|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-N-048 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | WG-CE | E-meeting, 15-17 February 2022 | |
| **DOCUMENT** | | | | |
| **Source:** | | Editors DEL7.4 | | |
| **Title:** | | DEL7.4 Update: Clinical evaluation of AI for health | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Co-chair, WG-CE Naomi Lee The Lancet United Kingdom | | Email: [naomi.lee@lancet.com](mailto:naomi.lee@lancet.com) |
| **Contact:** | | Co-chair, WG-CE Shubhanan Upadhyay Ada Health Germany | | Email: [shubs.upadhyay@ada.com](mailto:shubs.upadhyay@ada.com) |
| **Contact:** | | Co-chair, WG-CE Eva Weicken Fraunhofer Heinrich-Hertz-Institut Germany | | Email: [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".* 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 N, 15 – 17 February 2022:**
  + **Implementation of written & verbal WG-CE feedback**
    - Recommendations
    - summary table & executive summary
    - Discussions on cybersecurity, terminology
  + **Production of final version 1.1 summary of changes**
    - Simplify the executive summary
    - Restructuring/Condensing of sections, e.g., recommendations, ethics boxes
    - Revision of diagram
    - Reference section

The document is based on five earlier versions:

* **FG-AI4H meeting M, 28 – 30 September 2021**, [M-040](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7B9980F36F-6F9D-4898-8733-B3A9004D4927%7D&file=FGAI4H-M-040.docx&action=default)
  + 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)
* **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**: [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)
* Deliverable draft "*Clinical Evaluation of AI for Health”* J-053submitted 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 February 2022)**

* 08 November 2022: Follow-up meeting with all WG-CE members
* 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, and review rounds the writing group drafted the outline over time.

External expert group (in alphabetical order):

AbouElkhir Osama, Akogo Darlington, Allen Megan, Alsalamah Shada, Balachandran Pradeep, Bastawrous Andrew, Bathke Arne, Chiavegatto Filho Alexandre, Darkoh Ernest, Ehrenfeld Jesse, Fehr Jana, Fürstenau Daniel, Gaudin Robert, Gilbert Stephen, Greaves Felix, Gupta Saurabh, Gütter Zdenek, Hatton Grace, Ho Dean, Ibrahim Hussein, Islam Shariful, Jarral Reza, Jeon Jonghong, John Oommen, Kadam Rigveda, Kherif Ferath, Kuku Stephanie, Kurtys Michal, Lapão Luís, Linder Nina, Loh Irv, Loveys Kate, Magrabi Farah, Mahajan Arnav, Malpani Rohit, Mamun Khondaker, Masud Jakir Hossain Bhuiyan, Matin Rubeta, Matthew Fenech, McCradden Melissa, Menezes Audrey, Murchison Andrew, Murphy Lisa, Nakasi Rose, Oala Luis, Pankova Natalie, Piekut Agata, Porras Lina, Reddy Sandeep, Salim Ally, Schwendicke Falk, Sethi Tavpritesh, Sood Harpreet, Sousa Inês, Srivastava Manish, Starlinger Johannes, Wasswa William, Werneck Leite Alixandro

Writing group included (in alphabetical order):

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

**CONTENTS**

| Page |
| --- |
| [1 Scope 7](#_Toc95750323)  [2 References 7](#_Toc95750324)  [3 Terms and definitions 7](#_Toc95750325)  [3.1 Terms defined elsewhere 7](#_Toc95750326)  [3.2 Terms defined here 7](#_Toc95750327)  [4 Abbreviations 8](#_Toc95750328)  [5 Introduction and Background 9](#_Toc95750329)  [6 Model design and purpose 15](#_Toc95750330)  [7 Algorithmic Validation 19](#_Toc95750331)  [8 Clinical validation 25](#_Toc95750332)  [9 Deployment and ongoing evaluation 28](#_Toc95750333)  [10 Economic evaluation 33](#_Toc95750334)  [11 Communication of results 38](#_Toc95750335)  [12 Conclusions and recommendations for future action 40](#_Toc95750336)  [References 49](#_Toc95750337) |

**List of Tables**

| Page |
| --- |
| [Table 1: Summary of most important elements of the key components of clinical evaluation 38](#_Toc95750338)  [Table 2: Summarised guidance for clinical evaluation of AI systems in health 41](#_Toc95750339) |

**List of Figures**

| Page |
| --- |
| [Figure 1: Framework for evaluation of AI technologies in health 15](#_Toc95750340) |

FG-AI4H Deliverable DEL7.4

Clinical evaluation of AI for health

Executive Summary

Ensuring that AI systems that are adopted into health systems are effective, safe, cost effective, ethical, inclusive, and fair is a global concern and requires input from a wide range of stakeholders. Clinical evaluation of AI systems including their underpinning data, performance, and safety, and transparent communication of these results are critical to delivering this.

Artificial intelligence (AI) in healthcare could hold great promise to improve people’s health worldwide by transforming the screening, diagnosis, therapy and monitoring of diseases. The increasing amount and availability of digitized health data has facilitated the use of AI which can be used to analyse large datasets, provide new insights, and identify patterns which can be extrapolated to unseen data. There are already many potential applications for AI in medicine and considering factors such as the global shortage of healthcare professionals, changing population demographics worldwide, and ongoing global digital transformations there is huge interest in the potential of AI systems in both high- and low-resourced settings. Acheiving the potential beneficial impact requires frameworks for evaluating AI systems, in order to ensure that they are safe, effective, useful and that they do not cause unanticipated harm when applied into a complex clinical pathway or when used autonomously, and that the costs and ethics are adequately considered.

Working from the principles of evidence-based medicine but acknowledging the particular challenges and opportunities of artificial intelligence-based technologies, this document provides a framework for the evaluation for AI systems in health that can be used by clinicians, researchers, developers, regulators and policy makers to understand whether a particular AI system is likely to be effective and safe in their setting. It was developed by members of the World Health Organization (WHO) / International Telecommunications Union (ITU) Focus Group on AI for Health [FG-AI4H] (Wiegand, 2020) Working Group on Clinical Evaluation and is part of a series of guideline documents (deliverables) produced by FG-AI4H. In keeping with the WHO stated goal to ‚leave no one behind’ the group gave special considerations to low resourced settings when creating the framework and recommendations that draw on current best practices and also identify potential gaps for future research.

The framework for clinical evaluation divides evaluation in 4 phases of: evaluation of model purpose and suitability, algorithmic validation, clinical validation, and ongoing monitoring, while also drawing attention to the essential requirements of ethical and economic evaluation which cut across those 4 phases.

Evaluation of model purpose and suitability requires i) an understanding of the problem and intended use of the AI system ii) definition of the intended benefits iii) description of the potential risks and harms iv) documentation of interoperability and security v) user testing and user engagement reports. Algorithmic validation (used here to refer to evaluation of the AI system ‚in silico‘) requires i) a description of the data used for development, internal and external testing, and of the model type used ii) reporting of performance metrics in the internal and independent external testing data iii) benchmarking of system performance against standard of care, and where relevant, other AI systems. Clinical validation (for the purposes of this document this is evaluation of the AI system through interventional or clinical studies) requires a clinical study with a relevant comparator and a meaningful endpoint, and steps taken to minimise bias. Finally, deployment and ongoing evaluation requires i) monitoring or performance (safety and effectiveness) ii) algorithmic audits to understand how adverse events or algorithmic errors occur.

# 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

*Please note that the references are listed at the end of this document*

# Terms and definitions

## Terms defined elsewhere

This document uses the following terms defined elsewhere: **[tbc]**

## Terms defined here

This document uses the following terms defined here: **[tbc]**

# 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 healthcare professionals, a rapid growth in health data, and an expansion in the usage of artificial intelligence (AI) systems in other sectors has contributed to an increasing interest in the use of AI for health and clinical practice. As a growing number of health AI models, methods, and tools become available for use, researchers, patients, clinicians, and policy makers require a framework to understand whether they 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 technologies in clinical use is complex 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 AI models for health, that include an evaluation of the underlying data, the potential for bias, and the contextual nature of the AI model performance (Djulbegovic & Guyatt, 2017). 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 (The Lancet, 2018).

This document is produced by the World Health Organization (WHO) and members of the WHO/ITU Focus Group on AI for Health [FG-AI4H] (Wiegand, 2020) 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 or national health agencies, developers including 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 other stakeholders including patients, the public, and developers of AI technologies. A shared understanding of evaluation best practice could help reduce uncertainty and 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.

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 [SDGs] (United Nations, 2015), particularly SDG 3 ‘Ensure healthy lives and promote 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 uses consistent nomenclature with other WHO - and FG-AI4H documentation to encourage a shared understanding and facilitate communication between stakeholders. Other working groups within the FG-AI4H are considering the topics of ethics, regulation, and data handling - and assessment methods including an Open Code Initiative (Open Code Initiative, 2020) which is developing a software platform for the assessment, and an associated ML auditing (*aiaudit.org*, 2020) process to test the applicability 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 Working Group (WG) on ethics of AI for health which considered the ethical, human rights, legal, and social concerns that should be taken into account when evaluating an AI system for health (including clinical medicine, public health, and operational management) has already reported. The full report “Ethics and governance of AI for health” produced by the WHO ethics expert group including members of the WG-Ethics can be found here (World Health Organization, 2021). A summary of the ethical principles relating to clinical evaluation, which has informed the framework for clinical evaluation is included here in Box 1.

The working group on regulatory considerations for artificial intelligence in health reports alongside this clinical evaluation working group and the clinical evaluation discussed in that report is based on the framework outlined here. The interplay between the two reports makes clear that clinical evaluation does not replace the system for regulatory approval but equips stakeholders with the framework necessary to evaluate the safety, effectiveness, cost, and performance of health AI models, in keeping with the principles of evidence-based medicine for their clinical problem, setting and population. These principles can be used by regulators, but also for AI systems that do not require regulatory approval, and to complement regulatory approval when considering the use of AI systems in individual settings. The framework seeks to address some concerns that have been voiced around existing regulatory frameworks, including key areas specific to AI systems and their context, that may not be adequately addressed, or may be less stringently applied such as: evaluation of frequently updated (and potentially continuously updating) models; detection and accounting for bias or poor quality underlying data; consideration of wider impact, notably the effect of the technology in the context of the care pathway and health system. The proposed evaluation framework strengthens transparency in those areas. For AI systems that are classified as medical devices, existing guidance by the IMDRF and from the WHO Digital Innovation and Health group (WHO et al., 2021) provides specific guidance and links the extent and depth of evaluation may relate to the estimated level of risk to patients (Larson et al., 2021). This document draws on, and dovetails with, these existing evaluation frameworks that are tailored to the requirements of evaluating AI health technologies of particular classes or types to provide an overarching framework that can be used to evaluate the breadth of AI systems in health [(Liu et al., 2020),(Cruz Rivera et al., 2020), (Sounderajah et al., 2020)].

The work of FG-AI4H is closely aligned to the principle that health care should be equitable and fair. The framework and best practice recommendations provided here looks to uphold this principle providing 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(Wahl et al., 2018). 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 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 many settings (e.g., rural areas, marginalised populations).

The framework for clinical evaluation also responds to ‚The Lancet and Financial Times Commission on governing health futures 2030‘ (Kickbusch et al., 2021) recommendation that in order for digital health innovation to deliver meaningful benefits in health for all, requires development of strong accountability frameworks. Calling for the development of a trust architecture, which leverages multilateral forum, such as the WHO and transnational, mutlistakeholder coalitions the Commission recommended the promotion of strong transparency and accountability requirements for emerging AI and machine learning applications in health. A lack of trust within society, and on an individual level, of digital health and AI systems, could undermine the willingness to use tools or to share personal health data, a key component of the Commission’s recommendation for the concept of health data solidarity. Delivering benefits for all and moving away from the current digital health ecosystem, which the Commission described as based on extractive or exploitative business models, and at risk of potential misuse and discrimination, particularly in politically unstable and autocratic circumstances. The Commission identified that many people have insufficient control of what purposes the data can be used for if shared and lack a system to assure quality of digital and AI health tools, requires a comprehensive framework for evaluation.

The WG-CE recommends that clinical evaluation of AI systems in health encompasses four key phases with two cross cutting themes (Figure 1). The phases of evaluation are: evaluation of model design and purpose, analytical validation, clinical validation and ongoing monitoring. The cross-cutting themes are ethical consideration and evaluation of cost effectiveness. These may be performed sequentially as the system is developed, but also can have steps performed together if a tool is already in developed and in use, for example when a system is considered for a new setting, problem or population, or changes are made in the AI system.

|  |
| --- |
| *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 which are described in their report “*Ethics and governance of AI4H”*:   * Protecting human autonomy * Promoting human well-being and safety and the public interest * Ensuring transparency, intelligibility and considering the requirements for explainability * Fostering AI technologies responsibility and accountability * Ensuring inclusiveness and equity * Promoting AI that is responsive and sustainable.     The report emphasizes the need for 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. The following ethical principles are incorporated into the recommendations for clinical evaluation in this report:  The AI system should meet standards of scientific validity, accuracy, and explainability/reproducibility 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.   * Consideration should also be given as to whether or not the use of an AI technology contributes to the decommissioning of existing services, without replacing the previous services to the same or better level. (I.e., policy makers and politicians often use the introduction of emerging technology as a means to decommission existing face-to-face services, but the new technology does not replace like for like and may not (initially) have the same reach) * 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, and biased (or selective) training data may preclude the use of AI technology for certain constituent groups (see also below). * 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 how barriers of language and availability of data (for model training, validation, and maintenance) may be addressed.   When performing a clinical evaluation of an AI systems, stakeholders may wish to consider the following questions:  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 healthcare professionals that make use of such technologies? * How was the data collected (was there misuse of data; was there appropriate consent for the collection of such data – however consent is defined)? * How should clinical evaluation assess 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)? * Does the examination and collection and use of health data follow the relevant governance structures of the reviewing body Is there a review and clearance of ‘inclusion and exclusion criteria’ for test data? * Does the examination and collection and use of health data follow the relevant governance structures of the reviewing body (PAHO, 2019) (Finddx.org, 2021) (World Health Organisation, Open Data Institute, 2021)? |

Diagram

Description automatically generated

Figure 1: Framework for evaluation of AI technologies in health

# Model design and purpose

Evaluating the clinical utility 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 problem, and demonstration that AI is the most suitable option to solve it (IDEO, n.d.) (*People + AI Guidebook*, n.d.). 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 (theiet, n.d.).

Similar to other digital health products, there are concerns that AI technologies may be applied indiscriminately and over optimistically as technological solutions, where addressing deeper social,

structural, economic, and institutional issues would have greater effect. 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 (box 1).

*1.* *Understanding the problem and intended use*

AI health technologies are 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 algorithmic and clinical validation steps taken are appropriate and sufficient for the intended use case.

The description of the problem and intended use should include:

I) 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 alone required, or does the model also need lab test results? For continuous variables, what specific cut offs are used and how are these justified?

II) A description of how and where the model would fit in the patient journey or clinical workflow. Who are the intended users of the model and who are the intended beneficiaries? What could the interaction between the technology and user look like? What effect would adoption of the AI technology have on the workflow and workload? What will the interaction between technology and user look like, and what is the level of autonomy (Lyell et al., 2021)?

III) A consideration of any special circumstances related to the intended users or context. For example, in pediatric age-groups there may be a need to consider child protection issues; in rural settings there may be a need to consider issues such as little or no internet provision, variations of clinical pathways in different regions, socio-cultural variations around data and technology affecting the willingness to design and implement AI tools.

*2. 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 of this may include:

* 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), quicker linkage from diagnosis to care, reduced out of pocket expenditure.
* Clinical workflow benefits could be a reduced administrative burden on health care professionals (HCPs), increased time to care, and provide a better HCP experience.
* Health system benefits could be efficiencies found or created in pathways, improved detection of cases, better allocation of resources, cost savings, addressing shortages of skilled HCPs.

*3. Describing potential risks and harms*

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

* Patient level risks might include harmful consequences due to misclassification, misdiagnosis, delayed care, under- or overdiagnosis or unnecessary treatment, or consequences of bias in the AI technology
* Clinical workflow risks, including removing safeguards, additional time, administrative or cognitive task burden for clinicians
* System level risks, for example the health economic costs of expensive technology, or the potential for technologies to direct people to expensive and unnecessary care to be replicated at scale across large groups of people

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, FDA, EU MDR). In general, the higher the risk class, the greater the requirement to demonstrate that a technology’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 (IMDRF, 2013) and its associated risk classification (IMDRF) that can inform subsequent regulatory, clinical requirements to a certain extent.

*4. Interoperability and security*

Interoperability requirements (such as minor and significant hardware and software upgrades) of AI technologies with other devices and IT systems are often overlooked but are an essential component of evaluation. Recognising that 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 compliance with standards, where developed, offer some assurances, but full evaluation depends on local requirements and the local context (DICOM Standard.Org, n.d.) (hl7FHIR, n.d.). There remains much work and implementation to be done to create common standards that would bring interoperability by default.

*5. User-testing and stakeholder engagement*

AI technologies that have had stakeholders engaged in the design following a user centered approach may be more 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 AI health technologies throughout the design and development process, and clinical evaluation. A mixed methods approach can be used, 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) (Uk, n.d.)

Privacy and Security:

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

# Algorithmic Validation

A key part of evaluating a potential AI health technology is to understand the profile of the training data used to develop a model, to assess the technical performance of the AI model when confronted with new data representative of its intended use and understand this performance in comparison to the current standard for that use case. For the purposes of this document, we use the term 'algorithmic validation' to describe this evaluation of the adequacy of the AI model ‚in silico‘ contrast to 'clinical validation' in which the whole AI health technology is evaluated in the context of the clinical pathway.

*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 (usually a health professional doing the same task), and potentially against other relevant comparators which may include other AI technologies. Standard of care will change over time, and in the future may no longer be the human alone but may be human with AI (augmented intelligence) or AI alone (autonomous AI) if these become validated and adopted. Performance metrics should be appropriate to the application. For example, for a diagnostic test this would include sensitivity and specificity, ideally with a full confusion matrix (true positive, false positive, true negative, false negative). Other measures such as area under the receiver operator curve (AUC) and area under the precision-recall curve (AUPRC) may also be helpful.

*Description of internal and external testing datasets*

AI models are highly dependent on the training data used to develop them. It is important to evaluate the data that has been used for training, tuning, internal and external validation and assessing the extent to which these datasets align to the intended use and the clinical outcome, 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, which is a major limitation.

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. The ratio of training and testing data should be described and justifiable.
  + A description of the input data type, and source, including where, when, and how 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. Where diagnosis of a clinical condition is difficult, the robustness of the labels may be poor, and a knowledge of the clinical condition and strengths and limitations of diagnostic pathways is important for context.

*External Validation*

For the purposes of this, and other FG AI4H documents, external validation refers to the process of evaluating the performance of the AI model using previously unseen, and independent data ‚in silico‘. This is in contrast to clinical validation through interventional or clinical studies.

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 reasons underlying these require investigation.

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(Macdonald et al., 2021) possible resulting degradations in predictive confidence or its learned decision heuristics and more (Hägele et al., 2020).

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 technology on an unseen dataset, it is possible to perform comparative benchmarking of AI technologies. This has been performed in a limited number of settings, but as the number of AI technologies increases, this may become increasingly important. Benchmarking against unseen datasets also has a number of potential uses beyond comparison of alternative AI technologies. 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 evaluate dynamic AI technologies constantly and quickly, without requiring full clinical evaluation for each iteration. Further, where clinical validation has been performed for an AI technology or a class of technologies, 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 commercial nature of many AI health technologies could pose barriers to this, requiring the evaluation by independent entities. It would also be important to demonstrate that the developers did not have access to the environment/platform on which the evaluations were run. In the benchmarking study, performance evaluation should also take into account if a black box AI technology is compared to an open source reproducible predictive or inferential tool. Technologies that are not being fully open and reproducible, may be required to have better performance to be attractive over the more open technology.

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 technology. Where local, regional, and national bodies are interested in evaluating AI technologies, they could hold their own hidden dataset to enable this external validation set (for example, an initiative underway by NHS, in the UK, to develop nationally representative datasets for some common AI use cases). Prioritising data collection could be an example of driving ‘needs based’ innovation as recommended by the 2020 Global Digital Health Partnership policy document (GDHP et al., 2020).

Some groups, including a group within the Focus Group are considering the requirements of stakeholders with regard to algorithmic validation of AI models in health and actively developing software tools for the production and consumption of such results [(Open Code Initiative, 2020) (*aiaudit.org*, 2020)].

*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 AI model alone on a dataset, would provide the AI model to the intended user and ask them to perform the intended task on test data with and without the AI model. This enables an understanding of the technology’s performance in the hands of the user.

***Special considerations***

*Building high-quality, representative datasets*

Quality algorithmic validation depends on the existence of and access to datasets that are sufficiently representative of the local population, and of sufficient quality with the required labels. This can be logistically difficult and requires appropriate consideration of the ethical aspects of data collection. Box 2 summarises the ethical considerations on data collection, data use, bias and discrimination which were considered by the WG-Ethics as part of the focus group.

This lack of health data for certain people, groups or even whole nations 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 and clear, targeted action.

Within populations, under-representation of people groups may lead to many harms, including exclusion (the AI health technology is recognised not to perform reliably in that group) or exposure to under-performance (which may be recognised 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 a growing pool of data) but which may only be present in more affluent populations, regions and 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, or where cultural beliefs exist around data collection and use (Rainie et al., n.d.). It also results in the potential that it will only be these populations that stand to benefit. Health data poverty – the inability for individuals, groups, or populations to benefit from a discovery or innovation due to insufficient data that are adequately representative - can play out at both ends of the economic spectrum (Mbchb et al., 2021). 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.

|  |
| --- |
| *Box 2:* *Ethical considerations on data collection, data use, bias, and discrimination*  In their report “*Ethics and governance of AI4H”* the WHO/ITU FG-AI4H working group on “Ethical considerations on AI4H” highlighted that ethical considerations around data use should recognise both, the potential benefits and the risks. 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 characteristics. The data might also not be the right data, e.g., historically, and might have been collected for another purpose. A lot of "health data" is actually collected for the purposes of insurance, and often reflects insurance priorities. This means it might be unsuitable data for developing and training AI for clinical purposes, leading to additional data collection burden. 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 valid 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 ‘behavioural data surplus’ for uses that raise ethical, regulatory, legal and human rights concerns. * There are concerns about 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 marginalised 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. * 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. |

# 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 of a model *in silico* may not translate into performance of an AI system *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. Depending on the risk profile of the AI system, clinical evaluation may be done before, on in parallel with deployment.

The overarching aim should be to design studies that give confidence in results by minimising 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 (Liu et al., 2019) , (Nagendran et al., 2020). Randomised controlled trials (RCT) remain the benchmark of clinical studies, in which key elements help to minimise bias and increase confidence in the findings (Sibbald & Roland, 1998), (Moher et al., 2012). 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 optimise 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 (Rivera et al., 2020)(for reporting of study protocols) and CONSORT-AI (Liu et al., 2020) (for reporting of trial reports); additional EQUATOR guidelines are currently in development for diagnostic test accuracy studies (STARD-AI (Sounderajah et al., 2020)) and studies of prediction models [TRIPOD-AI] (Collins & Moons, 2019).

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 effectiveness, 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 accurately to ensure results are ascribed to a specific AI system (including version) and would enable replication of the study. This should include product details including version number, supplier and contact details.
* *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 are defined in advance and include those that are the most important to patients, and the key stakeholder groups; use of core outcome sets are recommended where they exist for the condition of interest; pre-specification avoids bias through retrospective selection of most favorable outcome or of positive result arising through chance and multiple testing.
* *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 (including the metric that will be used), and should include sufficient consideration of subgroups to ensure that any deviations of performance and potential risk of harm is detected; errors should be analysed at the individual error level to identify the reasons for failure where possible.
* *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 (Liu et al., 2020), adapted from the CONSORT 2010 flow diagram (Schulz et al., 2010)).

It is encouraging to see the emergence of well-designed clinical studies of AI interventions. RCT remains the ideal trial design, although well designed, 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.

# 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 real world 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 or severity. 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 that may permit earlier deployment under 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 Software as a Medical Device (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 “Good Machine Learning Practice '' (GMLP) principles (Center for Devices & Radiological Health, 2021), which aim to “help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML)”. 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). The EU published a draft of a new EU Artificial Intelligence Act in April 2021 (European Parliament, 2021), which sets out similar principles to the SPS and ACP, albeit less comprehensively described than the US FDA proposals.

|  |
| --- |
| *Box 3: 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. A key role for clinical evaluation is to identify short, medium, and long-term risks to patient safety. Risks may be wide-ranging and may relate to failures within the technology itself (including issues with algorithm design and the data used for training) or with how the technology is used by humans (intentional or unintentional misuse) or with issues relating to the deployment setting (including deviations from the inputs, outputs and supporting infrastructure anticipated). In addition to individual errors, there may be systematic errors or biases. Both individual and systematic errors may result in patient harm and should be actively looked for at all stages of development and deployment of AI health technologies. Approaches such as the 'Medical Algorithmic Audit' actively look for and analyse these errors and failure modes and provide a framework for early detection and mitigation [REF Liu et al Lancet Digital Health [in press]].  In terms of patient safety, the consequences of errors will vary according to context but could include incorrect outputs in relation to diagnostic classification (e.g., labelling a potentially malignant lesion as benign), or prediction of outcome (e.g., providing a risk estimate that leads to a patient not receiving a kidney transplant). There is particular ethical concern around systematic performance deviations (bias) relating to certain characteristics such as ethnicity or gender that may result in negative consequences; this may arise due to the unintended translation of human biases into the data and a lack of data collection or inadequate data collection in some areas such as gender, ethnicity used to train the model or potentially be introduced in the later stages of design. Finally, there remains the issue that models may be developed that simply are not sufficiently trained or tested on certain groups, resulting in those groups being unable to benefit from this technology (notably the use of some skin cancer AI diagnostics being limited to paler skin types). |

*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 and changes 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*

*Most* regulatory authorities stipulate that manufacturers of medical devices, including AI included in the 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] (FDA, n.d.) database and the UK Medicines and Healthcare products Regulatory Agency [MHRA] (MHRA, n.d.) 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 that 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 (including professionals and patients) and developers of AI systems are the most active stakeholder groups engaging in post-deployment monitoring, with users usually being the first to discover problems arising at this stage. 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 a ‘medical algorithmic audit’ of the AI health technology. Through the audit, existing and potential risks can be assessed and prioritised, risk mitigation plans can be put in place, and future audits can monitor whether risk mitigation measures were successful in avoiding harm. 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 comparative measurement of the expected costs relative to its expected impacts when implemented in a particular context. AI is often being adopted to manage cost pressures - it’s not always about increasing quality, but often driven by a desire for automation and cost reduction of repetitive tasks. As such overall cost of a conventional care parthway vs an AI enabled pathway is often the critical cost effectiveness data that decision makes are looking for. Conventional health economic analyses often fail to look an entire existing pathway cost, and comparison with the disrupted and often very different pathway in the AI enabled route.

Economic evaluation is defined as a comparative analysis of two or more interventions in terms of their costs and consequences (M. F. Drummond et al., 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 (including quantifying the extent of uncertainty) 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, and quantifying uncertainty associated with estimates, the process can inform the best course of action within budget constraints; 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).

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 interventions 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. A common motivation for the application of economic evaluation of AI health technologies is to assess the extent of cost savings within a system as a result of the technology. This requires substantive system modelling of an often complex patient and administrative pathway, whereas economic evaluation of traditional interventions often are limited to a specific element of the patient pathway to identify economic impact.

AI-health technologies have a distinct cost profile, where innovation or development costs are substantial and, in some scenarios, would have an 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 generalised approach to economic evaluation (i.e., across a region or grouping of differing contexts) introduces substantial uncertainty. In addition, an AI-health systems 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 guidelines and texts on methods for economic evaluation of non-digital health interventions, there is relatively limited research on methods for the economic evaluation of AI health technologies. As the body of literature of applied economic evaluation of AI health technologies develops, more definitive and specific normative guidance on methods and approaches will be possible.

*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 maximises outcomes. This has given rise to the common use of a form of cost effectiveness analysis (CEA) termed cost utility analysis (CUA), where the net incremental costs of an intervention are presented as a ratio to net incremental health’s outcomes, and health is a generalised 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. While a CUA employs utility as the effectiveness measure, a CEA is simply a representation of costs to any relevant outcome, be that lives saved, cases of disease averted, or clients reached. In the broader assessment of health technologies, there are commonly multiple outcomes of interest that span health, health system, non-health (e.g., cross-sectorial) and wider considerations such as equity and patient experience of care. Benefit cost analysis (BCA) is a form of economic evaluation that represents all outcomes in a monetary form based on a welfarist framework allowing aggregation of results into a representation of net benefit and benefit cost ratio. In instances where the health impacts of a technology is expected to be equal or similar, the use of cost minimization analysis (CMA) is also employed where the evaluation will primarily focus on representing differences in cost of competing interventions where the lowest cost option is considered the favoured option. Cost-consequence-analysis (CCA) is a disaggregated form of economic evaluation where the range of costs and impacts are listed and described in impact inventories following a common structure to provide decision makers with a full representation of the consequences of the competing courses of action (for example in the case a new diagnostic method is able to reduce the need and/or frequency of other tests). Ultimately, the choice of economic evaluation method will depend on the context of the proposed implementation of the digital technology and the needs of the investor or decision maker.

Increasingly, budget impact analysis (BIA) forms an essential part of an economic evaluation. To determine affordability in the short term, 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 (YHEC, 2020), (*Budget impact analysis*, 2018)). BIA entails addressing the estimated use and costs of the proposed health technology, an estimation in the changes in use and cost of other health interventions or 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 minor and significant upgrades, to support the adoption of new SaMD technologies) and addressing uncertainty in terms of model parameter inputs and structural uncertainty underpinned by certain assumptions (ISPOR, n.d.). 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.

Regardless of the economic evaluation methodology chosen, 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, and the comparator (the intervention would be displaced if the new intervention was adopted) 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 once implemented and at scale, the cost per additional patient diagnosed is likely to be negligible, with ongoing development costs related to refinement of the algorithm and adjustments to apply the algorithm to additional patient populations.

*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 et al., 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, 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 (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 (M. 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 or fee-per-use basis. Reimbursement of digital health technologies is limited, although an active area of exploration and change led for example by countries like Germany (made possible by the German Digital Healthcare Act, DVG) and the UK (Gerke et al., 2020).

(Innovators Guide to the NHS, 2020) (Kaltenthaler et al., n.d.)

# Communication of results

Communicating the results of the steps of the clinical evaluation process transparently is fundamental to the safe and effective use of AI health technologies, to enable clinicians, patients, regulators, and other stakeholders have the evidence they need to assess the safety, effectiveness and likely value of the technology and its performance in their setting. Key principles include: transparent reporting of the datasets used in the training, testing and validation of the model at all stages (such as using the 'datasheets for datasets' approach (Gebru et al., 2018)); transparent description of the model (such as using the 'model fact cards' approach (Sendak et al., 2020) ); transparent reporting of all clinical studies with standard metrics (using the relevant EQUATOR guideline); and transparent reporting of all post-deployment audits (using a appropriate algorithmic audits [ref :Liu; (*aiaudit.org*, 2020)].

The key components of clinical evaluation have been discussed in detail throughout this WG-CE report, but some of the most important elements are summarised in Table 1.

Table 1: Summary of most important elements of the key components of clinical evaluation

|  |  |
| --- | --- |
| **Evaluation Phase** | **Content of report** |
| Evaluation of AI system purpose and suitability | A description of the clinical problem and intended clinical pathway with intended benefits and potential risks of AI use.  Evaluation of interoperability and security.  Description of stakeholder engagement and user testing. |
| Algorithmic validation | Description of data used for development and testing including size of dataset, demographics of population (age, gender/sex, race, ethnicity), setting data was collected (type of facility, date of collection).  Type of model used, and performance metrics obtained at internal and independent evaluation with a comparison against the current standard of care.  Results of any benchmarking against either current standard of care of other models. |
| Clinical Validation | Description of the clinical evaluation including the study design (registration and location of report), population, setting, intervention (including inputs and output of AI), comparator and pre specified outcomes.  Transparent reporting to describe the clinical evidence (using the appropriate EQUATOR guideline and standard performance metrics).  Details of any regulatory approval, if required. |
| Ongoing Monitoring | Description of ongoing monitoring in place. How will adverse events be collected? When will the model be audited and by whom?  Clear reporting of audits to describe the post-deployment performance (e.g. Medical Algorithmic Audit [REF Liu et al] and the WHO/ITU AI4H Audit template (Verks et al., 2020). |
| Economic Evaluation | A description of the economic evaluation carried out. |

# Conclusions and recommendations for future action

This document provides a framework for current best practice evaluation of AI systems in health, primarily aimed at clinicians, researchers, and policy makers, but that may also be useful for other stakeholders including patients, the public, and developers of AI technologies. Table 2 summarises the components of this evaluation framework. The working group identified the clinical evaluation of AI systems as an urgent global priority that must be prioritised in order to ensure that AI systems that are adopted into use in health are effective, cost-effective, safe, ethical, inclusive and fair. Undertaking this evaluation requires input from a range of stakeholders in order to understand the range of considerations required. The working group identified that while some elements of evaluation are triggered by regulatory or health technology assessment requirements, neither of these processes may be done routinely for a number of systems, which nevertheless require evaluation. The working group considers that evaluation must be transparent, and the results made open and accessible, in order to build trust.

The working group identified a number of areas for future action. The first is around the requirement for high quality datasets to train and evaluate the performance of health AI systems. The working group recommends that clincians, researchers and those responsible for the local or national governance of AI systems prioritise the collection of data based on the most urgent clinical problems, including 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. The current development of models is often driven by data availability rather than clinical need and equity. Targeting resources according to these priorities, focuses on the data required for algorithmic validation, including validation on external test sets and benchmarking AI models to local or national datasets, is required [recommended] in order to provide evidence of the extent to which the model generalises and its likely performance in the setting and population under consideration.

The working group also identified that despite the expanding number of AI systems in health, there is still a paucity of clinical studies, especially those with long term evaluation, clinical endpoints and rigorous safety analyse. This may be holding back the potential of these tools and affect trust. Clinical studies are essential in order to assess whether in silico performance translates into measurable benefit to patients and health systems in the real world. The design of clinical studies of AI health technologies should use established principles of good methodological design to minimise bias and be transparently reported according to established international guidelines. The clinical validation process should recognise the risk that the AI health technology may not generalize and ensure that the evaluation considers sufficiently diverse populations and settings to support the intended use. Collaborative studies that support evaluation in diverse settings are critical to enabling the potential benefits of AI health technologies to be disseminated more widely and equitably.

Finally, the working group recognized that the evaluation process should include acquisition of relevant health economic data to support decisions and underpin public trust in procurement. While it is possible to find examples of where digital evidence frameworks explain what kind of economic evaluation should be done and level of evidence required, it is rare to find the results of these reported transparently and openly, often even for widely used digital tools. This evaluation, and the 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. Health technology assessment had received less attention than regulation in this area, but the working group recommends that this is a key focus for future work. Agencies responsible for health technology assessment may need to expand the workforce skills to include evaluation of AI technologies and in many countries, particularly where HTA capacity is low, this should be a key focus alongside expanding digital health tools.

In conclusion, the working group provides a framework for clinical evaluation of AI systems in health that if applied, will work towards ensuring AI systems are effective, safe, cost effective, ethical, and fair. Transparent communication of the results of evaluation will build trust and increase the appropriate use of AI systems in health, which could bring benefits to all.

Table 2: Summarised guidance for clinical evaluation of AI systems in health

|  |  |
| --- | --- |
| Introduction and background | This section provides the context to AI for health and the role of clinical evaluation. |
| *Ensuring that AI systems that are adopted into health systems are effective, safe, cost-effective, 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 are critical to delivering this.* | |
| Requirement 1) Model design and suitability | This section tackles the evaluation of the design and the useability of the AI model, in cooperation with users and key stakeholders (e.g., clinicians, patients, public). |
|  | Key findings and considerations:  Understanding the problem and intended use   * Evaluation of the purpose and demonstration that AI is the most suitable option for a particular clinical problem * Explanation of the clinical problem and the setting it was designed for (special considerations to the context/users, how it fits best in the clinical workflow, patient journey, etc.)     Defining intended benefits   * Understanding of the intended benefits to the individual patient, clinical workflow, or health system (or a combination of these)     Describing potential risks and harms   * Patient level risks (e.g., misclassification, misdiagnosis, automation bias, delayed care, under-or overdiagnosis, unnecessary treatment) * Clinical workflow risks (e.g., additional administrative or cognitive task burden for clinicians) * System level risks (e.g., health economic costs of the clinical impacts)     Interoperability and security   * Consideration of interoperability requirements of AI technologies with other devices and IT systems (e.g., hardware and software upgrades) * Consideration of security aspects (e.g., regarding data collection, privacy)     User-testing and user engagement   * Getting involved with stakeholders; understanding and description of the interaction of the model with real-world scenarios through a “mixed methods approach” (e.g., user feedback, interviews, usability testing, focus groups, Delphi studies, and ethnographic studies) |
| Requirement 2) Algorithmic validation | The term “algorithmic validation” is used in this document to refer to the evaluation of the AI model “in silico”. |
|  | Key findings and considerations:  Description of internal and external testing dataset   * Evaluating data used for training, tuning, internal and external validation and assessing the extent to which these datasets align to the intended use, including a specific use case, population and setting     Training/Testing data reporting should include   * Description of the demographic spread (including gender/sex, age, race/ethnicity) * Performance metrics for the population as a whole and for key groups that might be vulnerable for under-representation in the training data set * Description of data type, source, when, how it was collected * Quality of the training data, robustness of the labels, understanding what was used as ‘ground truth’     Algorithmic validation   * Purpose: Demonstration of the robustness of the model and the performance on an acceptable level in the intended setting * Requires: 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 has been used in those stages * External validation: on a representative high-quality dataset with accurate labels; Transparent reporting of the performance metrics     Benchmarking of performance   * Evaluation against an accepted standard should be made * Comparative benchmarking of AI models against an unseen dataset (similar process as external validation) can improve algorithmic validation * Foster the availability of unseen external datasets by initiatives, local, regional, governmental bodies for creating/to supply commonly available datasets     Building high-quality datasets   * Obstacle of availability of sufficient quality datasets with required labels, and representative of local population results in a lack of health data for certain people groups or populations (‘health data poverty’) * May result in poor safety and performance and lead to various harms including exclusion or exposure to under-performance * Major problem: availability of the technology (e.g., EHR, national screening programs) or even exclusion from the health system especially in non-affluent regions / countries * Issue: Low internet connectivity and IT infrastructure for AI training and implementation |
| Requirement 3) Clinical validation | The term “clinical validation” in this document refers to the evaluation of AI technologies through interventional or clinical studies. |
|  | Key findings and considerations:   * Purpose: Clinical studies provide the necessary evidence as to whether an AI system is effective and safe when deployed in a clinical pathway. It provides confidence in results by minimizing bias and the risk of harm (builds trust in AI technology) * Principles of good clinical study design are equally applicable for AI technologies, but are currently not yet established * Randomized controlled trials (RCTs) remain the ideal trial design, but in some cases observational studies with a relevant comparator may be adequate for some AI technologies * Several guidance documents help optimize specific study designs when evaluating an AI intervention (through the EQUATOR network) * Specific elements that should be considered in clinical studies of an AI evaluation include *Study design, population, setting, intervention(s), intervention inputs and outputs, comparator, pre-specific outcomes relevant to all stakeholders, process measures, balancing measures, protocol deviations, analysis, reporting of study protocol, reporting of study conduct and results.* |
| Requirement 4) Deployment and ongoing evaluation | This section addresses the key considerations for the deployment phase, regulatory requirements, and the ongoing monitoring and surveillance of the AI technology after implementation including algorithmic audits. |
|  | Key findings and considerations:  Deployment   * Evaluation needs to be continued into the deployment phase and for as long as the product continues to be used; concern of “generalisability” (interventions under-performing or even catastrophically failing when moved from one population or setting into another) * Any need for further training or local tuning should be actively sought, as a critical part of ongoing evaluation for efficacy and safety * Challenge (also for regulators): determine the level of additional evaluation required to appropriately assure version updates of AI products and continuously learning or adaptive algorithms * One approach: “Good Machine Learning Principles” (algorithm changes must be transparently labelled in specific protocols for change control plans)     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     Regulatory requirements   * Manufacturers should consider regulatory requirements: systematically carry out post-deployment monitoring of safety and performance and taking a corrective action when required * This requires an ongoing monitoring of the device’s performance, or for an adverse event to be detectable and attributable to the AI device in question * Manufacturer’s post market clinical follow-up plan should consider collection of post-deployment data due to potential harms that might occur within the clinical pathway     Relevant stakeholders   * Crystallization and definition of the most active stakeholders engaging in post deployment monitoring: users and developers; regulators, etc. * Should be part of the indications for use statement and can inform the level of pot deployment surveillance     Algorithmic audits   * Purpose: detailed analysis may be performed in an algorithmic audit to determine what, how and why adverse events or algorithmic errors occurred * Variable applicable: Local performance monitoring, establish a “baseline performance” by e.g., assessment and prioritization of existing and potential risks |
| Requirement 5) Economic evaluation | Economic evaluation is defined as a comparative analysis of two or more interventions in terms of their costs and effect/impact when implemented in a particular context. Such an evaluation enables the assessment not only of the comparative effectiveness of a health intervention, but the incremental costs (or costs savings) of achieving the effect |
|  | Key findings and considerations:  Types of economic evaluation for AI-enabled digital interventions   * Quantification of the costs (implementation and running costs) relative to the outcomes enables decision makers (e.g., funders) to identify the best course of action (also with respect to the best potential to improve patient’s lives) * Important considerations include the definition of the objective and the users of the economic evaluation * Costs and effects highly dependent on local digital architecture and infrastructure   Reimbursement   * A reasonable level of reimbursement is required (limited national opportunities so far) |
| Requirement 6) Ethical Evaluation | Ethical evaluation for the design, development, and deployment of AI technology can be guided by the 6 key principles identified by the WHO “Ethics and Governance of AI4H” report: Protecting human autonomy; promoting human well-being and safety and the public interest; ensuring transparency, intelligibility and considering the requirements for explainability; fostering AI technologies responsibility and accountability; ensuring inclusiveness and equity; promoting AI that is responsive and sustainable. |
|  | Key findings and considerations:  To meet the appropriate governance, evaluation, and regulation, AI technology should:   * meet standards of scientific validity, accuracy, and explainability/reproducibility applied to medical technologies (in consideration of the infrastructure and institutional context) * consider the decommissioning of existing services, without replacing the previous services to the same or better level * consider both, potential benefits and risks, e.g., with regards to biased (or selective) training data * take full account of the total cost and investment required for its use, including digital infrastructure, training, maintenance, and monitoring costs * consider if it is appropriate and adaptable to the context of LMICs, including barriers of language and availability of data (for model training, validation, and maintenance) |

References

*aiaudit.org*. (2020). Aiaudit.Org. https://aiaudit.org/projects/assessmentplatform/

Arrow. (1963). *Uncertainty and the Welfare Economics of Medical Care*. https://web.stanford.edu/~jay/health\_class/Readings/Lecture01/arrow.pdf

Barros, P. (2020). *Incentives for R&D: Payment Options and Pricing Challenges* (No. 002289). Office of Health Economics. https://ideas.repec.org/p/ohe/sembri/002289.html

*Budget impact analysis*. (2018, September 26). YHEC - York Health Economics Consortium. https://yhec.co.uk/glossary/budget-impact-analysis/

Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. *Lancet*, *393*(10181), 1577–1579. https://doi.org/10.1016/s0140-6736(19)30037-6

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, & SPIRIT-AI and CONSORT-AI Consensus Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

DICOM Standard.Org. (n.d.). *DICOM*. dicomstandard.org. Retrieved February 8, 2022, from https://www.dicomstandard.org/

Djulbegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. *Lancet*, *390*(10092), 415–423. https://doi.org/10.1016/s0140-6736(16)31592-6

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes* (4th ed.). Oxford University Press. https://books.google.at/books?id=lvWACgAAQBAJ

Drummond, M., Jönsson, B., & Rutten, F. (1997). The role of economic evaluation in the pricing and reimbursement of medicines. *Health Policy (Amsterdam, Netherlands)*, *40*(3), 199–215. https://doi.org/10.1016/s0168-8510(97)00901-9

European Parliament. (2021). *Artificial Intelligence Act EPRS\_BRI(2021)698792\_EN.pdf*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS\_BRI(2021)698792\_EN.pdf

FDA. (n.d.). *Manufacturer and User Facility Device Experience Database - (MAUDE)*. MAUDE. Retrieved February 10, 2022, from https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/manufacturer-and-user-facility-device-experience-database-maude

FDA, Good Machine Learning Practice for Medical Device Development. (2021, October 27). *Good Machine Learning Practice for Medical Device Development*. U.S. Food and Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

Finddx.org. (2021). *Access to COVID-19 Tools (ACT) Accelerator*. Finddk.org. https://www.finddx.org/wp-content/uploads/2021/01/ACT-A-Dx-data-governance-framework\_15.01.2021.pdf

GDHP, Cyber Security, Interoperability, Evidence and Evaluation, & Clinical and Consumer Engagement. (2020). *GDHP*. https://gdhp.nhp.gov.in/home/index/white-paper-2020

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2018). Datasheets for Datasets. In *arXiv [cs.DB]*. arXiv. http://arxiv.org/abs/1803.09010

Gerke, S., Stern, A. D., & Minssen, T. (2020). Germany’s digital health reforms in the COVID-19 era: lessons and opportunities for other countries. *Npj Digital Medicine*, *3*(1), 94. https://doi.org/10.1038/s41746-020-0306-7

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.-R., & Binder, A. (2020). Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific Reports*, *10*(1), 6423. https://doi.org/10.1038/s41598-020-62724-2

hl7FHIR. (n.d.). *Overview - FHIR v4.0.1*. Retrieved February 8, 2022, from https://www.hl7.org/fhir/overview.html

IDEO. (n.d.). *Design Kit*. Retrieved February 8, 2022, from https://www.designkit.org/methods

IMDRF. (2013). *imdrf-tech-131209-samd-key-definitions-140901.pdf*. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf

Ingenbleek, P. T. M., Frambach, R. T., & Verhallen, T. M. M. (2013). Best practices for new product pricing: Impact on market performance and price level under different conditions. *The Journal of Product Innovation Management*, *30*(3), 560–573. https://doi.org/10.1111/jpim.12008

Innovators Guide to the NHS. (2020). *An innovator’s Guide to the NHS\_Navigating the barriers to digital health*. Boehringer Ingelheim. https://www.boehringer-ingelheim.co.uk/sites/gb/files/documents/innovators\_guide.pdf

ISPOR. (n.d.). *ISPOR - Home*. Retrieved February 9, 2022, from https://www.ispor.org/

Kaltenthaler, E., Tappenden, P., Paisley, S., & Squires, H. (n.d.). *REPORT BY THE DECISION SUPPORT UNIT*. http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD-13-model-parameters.pdf

Kickbusch, I., Piselli, D., Agrawal, A., Balicer, R., Banner, O., Adelhardt, M., Capobianco, E., Fabian, C., Singh Gill, A., Lupton, D., Medhora, R. P., Ndili, N., Ryś, A., Sambuli, N., Settle, D., Swaminathan, S., Morales, J. V., Wolpert, M., Wyckoff, A. W., … Secretariat of the Lancet and Financial Times Commission. (2021). The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. *Lancet*, *398*(10312), 1727–1776. https://doi.org/10.1016/S0140-6736(21)01824-9

Larson, D. B., Harvey, H., Rubin, D. L., Irani, N., Tse, J. R., & Langlotz, C. P. (2021). Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: Summary and recommendations. *Journal of the American College of Radiology: JACR*, *18*(3 Pt A), 413–424. https://doi.org/10.1016/j.jacr.2020.09.060

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/s2589-7500(19)30123-2

Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. *BMJ Health & Care Informatics*, *28*(1). https://doi.org/10.1136/bmjhci-2020-100301

Macdonald, J., März, M., Oala, L., & Samek, W. (2021). Interval neural networks as instability detectors for image reconstructions. In *Bildverarbeitung für die Medizin 2021* (pp. 324–329). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-33198-6\_79

Mbchb, I., Mbchb, L., Denniston, A. K., Ibrahim, H., Liu, X., Zariffa, N., & Morris, A. D. (2021). *Health data poverty: an assailable barrier to equitable digital health care*. https://www.thelancet.com/action/showPdf?pii=S2589-7500(20)30317-4

MHRA. (n.d.). *Medicines and Healthcare products Regulatory Agency*. GOV.UK. Retrieved February 10, 2022, from https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., Altman, D. G., & CONSORT. (2012). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery (London, England)*, *10*(1), 28–55. https://doi.org/10.1016/j.ijsu.2011.10.001

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ (Clinical Research Ed.)*, *368*, m689. https://doi.org/10.1136/bmj.m689

Open Code Initiative. (2020). *Focus Group on Artificial Intelligence for Health*. https://github.com/FG-AI4H

PAHO. (2019). *Data Governance in Public Health; Information systems for Health toolkit; IS4H-KCDG-EN.pdf*. https://www3.paho.org/ish/images/toolkit/IS4H-KCDG-EN.pdf

*People + AI Guidebook*. (n.d.). Google’s People and AI Guidebook on How to Decide If AI Is the Right Solution. Retrieved February 8, 2022, from https://pair.withgoogle.com/guidebook/patterns

Rainie, S. C., Kukutai, T., Walter, M., Figueroa-Rodriguez, O. L., Walker, J., & Axelsson, P. (n.d.). *Issues in Open Data - Indigenous Data Sovereignty*. African Minds; International Development Research Center. Retrieved February 8, 2022, from https://www.stateofopendata.od4d.net/chapters/issues/indigenous-data.html

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ (Clinical Research Ed.)*, m3210. https://doi.org/10.1136/bmj.m3210

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*(1), 18. https://doi.org/10.1186/1741-7015-8-18

Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *Npj Digital Medicine*, *3*(1), 41. https://doi.org/10.1038/s41746-020-0253-3

Sibbald, B., & Roland, M. (1998). Understanding controlled trials. Why are randomised controlled trials important? *BMJ (Clinical Research Ed.)*, *316*(7126), 201. https://doi.org/10.1136/bmj.316.7126.201

Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A. K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., Markar, S. R., McInnes, M. D. F., Panch, T., Pearson-Stuttard, J., Ting, D. S. W., Golub, R. M., Moher, D., Bossuyt, P. M., & Darzi, A. (2020). Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nature Medicine*, *26*(6), 807–808. https://doi.org/10.1038/s41591-020-0941-1

The Lancet. (2018). Is digital medicine different? *Lancet*, *392*(10142), 95. https://doi.org/10.1016/s0140-6736(18)31562-9

theiet. (n.d.). *Design and evidence*. Https://Www.Theiet.Org/Impact-Society/Factfiles/Healthcare-Factfiles/Design-and-Evidence/. Retrieved February 8, 2022, from https://www.theiet.org/impact-society/factfiles/healthcare-factfiles/design-and-evidence/

Uk, G. (n.d.). *Ethnographic study: qualitative studies*. GOV.UK. Retrieved February 8, 2022, from https://www.gov.uk/guidance/ethnographic-study-qualitative-studies

United Nations. (2015, September). *The 17 sustainable development goals*. Https://Sdgs.Un.Org/Goals. https://sdgs.un.org/goals

Verks, B., Switzerland, D. A., Oala Fraunhofer, L., & Germany, H. (2020). https://aiaudit.org/assets/pdf/standards/FGAI4H-J-048.pdf

Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Global Health*, *3*(4), e000798. https://doi.org/10.1136/bmjgh-2018-000798

WHO, Digital Health, & Innovation. (2021, November 17). *Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training Validation and Evaluation*. World Health Organization. https://www.who.int/publications/i/item/9789240038462

Wiegand, T. et al. (2020). *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for health*. https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H\_Whitepaper.pdf

World Health Organisation, Open Data Institute. (2021). *Data Governance maturity and best-practices*. WHO, ODI. https://cdn.who.int/media/docs/default-source/world-health-data-platform/events/health-data-governance-summit/preread-2-who-data-governance-summit\_health-data-as-a-public-good.pdf?sfvrsn=2d1e3ad8\_8

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. World Health Organization. https://www.who.int/publications/i/item/9789240029200

YHEC. (2020, September 15). *YHEC*. YHEC - York Health Economics Consortium. https://yhec.co.uk/

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