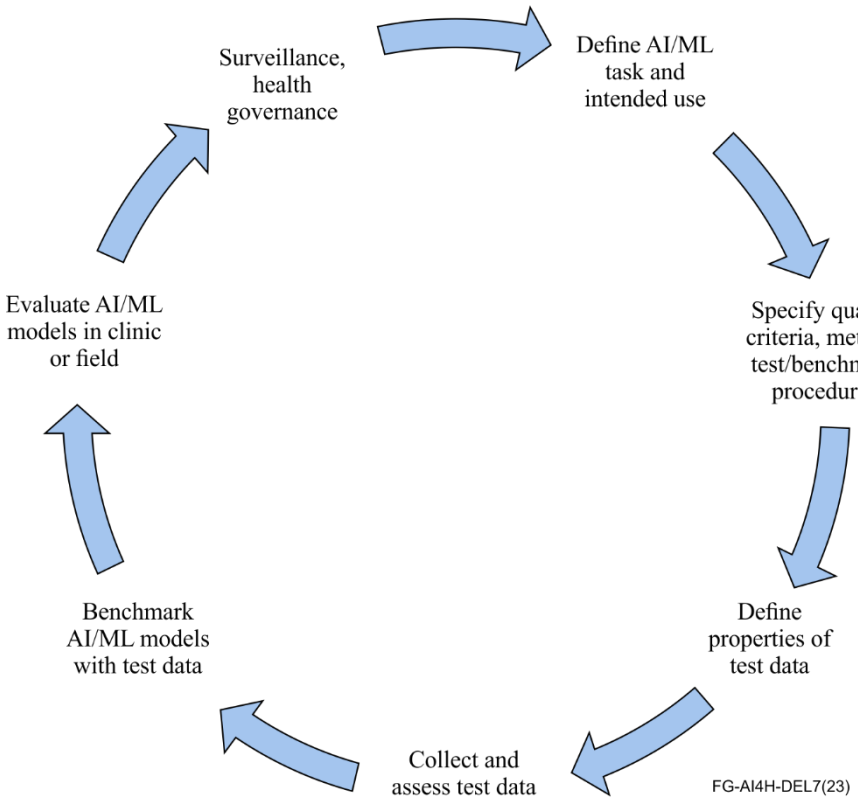

Figure 2 – Cyclical evaluation process

The cyclical evaluation process depicted above in Figure 2, with its distinct components, might present a robust path to assess the application of AI models in the healthcare context. The process commences with a clear definition of the respective task that the AI/ML models under assessment are expected to perform and of the intended use. Suitable quality criteria with corresponding meaningful procedures and metrics are specified for the initial technical test / benchmarking steps. The required test data properties are defined. The test data set is collected according to these requirements and assessed for quality, realism, representativeness, and fidelity to the target population. Then, the AI/ML models are automatically tested and pre-assessed on a benchmarking platform using the previously defined procedures, metrics, and test data sets. Ideally, these assessments are conducted in a standardized way on a benchmarking platform operated by a trusted third party. If the results of the benchmarking procedure indicate a sufficient level of quality, the next step can be initiated. While technical testing and benchmarking are important steps in the evidence-generating procedure, it should be followed up by clinical evaluation (e.g., with randomized controlled trials) or field / scientific tests and post market surveillance procedures to assess the efficacy, safety, usability, and cost-effectiveness of the technology in practice. Pre- and post-deployment of the model, health governance mechanisms, which monitor for safety, quality, risks, and ethical and privacy issues, 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-institutions and if necessary relevant authorities. The evaluation process is meant to be connected and cyclical to ensure continuous quality improvement is enshrined.

6.1 Reference to related deliverables

Best practices and particular requirements for testing the AI models are addressed in detail in the separate deliverable "AI technical test specification" [DEL07.2], which covers insights from software

verification and validation as well as from model testing in machine learning (ML) research. The deliverable "Clinical evaluation of AI for health" [DEL07.4] considers how AI models should be evaluated for a safe use in a complex clinical environment. As growing numbers of AI models become available for use researchers, patients, clinicians, and policy makers require a framework to understand whether the models are safe, purpose-fit and cost effective, and to compare model performance with current standards of care, and between each other. [DEL07.4] describes current best practices for clinical evaluation of AI models in health both, pre- and post- deployment. It also identifies gaps in the current evaluation framework for future work and addresses how it can be assured that prior technical validation actually considers relevant, correct, meaningful objectives and defines clinically meaningful endpoints. Topic-specific benchmarking procedures are discussed in the topic description documents of [DEL10]. Ethical, regulatory, and data-related aspects must be considered for the evaluation of AI for health as well. These aspects are considered in much detail in the separate deliverables [DEL01], [DEL02], and [DEL05].

7 Novelty

Fortunately, we can build on a large corpus of previous work about the assessment of digital health technologies and interventions and of AI/ML models. We can profit from experience gained from the evaluation and regulation of software as a medical device (SaMD), and computer-aided detection (CAD) in radiology, as well as from experience gained in comparing different AI/ML models systematically in scientific research. A non-comprehensive collection of best practices, recommendations, insights, and perspectives (from the scientific literature and other documents concerned with the validation and evaluation of health technology and AI/ML models) is presented in clause 10.

Considering this prior knowledge, we should think about characteristics of health AI/ML model validation and evaluation that are novel and unique. Which characteristics have not been dealt with before when assessing other digital health technologies? What do well-established standard assessment methods not capture? How can these gaps be addressed best? Below, we discuss several major points.

Table 1 – Novel or unique aspects of health AI model validation and evaluation encountered during different phases of the evaluation cycle

N°	Aspect (key word; full description below)
1	Appropriate test data sets
2	Internal versus external validation
3	Technical versus clinical criteria
4	Proper benchmarking
5	Complex, opaque deliverable (DL) models and explainable AI (XAI) solutions to make them transparent and interpretable
6	Self-learning algorithms
7	Human factors
8	Misuse of accurate tools
9	Design of clinical trials with AI
10	Agreement on/interoperable encoding of benchmarking cases
11	Testing existing AI products versus ML-challenges

7.1 Appropriate test data sets

The significance of the technical validation of an AI model depends on appropriate test data sets.

a) 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 test sites, presumably due to the domain gap between medical centres 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 choices of test cases. A proper sampling paradigm / scheme (that says we need exactly more of, e.g., "male; 10-15 year-old", "female; 70-80 year-old; 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 correct data. Community efforts to gather standardized test data sets from around the world are a possible solution. This test data set collection could either be organized on a central validation platform or in a federated fashion (see clause 8 below).

b) It can be perceived that "proper sampling" is the solution to collecting the right test data. However, this implies that we need to know the influencing parameters a priori. At the same time, the risk of "unknown unknowns" exists that cannot be taken into account for the sampling strategy. Therefore, we propose a cyclical evaluation process (see Figure 2) where post-deployment monitoring procedures inform the validation and test data sampling scheme.

7.2 Internal versus external validation

In-house technical validation procedures are characterized by a lack of transparency and might suffer from a limited test data coverage. Moreover, the validation results of different AI developers might not be comparable, e.g., due to different collection / curation / selection of the test data or different implementation of the procedure. A possible solution being an external validation (through independent benchmarking by trusted third parties), using standard technical test procedures designed by a multidisciplinary expert team.

7.3 Technical versus clinical criteria

Technical validation criteria for AI models are potentially clinically irrelevant. Setting clinical objectives for the technical assessments (and involving health domain experts in the test design), and the subsequent evaluation with patient outcomes are possible solutions.

7.4 Proper benchmarking

Aiming at becoming close to a technical equivalent to clinical trials, benchmarking challenges / competitions are applied to assess the technical performance of AI algorithms.

NOTE – Challenges can also be seen as collaborative efforts in which researchers work together on the best solution of a specific problem and not only as competitions.

Benchmarking challenges have a very high impact on the research field but there is almost no quality control. On the other hand, clinical trials take time, put test subjects at risk, 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 properly validate the models *in silico* first, check them for different quality criteria, and then follow up with clinical trials. Therefore, the benchmarking design should be peer-reviewed and published to ensure transparency and reproducibility. Standardized guidelines should be integrated as well.

7.5 Complex, opaque DL models and XAI solutions to make them transparent and interpretable

Concerns are raised that the unprecedented model complexity applied in complex settings makes it difficult to assess the AI models. However, first, performance tests (both *in silico* tests and clinical trials) can be conducted for highly complex AI models too. Accordingly, a group of multidisciplinary experts should carefully design appropriate testing procedures with meaningful metrics, in a community effort. Second, methods of explainable AI (XAI) can shed some light on the neural network models' decision-making process to some degree and make them interpretable.

7.6 Self-learning algorithms

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" or "frozen" and not necessarily "self-learning".

If assuming that a self-learning model might also perform worse over time, is 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. Then, in turn, we do not realize the potential for increased accuracy of self-learning systems. Therefore, the patient and healthcare system are not leveraging AI to its potential (*cf.* [b-Gerke et al, 2020]).

Automated pre-assessment via a platform could be a solution to this problem. The benchmarking platform can frequently assess the updated model versions and ensure that the performance on the test data does not deteriorate substantially. This check could support post-market surveillance.

7.7 Human factors

The human factor, as in human in the loop (HIL), needs to be considered in a systemic view (*cf.* [b-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 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, maybe technology-critical radiologists might more often overrule the AI output; that might be correct or wrong). Models tuned for high sensitivity might have too many false positives. Hence, they are ignored by HCPs after a while (considering the models as not trustworthy). Therefore, HIL is an important consideration in the assessment of AI models in clinical care delivery.

7.8 Misuse of accurate tools

Similar to clause 7 if the reimbursement or legal frameworks either prefer or discourage the use of AI, the HCPs could subconsciously be biased to use an accurate tool in the wrong way (training might be needed). As an example, from [b-Gerke et al, 2020] if payers only reimburse if the recommendation is according to the AI system, one gets a very strong emphasis on the AI although the system was designed as a HIL setup.

7.9 Design of clinical trials with AI

There are AI systems (sys_1) that identify patients and design clinical trials. If these trials are meant to assess AI systems (sys_2), then AI is assessing AI. If sys_1 is built on false data, then sys_2 is also erroneous, is it not? 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".

7.10 Agreement on/interoperable encoding of benchmarking cases

The agreement on the benchmarking cases and the interoperable encoding of these benchmarking cases requires careful thought, in particular, if we go beyond the typical image analysis AI tasks. For images, an agreement on the input data format often already exists (e.g., with DICOM), and mapping friction between benchmarking data and the inner workings of the AI model (e.g., ontologies) is usually not to be expected.

7.11 Testing existing AI products versus ML-challenges

Testing existing commercial AI products differs from the setting of "ML-challenges", where the AI models are typically purpose-built for the specific task. Moreover, commercial AI products cannot always be submitted for testing, since they can be IP- and capital-intensive and cannot be easily containerized.

8 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 validation step, pursued by the ITU/WHO Focus Group on "AI for health" (*cf.* Figure 3). This approach could be a valuable complement to in-house or local technical tests and subsequent clinical trials. It does not put test subjects at risk, can be repeated in the case of model / software updates, can be based on large amounts of high-quality test data from different sources and sites, 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 research and development (R&D) community and can serve as an incentive to shift R&D resources in the desired directions.

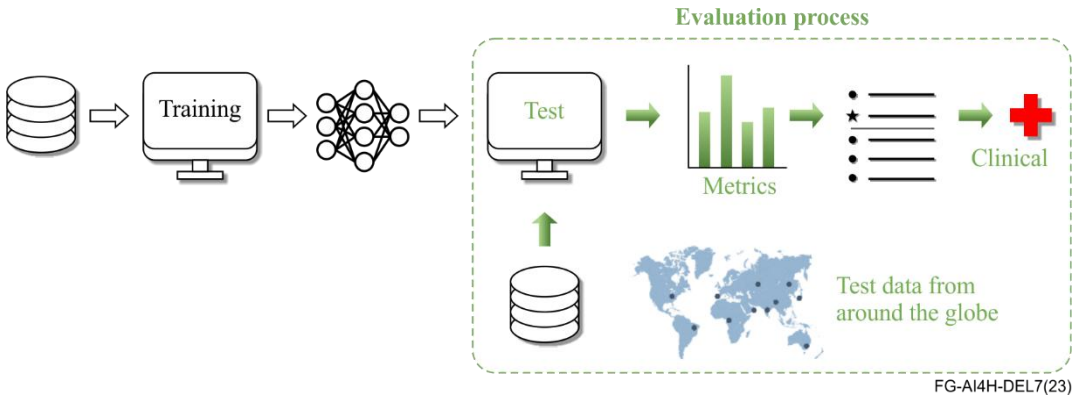

Figure 3 – Independent model validation with standard procedures on separate test data and subsequent clinical evaluation

Standard benchmarking tasks with corresponding standard validation 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 such as malaria detection, symptom checkers, histopathology, ophthalmology, radiology, and many more. The topic groups develop these standards in a dialogue (see Figure 4) with the working groups (that again are writing down their insights and recommendations in their deliverables).

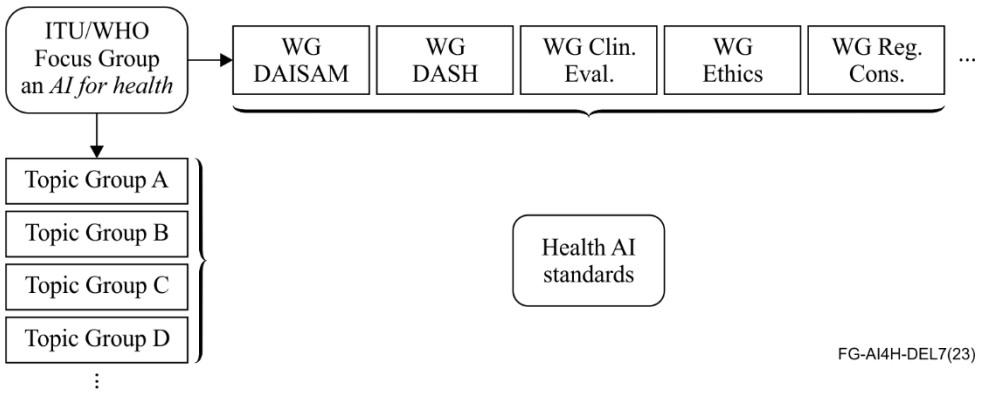

Figure 4 – Structure of the focus group with topic groups and working groups (WGs)

The *topic description documents* [DEL10] 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 the generalizability capabilities of submitted algorithms (e.g., data from another procedure / organ / hospital)? The topic description documents also specify the metrics (accuracy, F1 score, precision, recall, ROC/AUC, Jaccard index, etc.) for every benchmarking task. With these metrics, the predictive performance of a classification or regression model can be assessed. In addition, attribution methods from explainable AI (XAI) can be automatically evaluated *in silico*, e.g., by comparison of the resulting "heatmaps" with ground truth annotations (in the spirit of, e.g., [b-Arras et al., 2019], [b-Arras et al., 2022], [b-Hedström et al., 2022], [b-Saporta et al., 2022] and [b-Samek et al., 2016]). These automatic assessments of XAI methods can be enhanced with complementary user studies investigating user satisfaction with the given explanations.

In the long-term, a clear qualification process is needed for the recruitment of experts that construct the test data and metrics, 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".

8.1 Benchmarking platform with validation in a closed environment

A benchmarking platform for 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 fraud as good as possible.

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

A model that has been over-fitted 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 are 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 validation 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 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 validation results are returned to the AI/ML developer and the benchmark organizer.

Clause 9 explains this concept in detail.

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, kaggle.com, https://eval.ai/, grand-challenge.org, ichallenge.baidu.com, https://www.compression.cc/, codalab.org, https://ramp.studio/, carpl.ai/). Learnings from conducting the aforementioned ML challenges are of paramount importance for the conceived health AI benchmarking platform, in order 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 [b-Reinke et al., 2018], and "why rankings of biomedical image analysis competitions should be interpreted with care" [b-Maier-Hein et al., 2018].

Two modifications of this centralized approach are presented in clauses 8.2 and 8.3.

8.2 Benchmarking platform with validation via interface

The concept for a benchmarking platform with validation *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 is in this case not uploaded to the benchmarking platform but remains on the computer of the AI developer. The model connects to the benchmarking platform via an interface, i.e., over the Internet. The platform sends test data points to the AI model, which processes the data, computes the corresponding output (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 *topic group on symptom assessment* follows this concept and describes it in more detail in their topic description document [DEL10.14].

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 above.

However, this concept requires that new test data are created every time a benchmark is conducted, in order to obtain meaningful results in a fair validation procedure, where cheating is prohibited. 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, for instance, by letting humans label the data by hand and then re-training (i.e., overfitting) the model on these "test" data. Unsupervised approaches (without label information) are conceivable too 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 of the closed environment described earlier. (A possible solution: Generative adversarial networks or other synthesiser algorithms might be able to create 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 validation 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.)

8.3 Federated benchmarking platform

Benchmarking in a *closed environment* can potentially also be conducted in a federated fashion (unlike the centralized concept described in clause 8.1 and without transferring samples such as in clause 8.2).

In the federated approach to benchmarking, the test data sets remain where they had been acquired, for instance, 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 validation *via interface* by:

– Lowering security risks related to data transfer, in that sensitive data remains where the acquisition took place

– Lowering overhead related to curation for transfer, in that the test data is not transferred

– Lowering communication load, in that less data is transferred overall.

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 validation 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 apache 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.* [b-Juuti et al, 2019]). Moreover, the security mechanisms need to assure that all models have processed the very same data. Only then, the results are comparable, and the procedure is trustworthy. Of course, federation does not solve all potential privacy and security issues and will always carry some risks that are similar to the other benchmarking approaches. 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.

9 Requirements of a benchmarking platform with validation in a closed environment

9.1 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 validate 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 validation.

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 that 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 5). 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*. Together with the aforementioned documentation, the participants should be provided a minimal example simulating the execution environment, with the exact syntax that is later used in the validation but using exemplary input data only (and not the actual test data). In this way, the AI developers can check themselves whether their implementation works because it adheres to the standardized interface, which would greatly ease any kind of troubleshooting.

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). Further, all access and user rights of the aforementioned usage scenarios require authentication and authorization via an overarching policy management system. This guarantees traceability of the process and informs administrators of the platform, where which AI developer has requested benchmarking with a certain frequency. The following subclauses provide detailed information about each of these components.

9.1.1 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. The overall architecture follows a best-practise modular paradigm (rather than monolithic, hard-coded solution). This allows more efficient error handling, easier maintenance and expandability for future, novel benchmark scenarios.

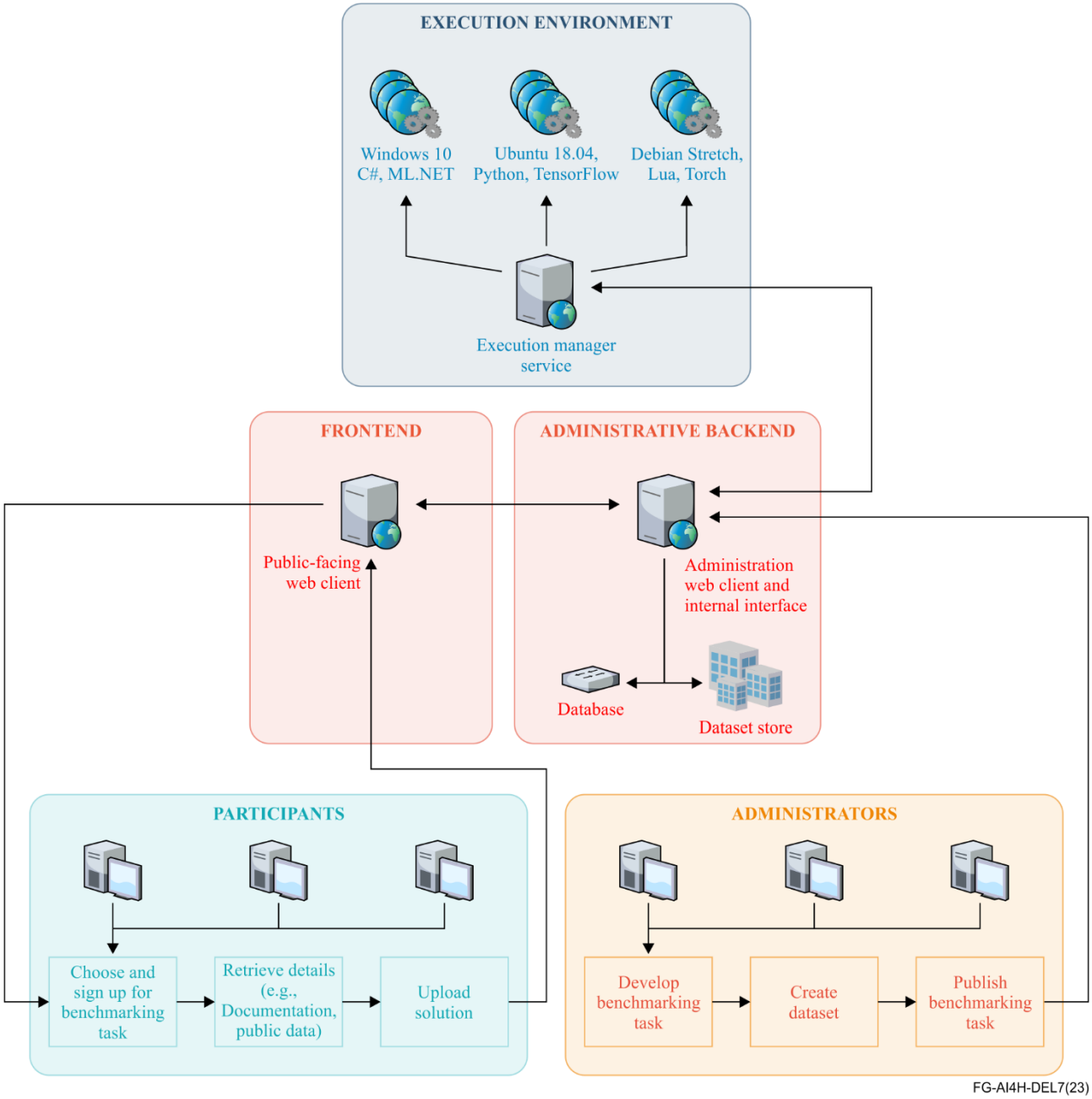

Figure 5 – Overview of the architecture of the system

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 validate the submitted solutions. It represents the "gold standard" with the included true labels or annotations as "ground truth". In contrast, public data (if applicable) 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 the model 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* (viewing can be with or without authorization).

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 access 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 catalogue 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 a 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 needs to be stored and 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 and version control are 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" must 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.

9.1.2 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 (if available). 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 validation 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.

9.1.3 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 as well as the calculation of the benchmarking metrics, it will report the results back to the *execution manager service*, which will in turn 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 (ZIP, TAR GZIP etc.), which contains all files necessary to execute the solution. The package should contain a top-level executable (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 the 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 (Docker, Mesos, Singularity/Apptainer etc.) or through virtualization (VMWare, Virtual Box, HyperV, etc.). 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 20.04.6 LTS, 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 must be strictly withheld from the AI model that must have access only to the unlabelled / unannotated data points (i.e., "features"). 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. A strictly separate functionality, that the to-be-tested AI model cannot interfere with (!), then calculates the benchmarking metrics by comparing the reported results with the true labels or annotations using the respective statistical metrics. These metrics need to be specified for each benchmarking task (accuracy, F1 score, precision, recall, ROC/AuC, Jaccard index, etc.). These results are reported back to the *execution manager service.*

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 validation 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.

9.2 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.

9.2.1 Security

Test data and submitted AI solutions to-be-validated 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.

9.2.2 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 to 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 seems to be 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. Nevertheless, managing a custom data centre takes a lot of effort. The alternative is to run the system in the public cloud, for instance 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 centre, an adequate number of servers must be managed and kept available at all times while from a cloud service, virtual machines may be rented by the minute.

9.2.3 Computing resources

The computing resources required for the benchmarking platform depend on the number and the difficulty of the offered benchmarking tasks and on the quantity and properties of the future AI submissions and test data. 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.

9.2.4 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, including 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.

10 Best practices for the testing, validation, benchmarking and evaluation of health AI/ML models from the scientific literature and other documents

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

Boldface type highlights text passages that were considered as particularly relevant for the evaluation of AI for health (boldface in this Technical Report; not in the original source).

Digital health: a path to validation.
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*. [b-Mathews et al, 2019]. 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)."*

Comments:

- Especially interesting is Figure 2 where the proposed "independent evaluator" matches well with the independent benchmarking proposed in the present report [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 clause 7 in the present report [DEL07].

Consort-AI and Spirit-AI steering group: Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed.

Liu, X., Rivera, S. C., Faes, L., di Ruffano, L. F., Yao, C., Keane, P. A., Ashrafian, H., Darzi, A., Vollmer, S. J., Deeks, J., Bachmann, L., Holmes, C., Chan, A. W., Moher, D., Calvert, M. J., and Denniston, A, K. (2019), Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nature medicine 25*, 1467–1468. [b-Liu et al., 2019]. https://doi.org/10.1038/s41591-019-0603-3

(Additional publication: Liu, X., Faes, L., Calvert, M. J., and Denniston, A. K. (2019). Extension of the CONSORT and SPIRIT statements. *The lancet*, *394*(10205), 1225. [b-Liu X. et al, 2019]. https://doi.org/10.1016/S0140-6736(19)31819-7)

"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 outcomes 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 development. 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 validation."*

Comment:

– The independent benchmarking proposed in the present report [DEL07] is complementary to CONSORT-AI, SPIRIT-AI and TRIPOD-ML and provides chances for thorough technical tests prior to clinical trials.

Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension.

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., and The SPIRIT-AI and CONSORT-AI Working Group. (2020), Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* **26,** 1364–1374. [b-Liu et al., 2020]. https://doi.org/10.1038/s41591-020-1034-x

"The CONSORT-AI (Consolidated standards of reporting trials–artificial intelligence) extension is a new reporting guideline for clinical trials evaluating interventions with an AI component. It was developed in parallel with its companion statement for clinical trial protocols: SPIRIT-AI (standard protocol items: Recommendations for interventional trials–artificial intelligence). [...] The CONSORT-AI extension includes 14 new items that were considered sufficiently important for AI interventions that they should be routinely reported in addition to the core CONSORT 2010 items. CONSORT-AI recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human–AI interaction and provision of an analysis of error cases. [...] Randomized controlled trials (RCTs) are considered the gold-standard experimental design for providing evidence of the safety and efficacy of an intervention. Trial results, if adequately reported, have the potential to inform regulatory decisions, clinical guidelines and health policy. It is therefore crucial that RCTs are reported with transparency and completeness so that readers can critically appraise the trial methods and findings and assess the presence of bias in the results. [...] However, in the most recent cases, published evidence has consisted of in silico, early-phase validation."

Comment

– See Consort-AI and Spirit-AI (2019) above.

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension.

Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nature medicine **26,** 1351–1363 (2020). [b-Rivera et al., 2020]. https://doi.org/10.1038/s41591-020-1037-7

"The SPIRIT-AI extension includes 15 new items that were considered sufficiently important for clinical trial protocols of AI interventions. These new items should be routinely reported in addition to the core SPIRIT 2013 items. SPIRIT-AI recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention will be integrated, considerations for the handling of input and output data, the human–AI interaction and analysis of error cases."

TRIPOD-ML: Reporting of artificial intelligence prediction models.
Collins, G. S., and Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. *The lancet*, Volume *393*, Issue 10181, pp1577-1579. [b-Collins et al., 2019]. https://doi.org/10.1016/S0140-6736(19)30037-6
Comment: – This is a comment about the reporting mechanisms for AI prediction models.

Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.
Collins, G. S., Reitsma, J. B., Altman, D. G., Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. [b-Collins G. S. et al., 2015]. PMID: 25569120
*"Reporting of studies developing, **validating**, or updating a **prediction model**, whether for diagnostic or prognostic purposes."* [b-Collins G. S. et al., 2015]. (Quote from https://www.equator-network.org/reporting-guidelines/tripod-statement/)
Comment: – The basis for the extension towards machine learning "TRIPOD-ML", cited above.

A machine learning process model with quality assurance methodology.
Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and Müller, K. R. (2020). Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology. *arXiv preprint arXiv:2003.05155*. [b-Studer et al., 2020]. 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 analyse 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."*

Comments:

– Excellent summary concerning ML model evaluation methods.

– This cross-industry perspective could be adapted to the health domain.

– Clinical evaluation needs to be added / integrated in this technical perspective.

Machine learning testing: Survey, landscapes and horizons.

Zhang, J. M., Harman, M., Ma, L., and Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE transactions on software engineering*. [b-Zhang et al., 2020]. 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)."

Comments:

– Valuable recent overview of the state of the art of machine learning model testing.

– The report is a general survey. Hence, the specifics for the health/medicine domain could deserve additional attention, e.g., the clinical perspective (Do the tests 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. [b-IMDRF] 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."*

Comments:

– In-house validation by the manufacturer is not as trustworthy as the external validation by an independent third party. Expert knowledge and large volumes 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 the training data. Hence, the 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) [b-Regulation EU] 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****. […]"*

Comment:

– The regulatory perspective is discussed in more detail in the *AI4H regulatory considerations* [DEL02].

BIAS: Transparent reporting of biomedical image analysis challenges.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A.,

Landman, B. A. (2020), BIAS: Transparent reporting of biomedical image analysis challenges. *Medical image analysis*, Volume 66, 101796. [b-Maier-Hein et al., 2020]. Doi: https://doi.org/10.1016/j.media.2020.101796

Also 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**."*

Comment:

– Standardized guideline for challenge design.

Benchmarking visualization toolkit.

Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A. (2021). Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific reports*. [b-Wiesenfarth et al., 2021]. Doi: https://doi.org/10.1038/s41598-021-82017-6

*"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**."*

Comment:

– Could be valuable for the benchmarking platform as well in order to visualize the results.

Causality matters in medical imaging.

Castro, D. C., Walker, I. and Glocker, B. (2020). Causality matters in medical imaging. *Nature communications 11*, 3673. [b-Castro et al., 2020]. https://doi.org/10.1038/s41467-020-17478-w

*"Causal reasoning can shed new light on the major challenges in machine learning for medical imaging: scarcity of high-quality annotated data and **mismatch between the development dataset and the target environment**. A causal perspective on these issues allows decisions about data collection, annotation, preprocessing, and learning strategies to be made and scrutinized more*

transparently, while providing a detailed categorisation of potential biases and mitigation techniques. Along with worked clinical examples, we highlight the importance of establishing the causal relationship between images and their annotations, and offer step-by-step recommendations for future studies.

(...) Importantly, the ***assumption that the performance of a trained model on the development test set is representative of the performance on new clinical data after deployment in varying environments is often violated due to differences in data characteristics****, as discussed earlier. It is therefore absolutely critical to be able to clearly formalise and communicate the underlying assumptions regarding the data-generating processes in the lab and real-world environments, which in turn can help anticipate and mitigate failure modes of the predictive system.*

(...) ***The recurrent issue of mismatch between data distributions, typically between training and test sets or development and deployment environments, tends to hurt the generalisability of learned models****. In the generic case when no assumptions can be made about the nature of these differences, any form of learning from the training set is arguably pointless, as the test-time performance can be arbitrarily poor. Nonetheless, causal reasoning enables us to recognise special situations in which direct generalisation is possible, and to devise principled strategies to mitigate estimation biases. In particular, two distinct mechanisms of distributional mismatch can be identified:* ***dataset shift*** *and* ***sample selection bias****. Learning about their differences is helpful for diagnosing when such situations arise in practice."*

Comment:

– From a causality point of view, the authors analyse the relationship between images and annotations as well as the generalisation from training data to test data, which is of paramount importance for the design of the benchmarking procedure discussed in this deliverable [DEL07].

MINIMAR (MINimum information for medical AI reporting).

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., & Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association, 27(12). [b-Hernandez-B. et at., 2020]. https://doi.org/10.1093/jamia/ocaa088

"[...] describing the minimum information necessary to understand intended predictions, target populations, and hidden biases, and the ability to generalize these emerging technologies. We call for a standard to accurately and responsibly report on AI in health care. This will facilitate the design and implementation of these models and promote the development and use of associated clinical decision support tools, as well as manage concerns regarding accuracy and bias."

"MINIMAR will also promote external validation, encouraging the use of secondary resources."

"Model evaluation strategies should be defined in detail, in terms of data used for both internal and external validation as well as the adopted approach adopted for evaluation (e.g., 5-fold cross-validation or 80/20 split). The choice of validation metrics, such as sensitivity, specificity, positive predictive value, or area under the receiver-operating characteristic curve, also needs to be defined. [...] Finally, as part of model evaluation, transparency is necessary for broad AI application in health care in order to achieve and retain confidence and trust from all the stakeholders. Indeed, recent studies show an alarming difficulty in reproducing models developed in research studies and suggest that even if the training data cannot be shared due to privacy issues, the source code of the model should be shared publicly.26 Therefore, in order to demonstrate the provenance and authenticity of the data and knowledge used to make decisions by AI models, promoting access to training data and source code is crucial to ensure that ML in biomedicine can

be broadly applied and generalized. This is essential not only for choosing the best model for the given setting, but also for the unbiased comparison of different models or different settings."

Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., ... & Obermeyer, Z. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature medicine*, **26**(9), 1320-1324. [b-Norgeot]. https://www.nature.com/articles/s41591-020-1041-y

"MI-CLAIM checklist, a tool intended to improve transparent reporting of AI algorithms in medicine."

Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers.

Mongan, J., Moy, L., and Kahn Jr, C. E. (2020). Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology: artificial intelligence, Volume 2, No.2. [b-Mongan et al., 2020]. https://doi.org/10.1148/ryai.2020200029

"To aid authors and reviewers of AI manuscripts in medical imaging, CLAIM, the checklist for AI in medical imaging is proposed. CLAIM is modeled after the STARD guideline and has been extended to address applications of AI in medical imaging that include classification, image reconstruction, text analysis, and workflow optimization. The elements described here should be viewed as a "best practice" to guide authors in presenting their research."

STARD-AI (Standards for reporting of diagnostic accuracy studies - AI).

Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A. K., Greaves, F., ., Karthikesalingam, A., King, D., Liu, X., Markar, S. R., Mclnnes, M. D. F., Panch, T., Pearson-Stuttard, J., Ting, D. S. W., Golub, R. M., Moher, D., Bossuyt, P. M., and Darzi, A. (2020). Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI steering group. Nature medicine, 1-2. [b-Sounderajah et al., 2020]. https://doi.org/10.1038/s41591-020-0941-1

"The STARD (Standards for reporting of diagnostic accuracy studies) 2015 statement remains the most widely accepted set of reporting standards for diagnostic accuracy studies. In particular, STARD was developed to improve the completeness and transparency of studies investigating diagnostic accuracy. However, STARD was not designed to address the issues and challenges raised by AI-driven modalities. Such issues include unclear methodological interpretation (e.g., the use of external validation datasets, complexities of datasets and comparison to human performance) and the lack of standardized nomenclature (e.g., the definition of a 'validation dataset'), as well as the heterogeneity of outcome measures (e.g., area under the receiver operating characteristics (AUROC), sensitivity, positive predictive value and F1 score). Until these issues are overcome at a validation stage, downstream evaluation of these technologies and their potential real-word benefits will remain limited. Journal editors have also commented that approximately 25% of all manuscript submissions in leading journals now center on the diagnostic accuracy of AI algorithms. In summation, there is a clear multi-faceted need to establish guidelines on the conduct and reporting of such projects. In order to tackle these problems, the STARD-AI Steering Group is preparing an AI-specific extension to the STARD statement (STARD-AI) that aims to focus upon the specific reporting of AI diagnostic accuracy studies."

Trusted artificial intelligence: towards certification of machine learning applications.

Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., Hochreiter, S., and Nessler, B. (2021). Trusted artificial intelligence: towards certification of machine learning applications. arXiv preprint arXiv:2103.16910. [b-Winter et al., 2021]. https://arxiv.org/abs/2103.16910v1

"TÜV AUSTRIA group in cooperation with the institute for machine learning at the Johannes Kepler University Linz, proposes a ***certification process and an audit catalog for machine learning applications****."*

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals.

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nature medicine 27, 582-584. [b-Wu et al., 2021]. https://doi.org/10.1038/s41591-021-01312-x

"A comprehensive overview of medical AI devices approved by the US Food and drug administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients."

Comment:

– Summary: Multi-site evaluation of sufficient size and quality is recommended, because single-site models often do not generalize.

Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan.

FDA (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. [b-FDA]. https://www.fda.gov/media/145022/download

"What type of reference data are appropriate to utilize in measuring the performance of AI/ML software devices in the field? How much of the oversight should be performed by each stakeholder? How much data should be provided to the Agency, and how often? How can the algorithms, models, and claims be validated and tested?"

"Strengthen FDA's encouragement of the harmonized development of Good Machine Learning Practice (GMLP) through additional FDA participation in collaborative communities and consensus standards development efforts."

"Support regulatory science efforts on the development of methodology for the evaluation and improvement of machine learning algorithms, including for the identification and elimination of bias, and on the robustness and resilience of these algorithms to withstand changing clinical inputs and conditions."

Common limitations of image processing metrics.

Reinke, A., Tizabi, M. D., Sudre, C. H., Eisenmann, M., Rädsch, Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Blaschko, M., Büttner, Cardosa, M. J., Chen, J., Cheplygina, V., Christodoulou, E., Cimini, B., Collins, G. S., Engelhardt, S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Glocker, B., Godau, P., Haase, R., Hamprecht, F., Hashimoto, D. A., Heckmann-Nötzel, D., Hirsch, P., Hoffman, M. M., Huisman, M., Isensee, F., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kavur, A. E., Kenngott, H., Kleesiek, J., Kleppe, A., Kohler, S., Kofler, F., Kopp-Schneider, A., Kooi, T., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson. P., Meijering, E., Menze, B., Moher, D., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Noyan, M. A., Petersen, J., Polat, G., Rafelski, S. M., Rajpoot, N., Reyes, M., Rieke, N., Riegler, M., Rivaz, H., Saez-Rodriguez, J., Sánchez, C. I., Schroeter, J., Saha, A., Selver, M. A., Sharan, L., Shetty, S., van Smeden, M., Stieltjes, B., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Calster, B. V., Varoquaux, G., Wiesenfarth, M., Yaniv, Z. R., Jäger, P., and Maier-Hein, L (2021). Common limitations of image processing metrics: a picture story. [b-Reinke et al., 2021]. https://arxiv.org/abs/2104.05642

"While the importance of automatic image analysis is increasing at an enormous pace, recent meta-research revealed major flaws with respect to algorithm validation. Specifically, performance metrics are key for objective, transparent and comparative performance assessment, but relatively little attention has been given to the practical pitfalls when using specific metrics for a given image analysis task. A common mission of several international initiatives is therefore to provide researchers with guidelines and tools to choose the performance metrics in a problem-aware manner. This dynamically updated document has the purpose to illustrate important limitations of performance metrics commonly applied in the field of image analysis. The current version is based on a delphi process on metrics conducted by an international consortium of image analysis experts."

DECIDE-AI.

[b-DECIDE-AI steering group] (2021). DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature medicine.* **27** pp186-187. https://doi.org/10.1038/s41591-021-01229-5

"Delphi process to reach expert consensus on the key information items that should be reported during 'developmental and exploratory clinical investigation of DEcision-support systems driven by artificial intelligence' (DECIDE-AI)."

A governance model for the application of AI in health care.

Reddy, S., Allan, S., Coghlan, S., Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the american medical informatics association*. Volume 27(3):491-497. [b-Reddy et al., 2020]. https://doi.org/10.1093/jamia/ocz192

"A governance model that aims to not only address the ethical and regulatory issues that arise out of the application of AI in health care, but also stimulate further discussion about governance of AI in health care."

Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues.

Rogers, W. A., Draper, H., Carter, S. M. (2021). Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. *Bioethics*. [b-Rogers et al., 2021]. https://doi.org/10.1111/bioe.12885

"Our ethical evaluation identifies context- and case-specific healthcare ethical issues for two applications, and investigates the extent to which the general ethical principles for AI-assisted healthcare expressed in existing frameworks capture what is most ethically relevant from the perspective of healthcare ethics. We provide a detailed description and analysis of two AI-based systems for clinical decision support [...]"

"[...] our analysis points to the close connection between ethical evaluation of healthcare and technical reporting. Details about the training set, the way the algorithm is constructed, sensitivity and specificity, data storage and so forth are essential for making ethical evaluations."

"[...] ethicists must contribute to multidisciplinary evaluations of healthcare AIs, to ensure accuracy in interpreting the ethical implications of technical specifications."

"Our paper provides a potential blueprint for further use-case analyses. It is a contribution towards developing a robust ethical evaluation of healthcare AI that can be integrated with other appraisal tools."

ECLAIR guidelines.

Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn Jr., C. E., Louvet-de Verchère, F., Santos, D. P. D., Kober, T., and Richiardi, J. (2021). To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *European radiology* **31,** 3786–3796. [b-Omoumi et al. 2021]. https://doi.org/10.1007/s00330-020-07684-x

"While several guidelines describing good practices for conducting and reporting AI-based research in medicine and radiology have been published, fewer efforts have focused on recommendations addressing the key questions to consider when critically assessing AI solutions before purchase. Commercial AI solutions are typically complicated software products, for the evaluation of which many factors are to be considered. In this work, authors from academia and industry have joined efforts to propose a practical framework that will help stakeholders evaluate commercial AI solutions in radiology (the ECLAIR guidelines) and reach an informed decision. Topics to consider in the evaluation include the relevance of the solution from the point of view of each stakeholder, issues regarding performance and validation, usability and integration, regulatory and legal aspects, and financial and support services."

Clinician checklist for assessing suitability of machine learning applications in healthcare.

Scott, I., Carter, S., Coiera, E. (2020). Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ health & care informatics Volume 28. [b-Scott et al., 2020]. http://dx.doi.org/10.1136/bmjhci-2020-100251

"Hundreds of new algorithms are being developed, but whether they improve clinical decision making and patient outcomes remains uncertain. If clinicians are to use algorithms, they need to be reassured that key issues relating to their validity, utility, feasibility, safety and ethical use have been addressed. We propose a checklist of 10 questions that clinicians can ask of those advocating for the use of a particular algorithm, but which do not expect clinicians, as non-experts, to demonstrate mastery over what can be highly complex statistical and computational concepts. The questions are: (1) What is the purpose and context of the algorithm? (2) How good were the data

used to train the algorithm? (3) Were there sufficient data to train the algorithm? (4) How well does the algorithm perform? (5) Is the algorithm transferable to new clinical settings? (6) Are the outputs of the algorithm clinically intelligible? (7) How will this algorithm fit into and complement current workflows? (8) Has use of the algorithm been shown to improve patient care and outcomes? (9) Could the algorithm cause patient harm? and (10) Does use of the algorithm raise ethical, legal or social concerns? We provide examples where an algorithm may raise concerns and apply the checklist to a recent review of diagnostic imaging applications. This checklist aims to assist clinicians in assessing algorithm readiness for routine care and identify situations where further refinement and evaluation is required prior to large-scale use."

Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning (DL) studies.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., Maruthappu, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**:m689. [b-Nagendran et al., 2020]. http://dx.doi.org/10.1136/bmj.m689

"Adherence to reporting standards was assessed by using CONSORT (consolidated standards of reporting trials) for randomised studies and TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) for nonrandomised studies. Risk of bias was assessed by using the Cochrane risk of bias tool for randomised studies and PROBAST (prediction model risk of bias assessment tool) for non-randomised studies."

"Few prospective deep learning studies and randomised trials exist in medical imaging. Most nonrandomised trials are not prospective, are at high risk of bias, and deviate from existing reporting standards. Data and code availability are lacking in most studies, and human comparator groups are often small. Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions."

Evaluating artificial intelligence in medicine: phases of clinical research.

Park, Y., Jackson, G. P., Foreman, M. A., Gruen, D., Hu, J., Das, A. K. (2020). Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA open Volume 3(3) pp326-331. [b-Park et al., 2020]. http://dx.doi.org/10.1093/jamiaopen/ooaa033

"Increased scrutiny of artificial intelligence (AI) applications in healthcare highlights the need for real-world evaluations for effectiveness and unintended consequences. The complexity of healthcare, compounded by the user- and context-dependent nature of AI applications, calls for a multifaceted approach beyond traditional in silico evaluation of AI. We propose an interdisciplinary, phased research framework for evaluation of AI implementations in healthcare. We draw analogies to and highlight differences from the clinical trial phases for drugs and medical devices, and we present study design and methodological guidance for each stage."

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., and Schönlieb, C-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature machine intelligence* **3,** 199–217. [b-Roberts et al., 2021]. https://doi.org/10.1038/s42256-021-00307-0

"Many papers (16/62) used the pneumonia dataset of Kermany et al. as a control group. They commonly failed to mention that this consists of paediatric patients aged between one and five. Developing a model using adult patients with COVID-19 and very young patients with pneumonia is likely to overperform as it is merely detecting children versus adults."

Comment:

– The report deals with biases and confounders in the data and shows how severe this may be.

Evaluation framework to guide implementation of AI systems into healthcare settings.

Reddy, S., Rogers, W., Makinen, V-P., Coiera, E., Brown, P., Wenzel, M., Weicken, E., Ansari, S., Mathur, P., Casey, A., and Kelly B. (2021). Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health & care informatics*, Volume 28(1). [b-Reddy S. et al 2021]. http://dx.doi.org/10.1136/bmjhci-2021-100444

"[…] many artificial intelligence (AI) systems have been developed in healthcare, but adoption has been limited. This may be due to inappropriate or incomplete evaluation and a lack of internationally recognised AI standards on evaluation. To have confidence in the generalisability of AI systems in healthcare and to enable their integration into workflows, there is a need for a practical yet comprehensive instrument to assess the translational aspects of the available AI systems. Currently available evaluation frameworks for AI in healthcare focus on the reporting and regulatory aspects but have little guidance regarding assessment of the translational aspects of the AI systems like the functional, utility and ethical components."

"To address this gap and create a framework that assesses real-world systems, an international team has developed a translationally focused evaluation framework termed 'Translational evaluation of healthcare AI (TEHAI)'. A critical review of literature assessed existing evaluation and reporting frameworks and gaps. Next, using health technology evaluation and translational principles, reporting components were identified for consideration. These were independently reviewed for consensus inclusion in a final framework by an international panel of eight experts."

"TEHAI includes three main components: capability, utility and adoption. The emphasis on translational and ethical features of the model development and deployment distinguishes TEHAI from other evaluation instruments. In specific, the evaluation components can be applied at any stage of the development and deployment of the AI system."

Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation.

World Health Organization. (2021). Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation. World Health Organization. [b-WHO 2021]. https://apps.who.int/iris/handle/10665/349093. License: CC BY-NC-SA 3.0 IGO

From the executive summary: *"The development of artificial intelligence and machine learning based software as a medical device (we will refer to these as AI-SaMD in this document) is rapidly evolving. However, there is currently a lack of globally recognised benchmarking frameworks to assess evidence generated by the use of these devices. International regulatory frameworks for digital health products are also evolving. The framework provides an overview of considerations used in evaluating clinical evidence regarding AI-SaMD, aiming to help formulate a consensus for guiding validation, evidence generation and reporting across the total product life-cycle within a global health context."*

"As well as reviewing the current literature, this framework provides a real-world example in the form of a use-case for AI-SaMD applied to a WHO priority: cervical cancer screening. (Use-case methodology is a systems analysis approach to clarifying product requirements, describing the complete sequence of steps required to reach a specified goal). Several chapters also list minimum standards for different aspects of evidence generation."

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review.

de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L. Kant, I. M. J., Nijman, S. W. J., van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., van Smeden, M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., and Moons, K. G. M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital medicine* **5,** 2. [b-de Hond et al., 2022]. https://doi.org/10.1038/s41746-021-00549-7

From the abstract: *"While the opportunities of ML and AI in healthcare are promising, the growth of complex data-driven prediction models requires careful quality and applicability assessment before they are applied and disseminated in daily practice. This scoping review aimed to identify actionable guidance for those closely involved in AI-based prediction model (AIPM) development, evaluation and implementation including software engineers, data scientists, and healthcare professionals and to identify potential gaps in this guidance. We performed a scoping review of the relevant literature providing guidance or quality criteria regarding the development, evaluation, and implementation of AIPMs using a comprehensive multi-stage screening strategy."*

Algorithmic impact assessment: a case study in healthcare

[b-Ada Lovelace Institute] (2022). Algorithmic impact assessment: a case study in healthcare. [Online:] https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/

"Governments, public bodies and developers of artificial intelligence (AI) systems are becoming interested in algorithmic impact assessments (referred to throughout this report as 'AIAs') as a means to create better understanding of and accountability for potential benefits and harms from AI systems. At the same time – as a rapidly growing area of AI research and application – healthcare is recognised as a domain where AI has the potential to bring significant benefits, albeit with wide-ranging implications for people and society. This report offers the first-known detailed proposal for the use of an algorithmic impact assessment for data access in a healthcare context – the UK national health service (NHS)'s proposed national medical imaging platform (NMIP). It includes actionable steps for the AIA process, alongside more general considerations for the use of AIAs in other public and private-sector contexts."

The medical algorithmic audit
Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., and Oakden-Rayner, L. (2022). The medical algorithmic audit. *The lancet digital health*. [b-Liu et al., 2022]. https://doi.org/10.1016/S2589-7500(22)00003-6
"Artificial intelligence systems for health care, like any other medical device, have the potential to fail. However, specific qualities of artificial intelligence systems, such as the tendency to learn spurious correlates in training data, poor generalisability to new deployment settings, and a paucity of reliable explainability mechanisms, mean they can yield unpredictable errors that might be entirely missed without proactive investigation. We propose a medical algorithmic audit framework that guides the auditor through a process of considering potential algorithmic errors in the context of a clinical task, mapping the components that might contribute to the occurrence of errors, and anticipating their potential consequences. We suggest several approaches for testing algorithmic errors, including exploratory error analysis, subgroup testing, and adversarial testing, and provide examples from our own work and previous studies. The medical algorithmic audit is a tool that can be used to better understand the weaknesses of an artificial intelligence system and put in place mechanisms to mitigate their impact."

The importance of being external. methodological insights for the external validation of machine learning models in medicine
Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L. G., Challa, F., Sulejmani, A., Seghezzi, M., and Carobene, A. (2021). The importance of being external. Methodological insights for the external validation of machine learning models in medicine. *Computer methods and programs in biomedicine*, Volume 208, 106288. [b-Cabitza et al., 2021]. https://doi.org/10.1016/j.cmpb.2021.106288
"Background and objective: Medical machine learning (ML) models tend to perform better on data from the same cohort than on new data, often due to overfitting, or co-variate shifts. For these reasons, external validation (EV) is a necessary practice in the evaluation of medical ML. However, there is still a gap in the literature on how to interpret EV results and hence assess the robustness of ML models. *Methods: We fill this gap by proposing a meta-validation method, to assess the soundness of EV procedures. In doing so, we complement the usual way to assess EV by considering both dataset cardinality, and the similarity of the EV dataset with respect to the training set. We then investigate how the notions of cardinality and similarity can be used to inform on the reliability of a validation procedure, by integrating them into two summative data visualizations. [...]* *Conclusions: In this paper, we propose a novel, lean methodology to:1) study how the similarity between training and validation sets impacts the generalizability of an ML model; 2) assess the soundness of EV evaluations along three complementary performance dimensions: discrimination, utility and calibration; 3) draw conclusions on the robustness of the model under validation. We applied this methodology to a state-of-the-art model for the diagnosis of COVID-19 from routine blood tests and showed how to interpret the results in light of the presented framework."*

Many more publications can be added to this non-comprehensive collection of best practices. Examples include [b-Dustler 2020], [b-Kim et al., 2020], [b-McKinney at al., 2020], [b-Talmon et al, 2009], (The English National Institute for Health and Care Excellence, 2019 [b-NICE]), [b-Van Calster et al., 2019], [b-Gerke et al., 2020] [b-Reinke et al., 2022], [b-Brundage et al., 2020], [b-Ada Lovelace Institute, 2020], [b-McGregor, 2020], and [b-Mincu et al, 2022].

Bibliography

[b-Ada Lovelace Institute, 2020] Ada Lovelace Institute (2020), *Examining the Black Box: Tools for assessing algorithmic systems*. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

[b-Ada Lovelace Institute] Ada Lovelace Institute (2022), *Algorithmic impact assessment: a case study in healthcare*. [Online:] https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/

[b-Arras et al., 2019] Arras, L., Osman, A., Müller, K-R., and Samek, W. (2019), *Evaluating recurrent neural network explanations*. arXiv preprint arXiv:1904.11829. https://arxiv.org/abs/1904.11829

[b-Arras et al., 2022] Arras, L., Osman, A., and Samek, W. (2022), *CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations*. Information Fusion, Volume 81, 14-40. https://doi.org/10.1016/j.inffus.2021.11.008

[b-Brundage et al., 2020] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Hass, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S. Ó., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T. K., Dyer, L., Khan, S., Bengio, Y., and Anderljung, M. (2020), *Toward trustworthy AI development: mechanisms for supporting verifiable claims*. arXiv preprint arXiv:2004.07213. https://arxiv.org/abs/2004.07213

[b-Cabitza et al., 2021] Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L. G., Challa, F., Sulejmani, A., Seghezzi, M., Carobene, A. (2021), *The importance of being external. methodological insights for the external validation of machine learning models in medicine*. Computer methods and programs in biomedicine, Volume s208, 106288. https://doi.org/10.1016/j.cmpb.2021.106288

[b-Collins G. S. et al., 2015] Collins, G. S., Reitsma, J. B., Altman, D. G., Moons, K. G. M. (2015), *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement*. BMJ. https://pubmed.ncbi.nlm.nih.gov/25569120/

[b-Collins et al., 2019] Collins, G. S., and Moons, K. G. M. (2019), *Reporting of artificial intelligence prediction models*. The Lancet, Volume 393, Issue 10181, pp1577-1579. https://doi.org/10.1016/S0140-6736(19)30037-6

[b-Castro et al., 2020] Castro, D. C., Walker, I., and Glocker, B. (2020), *Causality matters in medical imaging*. Nature Communications. 11, 3673. https://doi.org/10.1038/s41467-020-17478-w

[b-DECIDE-AI steering group] DECIDE-AI Steering Group. (2021), *DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence*. Nature Medicine. **27** pp. 186-187. https://doi.org/10.1038/s41591-021-01229-5

[b-Dustler 2020] 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

[b-de Hond et al., 2022] de Hond, A. A. H., Leeuwenberg, A. M., Hooft, L. Kant, I. M. J., Nijman, S. W. J., van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., van Smeden, M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., and Moons, K. G. M. (2022), *Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review*. npj Digital Medicine. **5,** 2.

https://doi.org/10.1038/s41746-021-00549-7

[b-EU Regulation] 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

[b-FDA] FDA (2021), *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. https://www.fda.gov/media/145022/download

[b-Gerke et al., 2020] Gerke, S., Babic, B., Evgeniou, T. and Cohen, G. I. (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

[b-Hedström et al., 2022] Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M-C. (2022), *Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond*. arXiv preprint arXiv:2202.06861.
https://arxiv.org/abs/2202.06861

[b-Hernandez-B. et at., 2020] Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A., and Shah, N. H. (2020), *MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care*. Journal of the American Medical Informatics Association, 27(12). https://doi.org/10.1093/jamia/ocaa088

[b-IMDRF] IMDRF/SaMD WG/N41FINAL:2017, *Software as a Medical Device (SaMD): Clinical Evaluation*. Software as a Medical Device Working Group. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-170921-samd-n41-clinical-evaluation_1.pdf

[b-Juuti et al, 2019] Juuti, M., Szyller, S., Marchal, S., and Asokan, N. (2019), *PRADA: Protecting against DNN model stealing attacks*. 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE.
https://doi.org/10.1109/EuroSP.2019.00044

[b-Kim et al., 2020] Kim, H. E., Kim, H. H., Han, B. K., Kim, K. H., Han, K., Nam, H., Lee, E. H., and 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

[b-Liu X. et al, 2019] Liu, X., Faes, L., Calvert, M. J., and 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

[b-Liu et al., 2019] Liu, X., Rivera, S.C., Faes, L., di Ruffano, L. F., Yao, C., Keane, P. A., Ashrafian, H., Darzi, A., Vollmer, S. J., Deeks, J., Bachmann, L., Holmes, C., Chan, A. W., Moher, D., Calvert, M. J., and Denniston, A, K. (2019), *Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed*. Nature Medicine *25*, 1467–1468 (2019). https://doi.org/10.1038/s41591-019-0603-3

[b-Liu et al., 2020] Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K. and The SPIRIT-AI and CONSORT-AI Working Group. (2020), *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. BMJ. https://www.nature.com/articles/s41591-020-1034-x

[b-Liu et al., 2022] Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., and Oakden-Rayner, L. (2022), *The medical algorithmic audit*. The Lancet Digital Health. https://doi.org/10.1016/S2589-7500(22)00003-6

[b-Maier-Hein et al., 2018] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S. Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., Ginneken, B. V., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B., H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., Sommen, F. V. D., Wang, C-W., Weber, M-A., Zheng, G., Jannin, P., and Kopp-Schneider, A. (2018), *Why rankings of biomedical image analysis competitions should be interpreted with care*. Nature communications, 9, 5217. https://doi.org/10.1038/s41467-018-07619-7

[b-Maier-Hein et al., 2020] Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A. L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B. A. (2020), *BIAS: Transparent reporting of biomedical image analysis challenges*. Medical image analysis, Volume 66, 101796. https://doi.org/10.1016/j.media.2020.101796

[b-Mathews et al, 2019] Mathews, S. C., McShea, M. J., Hanley, C. L., Ravitz, A., Labrique, A. B., and Cohen, A. B. (2019), *Digital health: a path to validation*. npj Digital Medicine, 2. https://doi.org/10.1038/s41746-019-0111-3

[b-McGregor, 2020] McGregor, S. (2020), *Preventing repeated real world AI failures by cataloging incidents: The AI incident database*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 15458-15463).
https://arxiv.org/abs/2011.08512

[b-McKinney at al., 2020] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D., and Shetty, S. (2020), *International evaluation of an AI system for breast cancer screening*. Nature 577, 89-94.
https://doi.org/10.1038/s41586-019-1799-6

[b-Mongan et al., 2020] Mongan, J., Moy, L., and Kahn Jr, C. E. (2020), *Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers*. Radiology: Artificial Intelligence, Volume 2, No. 2.
https://doi.org/10.1148/ryai.2020200029

[b-Mincu et al, 2022] Mincu, D., and Roy, S. (2022), *Developing robust benchmarks for driving forward AI innovation in healthcare*. Nature Machine Intelligence 4, 916–921.
https://doi.org/10.1038/s42256-022-00559-4

[b-Nagendran et al., 2020] Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H. Topol, E. J., Ioannidis, J. P. A., Collins, G. S., Maruthappu, M. (2020), *Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies*. BMJ, **368**:m689.
http://dx.doi.org/10.1136/bmj.m689

[b-Norgeot] Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., and Butte, A. J. (2020), *Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist*. Nature medicine, **26**, 1320-1324.
https://www.nature.com/articles/s41591-020-1041-y

[b-Omoumi et al. 2021] Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn Jr., C. E., Louvet-de Verchère, F., Santos, D. P. D., Kober, T., and Richiardi, J. (2021), *To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines)*. European Radiology 31, 3786–3796.
https://doi.org/10.1007/s00330-020-07684-x

[b-Park et al., 2020] Park, Y., Jackson, G. P., Foreman, M. A., Gruen, D., Hu, J., Das, A. K. (2020), *Evaluating artificial intelligence in medicine: phases of clinical research*. JAMIA Open, Volume 3(3) pp. 326-331.
http://dx.doi.org/10.1093/jamiaopen/ooaa033

[b-Reddy et al., 2020] Reddy, S., Allan, S., Coghlan, S., Cooper, P. (2020), *A governance model for the application of AI in health care*. Journal of the American Medical Informatics Association. Volume 27(3) pp. 491-497. https://doi.org/10.1093/jamia/ocz192

[b-Reddy S. et al 2021] Reddy, S., Rogers, W., Makinen, V-P., Coiera, E., Brown, P., Wenzel, M., Weicken, E., Ansari, S., Mathur, P., Casey, A., Kelly, B. (2021), *Evaluation framework to guide implementation of AI systems into healthcare settings*. BMJ health & care informatics, Volume 28(1). http://dx.doi.org/10.1136/bmjhci-2021-100444

[b-Reinke et al., 2018] Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P. M., Bogunovic, H., Landman, B. A., Maier, O., Menze, B., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Sommen, F. V. D., Zheng, G., Müller, H., Kozubek, M., Arbel, T., Bradley, A P., Jannin, P., Kopp-Schneider, A., and Maier-Hein, L. (2018), *How to exploit weaknesses in biomedical challenge design and organization*. 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

[b-Reinke et al., 2021] Reinke, A., Tizabi, M. D., Sudre, C. H., Eisenmann, M., Rädsch, Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Blaschko, M., Büttner, Cardosa, M. J., Chen, J., Cheplygina, V., Christodoulou, E., Cimini, B., Collins, G. S., Engelhardt, S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Glocker, B., Godau, P., Haase, R., Hamprecht, F., Hashimoto, D. A., Heckmann-Nötzel, D., Hirsch, P., Hoffman, M. M., Huisman, M., Isensee, F., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kavur, A. E., Kenngott, H., Kleesiek, J., Kleppe, A., Kohler, S., Kofler, F., Kopp-Schneider, A., Kooi, T., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson. P., Meijering, E., Menze, B., Moher, D., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Noyan, M. A., Petersen, J., Polat, G., Rafelski, S. M., Rajpoot, N., Reyes, M., Rieke, N., Riegler, M., Rivaz, H., Saez-Rodriguez, J., Sánchez, C. I., Schroeter, J., Saha, A., Selver, M. A., Sharan, L., Shetty, S., van Smeden, M., Stieltjes, B., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Calster, B. V., Varoquaux, G., Wiesenfarth, M., Yaniv, Z. R., Jäger, P., and Maier-Hein, L. (2021), *Common Limitations of Image Processing Metrics: A Picture Story*. Electrical Engineering and Systems Science > Image and Video Processing. https://arxiv.org/abs/2104.05642

[b-Reinke et al., 2022] Reinke, A., Maier-Hein, L., Godau, P., Tizabi, M. D., Buettner, Christodoulou, E., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., van

Ginneken, B., Haase, R., Hashimoto, D. A., Jaeger, P. F. (2022), *Metrics Reloaded – A new recommendation framework for biomedical image analysis validation*. Medical Imaging with Deep Learning 2022. https://mauricioreyes.me/Publications/ReinkeMIDL2022.pdf

[b-Rivera et al., 2020] Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group. (2020), *Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension*. Nature Medicine 26, 1351-1363. https://www.nature.com/articles/s41591-020-1037-7

[b-Roberts et al., 2021] Roberts, M., Driggs, D., Thorpe, M. Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., and Schönlieb, C-B. (2021), *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans*. Nature Machine Intelligence 3, 199-217. https://doi.org/10.1038/s42256-021-00307-0

[b-Rogers et al., 2021] Rogers, W. A., Draper, H, Carter, S. M. (2021), *Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues*. Bioethics. https://doi.org/10.1111/bioe.12885

[b-Samek et al., 2016] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K-R. (2016), *Evaluating the visualization of what a deep neural network has learned.* IEEE transactions on neural networks and learning systems, Volume 28(11), 2660-2673. https://doi.org/10.1109/TNNLS.2016.2599820

[b-Studer et al., 2020] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and 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

[b-Saporta et al., 2022] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T., Ngo, V-D., Seekins, J., Blakenberg, F. G., Ng, A. Y. Lungren, M. P. and Rajpurkar, P. (2022), *Benchmarking saliency methods for chest X-ray interpretation*. Nature Machine Intelligence 4, 867–878 (2022). https://doi.org/10.1038/s42256-022-00536-x

[b-Scott et al., 2020] Scott, I., Carter, S., Coiera, E. (2020), *Clinician checklist for assessing suitability of machine learning applications in healthcare*. BMJ Health & Care Informatics Volume 28. http://dx.doi.org/10.1136/bmjhci-2020-100251

[b-Sounderajah et al., 2020] Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A. K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., Markar, S. R., Mclnnes, M. D. F., Panch, T., Pearson-Stuttard, J., Ting, D. S. W., Golub, R. M., Moher, D., Bossuyt, P.

M., and Darzi, A. (2020), *Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group*. Nature Medicine. https://doi.org/10.1038/s41591-020-0941-1

[b-Talmon et al, 2009] Talmon, J., Ammenwerth, E., Brender, J., De Keizer, N., Nykänen, P., and Rigby, M. (2009), *STARE-HI—Statement on reporting of evaluation studies in Health Informatics*. International Journal of Medical Informatics, 78(1), p1-9. https://doi.org/10.1016/j.ijmedinf.2008.09.002

[b-NICE] 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

[b-Van Calster et al., 2019] Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., and 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

[b-WHO 2021] World Health Organization (2021), *Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation*. World Health Organization. https://apps.who.int/iris/handle/10665/349093.

[b-Wiesenfarth et al., 2021] Wiesenfarth, M., Reinke, A., Landman, B. A., Eisenmann, M., Saiz, L. A., Cardoso, M. J., Maier-Hein, L., Kopp-Schneider, A. (2021), *Methods and open-source toolkit for analyzing and visualizing challenge results*. Scientific reports. https://doi.org/10.1038/s41598-021-82017-6

[b-Winter et al., 2021] Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., Hochreiter, S., and Nessler, B. (2021), *Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications*. arXiv preprint arXiv:2103.16910. https://arxiv.org/abs/2103.16910v1

[b-Wu et al., 2021] Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., and Zou, J. (2021), *How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals*. Nature Medicine 27, 582-584.

https://doi.org/10.1038/s41591-021-01312-x

[b-Zhang et al., 2020] Zhang, J. M., Harman, M., Ma, L., and Liu, Y. (2020), *Machine learning testing: Survey, landscapes and horizons*. IEEE Transactions on Software Engineering. https://ieeexplore.ieee.org/document/9000651