FGAI4H-I-028-A01
E-meeting, 7-8 May 2020

Source: Editors

Title: *DEL07: AI for Health Evaluation Considerations*

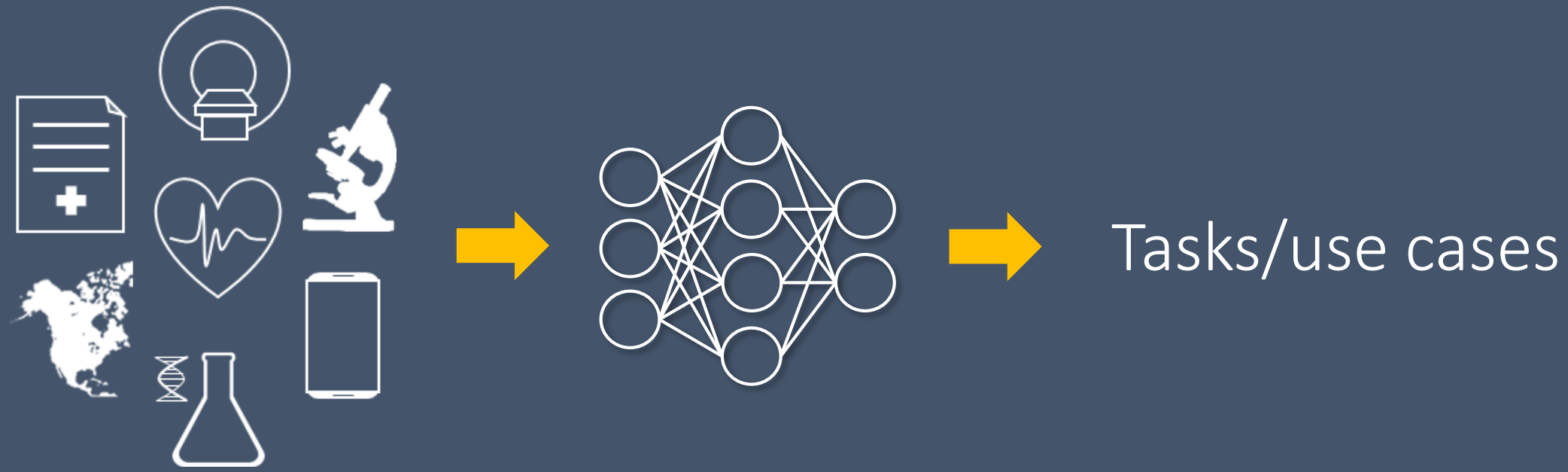
Purpose: Discussion

Contact: Markus Wenzel E-mail: markus.wenzel@hhi.fraunhofer.de

Abstract: These slides accompany document FGAI4H-I-028.

AI for Health

AI/ML Models for Health



Trustworthy? Accurate? Robust? Plausible? Effective? Safe?

AI for Health

AI4H Evaluation Considerations - Outline

- *Call for Participation*
- Intro · Background
- Novelty
- Independent standardized model benchmarking
- Benchmarking platform - closed environment
- Evaluation process (preliminary consider.)
- Best practices from literature
- Overview evaluation deliverables

AI for Health

Novelty

Which aspects of health AI model evaluation are novel/unique/essential?

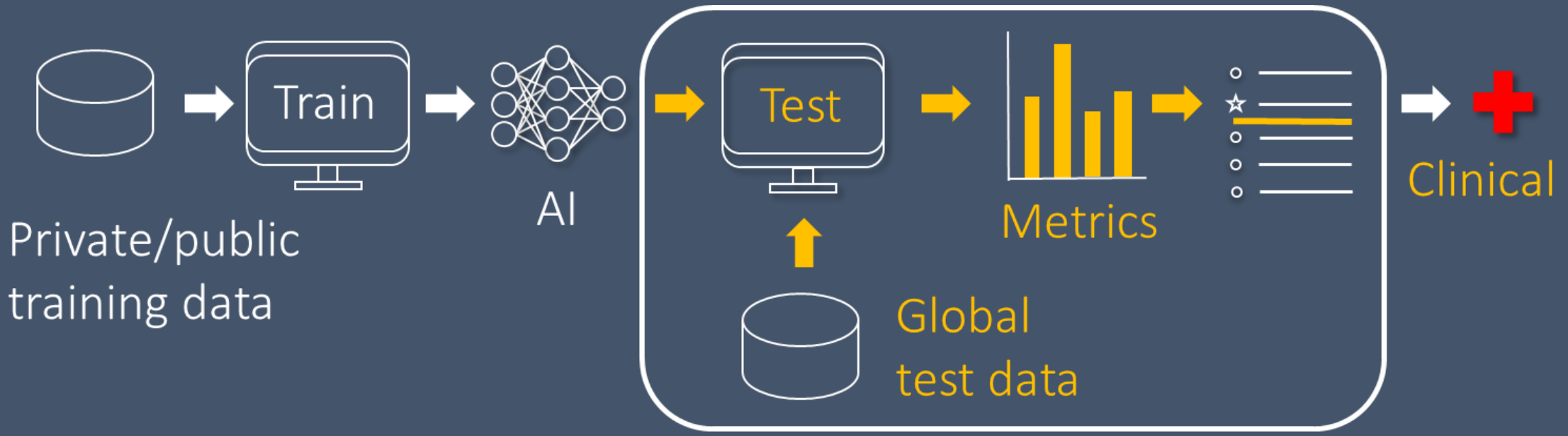
Ten initial ideas in Table 1.
Please send me yours.

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

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

AI for Health

Independent model benchmarking



1) Closed environment
2) Via interface
3) Federated

AI for Health

Appendix I

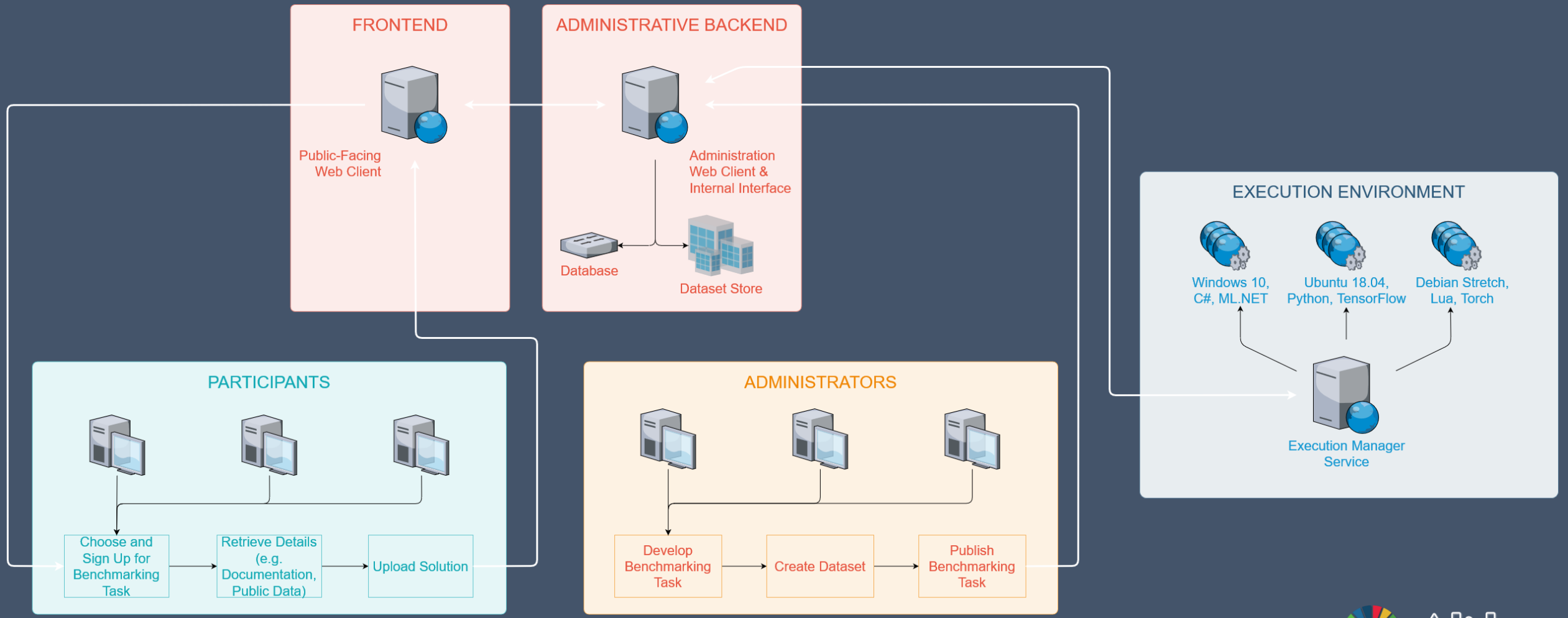
Benchmarking platform - closed environment

System overview (administrative backend · frontend · execution environment)

General considerations (security · hosting · computing resources · availability)

AI for Health

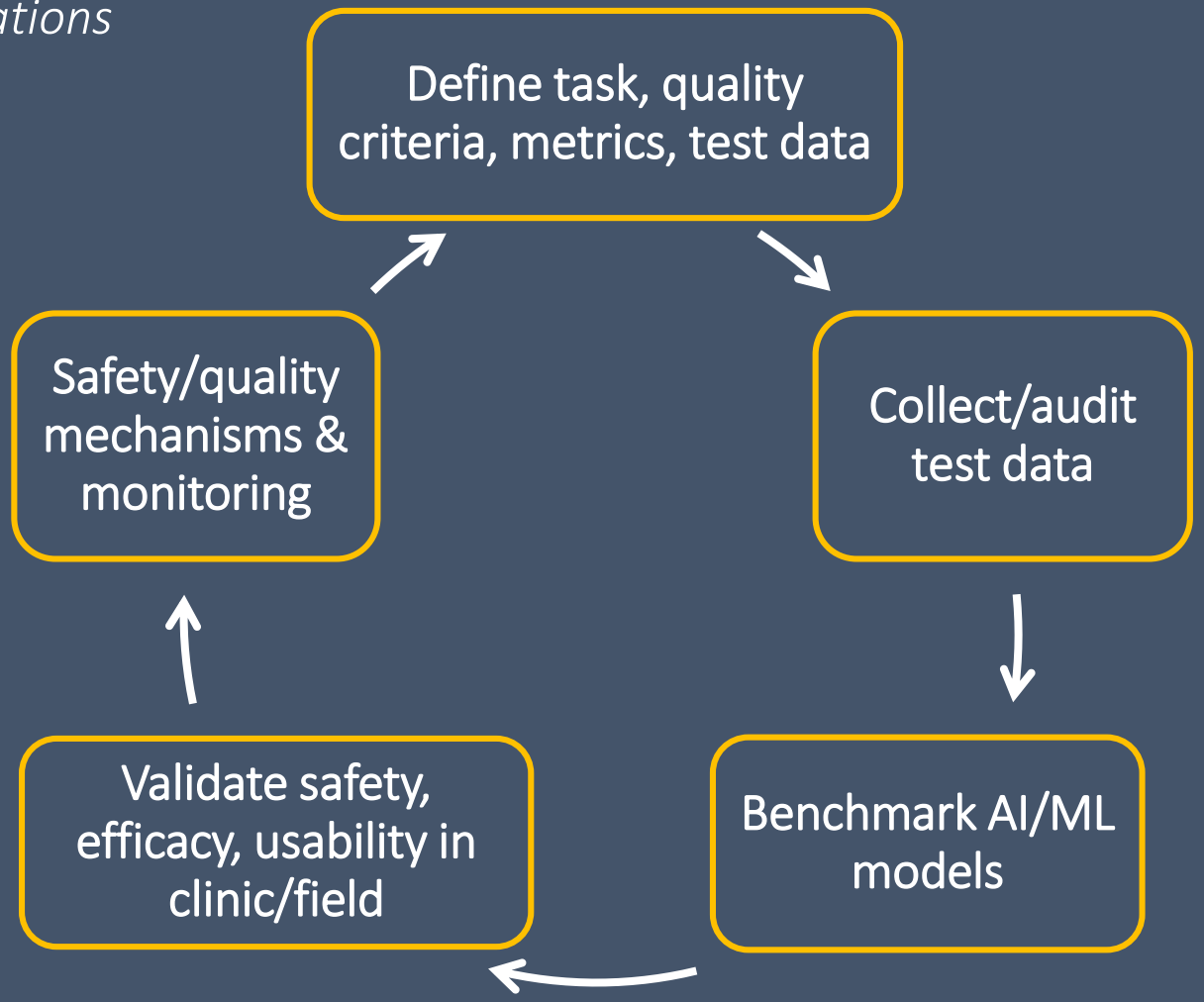
System architecture



AI for Health

Evaluation process

Preliminary considerations



AI for Health

Appendix II

Best practices from the scientific literature and other documents

From 13 important sources with brief discussion.
Please send me literature you find relevant.

AI for Health

Overview evaluation deliverables

N°	Title	Editor
7.1	AI for health evaluation process description	Sheng Wu *World Health Organization*
7.2	AI technical test specification	Auss Abbood *Robert-Koch Institut*
7.3	AI technical test metric specification	Luis Oala *Fraunhofer*
7.4	Clinical validation	Naomi Lee, Rupa Sarkar *The Lancet*

AI for Health

Thanks to all contributors!

David Neumann *Fraunhofer* Sandeep Reddy *Deakin University*
Annika Reinke/Manuel Wiesenfarth *German Cancer Research Center*
Alberto Merola *AICURA medical* Steffen Vogler *Bayer*

Next steps

Feedback topic/working groups + new contributors (join!)

→ markus.wenzel@hhi.fraunhofer.de

Read DEL07: https://www.itu.int/go/fgai4h/collab

AI for Health