|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-M-040 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | WG-CE | E-meeting, 28-30 September 2021 | |
| **DOCUMENT** | | | | |
| **Source:** | | Editors DEL7.4 | | |
| **Title:** | | Updated DEL7.4: Clinical evaluation of AI for health | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Co-chair, WG-CE Naomi Lee The Lancet United Kingdom | | E-Mail: [naomi.lee@lancet.com](mailto:naomi.lee@lancet.com) |
| **Contact:** | | Co-chair, WG-CE Shubhanan Upadhyay Ada Health Germany | | E-Mail: [shubs.upadhyay@ada.com](mailto:shubs.upadhyay@ada.com) |
| **Contact:** | | Co-chair, WG-CE Eva Weicken Fraunhofer Heinrich-Hertz-Institut Germany | | E-Mail: [eva.weicken@hhi.fraunhofer.de](mailto:eva.weicken@hhi.fraunhofer.de) |

|  |  |
| --- | --- |
| **Abstract:** | This document provides an overview of the current challenges of *"Clinical Evaluation of AI for Health"*. It is part of the deliverable-series 7.1-7.4 that are outlined by deliverable No.7 "*AI for Health Evaluation considerations".*  Although the performance of AI models in health is often measured by their accuracy, establishing confidence among clinicians, patients, researchers and policy makers in the safety, efficacy, and cost-effectiveness of AI solutions in health requires a more comprehensive evaluation.  The purpose of the deliverable No.7.4 is to outline the current best practices, the principles and outstanding issues for further considerations related to clinical evaluation of AI health technologies. It serves as the output document of the WHO/ITU Focus Group on AI for Health (FG-AI4H) Working group on Clinical Evaluation of AI for Health (WG-CE). |

**Call for participation**

If you are interested in contributing to this deliverable "*Clinical Evaluation of AI for Health"*, please contact Eva Weicken ([eva.weicken@hhi.fraunhofer.de](mailto:eva.weicken@hhi.fraunhofer.de)) and the secretariat of the ITU/WHO Focus Group on AI for Health ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) with "AI4H – Clinical Evaluation" as email subject and a brief introduction to yourself and your relevant expertise.

More background information are provided in the *Terms of reference* (<https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/wg.aspx>) and the *WG-CE collaboration site* ([https://extranet.itu.int/sites/‌itu-t/focusgroups/ai4h/wg/SitePages/WG-CE.aspx](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/wg/SitePages/WG-CE.aspx)).

Keywords

clinical evaluation, clinical validation, economic evaluation, generalisability, reporting, bias

Change Log

* **FG-AI4H meeting M, 28 – 30 September 2021:**
  + The feedback and comments of review round 1 have been converted into the outline draft/deliverable 7.4 respectively
  + The review and editing resulted in:
    - Redrafting of the first section, which is now titled “model design and suitability”, to refocus on the evaluation and reduce the duplication
    - Addition of pieces on ethics considerations relevant to this work, provided by Rohit Malpani from WG-Ethics
    - Addition of a figure and table to summarize what we say
    - Improved consistency of message and language, and reduction of duplication
  + Next steps:
    - External review round 2 started 21 September 2021 to receive feedback and comments on the current outline version (2.1) of all members of WG-CE and WHO-experts by 12 October 2021
    - Particularly, the group is asked for their thoughts on security of AI technologies and if this needs to be addressed here, and on the terminology (whether we use new terms as ‘Algorithmic validation’ and ‘Clinical validation’ that are used in the WG-RC document, use established terms or a longer definition of these terms)
* The document is based on four earlier versions:
* **FG-AI4H meeting L, 19 – 21 May 2021**, [L-040](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BB2C7C1C9-2315-43C9-9942-B10DC9CBE4CD%7D&file=FGAI4H-L-036.docx&action=default):
  + Initial formatting clean-up applied by secretariat.
  + Deliverable 7.4 includes the status update of the current WG-CE outline draft (version 1.1). ***Please* *note that this iteration of the outline draft (version 1.1.) is not finalized yet and corresponds to the document version that was shared with the WG-CE members for review-round 1 in beginning of April 2021. Therefore, the member's comments and feedback (in verbal and written form) have not yet been included***. The work on the implementation of the feedback of review-round 1 is currently in progress.
  + This iteration of DEL. 7.4 (outline draft version 1.1.) provides an overview of the outcome of the WG-CE activities until the time before the last WG-CE meeting in April 2021(also see chronologic timeline below). Compared to meeting K, the document has been completely updated and restructured following the insights of the WG-CE meetings and discussions.
  + Next steps:
    - Inclusion of the comments and feedback of review round 1, synchronization of the language within other FG-AI4H working groups and production of outline draft version 1.2
    - circulate outline draft version 1.2 with WC-CE members for review round 2
    - According to need: Follow-up meeting for final comments/review
    - Develop final outline draft > Deliverable 7.4
* **FG-AI4H meeting K, 27 – 29 January 2021**: [K-041](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BBC55470D-455E-4362-AA19-96990C9702A5%7D&file=FGAI4H-K-041.docx&action=default&CT=1621348538619&OR=DocLibClassicUI) – revised version with additions in some sections based on the output of the inaugural WG-CE workshop in October 2020 and the follow-up meetings in subgroups on pre- and post-deployment on clinical evaluation 15&16 December 2020
* **FG-AI4H meeting J, 30 September – 2 October 2020**: Deliverable draft "*Clinical Evaluation of AI for Health*" [J-053](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BCBB4539C-36CA-46A4-A015-0A56267FDD77%7D&file=FGAI4H-J-053.docx&action=default), submitted to the secretariat of the ITU/WHO Focus Group on "AI for Health" ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) and presented by the authors.
* Working draft sent as an update on 24 March 2020 to [tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int), titled DEL7.4: "Clinical Evaluation of AI for Health"

**WG-CE activities – chronologic timeline (stand of September 2021)**

* 21 October 2021: Starting review round 2 (version 2.1)
* 21 April 2021: WG-CE meeting for feedback on review round 1
* 10 March 2021: LMIC considerations meeting
* 15 & 16 December 2020: Pre/Post deployment subgroup meetings
* 14 October 2020: Inaugural WG-CE Workshop

**Contributors**

This document was developed in joint collaboration with all members of the FG-AI4H Working group on Clinical Evaluation. Based on the inputs of verbal and written feedback provided by the WG-CE members during the inaugural workshop and the follow-up meetings, the writing committee drafted the outline over time. The writing committee included (names in random order):

Alastair Denniston, Kassandra Karpathakis, Jane Carolan, Tommy Wilkinson, Xiao Liu, Naomi Lee, Shubs Upadhyay, Eva Weicken

External experts (in alphabetical order):

Table

Description automatically generated

**CONTENTS**

| Page |
| --- |
| [1 Scope 6](#_Toc83386135)  [2 References 6](#_Toc83386136)  [3 Terms and definitions 6](#_Toc83386137)  [3.1 Terms defined elsewhere 6](#_Toc83386138)  [3.2 Terms defined here 6](#_Toc83386139)  [4 Abbreviations 6](#_Toc83386140)  [5 Introduction and Background 8](#_Toc83386141)  [6 Model design and suitability 11](#_Toc83386142)  [6.1 Understanding the problem and Intended use 11](#_Toc83386143)  [6.2 Defining Intended benefits 12](#_Toc83386144)  [6.3 Defining potential risks 13](#_Toc83386145)  [6.4 Interoperability and security 13](#_Toc83386146)  [6.5 User testing and stakeholder engagement 14](#_Toc83386147)  [7 Algorithmic Validation 15](#_Toc83386148)  [8 Clinical validation 22](#_Toc83386149)  [9 Deployment and ongoing evaluation 26](#_Toc83386150)  [10 Economic evaluation 32](#_Toc83386151)  [11 Communication of results 36](#_Toc83386152)  [12 Recommendations 38](#_Toc83386153)  [Bibliography 40](#_Toc83386154) |

**List of Tables**

| Page |
| --- |
| [Table 1: Overview of required information for communication of results 37](#_Toc83386155) |

**List of Figures**

| Page |
| --- |
| [Figure 1: Framework for evaluation of AI models in health 10](#_Toc83386156) |

FG-AI4H Deliverable DEL7.4

Clinical evaluation of AI for health

Summary

[TBD]

# Scope

This document describes considerations on clinical evaluation of AI for health and aims to produce guidance for current best-practice evaluation of AI technologies in health. Iterations of the document are produced in collaboration with the contributors of this deliverable and presented at each FG-AI4H meeting. It serves as the output document of the FG-AI4H Working Group on Clinical Evaluation and is part of a series of deliverables as listed in FG-AI4H-K-200 [https://itu.int/en/ITU-T/focusgroups/ai4h/Documents/‌listdeliverables.pdf](https://itu.int/en/ITU-T/focusgroups/ai4h/Documents/listdeliverables.pdf)

# References

*Note: literature references are listed in the bibliography.*

# Terms and definitions

## Terms defined elsewhere

This document uses the following terms defined elsewhere:

**3.1.1 [tbc]**

## Terms defined here

This document does not define any new terms.

# Abbreviations

[tbc]

|  |  |
| --- | --- |
| AI | Artificial Intelligence |
| AI4H | Artificial Intelligence for Health |
| BIA | Budget Impact Analysis |
| CBA | Cost Benefit Analysis |
| CONSORT | Consolidated Standards of Reporting Trials |
| CUA | Cost utility analysis |
| DALY | Disability adjusted life year |
| DHI | Digital Health Intervention |
| DHT | Digital Health Technologies |
| EU | European Union |
| EQUATOR | Enhancing the quality and transparency of health research |
| FDA | Food and Drug Administration |
| FDR | Food and Drug Regulations |
| FHIR | Fast Healthcare Interoperability Resources |
| FG-AI4H | Focus Group on Artificial Intelligence for Health |
| GDPR | General Data Protection Regulation |
| HCP | Health Care Providers |
| I-DAIR | International Digital Health & AI Research Collaborative |
| IEEE | Industrial Electronics and Electrical Engineers |
| IMDRF | International Medical Device Regulators Forum |
| ITU | International Telecommunication Union |
| LMIC | Low -and middle-income countries |
| MDR | Medical Device Regulation |
| MI-CLAIM | Minimal information about clinical artificial intelligence modelling |
| ML | Machine learning |
| ML-OPS | Machine Learning Operations |
| ML4H | Machine Learning for Health |
| NICE | National Institute for Health and Care Excellence |
| NGO | Non-Government Organization |
| NHSX | National Health Service User experience |
| QALY | Quality adjusted life year |
| RWE | Real-World Evidence |
| SaMD | Software as a medical device |
| SPIRIT | Standard Protocol Items: Recommendations for interventional trials |
| STARD | Standards for the Reporting of Diagnostic accuracy studies |
| TDD | Topic Description Document |
| TG | Topic Group |
| TRIPOD | Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis |
| UHC | Universal Health coverage |
| UK | United Kingdom |
| UN | United Nations |
| WG-CE | Working Group on Clinical Evaluation |
| WHO | World Health Organization |

# Introduction and Background

Globally a growing shortage of clinicians, a rapid growth in digital health data, and an expansion in the usage of artificial intelligence (AI) technologies in other sectors has contributed to an increasing interest in the use of AI for health and clinical practice. As a growing number of models become available for use, researchers, patients, clinicians, and policy makers require a framework to understand whether the models are safe, effective, and cost-effective, and also to compare the performance of different models, including a comparison with the current clinical standard of care. The adoption of AI models in clinical use is complicated and may be hampered by a lack of trust and concerns about generalisability of models and bias. The current framework for evaluating health innovation, centered on evidence-based medicine, requires special considerations in order to evaluate artificial intelligence models for health, that include an evaluation of the underlying data, the potential for bias, and the contextual nature of the AI model performance[i]. Appropriate evaluation of models is key to safe adoption, and informing decisions on where and when AI health models can deliver meaningful improvements over current practice[ii].

This document is produced by the World Health Organization and members of the WHO/ITU Focus Group on AI for Health[iii] (FG-AI4H) Working Group on Clinical Evaluation (WG-CE) in a joint effort with other expert groups and global stakeholders stemming from various fields (clinicians, academia, research, commissioning, health-startups, NGOs, etc.). The aim is to produce guidance for current best practice evaluation of AI technologies in health, primarily aimed at researchers, clinicians, and policy makers, but that may also be useful for patients, the public, and developers of AI technologies. A shared understanding of evaluation best practice could help facilitate adoption of tools that are safe and effective, and therefore have the potential to improve health outcomes for all. WG-CE held its inaugural workshop in October 2020 followed by regular meetings, including a dedicated workshop for clinical evaluation in low-and middle-income (LMIC) settings. The outline document was shared and reviewed by all WG-CE members (NL: Need to add review process by WHO and by external reviewers).

FG-AI4H is formed from a collaboration between the World Health Organization (WHO) and International Telecommunication Union (ITU). As such, it has a global scope and interest in evaluation that supports the Sustainable Development Goals[iv] (SDGs), particularly SDG 3 on good health and well-being for all at all ages. The emphasis throughout this document is on principles of evaluation to ensure that it is relevant across all countries with minimal assumptions around particular health systems, or the agencies involved. This document will use a consistent nomenclature with other WHO and FG-AI4H documentation to ensure a universal understanding. Other working groups within the FG-AI4H are considering the topics of ethics, regulatory considerations and data handling - and assessment methods including an open code initiative and auditing trials with respect to AI models for health. This document will draw on and reference their considerations and recommendations, and the documents are intended to be used together.

The use of AI for health – including clinical medicine, public health, and operational management - raises a number of ethical, human rights, legal, and social concerns that should be taken into account when evaluating an AI technology. The full range of ethical and human rights related concerns are discussed in deliverable [DEL01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B0505B020-362C-45B2-94BF-215D2EBBD8F5%7D&file=DEL01.docx&action=default).[v]

|  |
| --- |
| *Box 1. Ai4H Ethics Principles*  The WHO-ITU FG-AI4H working group on “Ethical considerations on AI4H” has identified six principles that should guide the design, development, and deployment of any AI technology for health:  · Protecting human autonomy  · Promoting human well-being and safety and the public interest  · Ensuring transparency, explainability, and intelligibility  · Fostering responsibility and accountability  · Ensuring inclusiveness and equity  · Promoting AI that is responsive and sustainable.  The “*ethics and governance of AI4H”* report also provides recommendations to facilitate the appropriate governance of AI technologies for health, including the appropriate evaluation and regulation of AI technologies. Such legal and non-legal governance of AI technologies can help to balance competing influences and demands and maximise the benefits of these technologies while addressing or mitigating ethical and human rights related concerns. |

This clinical evaluation document will consider current best practice evaluation of AI models for health, in keeping with those ethical and regulatory principles. The best practice evaluation described does not replace the system for regulatory approval, but equips stakeholders with the framework necessary to evaluate the safety, cost and performance of health AI models, in keeping with the principles of evidence-based medicine for their clinical problem, setting and population. This existing framework has limitations when applied to evaluate medical innovations, and AI based interventions, as, among other issues, it does not facilitate evaluation of rapidly changing models, account for bias or poor quality underlying data, and does not acknowledge the contextual performance of AI models. An overreliance has developed on analytical performance measures such as accuracy, which is insufficient as a framework to evaluate health AI models, especially as such technologies are often used in a complex clinical environment. In this context, the document will also consider how the extent and depth of evaluation may relate to the estimated level of risk to patients; for the AI systems that are classified as medical devices this is indicated by the regulatory class of device[vi]. This document draws on existing evaluation frameworks and reporting guidelines tailored for special requirements on the evaluation of health AI. The EQUATOR network provides reporting guidelines for trial reports and study protocol involving AI models, for example, CONSORT-AI[vii] and SPIRIT-AI[viii] and further reporting guidelines are under development. This document proposes a framework for clinical evaluation of AI models in health based on these current best practices (Figure 1). It also identifies potential gaps in the existing evaluation frameworks and produces a set of recommendations for future work, as well as highlighting key documents for reference.

Graphical user interface, diagram

Description automatically generated

Figure 1: Framework for evaluation of AI models in health

The work of FG-AI4H is closely aligned to the principle that health care should be equitable and that no one should be left behind. The best practice recommendations provided here look to uphold this principle. In this regard, this document will also provide special consideration on clinical evaluation of AI for health implementations in LMIC settings. When considering the applicability of AI-tools there are a number of potential barriers to equitable access, which may be particularly acute in low resourced settings[ix]. Availability of representative datasets, with quality annotation is a major challenge, and improving the availability of representative and diverse data including the presentation of underrepresented populations for key medical conditions is a key priority. Poor technical infrastructure and a lack of access to technology (e.g., stable internet provision) might be a basic obstacle, especially in low resourced settings, but also remains a major problem in high-income countries in certain settings (e.g., rural areas, marginalised populations).

# Model design and suitability

Evaluating the purpose of an AI health technology and understanding whether or not it is suitable for the clinical problem and the setting for which it is being considered requires an explanation of the clinical problem, and demonstration that AI is the most suitable option to solve it[x]. Best practice development employs co-design with users and key stakeholders (e.g., patients, public, clinicians or other health professionals). As when developing other types of health technology, co-design enables developers to have a better understanding of user needs and priorities, the clinical problem and workflows, and can improve relevance, usability and adoption of the resultant technology (<https://www.theiet.org/impact-society/factfiles/healthcare-factfiles/design-and-evidence/>).

Similar to digital health products, there are concerns that AI technologies may be applied blindly as technological solutions or considered a ‘magic bullet’ to address what are deeper social, structural, economic, and institutional barriers. Use of an AI technology should be preceded by rigorous analysis and evaluation to ensure it is suitable, appropriate, and will not unnecessarily divert resources from proven health interventions (technological or otherwise). This also includes ethical considerations about the appropriate use of AI technologies, summarized at the end of this section in box 2.

## Understanding the problem and Intended use

Health AI models are complex, dependent not only on the constituent code, but also on the training data, clinical setting and user interaction. They are most often deployed into a complex clinical pathway and may involve a complete pathway transformation. The description of the use case therefore has a substantial role, both to inform end users where the technology can safely and appropriately be utilised, and for regulated tools (the statement of intended use) to allow regulators to assess if the evidence of the analytical and clinical validation steps taken are appropriate and sufficient for the intended use case.

The description of the problem and intended use should include:

1. Identification and description of 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). Are clinical parameters required, or does the model also need lab test results?
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 model and user look like?
3. Are there any specific considerations 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.

## Defining Intended benefits

Evaluation of a model requires an understanding of the intended benefits to the individual patient, clinical workflow or health system (or a combination of these).

Examples:

* Patient level benefits could be an improvement of the patient experience, including reduced waiting times and better clinical outcomes (improved survival rates, reduced complications compared with current context relevant standard of care).
* Clinical workflow benefits could be reduced administrative burden on HCPs, increased time to care, and provide a better HCP experience
* Health system benefits could be efficiencies found or created in pathways, better allocation of resources, cost savings.

## Defining potential risks

Consideration of the potential risks of an AI model are an important component of model evaluation.

* Patient level risks might include misclassification, misdiagnosis, automation bias, delayed care, under-or overdiagnosis or unnecessary treatment
* Clinical workflow risks, including additional administrative or cognitive task burden for clinicians
* System level risks, for example the health economic costs of the clinical impacts, plus the potential for these to be replicated at scale across large groups of people

Depending on the task, the user and the context, the risk profile may vary. There are examples of frameworks to help AI developers define which risk class their tool might belong to (NICE evidence standards framework for digital health, FDA, EU MDR). In general, the higher the risk class, the greater the requirement to demonstrate that a technologiel's benefits outweigh any potential risks, and the greater the responsibility to show how those risks are mitigated. This also has implications for whether the technology might be classified as a medical device as per IMDRF/FDA definition and its associated risk classification (IMDRF) that can inform subsequent regulatory, clinical requirements to a certain extent.

## Interoperability and security

Interoperability requirements (such as minor and significant hardware and software upgrades) of AI technologies with other devices and IT systems is often overlooked. Being 'connected' is not the same as being 'integrated', and this may affect the technical performance of the AI health technology (for example due to unintended changes in the nature of input or output data arising from other IT systems around it), and the wider performance including the extent to which any potential efficiency gains are realised. International initiatives to provide communication standards (e.g., DICOM, FHIR) that support interoperability are essential and should be supported (<https://www.dicomstandard.org/>, <https://www.hl7.org/fhir/overview.html>).

Whilst some groundwork has been done in areas such as medical imaging, or with the attempts to create communication standards (e.g., FHIR), there remains much work and implementation to be done to create common standards that would bring interoperability by default.

Full consideration of security is out of scope of this document, but appropriate consideration should be undertaken of the security of the health AI tool, including the data collected.

## User testing and stakeholder engagement

AI technologies that have had stakeholders engaged in the design following a user centered approach are likely to have greater support and successful adoption. This can then be evaluated through user testing to understand the interaction with the model in real world situations.

A number of user testing and evaluation methods can be carried out with end users and stakeholders of health AI throughout both, the design and development process, and clinical evaluation.

A mixed methods approach is usually required, especially for different points of enquiry. These methods include, but are not limited to user feedback (quantitative or qualitative study), interviews (qualitative study), usability testing (qualitative study), focus groups (qualitative study, Delphi studies (quantitative study) and ethnographic study (qualitative study) [xv]

|  |
| --- |
| *Box 2: Ethical considerations about the appropriate use of AI*  Appropriate evaluation of whether an AI technology should be used, from an ethical standpoint, should consider the following:  · The AI technology should meet standards of scientific validity and accuracy that are currently applied to medical technologies.  · The benefits of AI should consider the infrastructure and institutional context in which the technologies will be used. In particular, the digital divide may undermine the ability of providers and health systems to make use of such an AI technology equitably and/or fully within a health system.  · Irrespective of whether an AI technology provides accurate, useful information and insights, there may be enough ethical concern about a use case to or specific AI technology to discourage a particular use.  · The use of an AI technology should take full account of the total cost and investment required for its use, including digital infrastructure, training, maintenance and monitoring costs.  · There should be sufficient consideration as to whether an AI technology is appropriate and adaptable to the context of LMICs, including diverse languages and scripts in a country or amongst countries.  Thus, clinical evaluation of AI technologies may wish to consider the following questions:  · How should benchmarking account for the context, whether it is the state of digital infrastructure, languages spoken, or capacity for use in the health system, when evaluating an AI technology? To what extent should an existing digital divide affect an evaluation and/or recommendation of use of an AI technology in a specific context?  · How can benchmarking consider ethical challenges, or human rights concerns that may arise with the use of an AI technology? |

# Algorithmic Validation

A key part of evaluating a potential AI health technology is to assess the technical performance of the AI model when confronted with new data representative of its intended use. For the purposes of this document, we use the term 'algorithmic validation' to describe this evaluation of the adequacy of the AI model in contrast to 'clinical validation' in which the whole AI health technology is evaluated in the context of the clinical pathway.

This algorithmic validation requires

1) an understanding of the performance of the model through development (training, tuning and internal validation stages) and an assessment of the suitability of the data that have been used in those stages;

2) assessment of performance against one or more unseen external datasets (external validation);

3) assessment of performance against the current standard of care, and potentially against other new AI technologies. Performance metrics should be appropriate to the application such as sensitivity and specificity for a diagnostic test, although other measures such as area under the receiver operator curve may also be helpful.

*Description of internal and external testing datasets*

AI models are highly dependent on the training data used to develop them. Evaluating data used for training, tuning, internal and external validation and assessing the extent to which these datasets align to the intended use, including specific use case, population and setting. AI technologies may not perform well in populations or contexts that are different to that in which the training data was collected.

External validation is critical for all health AI models to show that the model can be used beyond the data in which it was developed, and give some indication of the extent to which that model may generalise. Data for external validation must not have previously been seen by the AI model, and would commonly be from one or more new locations (e.g., different hospitals to those that provided data for training and internal validation stages). Testing on newly acquired data from the original location has value in providing assurance of ongoing stability of performance with the original population and setting, but does not demonstrate generalisability beyond the setting and population in which the model was developed.

Training/testing data reporting should include:

· Description of demographic spread of the data including gender/sex, age and race/ethnicity that may indicate how inclusive the data is, and how representative it would be for the target population for the intended use of the AI health technology. Performance metrics should be provided not only for the population as a whole but for key groups within the population in whom under-performance may occur due to their under-representation in the training dataset.

· A description of the input data type, and source, including where and when it was collected.

· Quality of the training data, and the robustness of the labels will also affect the AI model’s performance. Understanding what was used as the ‘ground truth’ for training data, and the steps that were taken to ensure the quality of these labels is important for evaluation. For example, where the ‘ground truth’ is diagnosed by an expert, understanding the training and experience of these experts, how many experts made a decision and how conflicts or variations were resolved, all provide information which underpins the quality of labelled data.

*Algorithmic Validation*

For the purposes of this, and other FG AI4H documents, algorithmic validation refers to the process of validating the AI tool using data in silico, in contrast to clinical validation through interventional or clinical studies.

As discussed above, appropriate algorithmic validation in an independent, quality, external dataset demonstrates that a model is robust and performs to an acceptable level in the intended setting. It may also provide evidence of areas of potential bias and risks around generalisability. For the data this can include, among other things, the assessments of bias and stratification or missingness. The AI tool may be examined for its behaviour under distribution shifts [xvi], possible resulting degradations in predictive confidence or its learned decision heuristics and more. [xvii]

After training, internal testing should be carried out on an unseen portion of the original dataset, and further tuning may be performed. An AI tool must then be externally validated in a dataset that is independent from that in which it was trained (not merely an unseen portion of the training dataset) in order to demonstrate external validity. External validation should be carried out in a dataset that is representative of the setting and population intended for use. This can be carried out several times in different settings and populations to demonstrate robust performance within the intended use across those settings and populations. The external validation dataset should be of adequate quality with accurate labels to provide assurance that the performance metrics achieved by the AI model during external validation can be trusted. Failure cases particularly those that are surprising or unusual should also be identified.

The performance metrics should be transparently reported including, for example accuracy, positive and negative predictive values, and the area under the receiver operator curve. Providing these for subsets of the data can demonstrate the extent to which performance is maintained across subgroups, for example men and women, or in different ages or ethnic groups, or whether there is systematic under-performance in one or more groups.

*Benchmarking against the current standard of care or other AI models*

In order to understand the performance of the tool, evaluation against an accepted standard should be made. The most appropriate standard for comparison may differ according to the intended use but common examples of standards are human performance in a similar task or other models (for example derived from logistic regression). Depending on the intended use, the performance requirements may vary depending on whether the intended use is for screening or for diagnosis.

Using a similar process as external validation, that of testing the AI tool on an unseen dataset, it is possible to perform comparative benchmarking of AI tools. This has been performed in a limited number of settings, but as the number of AI tools increases, this may become increasingly important. Benchmarking against unseen datasets also has a number of potential uses beyond comparison of alternative tools. For example, if clinical evaluation has been performed for a model, which is then improved or updated with either new training data or a code change, benchmarking could demonstrate the algorithmic performance had remained similar, and provide a way to constantly and quickly evaluate dynamic AI tools, without requiring full clinical evaluation for each iteration. Further, where clinical validation has been performed for an AI tool or a class of tools, it may be possible to undertake algorithmic validation and infer likely clinical performance based on the algorithm matching or exceeding the technical performance of an equivalent clinically validated algorithm.

The availability of external unseen datasets for analytical validation is a current challenge in many commercial, government and academic settings, requiring collaborations to be established for each tool. Where local, regional, and national bodies are interested in evaluating AI tools, they could hold their own hidden dataset to enable this external validation set (for example, an initiative underway by NHSX, the UK’s digital health body, to develop nationally representative datasets for some common AI use cases). Prioritizing data collection could be an example of driving ‘needs based’ innovation as recommended by the 2020 Global Digital Health Partnership policy document [ref].

Good practices in algorithm validation comprise the use of state-of-the-art machine learning operations (MLOps) software tools for workflow management (<https://airflow.apache.org/>,<https://oozie.apache.org/>), result tracking and reporting (https://eval.ai/,<https://mlflow.org/> ) should be used. New tools, tailored to the specific needs of health AIs, are being actively developed (<https://ai4h-audit.org/>,<https://github.com/aiaudit-org/fgai4h-evaluation-platform> ) and can help stakeholders to produce or consume analytic validation results.

*Reader study*

If the analytical performance of a tool is acceptable, and clinical evaluation is intended, many tools may first be evaluated by a reader study to understand the performance of the tool when used in practice. A reader study, rather than evaluating performance of the tool alone on a dataset, would provide the tool to the intended user and ask them to perform the intended task on test data with and without the AI tool. This enables an understanding of the tool’s performance in the hands of the user.

*Special considerations*

|  |
| --- |
| *Box 3: Ethical considerations on data collection and data use*  The development of a successful AI health technology for use in health care relies on high-quality data for both training the algorithm and validating the AI model. The potential benefits of big health data can be ethically important, as AI technologies based on high-quality data can improve the speed and accuracy of diagnosis, improve the quality of care and reduce subjective decision-making.  Ethical considerations should recognize both the potential benefits and the risks of this data usage. The potential benefits to the individual and society include the expectation of health gains through faster, more accurate diagnosis, prognosis, treatment decision-support and a range of other AI health technology applications. The potential challenges include:  · There are concerns with the inclusiveness and representativeness of the data, including systematic underperformance or biases because of under-representation of a gender, age, race, sexual orientation or other characteristic. The data might also not be the right data, e.g., historically and might have been collected for another purpose. These biases will emerge during modelling and subsequently diffuse through the resulting algorithm.  · There are concerns with the safeguarding of individual privacy. The collection, use, analysis and sharing of health data have consistently raised broad concern about individual privacy, and the risk that it may harm an individual or cause a wrong.  · There are concerns of the repurposing of data, or ‘function creep’, wherein data shared initially for health purposes may be used by other government agencies to exercise control or employ punitive measures against individuals, or that technology providers may collect and use excess data, or so-called ‘behavioral data surplus’ for uses that raise ethical, regulatory, legal and human rights concerns.  · There are concerns how data is collected as this is also a trans-national issue, and in particular a concern with the collection of data from under-represented or marginalized groups, especially individuals from LMICs by companies and entities that are based in high-income countries. It can result in the use of data for commercial or non-commercial purposes without due respect for consent, privacy or autonomy.  Clinical evaluation of AI technologies may therefore wish to consider the following questions:  · How should clinical evaluation account for how data is collected (was there misuse of data; was their appropriate consent for the collection of such data – however consent is defined)?  · How should clinical evaluation evaluate the privacy of personal health information (for example, in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?  · What governance structures should an authority that reviews an AI technology should introduce to examine the collection and use of health data? |

*Building high-quality, representative datasets*

Obtaining datasets that are sufficiently representative of the local population, and of sufficient quality with the required labels can be difficult; this lack of health data for certain people groups or even whole nations is sometimes termed 'health data poverty' and is a major risk to the development and deployment of equitable digital health. At a national level, for example some LMICs, this may not only limit development of AI health technologies within that setting, but also restrict their ability to safely import AI health technologies produced elsewhere due to a lack of local data on which to assess its performance (external validation for the local setting and population). The ability to produce robust datasets with high quality ground truth labels is likely to be affected by limitations elsewhere in the health setting affecting access to diagnosis and treatment. These major challenges have the potential to not only propagate inequality of access, but also to compromise safety and performance of AI tools, and is an area which requires ongoing scrutiny/observation/monitoring/review.

Within populations, under-representation of people groups may lead to various harms including exclusion (the AI health technology is recognized not to perform reliably in that group) or exposure to under-performance (which may be recognized or not). There is a general risk that these biases exacerbate entrenched health inequalities. Building representative datasets of sufficient quality for the validation (and ideally also for training) of AI health technologies is the foundation of equitable digital health. One of the major factors driving unequal availability of data is the differential availability of the technology (notably electronic health records and instruments such as imaging devices, also the existence of data generation pipelines (such as national screening programmes) and adoption of EHRs in affluent countries mean an infrastructure that's ready to generate an evergreen pool of data) which may only be present in more affluent regions or countries; a second more fundamental challenge is where representative data does not exist because those individuals are excluded from full engagement with the health system. It also results in the potential that it will only be these populations that stand to benefit. Health data poverty can play out at both ends of the economic spectrum. There can be under-representation in the data for those in contexts that do not systematically capture this, ranging from rural areas with poor infrastructure, migrant health services, to wealthy areas with a high number of individuals choosing private medical services in which the data is siloed. Proactive, priority driven representative data collection is fundamental to the ability to carry out quality algorithmic validation and address bias in AI models.

Care must also be taken to tackle the issues around low internet connectivity and the availability of IT infrastructure for AI training and implementation, and the related variability in experience with digital tools and technologies, and the implications of human decision making in that context. Ensuring that AI developers are working with local healthcare professionals and adapting to available ‘lower tech’ IT solutions can be part of ensuring that AI systems are available for the greatest unmet need.

|  |
| --- |
| *Box 4: Ethical considerations on bias and discrimination:*  Societal bias and discrimination are often replicated by AI technologies. The different forms of discrimination and bias that a person or a group of people suffer because of identities such as gender, race and sexual orientation must be considered. There are three types of bias that clinical evaluation may need to contemplate:  · The data sets used to train AI models are biased, as many exclude girls and women, ethnic minorities, elderly people, rural communities and disadvantaged groups. In general, AI is biased towards the majority data set (the populations for which there is most data), so that, in unequal societies, AI may be biased towards the majority and place a minority population at a disadvantage. Such systematic biases, when enshrined in AI, can become normative biases and can exacerbate and fix (in the algorithm) existing disparities in health care.  · Biases often depend on who funds and who designs an AI technology. AI-based technologies have tended to be developed by one demographic group and gender, increasing the likelihood of certain biases in the design. Bias can also arise from insufficient diversity of the people who label data or validate an algorithm.  · Bias can also be introduced during implementation of systems in real-world settings. If the diversity of the populations that may require use of an AI system, due to variations in age, disability, comorbidities or poverty, has not been considered, an AI technology may discriminate against or work improperly for these populations.  Some of the questions that clinical evaluation may wish to consider include the following:  · What are the ethical implications of applying the AI model in real-world scenarios?  · How can clinical evaluation ensure benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of ‘inclusion and exclusion criteria’ for test data)?  · How can clinical evaluation ensure that those who design and develop an AI technology are representative of the populations who will rely on such technologies and to reassure providers that make use of such technologies? |

# Clinical validation

Clinical studies seek to provide the necessary evidence as to whether an AI system is effective and safe when deployed in a clinical pathway. Performance *in silico* may not translate into performance *in vivo*, due to numerous technical and human factors. As such, clinical studies should be considered a tool for both pre- and post-deployment evaluation of AI systems, designed to answer questions pertinent to the relevant populations, comparators and outcomes. Prospective clinical studies also allow the downstream and collateral consequences of the intervention to be measured, and may reveal unintended consequences outside of the limited outcomes assessed in the development, testing and validation phases.

The overarching aim should be to design studies that give confidence in results by minimizing bias and therefore provide confidence for decision makers. An important aspect of this is reporting transparency of studies, including prospective analysis plans, and reporting which is in line with the protocol and statistical analysis plan. Clinical studies should be designed to evaluate the impact on the whole pathway, and to understand the outcome for an endpoint which is robust and meaningful either clinically or for the system; it is important to acknowledge that performance metrics of the device itself (e.g., knowing sensitivity and specificity for a novel AI diagnostic) do not necessarily automatically improve clinical outcomes.

Additionally, depending on the intended use of the AI system and its setting, there may be regulatory requirements which need to be considered when planning the clinical evaluation phase.

The principles of good clinical study design are equally applicable for AI systems. Systematic reviews and meta-analyses have drawn attention to the poor levels of design and reporting in published AI studies, across the whole development pathway[xviii], [xix]. Randomized controlled trials (RCT) remain the benchmark of clinical studies, in which key elements help to minimize bias and increase confidence in the findings[xx], [xxi]. Other forms of study may be undertaken where an RCT is not feasible, but require additional consideration of some of the potential biases that may arise. A higher standard of evidence may also give confidence to clinicians using a tool, where the algorithm itself is not explainable.

As an intervention, AI systems do raise a number of specific challenges and considerations, and this has led to a number of guidance documents to help optimize specific study designs when evaluating an AI intervention. This is being addressed through the publication of AI-specific guidance for different study designs through the EQUATOR network, notably the publication of SPIRIT-AI [xxii] (for reporting of study protocols) and CONSORT-AI [xxiii] (for reporting of trial reports); additional EQUATOR guidelines are currently in development for diagnostic test accuracy studies (STARD-AI [xxiv]) and studies of prediction models (TRIPOD-AI [xxv]).

Specific elements that should be considered in clinical studies of an AI evaluation include:

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

· *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;

· *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.

· *Intervention(s)* - ensure that the AI component of any intervention is described sufficiently precisely 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.

· *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).

· *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;

· *Pre-specified outcomes relevant to all stakeholders* - ensure that outcomes 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.

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

· *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).

· *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.

· *Analysis* - analysis should be pre-specified, and should include sufficient consideration of subgroups to ensure that any deviations of performance and potential risk of harm is detected; errors should be analyzed at the individual error level to identify the reasons for failure where possible.

· *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.

· *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 (adapted from the CONSORT 2010 flow diagram[xxvi]).

It is encouraging to see the emergence of well-designed clinical studies of AI interventions. RCT remains the ideal trial design, although in some cases prospective observational studies with a relevant comparator, a meaningful outcome and systematic safety reporting may be considered adequate for some AI tools. By drawing together good study methodology, an understanding of the strengths and limitations of AI systems, and awareness of the types and levels of evidence required by key stakeholders, clinical studies can be designed and delivered which will enable regulators and other gate-keepers make better decisions regarding AI systems, enabling their populations to benefit from these interventions, whilst also reducing the risk of harm.

|  |
| --- |
| *Box 5: Ethical considerations of autonomous decision-making*  While AI is used only to augment human decision-making in the practice of public health and medicine, epistemic authority has, in some circumstances, been displaced, whereby AI systems (such as with the use of computer simulations) are displacing humans from the center of knowledge production. There are signs of full delegation of routine medical functions to AI.  Yet the introduction of autonomous decision-making into the healthcare field can also raise several ethical concerns, including the following:   * There is a widely held convention that many algorithms, e.g., those based on artificial neural networks or other complex models, are black boxes that make inferences and decisions that are not understood even by their developers. There are questions as to whether those that evaluate such technologies for use in healthcare, or those that must use such technologies, can be asked to make decisions made by such black-box algorithms· * There may be the emergence of “peer disagreement” between two competent experts – an AI machine and a doctor. In such situations, there is no means of combining the decisions or of reasoning with the algorithm, as it cannot be accessed or engaged to change its mind. There are also no clear rules for determining who is right, and if a patient is left to trust either a technology or a physician, the decision may depend on factors that have no basis in the “expertise” of the machine or the doctor. Choosing one of the two leads to an undesirable outcome. * There is a widely held convention that many algorithms, e.g., those based on artificial neural networks or other complex models, are black boxes that make inferences and decisions that are not understood even by their developers. There are questions as to whether those that evaluate such technologies for use in healthcare, or those that must use such technologies, can be asked to make decisions made by such black-box algorithms.   Clinical evaluation of AI technologies for health may therefore wish to ask several questions:  · How can clinical evaluation of an AI technology account for the possibility of peer disagreement, and what guidance can clinical evaluation provide to facilitate appropriate use and to address any possible conflicting opinions?  · What standard of transparency and explainability should clinical evaluation of an AI technology require? What are the positive benefits of such transparency and explainability, and in what circumstances might such transparency and explainability be less desirable? |

# Deployment and ongoing evaluation

*Deployment*

AI systems may be deployed earlier in their evaluation process than some traditional interventions. First, there is a demand from health systems to accelerate technological solutions through development to address crisis points of serious health need for which the capacity of human resources is inadequate and worsening, such as in screening programmes. Second, is a recognition that some factors, notably the question of generalisability, will only be adequately evaluated during wide-scale deployment.

Generalisability is a significant concern in AI systems, whether examples of interventions under-performing or even catastrophically failing when moved from one population or setting into another. There is a need therefore to ensure that evaluation is continued into the deployment phase and to continue for as long as the product continues to be used. It is in this deployment phase that the limitations of generalisability and any need for further training or local tuning should be actively sought, as a critical part of ongoing evaluation for efficacy and safety. The datasets used to train and test the AI system must be well described, ensuring transparency as to the characteristics of the datasets including its diversity. If the characteristics of the test population are not representative of the population into which it is intended to be deployed, there is an increased risk that the AI system will exclude or not function appropriately on behalf of unrepresented people when it is deployed.

The risk of harm arising from poor generalisability and other performance issues can be considered in terms of a risk matrix of likelihood, and consequence. The likelihood of a reduction in performance of the AI system will be increased by the differences between the populations and settings of the deployment phase compared to the test population and setting. Very rapid scaling such as moving to a full nation-wide roll-out based on a successful single center study in a homogenous population would have a high risk of failure. Pre-deployment evidence of likely generalisability and associated risks should be actively sought through algorithmic validation.

Some regulators and health systems are exploring novel approaches which may permit earlier deployment under more limited approval, and then with permission for wider scale deployment under less stringent monitoring as increasing safety data becomes available across an ever-increasing, diverse group of subjects. The adoption of silent trials - where the AI system is present within the care pathway but not acted upon - may have some value in testing deployment aspects (and acquiring data in a real-world setting) as an intermediary step before full deployment.

The deployment phase also provides greater ‘real world’ information regarding many of the impacts discussed earlier (*Section 5: Clinical validation*), such as outcome measures, process measures and balancing measures, providing a fuller assessment of both intended benefits and unintended consequences. One of the challenging areas of the deployment phase - and of particular relevance to regulators - is to determine the level of additional evaluation required to appropriately assure version updates of AI products, and, by extension, continuously learning or adaptive algorithms. Regulators have recently responded to this challenge by launching consultations and announcing new guidance and legislation. The US FDA announced in January 2021 a comprehensive action plan on regulatory approval strategies for adaptive AI/ML-based SaMD. The overall approach is based on standard regulatory principles, as applied to non-AI medical devices, including the device risk categorization principles, the benefit-risk framework, guidance on software modifications guidance, and the organization-based total product lifecycle approach. This is enhanced through the enforcement of “Good Machine Learning Practice '' (GMLP) principles, which are to be used to ensure rigorous AI/ML-based SaMD development. Algorithm changes must be transparently labelled for users and methodologies for ensuring robustness and identification and elimination of bias will be incorporated. For each device, a two-component Predetermined Change Control Plan (PCCP) is envisioned. This will include a SaMD pre-specification (SPS) - a predetermined change control plan setting out the scope of the permissible modifications, and secondly an Algorithm Change Protocol (ACP), which sets out the methodology used in the AI/ML-based SaMD to implement the defined changes within the scope of the SPS. The ACP is a step-by-step delineation of procedures to be followed so that the modification achieves its goals and the AI/ML-based SaMD remains safe and effective. The action plan is notable for its strengths in harnessing the iterative improvement power of AI/ML-based SaMD, whilst, at the same time, ensuring patient safety through continuous Real-World Performance (RWP) monitoring (the principles of RWP are set out in the section on ongoing monitoring below). As yet, the US FDA approach is described in low level detail only and a complete Guidance on the PCCP will be published in 2021. The EU published a draft of a new EU Artificial Intelligence Act in April 2021, which sets out similar principles to the SPS and ACP, albeit less comprehensively described than the US FDA proposals.

|  |
| --- |
| *Box 6: Ethical considerations on risks of AI to patient safety*  Patient safety could be at risk from use of AI that may not be foreseen during regulatory review of the technology for approval. Errors in AI systems, including incorrect recommendations (e.g., which drug to use, which of two sick patients to treat) and recommendations based on false-negative or false-positive results, can cause injury to a patient or a group of people with the same health condition. Model resilience, or how an AI technology performs over time, is a related risk. Health-care providers also commit errors of judgement and other human errors, but the risk with AI is that such an error, if fixed in an algorithm, could cause irreparable harm to thousands of people in a short time if the technology is used widely.  An AI application, like any information technology system, could also provide the wrong guidance if it has code errors due to human programming mistakes. It is also possible that a developer (or an entity that funds or directs the design of AI technology) designs an AI technology unethically, to optimize an outcome that would generate profits for the provider or conceal certain practices. The design might in fact be more accurate than another modelling technique but generate unmerited sales revenue.  Clinical evaluation may present an opportunity to anticipate and address any possible concerns to patient safety. Some questions which clinical evaluation may wish to consider include the following:  · How can clinical evaluation best anticipate and address possible risks to patient safety?  · How can clinical evaluation consider risks to patient safety that may emerge over time, especially as an algorithm may evolve after real world use? |

*Ongoing Evaluation*

Monitoring of ongoing performance (both safety and effectiveness) is important to determine whether the AI product continues to deliver as expected. 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. Additionally, variations in clinical workflow may negatively impact on the overall intended benefits of an AI system.

A key change after deployment is that performance monitoring is no longer the sole responsibility of product developers and regulatory authorities, but users, patients and the public also become gatekeepers for discovering and acting upon potential risks.

*Regulatory requirements*

Regulatory authorities stipulate that manufacturer of medical devices, including AI included in the software as medical device (SaMD) category, should systematically carry out post-deployment monitoring of safety and performance and carry out necessary corrective action when required. A post market surveillance plan, such as that required by the Medical Device Regulation (EU) 2017/745 (MDR) and outlined in MEDDEV 2.7.1 Rev 4 guidance (clinical evaluation), states that manufacturers need to plan for monitoring expected and unexpected adverse events, contraindications and instances of misuse, throughout the AI system’s life cycle and in alignment with findings of the clinical evaluation report.

Reported adverse events (including suspected device-related deaths, injuries and malfunctions) are recorded in regulatory databases such as the US FDA Manufacturer and User Facility Device Experience (MAUDE) database and the UK Medicines and Healthcare products Regulatory Agency (MHRA) Alerts and Recall database for Medical Devices. It should be noted, however, that the registration of post-deployment performance issues in these databases are dependent upon either the manufacturer’s ongoing monitoring of the device’s performance, or for an adverse event to be detectable *and* attributable to the AI device in question. These points are worth specifically highlighting in the context of clinical evaluation, as post market surveillance and post market clinical follow up may only detect adverse events supported by attributable harm, and those where causality cannot be established may remain unreported. It is also important to be aware that AI as diagnostic or prediction tools may cause harms which only become apparent downstream in the clinical pathway (for example, where an incorrect diagnosis first results in incorrect treatment, which in turn results in a poor outcome). In such cases, it may be difficult to trace the mechanisms of causality back to the AI system. AI manufacturers may, as part of their Post Market Clinical Follow up plan, monitor residual risks by collecting post-deployment data to establish ongoing safety or performance issues which still need to be addressed.

*Relevant stakeholders for post deployment monitoring*

Users and developers of AI systems will be the two most active stakeholders engaging in post deployment monitoring. Whilst there is no responsibility on the users (particularly patients and the public) to participate in reporting adverse events, they are often the first-in-line to discover problems arising post-deployment. As such, systems and processes which enable direct and transparent reporting of adverse events should be in place and users should be supported and encouraged to report openly.

Definition of the user may straddle a wide range of groups, including patients, the public, medical professionals or other non-medically qualified healthcare professionals. This should be stated in the intended use and indications for use statement and can inform the level of post-deployment surveillance the user can feasibly contribute to. Other important stakeholders include regulators, auditors (including external independent auditors), health institutions, funders and commissioners. Developers of AI systems should support open reporting by creating mechanisms to facilitate error reporting and user feedback. Such feedback should be made openly available by developers to all users and stakeholders.

*Algorithmic audits*

As well as discovering the occurrence of adverse events, a further step should be taken to understand *why* the events happened. This is an important consideration for two reasons: 1) AI systems are highly sensitive to characteristics within its input data and have a tendency to learn spurious correlations during algorithm training (relationships within the data which appear useful in the training context but are unreliable when applied to real-world inputs). This means AI systems may perform exactly as predicted the majority of the time, yet fail in a few instances when encountering unusual, rare, or previously unencountered cases. In such cases, close interrogation of the error case may reveal previously unknown weaknesses of the AI system which require future systematic error-proofing (either through modification of its intended use statement or to the algorithm itself). 2) The error may have arisen not from the AI system itself, but from the *way* it was implemented. Variations in clinical workflows, user training and guidance for decision-making may impact upon the algorithm’s performance, and may arise due to intended/unintended misuse or a lack of specificity in the AI product’s instructions for use.

To determine what, how and why adverse events or algorithmic errors occurred, detailed analyses may be performed through an algorithmic audit (NL: add ref from Liu et all in press at TLDH). Through the audit, existing and potential risks can be assessed and prioritized, risk mitigation plans can be put in place, and future audits can monitor whether risk mitigation measures were successful in avoiding harm. Algorithmic audits are particularly well-suited for local performance monitoring (such as in a hospital) where clinical workflows and populations vary. They can be used to establish a baseline performance and repeatedly performed over time to measure deviation from the expected baseline. Aside from safety concerns, algorithmic audit may also be an appropriate method for monitoring performance across different population groups (such as those with protected identities or social determinants), cost effectiveness, health service delivery effectiveness and user experience.

# Economic evaluation

*Introduction*

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

Economic evaluation is defined as a comparative analysis of two or more interventions in terms of their costs and consequences [REF Drummond 2015] and enables the assessment not only of the effectiveness of a health intervention, but the costs of achieving the effect. Economic evaluation has therefore become an essential tool for generating information for funding decisions about health interventions, whether these investments are made by governments, individuals, companies or donors and development partners.

The fundamental concept guiding use of economic evaluation is opportunity cost – the foregone benefits of investing limited resources on one course of action rather than another. By quantifying the costs (including implementation and running costs) relative to the outcomes through economic evaluation, and quantifying uncertainty associated with estimates, the best course of action can be identified from within available funding; depending on the objectives of the decision maker whether these are to improve health outcomes or patient experience,, improve access and equity or other objectives.

The intent of economic evaluations is to help inform policy and clinical guidelines so that health interventions, including AI technologies, that have the potential to improve patients’ lives are recognized, and neither patients nor society overpays for care that does not offer such potential.

Opportunity cost is blind to the type of intervention being considered, and so decision making about investment in any health intervention can benefit from some approach to economic evaluation, whether that intervention is a simple once-a-day medication, a complex public health program or an AI health technology. However, the assessment of the costs and consequences of some health intervention are more straightforward than others. Much of the development of methods for economic evaluation of individual health interventions were centered around pharmaceuticals, driven by a need by country governments to make evidence-informed and definable decisions. Although economic evaluation of pharmaceuticals can be highly complex, there are a number of aspects to the generation of evidence related to pharmaceuticals that make it more amenable to economic evaluation, such as an established regulatory framework, a static product and therapeutic action, a more predictable life-cycle and that the physical product (tablets or injections) satisfy the notion of a private good- i.e., where consumption by one individual prevents consumption by another.

Conducting economic evaluation of AI health technologies is more complex than the economic evaluation of traditional non-digital health interventions such as pharmaceuticals which commonly have a more static clinical and cost profile. Of course, many health interventions will differ in marginal cost at scale and clinical effect may alter over time based on health professional experience in use in established clinical pathways, however, the dynamics of clinical effects and costs associated with AI health technologies is unique. AI-enabled DHIs have a distinct cost profile, where innovation or development costs are substantial and, in some scenarios, at-scale marginal cost that can approach zero. Conversely, the effect is not static and is likely to improve with more use of data. Costs and effects are also highly dependent on the local digital architecture and infrastructure, meaning that a generalized approach to economic evaluation (i.e., across a region or grouping of differing contexts) introduces substantial uncertainty. In addition, an AI-enabled DHI will, by definition, produce information through its use – this revolutionizes the real-world evidence (RWE) area of methodological research, enabling economic evaluations to be informed by evidence beyond the clinical trial setting and incorporate evidence from use in clinical practice. While there are multiple guidances and texts on methods for economic evaluation of non-digital health interventions, there is relatively limited research on methods for AI health technologies. To address this methodological challenge, the World Bank is engaging in a comprehensive and collaborative effort to develop a framework for economic evaluation of digital health technologies to inform investment, building on existing methodological advancements in the field.

*Types of economic evaluation for AI-enabled digital interventions*

As an economic evaluation is principally an information-generating activity, an important consideration is the objectives of the evaluation and who will be the recipients or users of the information produced. In the case of national health technology assessment agencies, the objective of an economic evaluation is to inform use of limited resources across the health system, commonly supported by overarching principles including universal health coverage (UHC) and health equity. In this scenario the concept of allocative efficiency is important where the country wishes to distribute resources in a way that maximizes outcomes. This has given rise to the common use of cost utility analysis (CUA), where the net incremental costs of an intervention are presented as a ratio to net incremental health’s outcomes, and health is a generalized measure such as the quality adjusted life year (QALY) or disability adjusted life year-averted (DALY). In this way a country can assess the likely impact that an intervention can have on “health” - where health is comparable across interventions and diseases, incorporates both positive and negative impacts and represents mortality and morbidity. As the summary output of an economic evaluation is commonly the ratio of costs relative to effects, it is important to ensure that approaches to estimating costs are as robust as the methods for assessing clinical impact.

The basic premise of costing for economic evaluation is that costs should reflect full net costs of the intervention, aligning to the specification of the intended decision maker. This requires therefore that when estimating costs, the decision problem, perspective of the decision maker, an understanding of what the intervention would displace in the context of the decision problem should be established. The costing approach for AI health technologies is complicated by a dynamic costing structure. Many AI health technologies will involve substantial up-front development and implementation costs, with decreasing marginal costs potentially close to zero at scale. For example, an AI-enabled diagnostic technology may involve high development, data training and validation costs, but one implemented and at scale, the cost per additional patient diagnosed is likely to be negligible.

Increasingly, budget impact analysis (BIA) forms an essential as part of the economic evaluation. BIA estimates the extent of health technology uptake and financial implications of investing in a health intervention in a particular context. BIA addresses anticipated expenditure changes (commonly usually over a 3-to-5-year period) to a specific budget holder that are coupled with a decision to reimburse a new health technology, (York Health Economics Consortium). BIA entails addressing the estimated use and costs of the proposed health intervention medical service, an estimation in the changes in use and cost of other health interventions medical services (from a budget holder perspective), possible off-label use (applicable to SaMD) of the new intervention, accounting for any pre-requisite interoperability requirements (e.g., MRI hardware and software upgrades, to support the adoption of new AI technologies, that can either be a minor or significant cost) and addressing uncertainty in terms of model parameter inputs and structural uncertainty underpinned by certain assumptions (ISPOR). Data sources to inform BIA include cost data from registries, real-world use and data from clinical trials specific to the budget holder population and expert opinion.

*Reimbursement*

After regulatory approval confirming the safety and effectiveness of a health technology permitting market access, and prior to establishing a significant market penetration, third-party coverage decisions and establishing a reasonable level of reimbursement is required. Pricing of digital health technologies, like other commodities, influences both affordability and access. Initial price setting approaches include price skimming or price penetration (Ingenbleek 2013) and patents enable a period of market exclusivity to recoup R&D costs by delaying the entry of competition, which can also have other undesirable impacts, including enabling those who hold intellectual property to assess unaffordable prices. Patent monopolies, to the extent that it encourages a patent holder to recuperate investments through high prices, discourages investments in AI technologies that could serve marginalised populations or communities too poor to pay for such technologies, since such target populations do not represent an attractive market.

Unlike other typical commodities, there is an imperfect market at play for medical technologies, including DHTs; in both a universal health care system scenario or private health insurance scenario, the consumer (patient) typically does not incur the full cost of the product and further special characteristics of the medical care are described by Arrow (1963). Subsequently, in most countries either the government and/or health insurers exercise a degree of influence to the price and utilisation (through coverage/restrictions of indications (Drummond et al. 1997). Payers often re-evaluate safety and effectiveness evidence as part of the deliberation process; with an objective to reward innovation whilst achieving optimal resource allocation (Barros 2020). Digital health technologies may come in the form of capital associated with a one-off payment (e.g., MRI/CT), or software as a medical device (SaMD), that may be associated with existing capital (for example CT scanner) and the SaMD product is paid on either a subscription or fee-for-service basis. For digital health technologies there are limited national reimbursement opportunities and, within the UK, local negotiations with clinical commissioning groups (CCGs) are required. In the UK, the Innovation and Technology Payment (ITP) is a national reimbursement mechanism for DHTs. To be eligible, products must have NICE support (i.e., MedTech Innovation Briefing or Guidance), demonstrate positive in-year return on investment and be used in at least three NHS organisations (Innovators Guide 2020). [enter here reimbursement informed by HTA findings, and that said, HTA need to have skill and expertise to evaluate. See comment above.

# Communication of results

Communicating the results of the steps of the clinical evaluation process transparently is fundamental to the safe and effective use of health AI. Clinicians and other stakeholders using models need access to information on which to understand the model and its performance.

1) Transparent reporting of datasets - those used in training, testing and external validation steps - enable full evaluation of the model and its performance. There is currently no standardised process for documenting datasets, but proposals include ‘Data sheets for Dataset’s to describe the datasets (Gebru Timnit, <https://arxiv.org/abs/1803.09010>). The development of standard methods of communication for medical datasets, would improve transparency and accountability, and national

2) Similarly, standardised and transparent descriptions of models, including input and output data, type of model and output of validation steps need to be transparently communicated to aid safe use of AI technologies. 'Model fact cards' to describe the models are an example of this reporting. (Sendak)l[xxvii]

|  |  |
| --- | --- |
| Description of dataset and model characteristics | Size of dataset, and demographics of population (age, gender/sex, race, ethnicity)  Setting data was collected (type of facility, date of collection).  Input and output data of model, type of model. |

3) Transparent reporting to describe the clinical evidence (using the appropriate EQUATOR guideline and standard performance metrics)

4) Clear reporting of audits to describe the post-deployment performance (e.g., Liu et al and the WHOITU AI4H Audit template) [xxviii]

Based on the framework proposed in this document for clinical evaluation, any communication of results should include the information required to.

|  |  |
| --- | --- |
| Evaluation of model purpose and suitability | A description of the clinical problem with intended benefits and potential risks of AI use.  Evaluation of interoperability and security.  Description of stakeholder engagement and user testing.  Consideration of ethical principles applicable. |
| Algorithmic Audit | Description of data used for development and performance metrics for internal testing.  Description of data used for external validation and performance metrics obtained. |
| Clinical Validation | Description of the clinical evaluation including the study design (registration and location of report), population, setting, intervention (including inputs and output of model), comparator and pre specified outcomes. |
| Ongoing Monitoring | Description of ongoing monitoring in place. How will adverse events be collected? When will the model be audited and by whom? |
| Regulatory considerations | Is regulatory approval required for the model? If so, where is it approved and what is the use case it is approved for? |
| Economic Evaluation | A description of the economic evaluation carried out. |

Table 1: Overview of required information for communication of results

# Recommendations

1. Ensuring that AI systems that are adopted into health systems are effective, safe, ethical, inclusive and fair is a global concern and requires input from a wide range of stakeholders. Clinical evaluation and transparent communication regarding the use of AI systems, their underpinning data, their performance and their safety systems is critical to deliver this.
2. All stakeholders must be encouraged to carry out effective clinical evaluation (including the steps of algorithmic validation, clinical evaluation and ongoing monitoring) and make the results open and accessible. Availability of the clinical evaluation plan is key to building trust and should be made publicly available.
3. Algorithmic validation, including benchmarking of AI models either by locally, nationally or more widely is important to ensure local performance is acceptable, and to compare performance of tools on particular datasets. The Open Code Initiative is part of the FG-AI4H that is actively developing software that can be used by stakeholders to do this on their own datasets.
4. Clinical studies, especially those with long term evaluation of AI tools is required- despite the expanding number of tools, there is a paucity of long-term studies with clinical endpoints and rigorous safety analyses and this is holding back the potential of these tools. There is a lack of evidence on how transferable tools are that are developed in one setting and then used in another. Collaborative studies would accelerate progress and should be considered a priority.
5. Needs based development of AI models requires a dedicated effort to collect data in populations that are currently underrepresented and for clinical problems where AI may be effective, but datasets are poor. Not only does this facilitate development but also evaluation. Organizations like I-DAIR have identified this as a key task.
6. Procurers of AI tools should be clear about the economic evaluation that is required for AI tools. Not only is this important to ensure comprehensive evaluation, but it also builds trust in the evaluation system. While it is possible to find examples of where digital strategies explain what kind of evaluation should be done, it is still rare to find publicly available examples of where this has been done, and this can undermine public trust in procurement.
7. Priority setting for digital tools in all country settings requires a much more active role for health technology assessment, in addition to the role of regulators. Where fulfilled by two different agencies, close dialogue should be maintained. Where HTA agencies already exist, these need to expand the workforce skills to include evaluation of AI technologies. In countries where HTA capacity is low, this should be a key focus alongside expanding digital health tool.

Bibliography

***References for clause 10 (Economic evaluation)***

Guidelines

1. CHEERS statement
2. ISPOR budget impact guidelines
3. NICE Budget impact template
4. JAMA: Recommendations for Conduct, Methodological Practices, and Reporting of Cost Effectiveness Analyses

References

Kaltenthaler, E., Tappenden, P., Paisley, S., Squires, H. (2011) NICE DSU Technical Support Document 13: Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models. Available from [http://www.nicedsu.org.uk](https://eur01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.nicedsu.org.uk%2F&data=04%7C01%7C%7Cff125943d935430175e508d8d5d2f294%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C637494449337639571%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=6i8lP8iLhvarjZAnD6LM5aX2h1rNYnrI3hu9kJM%2Bk4g%3D&reserved=0)

Budget Impact Analysis [online]. (2016). York; York Health Economics Consortium; 2016. <https://yhec.co.uk/glossary/budget-impact-analysis/>

An Innovator's Guide to the NHS: Navigating the barriers to digital health. 2020

Best Practices for New Product Pricing: Impact on Market Performance and Price Level under Different Conditions Ingenbleek et al. 2013

The role of economic evaluation in the pricing and reimbursement of medicines Drummond et al. 1997

Uncertainty and the Welfare Economics of Medical Care Arrow, 1963 (seminal article)

Incentives for R&D: Payment Options and Pricing Challenges Barros OHE, 2020

Methods for the Economic Evaluation of Health Care Programs. Fourth Edition. Drummond, Sculpher, Claxton, Stoddart and Torrance, 2015. Oxford University Press.

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

[i] Progress in evidence-based medicine: a quarter century on. Djulbegovic, Benjamin et al. The Lancet, Volume 390, Issue 10092, 415 – 423<https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31592-6/fulltext>

[ii] Is digital medicine different? The Lancet, Volume 392, Issue 10142, 95<https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31562-9/fulltext>

[iii] <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx> Accessed 2/8/21

[iv]<https://sdgs.un.org/goals> Accessed 2/8/21

[v]<https://www.who.int/publications-detail-redirect/9789240029200> Accessed 2/8/21

[vi] [David B. Larson](https://www.ncbi.nlm.nih.gov/pubmed/?term=Larson%20DB%5BAuthor%5D&cauthor=true&cauthor_uid=33096088), MD, MBA,a,∗ [Hugh Harvey](https://www.ncbi.nlm.nih.gov/pubmed/?term=Harvey%20H%5BAuthor%5D&cauthor=true&cauthor_uid=33096088), MBBS,b [Daniel L. Rubin](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rubin%20DL%5BAuthor%5D&cauthor=true&cauthor_uid=33096088) et al. [J Am Coll Radiol.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7574690/) 2021 Mar; 18(3): 413–424.<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7574690/>

[vii] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. https://doi.org/10.1038/s41591-020-1034-x

[viii] Cruz Rivera, S., Liu, X., Chan, AW. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26, 1351–1363 (2020). https://doi.org/10.1038/s41591-020-1037-7

[ix] Wahl B, Cossy-Gantner A, Germann S, et alArtificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?BMJ Global Health 2018;3:e000798.<https://gh.bmj.com/content/3/4/e000798>

[x] IDEO’s Design Kit for ways to identify and understand a problem and Google’s People and AI Guidebook on how to decide if AI is the right solution

[xi]<https://www.gov.uk/guidance/user-feedback-study-descriptive-studies>

[xii]<https://www.gov.uk/guidance/interview-study-qualitative-studies>

[xiii]<https://www.gov.uk/guidance/usability-testing-qualitative-studies>

[xiv]<https://www.gov.uk/guidance/focus-group-study-qualitative-studies>

[xv]<https://www.gov.uk/guidance/ethnographic-study-qualitative-studies>

[xvi] Macdonald J., März M., Oala L., Samek W. (2021) Interval Neural Networks as Instability Detectors for Image Reconstructions. In: Palm C., Deserno T.M., Handels H., Maier A., Maier-Hein K., Tolxdorff T. (eds) Bildverarbeitung für die Medizin 2021. Informatik aktuell. Springer Vieweg, Wiesbaden. https://doi.org/10.1007/978-3-658-33198-6\_79

[xvii] Hägele, M., Seegerer, P., Lapuschkin, S. *et al.* Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci Rep* **10,** 6423 (2020). https://doi.org/10.1038/s41598-020-62724-2

[xviii] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019. <https://doi.org/10.1016/S2589-7500(19)30123-2>.

[xix] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.

[xx] Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998;316(7126):201.

[xxi] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869

[xxii] Cruz Rivera, S., Liu, X., Chan, AW. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26, 1351–1363 (2020).

[xxiii] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. <https://doi.org/10.1038/s41591-020-1034-x>.

[xxiv] Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, Karthikesalingam A, King D, Liu X, Markar SR, McInnes MDF, Panch T, Pearson-Stuttard J, Ting DSW, Golub RM, Moher D, Bossuyt PM, Darzi A. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med. 2020 Jun;26(6):807-808. doi: 10.1038/s41591-020-0941-1. PMID: 32514173.

[xxv] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577–9.

[xxvi] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. <https://doi.org/10.1038/s41591-020-1034-x>.

[xxvii] Sendak, M.P., Gao, M., Brajer, N. *et al.* Presenting machine learning model information to clinical end users with model facts labels. *npj Digit. Med.* **3,** 41 (2020). https://doi.org/10.1038/s41746-020-0253-3

[xxviii] ITU/WHO FG-AI4H-J-048, DAISAM Audit Reporting Template, [https://aiaudit.org/assets/‌pdf/standards/FGAI4H-J-048.pdf](https://aiaudit.org/assets/pdf/standards/FGAI4H-J-048.pdf)

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