|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2022-2024 | | FG-AI4H-R-063 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | WG-CE | Cambridge, 21-24 March 2023 | |
| **DOCUMENT** | | | | |
| **Source:** | | FG-AI4H working group clinical evaluation, editors | | |
| **Title:** | | DRAFT: Checklist for clinical evaluation of AI for health | | |
| **Contact:** | | Nina Linder University Helsinki, Uppsala University,  Finland, Sweden | | Email: [nina.linder@helsinki.fi](mailto:nina.linder@helsinki.fi) |
| **Contact:** | | Johan Linder University Helsinki, Uppsala University,  Finland, Sweden | | Email: [johan.lundin@helsinki.fi](mailto:johan.lundin@helsinki.fi) |
| **Contact:** | | Eva Weicken Fraunhofer Heinrich-Hertz-Institute, Germany | | Email: [eva.weicken@hhi.fraunhofer.de](mailto:eva.weicken@hhi.fraunhofer.de) |

**DRAFT: Checklist for clinical evaluation of AI for health**

This checklist for clinical evaluation of AI technologies in health (draft status) is based on the framework for clinical evaluation developed by the working group and covers important aspects of evaluating an AI system in a clinical setting across all relevant phases (design and purpose, analytical and clinical validation, ongoing monitoring, and economic evaluation). The checklist was created to provide a harmonized and comprehensive approach for evaluating AI systems in healthcare. It has been successfully tested and applied by a research team from Helsinki University, Karolinska Institute and Uppsala University conducting a study on digital microscopy for cervix cancer screening in Kenya demonstrating its applicability but also showing gaps for future work. The checklist can be used as a guidance on the considerations of clinical evaluation by a wide range of stakeholders involved in the development, evaluation, and implementation of AI systems in healthcare, for instance, developers of AI, researchers, clinicians, and regulatory authorities.

**1. Model design and purpose**

***1.1* *Identify the problem and intended use***

1.1.1 *Identify and describe the specific problem to be solved (population, input data required, output data from model, setting).*

* + For example, in an AI health technology designed to identify high risk patients with sepsis, the intended use should include target age-groups for which it is suitable and the setting (e.g., intensive care units, ICU vs non-ICU).
  + Additionally, developers should consider the range of clinical information needed for the problem and the intended use.

1.1.2 *Describe how and where the model would fit in the patient journey or clinical workflow.*

* + Who are the intended users of the model and who are the intended beneficiaries?
  + What could the interaction between the technology and user look like?
  + What effect would adoption of the AI technology have on the workflow and workload?
  + What will the interaction between technology and user look like, and what is the level of autonomy (Lyell et al., 2021)?

1.1.3 *Consider and describe any special circumstances related to the intended users or context.*

* + For example, in pediatric age-groups there may be a need to consider child protection issues
  + In rural settings there may be a need to consider issues such as little or no internet provision
  + Variations of clinical pathways in different regions
  + Socio-cultural variations around data and technology affecting the willingness to design and implement AI tools.

***1.2. Define intended benefits***

1.2.1 *What are the patient level benefits that can be achieved?*

* + For example, improvement of the patient experience, including reduced waiting times and better clinical outcomes (​e.g., ​improved survival rates, reduced complications compared with current context relevant standard of care)
  + quicker linkage from diagnosis to care or​ ​reduced out of pocket expenditure.

1.2.2 *What are the clinical workflow benefits?*

* + ​​For example,​ reduced administrative burden on health care professionals (HCPs)
  + increased time to care for HCP
  + provision of a better HCP experience.

1.2.3 *What are the health system benefits?*

For example,

* + efficiencies found or created in pathways
  + improved detection of cases
  + better allocation of resources
  + cost savings, addressing shortages of skilled HCPs

***1.3. Describe potential risks and harms***

1.3.1 *What are the potential patient level risks,* ​*like* ​*harmful consequences due to misclassification, misdiagnosis, delayed care, under- or overdiagnosis or unnecessary treatment, or consequences of bias in the AI technology?*

1.3.2 *What are the potential clinical workflow risks, including removing safeguards, additional time, administrative or cognitive task burden for* ​​​​*HCPs?*

1.3.3 *What are the potential system level risks, for example the health economic costs of expensive technology, or the potential for technologies to direct people to expensive and unnecessary care to be replicated at scale across large groups of people?*

*Frameworks developed by NICE, FDA, MDR could support to determine the risk class of your tool and provide guidance on its appropriate classification, e.g., whether it might be classified as a medical device as per IMDRF/FDA definition.*

***1.4. Interoperability and security***

1.4.1 *Describe interoperability requirements (such as minor and significant hardware and software upgrades) of the AI technology in order to work with other devices and IT systems.*

1.4.2 *What consequences could for example unintended changes have in the nature of input or output data arising from other IT systems around it?*

1.4.3 *Does the novel AI technology comply and make use of existing communication standards (e.g., Digital imaging and Communications Medicine (DICOM), Fast Healthcare Interoperability Resources (FHIR)?*

***1.5. User-testing and stakeholder engagement***

1.5.1 *What stakeholders have been engaged in development of the AI technology?*

1.5.2 *Have stakeholders been engaged in the design following a user centered approach?*

1.5.3 *What kind of user testing has been conducted to understand the interaction with the model in real world situations? A mixed methods approach can be used, including*

* + for example, user feedback (quantitative or qualitative study)
  + interviews (qualitative study)
  + usability testing (qualitative study)
  + focus groups (qualitative study, Delphi studies, quantitative study)
  + ethnographic study (qualitative study) ​​(Potts et al.).

**1.6. *Privacy and Security***

Stakeholders should be ​aware that​ data privacy and security are both rapidly evolving fields and should be given full consideration when a particular AI system is being considered. However, consideration of the privacy and security of AI systems in health, and the evaluation of these important considerations is out of scope of this document, and usually is given separate consideration to the clinical performance of a system.

**2. Algorithmic validation**

For the purposes of this document, we use the term 'algorithmic validation' to describe this evaluation of the adequacy of the AI model ‚in silico‘ in contrast to 'clinical validation' in which the whole AI health technology is evaluated in the context of the clinical pathway.

2.1 *How has the performance of the model been evaluated through development (training, tuning and internal validation stages)?*

* + The performance metrics should be transparently reported including, for example accuracy, positive and negative predictive values, and the area under the receiver operator curve.

2.2 *How suitable is the data that has been used in those stages in relation to the intended use?*

2.3 *Has the model performance been evaluated against one or more unseen external datasets (external validation)?*

* + *External validation* refers to the process of evaluating the performance of the AI model using previously unseen, and independent data ‚in silico‘. This is in contrast to clinical validation through interventional or clinical studies.

2.4 *Has the model performance been assessed against the current standard of care?*

* + For example, for a diagnostic test this would include sensitivity and specificity, ideally with a full confusion matrix (true positive, false positive, true negative, false negative).
  + Other measures such as area under the receiver operator curve (AUC) and area under the precision-recall curve (AUPRC) may also be helpful.

2.5 *Describe internal and external testing datasets that have been used.*

* + ​​Describe the input data type, and source, including where, when, and how it was collected.​

2.6 *Describe the demographic spread of the data including gender/sex, age, and race/ethnicity.*

* + These data points help indicate how inclusive the data is, and how representative it *​*is of the*​* target population for the intended use of the AI health technology.  *​     ​*

2.7. *Has the performance of the model been assessed within a population in whom under-performance may occur due to their under-representation in the training dataset?*​​

2.8 *Describe the ratio of training and testing data and provide a justification for the split.*

2.9 *How was the ‘ground truth’ established?*

* + If the ground truth was established by an expert, describe the training and experience of these experts, how many experts made a decision and how conflicts or variations were resolved, in order to establish quality of the labelled data.

**3. Clinical validation**

For the purposes of this document clinical evaluation refers to the evaluation of the AI system through interventional or clinical studies. Depending on the risk profile of the AI system, clinical evaluation may be done before or on parallel with the deployment. AI-specific guidance for different study designs is being developed and published by the EQUATOR network, e.g., SPIRIT-AI, CONSORT-AI.

Considerations of specific elements important in clinical studies include:

*3.1 Describe the study design* - consider the optimal study design for this intervention that will provide sufficient high-quality evidence across key domains (including effectiveness, safety, and cost-effectiveness) to support decision-making by relevant gate-keepers (e.g., ​health tech assessors, ​regulators, payers, users).

*3.2 Describe the population* - ensure that the study population (1) reflects the population in which it is intended to be used, and (2) that it is sufficiently diverse to detect under-performance or failure in specific groups.

*3.3. Describe the setting* - ensure that the study setting reflects the setting (or range of settings) of the intended use; again, diversity of setting is relevant, to provide sufficient confidence of performance outside of ideal scenarios.

*3.4 Describe intervention(s)* - ensure that the AI component of any intervention is described accurately   to ensure results are ascribed to a specific AI system (including version) and would enable replication of the study. This should include product details including version number, supplier and contact details.

*3.5 Describe intervention inputs and outputs*- ensure that the following are sufficiently clearly described to enable replication in both trial and clinical deployment contexts (1) the nature of the inputs into the AI system including both human and data elements (such as any data pre-processing); and (2) the nature of the outputs and how this is translated into actions within the healthcare pathway (includes human-computer interaction elements).

*3.6 Define the comparator* - the comparator (whether parallel control group or other design) should be a relevant reference. This reference is commonly ‘standard practice’ or ‘best practice’ with a view to informing decision-makers as to whether the intervention reflects an improvement (or not) on current health delivery.

*3.7 Describe pre-specified outcomes relevant to all stakeholders* - ensure that outcomes are defined in advance and include those that are the most important to patients, and the key stakeholder groups; use of core outcome sets are recommended where they exist for the condition of interest; pre-specification avoids bias through retrospective selection of most favorable outcome or of positive result arising through chance and multiple testing.

*3.8 Process measures* - describe relevant impacts on the overall health pathway such as positive or negative changes in time to diagnosis or treatment.

*3.9 Balancing measures* - consider upstream, lateral, and downstream consequences including changes in behaviour, changes in resource requirements, and potential ethical implications (such as loss of autonomy).

*3.10 Protocol deviations* - all deviations from study protocol should be recorded and reported. First, such deviations may affect the interpretation of results in relation to pre-specified outcomes. Second, such deviations may provide important information regarding the feasibility and safety of deploying the intervention more widely.

*3.11 Define the analysis* - analysis should be pre-specified (including the metric that will be used) and should include sufficient consideration of subgroups to ensure that any deviations of performance and potential risk of harm is detected; errors should be analysed at the individual error level to identify the reasons for failure where possible.

*3.12 Describe reporting of study protocol* - the study design should be registered (e.g., on the WHO International Clinical Trials Registry Platform) in advance; additional submission of protocols for publication may enable helpful independent peer review prior to commencement of the study.

*3.13 Reporting of study conduct and results* - open and transparent reporting should align to the registered protocol, include any protocol deviations, and full analysis of planned outcomes according to their pre-specified hierarchy. Participant flow (including exclusions at participant level, exclusions at input data level and losses to follow-up) should be reported according to the CONSORT-AI diagram (Xiaoxuan Liu et al. 2020), adapted from the CONSORT 2010 flow diagram (Schulz et al. 2010)).

**4. Deployment and ongoing monitoring, regulatory requirements, AI audits**

4.1 *Determine the level of additional evaluation required to appropriately assure version updates of AI systems and continuously learning or adaptive algorithms.*

4.2 *Identify short, medium, and long-term risks to patient safety.*

* + Risks may be wide-ranging and may relate to failures within the technology itself (including issues with algorithm design and the data used for training) or with how the technology is used by humans (intentional or unintentional misuse) or with issues relating to the deployment setting (including deviations from the inputs, outputs and supporting infrastructure anticipated).

4.3 *Describe potential individual errors, systematic errors or biases related to use of the AI technology.*

* + Both individual and systematic errors may result in patient harm and should be actively looked for at all stages of development and deployment of AI health technologies. There is particular ethical concern around systematic performance deviations (bias) relating to certain characteristics such as ethnicity or gender that may result in negative consequences

4.4 *Has the AI technology achieved regulatory approval?*

4.5 *Describe any monitoring of ongoing performance (both safety and effectiveness) of the AI product.*

* + AI systems are known to show poor generalisability when encountering new data and unexpected failure in spurious edge cases. Even in the presence of evidence supporting good performance across an aggregate population, it is important to be prepared for unexpected algorithmic outputs and potential adverse outcomes.

4.6 *What stakeholders are monitoring the AI product?*

* + In addition to product developers and regulatory authorities, *​*HCPs, *​*users, patients and the public also become gatekeepers for discovering and acting upon potential risks.

4.7 *How are adverse events reported (including suspected device-related deaths, injuries, and malfunctions)?*

* + It is important to be aware that AI as diagnostic or prediction tools may cause harms that only become apparent downstream in the clinical pathway and in some cases over extended time periods (for example, where an incorrect diagnosis first results in incorrect treatment, which in turn results in a poor outcome).

4.8 *Has the AI product been subject to algorithmic audits?*

* + AI audits can help discovering the occurrence of adverse events but also help understand why these happened. Through the AI audit, existing and potential risks can be assessed and prioritised, risk mitigation plans can be put in place, and future audits can monitor whether risk mitigation measures were successful in avoiding harm. Detailed analysis may be performed through a “medical algorithmic audit”, like in Liu et al., 2022 (10.1016/S2589-7500(22)00003-6).

**5. Economic ​​evaluation & reimbursement**

*5.1 Has the AI model been subject to an economic evaluation?*

* + An important aspect of evaluation for any health intervention, including AI health technologies, is the comparative measurement of the expected costs relative to its expected impacts when implemented in a particular context.

5.2 *Define potential opportunity cost related to the AI model* – the foregone benefits of investing limited resources on one course of action rather than another.

*5.3 Describe the types of economic evaluation that has been conducted related to AI-enabled digital interventions.*

* + These include for examplecost effectiveness analysis (CEA) termed cost utility analysis (CUA), where the net incremental costs of an intervention are presented as a ratio to net incremental health’s outcomes. Health is a generalised measure such as the quality adjusted life year (QALY) or disability adjusted life year-averted (DALY).

​​​​5.4 *Describe the*​ *outcomes of interest in the economic evaluation*. ​​

5.5 *Has a level of reimbursement for the AI technology been established*?

* + Pricing of digital health technologies, like other commodities, influences both affordability and access.

5.6. *Has a pricing been established for the AI product?*

* + Describe the pricing model i.e., is the product paid on either a subscription or fee-for-service or fee-per-use basis?

​​​​​​​​​​​​​​​​​**6. Communication of results**

6.1 *Describe how results of the clinical evaluation have been communicated*.

* + Communicating the results of the steps of the clinical evaluation process transparently is fundamental to the safe and effective use of AI health technologies​. It​ enable​s​ clinicians, patients, regulators, and other ​stakeholders to have​ the evidence they need to assess the safety, effectiveness and likely value of the technology and its performance in their setting.

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