ITUPublications

International Telecommunication Union
Standardization Sector

# ITU-T Focus Group Deliverable

(03/2023)

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

# FG-AI4H DEL7.4

# Clinical evaluation of AI for health

# ITU-T FG-AI4H Deliverable DEL7.4

# Clinical evaluation of AI for health

## Summary

Artificial intelligence (AI) in healthcare could hold great promise to improve people's health worldwide by transforming 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 in seen and unseen data. There are already many potential applications for AI in medicine and considering the factors such as the global shortage of healthcare professionals, changing population demographics worldwide, and the ongoing global digital transformations there is huge interest in the potential of AI systems in both high- and low-resourced settings. Achieving the potential beneficial impact requires frameworks for evaluating AI systems, in order to ensure that they are safe, effective, and useful and that they do not cause unanticipated harm when applied to a complex clinical pathway or when used autonomously, and that the costs and ethics are adequately considered.

The adoption of effective, safe, ethical, inclusive, and fair AI systems into health systems is a global concern that requires input from a wide range of stakeholders. Clinical evaluation of AI systems including their underpinning data, performance, safety, and transparent communication of these results are critical for delivery.

Working from the principles of evidence-based medicine but acknowledging the particular challenges and opportunities of AI-based technologies, this report provides a framework for the evaluation of AI systems in health that can be used by clinicians, researchers, developers, regulators, health systems, and policymakers 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 Telecommunication Union (ITU) Focus Group on AI for Health [FG-AI4H] [Wiegand et al., 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 into four phases: 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 that cut across the four phases.

Evaluation of model purpose and suitability requires:

- an understanding of the problem and the intended use of the AI system
- a definition of the intended benefits
- a description of the potential risks and harms
- documentation of interoperability and security
- user testing and user engagement reports.

Algorithmic validation (used here to refer to the evaluation of the AI system '*in silico*' requires:

- a description of the data used for development
- internal and external testing, and of the model type used
- reporting of performance metrics in the internal and independent external testing data
- benchmarking of system performance against standard of care, and where relevant, other AI systems.

Clinical validation (for the purposes of this Technical Report this is the evaluation of the AI system through interventional or clinical studies) requires:

- a clinical study with a relevant comparator and a meaningful endpoint, and the steps taken to minimise bias.

Finally, deployment and ongoing evaluation requires:

- monitoring of performance and impact (including safety and effectiveness) to understand the anticipated and unanticipated outcomes
- algorithmic audits [Liu et al., 2022] to understand how adverse events or algorithmic errors occur.

Annex A summarizes the key findings as a checklist to facilitate the application of this Deliverable.

**Keywords**

Artificial intelligence, bias, clinical evaluation, clinical validation, digital health, economic evaluation, generalisability, reporting.

## Note

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

## Change Log

This document contains Version 1 of the Deliverable DEL7.4 on "*Clinical evaluation of AI for health*" approved on 24 March 2023 at the ITU-T Focus Group on AI for Health (FG-AI4H) meeting held in Cambridge, Massachusetts, USA, 21-24 March 2023.

| | | |
|---|---|---|
| **Editing team**: | Co-chair & writing group, WG-CE<br>Naomi Lee<br>NICE - National Institute for Health and Care Excellence<br>United Kingdom | E-mail: naomi.lee@nice.org.uk |
| | Co-chair & writing group, WG-CE<br>Shubhanan Upadhyay<br>Ada Health<br>Germany | E-mail: shubs.upadhyay@ada.com |

Co-chair & writing group, WG-CE
Eva Weicken
Fraunhofer Heinrich Hertz Institute (HHI)
Germany
E-mail: eva.weicken@hhi.fraunhofer.de

Writing Group, WG-CE
Alastair Denniston
University of Birmingham
United Kingdom
E-mail: a.denniston@bham.ac.uk

Writing Group, WG-CE
Xiao Liu
University of Birmingham
United Kingdom
E-mail: xiao.liu@insight.hdrhub.org

Writing Group, WG-CE
Kassandra Karpathakis
Harvard University
USA
E-mail: kkarpathakis@hsph.harvard.edu

Writing Group, WG-CE
Thomas Wilkinson
Ministry of Health
New Zealand
E-mail: tommy.d.wilkinson@gmail.com

Writing Group, WG-CE
Jane Carolan
UCL Institute of Health Informatics
United Kingdom
E-mail: jc.carolan@gmail.com

## Contributors

This document was developed in 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, the follow-up meetings, and review rounds the writing group drafted the outline over time.

***External expert group*** *(in alphabetical order):* AbouElkhir Osama (Tachy Health, Dubai), Akogo Darlington (minoHealth AI Labs, Ghana), Allen Megan (Inspired Ideas, Tanzania), Alsalamah Shada (WHO, Switzerland), Arentz, Matthew (FIND, Switzerland), Baird Pat (Philips, USA), Balachandran Pradeep (Freelancer E-Health, India), Bastawrous Andrew (Peek Vision, Global Eye Health, UK), Bathke Arne (University of Salzburg, Austria), Chiavegatto Filho Alexandre (São Paulo University, Brazil), Cresswell Kathrin (University of Edinburgh, UK), Darkoh Ernest (BroadReach Healthcare, South Africa), Ehrenfeld Jesse (American Medical Association (AMA), USA), Fehr Jana (Hasso-Plattner-Institute, Germany), Fenech Matthew (Una Health, Germany), Fürstenau Daniel (Copenhagen Business School, Denmark), Gaudin Robert (Charitè, Germany), Gilbert Stephen (Technische Universität Dresden, Germany), Glod Mateusz (Infermedica, USA), Greaves Felix (National Institute for Health and Care Excellence (NICE), UK), Gupta Saurabh (Department of Cardiology, All India Institute of Medical Science, India), Gütter Zdenek (Ministry of Health of the Czech Republic , Czechia), Hatton Grace (Sensyne Health, UK), Ho Dean (National University of Singapore, Singapore), Ibrahim Hussein (Doctors.net.uk, UK), Islam Shariful (Deakin University, AUS), Jarral Reza (Pro Care, NZ), Jeon Jonghong (Electronics and Telecommuncations Research Institute, South Korea), John Oommen (George Institute for Global Health, India), Kadam Rigveda (Head of digital access, FIND, Switzerland), Kherif Ferath (LREN, Switzerland), Kuku Stephanie (WHO, UCL, Switzerland, UK), Kurtys Michal (Infermedica, Poland), Lapão Luís (Velez) (University of Lisbon, Portugal), Linder Nina (University of Helsinki, Uppsala University, Finland & Sweden), Loh Irving (Infermedica, USA), Loveys Kate (University of Auckland, NZ), Lundin

Johan (University of Helsinki & Karolinska Institute, Finland & Sweden), Magrabi Farah (Macquarie University, Sydney, Australia), Mahajan Arnav (Department of Medicine, University College Cork, Ireland), Malpani Rohit (WHO, Switzerland), Mamun Khondaker (CMED Health, Bangladesh), Masud Jakir Hossain Bhuiyan (CTS, Bangladesh), Matin Rubeta (Oxford University Hospitals NHS Foundation Trust, UK), McCradden Melissa (The Hospital for Sick Children, Canada), Menezes Audrey (Healthily, UK), Murchison Andrew (Oxford University Hospitals NHS Trust, UK), Murphy Lisa (Centre for Improving Data Collaboration at NHSX, UK), Essa Mohamedali (Tanzania AI Lab, Tanzania), Nakasi Rose (Makerere University, Uganda), Oala Luis (Fraunhofer Heinrich Hertz Institute (HHI), Germany), Pankova Natalie (Metadvice, UK), Piekut Agata (Health Action Tank, Poland), Porras Lina (1Doc3, Colombia), Reddy Sandeep (Deakin University, Australia), Salim Ally (Inspired Ideas, Tanzania), Schwendicke Falk (Charitè, Germany), Sethi Tavpritesh (Indraprastha Institute of Information Technology Delhi, India), Sood Harpreet (National Health Service (NHS), UK), Sousa Inês (Fraunhofer Portugal, Portugal), Srivastava Manish (Virginia Polytechnic Institute & State Univ., USA), Starlinger Johannes (Howto Health Digital Business Solutions, Germany), Wasswa William (Global Auto Systems LTD, Uganda), Werneck Leite Alixandro (Tribal, Mexico), Wiegand Thomas (Fraunhofer Heinrich Hertz Institute (HHI), Germany)

***Writing group*** *(in alphabetical order)*: Carolan Jane (UCL Institute of Health Informatics, UK), Denniston Alastair (University of Birmingham, UK), Karpathakis Kassandra (Harvard University, USA), Lee Naomi (National Institute for Health and Care Excellence (NICE), UK), Liu Xiaoxuan (University of Birmingham, UK), Upadhyay Shubhanan (Ada Health, Germany), Weicken Eva (Fraunhofer Heinrich Hertz Institute (HHI), Germany), Wilkinson Thomas (Ministry of Health, NZ)

## Table of Contents

# ITU-T FG-AI4H Deliverable DEL7.4

# Clinical evaluation of AI for health

## 1 Scope

This Technical Report describes considerations on clinical evaluation of artificial intelligence (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 – Artificial intelligence for health (AI4H) meeting. It serves as the output document of the Focus Group on artificial intelligence for health (FG-AI4H) Working Group on Clinical Evaluation and is part of a series of FG-AI4H deliverables.

## 2 References

[IMDRF CE] IMDRF, Medical Device Clinical Evaluation Working Group (MDCE WG) N56 FINAL: 2019, Clinical Evaluation.
https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf

Additional references are found at the end of this report.

## 3 Definitions

### 3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

**3.1.1 effectiveness** [IMDRF CE]: The ability of a medical device to achieve clinically meaningful outcome(s) in its intended use as claimed by the manufacturer.

**3.1.2 intended use/purpose** [IMDRF CE]: The objective intent of the manufacturer regarding the use of a product, process or service as reflected in the specifications, instructions and information provided by the manufacturer.

**3.1.3 safety** [IMDRF CE]: Acceptability of risks as weighed against benefits, when using the medical device according to the manufacturer's labelling.

### 3.2 Terms defined in this Technical Report

This Technical Report defines the following terms:

**3.2.1 analytical validation**: For the purposes of this report, this refers to the evaluation of the adequacy of the AI model *in silico* before being implemented *in vivo* in the clinical pathway.

**3.2.2 clinical validation**: For the purposes of this report, this is an evaluation of the AI system through interventional or clinical studies in which the whole AI health technology is evaluated in the context of the clinical pathway.

**3.2.3 external validation**: Refers to the process of evaluating the performance of the AI model using previously unseen and independent data *in silico*.

## 4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

| | |
|---|---|
| AI | Artificial Intelligence |
| AI4H | Artificial Intelligence for Health |
| BIA | Budget Impact Analysis |

| | |
|---|---|
| CBA | Cost Benefit Analysis |
| CCA | Cost Consequence Analysis |
| CEA | Cost Effectiveness 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 |
| EHR | Electronic Health Record |
| 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 |
| HTA | Health Technology Assessment |
| 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 |
| ML | Machine Learning |
| ML4H | Machine Learning for Health |
| NICE | National Institute for Health and Care Excellence |
| NGO | Non-Government Organization |
| NHS | National Health Service |
| NHSX | National Health Service |
| 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 |
| TRIPOD | Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis |
| UHC | Universal Health Coverage |

|  |  |
|---|---|
| UK | United Kingdom |
| UN | United Nations |
| US | United States (of America) |
| WG-CE | Working Group on Clinical Evaluation |
| WHO | World Health Organization |

## 5 Conventions

None.

## 6 Background

Globally a growing shortage of healthcare professionals, a rapid growth in health data and an expansion in the usage of AI systems in other sectors have contributed to an increasing interest in the use of AI for health and clinical practice. As a increasing number of health AI models, methods, and tools become available for use, researchers, patients, clinicians, and policymakers require a framework to understand whether they are safe, effective, and cost-effective, and to also 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 the generalisability of models and bias. The current framework for evaluating health innovation, centred 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 at al., 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 Technical Report is produced by the World Health Organization (WHO) and members of the ITU/WHO Focus Group on AI for Health (FG-AI4H) [Wiegand et al., 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-start-ups, non-government organizations (NGOs), etc.). The aim is to produce guidance for current best practice evaluation of AI technologies in health, primarily aimed at researchers, clinicians and policymakers but may also be useful for other stakeholders including patients, the public, and developers of AI technologies. A shared understanding of evaluation best practices could help reduce uncertainty and facilitate the 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 report was shared and reviewed by all WG-CE members.

FG-AI4H is formed from a collaboration between the World Health Organization [WHO Evaluation] and the 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 report is on the principles of evaluation to ensure that it is relevant across all countries with minimal assumptions around particular health systems or the agencies involved. This report uses consistent nomenclature with other WHO and FG-AI4H documentation (ITU/WHO FG-AI4H Del 0.1, 2022) [FG-AI4H Del 0.1] 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 (OCI) ITU/WHO FG-AI4H [OCI 2020] which is developing a software platform for the assessment and an associated ML auditing [aiaudit.org] process to test the applicability with respect to AI models for health. This report 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 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). 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 in (WHO, *Ethics and governance of artificial intelligence for health*, 2021). A summary of the ethical principles relating to clinical evaluation that informed the framework for clinical evaluation is included in Box 1 given in this Technical Report.

The working group on regulatory considerations for artificial intelligence in health reports alongside this WG-CE and the clinical evaluation discussed in that report is based on the framework outlined here. The regulatory DEL02 publication (ITU/WHO FG-AI4H Deliverable 02: Overview of regulatory concepts on artificial intelligence for health, 2023) [FG-AI4H D02] is led by the Working Group on Regulatory Considerations on Artificial Intelligence for Health. It aims to deliver an overview of regulatory concepts on AI for health that covers the following six general topic areas: Documentation and transparency, risk management and AI systems development life cycle approaches, intended use and analytical and clinical validation, data quality, privacy and data protection, and engagement and collaboration. This overview is not intended as a guidance, a regulatory framework, or policy. Rather, it is a discussion of key regulatory concepts and a resource that can be considered by all relevant stakeholders in medical device ecosystems.

The interplay between the two reports makes it 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 by developers to guide their thinking before and/or during the regulatory approval process and to complement the regulatory approval when considering the use of AI systems in individual settings. The framework seeks to address some concerns that have been voiced around the existing regulatory frameworks, including key areas that are specific to AI systems and their context, that may not be adequately addressed or maybe less stringently applied such as the 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 the transparency in those areas.

For AI systems that are classified as medical devices, existing guidance by the International medical device regulators forum (IMDRF) and from the WHO digital innovation and health group (WHO, *Generating evidence for artificial intelligence based medical devices: a framework for training validation and evaluation*, 2021) [WHO Framework] provides specific guidance and links the extent and depth of evaluation to the estimated level of risk to patients (b-Larson et al., 2021). This Technical Report draws on, and dovetails with, 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 with 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: growing up in a digital world [Kickbusch et al., 2021] recommendation that in order for digital health innovation to deliver meaningful benefits in health for all, development of strong accountability frameworks is a must.

- Calling for the development of a trust architecture, which leverages multilateral forums, such as the WHO and transnational, multi stakeholder coalitions the commission recommended the promotion of strong transparency and accountability requirements for emerging AI and machine learning (ML) 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, which is 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 is 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 the quality of digital and AI health tools, requiring 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 the 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 can also have steps performed together if a tool is already in development 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 explainability
- Fostering 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 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 the 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 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., policymakers 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 have the same reach).
- Irrespective of whether an AI technology provides accurate, useful information and insights there may be enough ethical concerns about a use case or a 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 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 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, collection and use of health data follow the relevant governance structures of the reviewing body [PAHO 2019], [finddx.org 2021], [WHO Governance], [Open Data Institute 2021])?

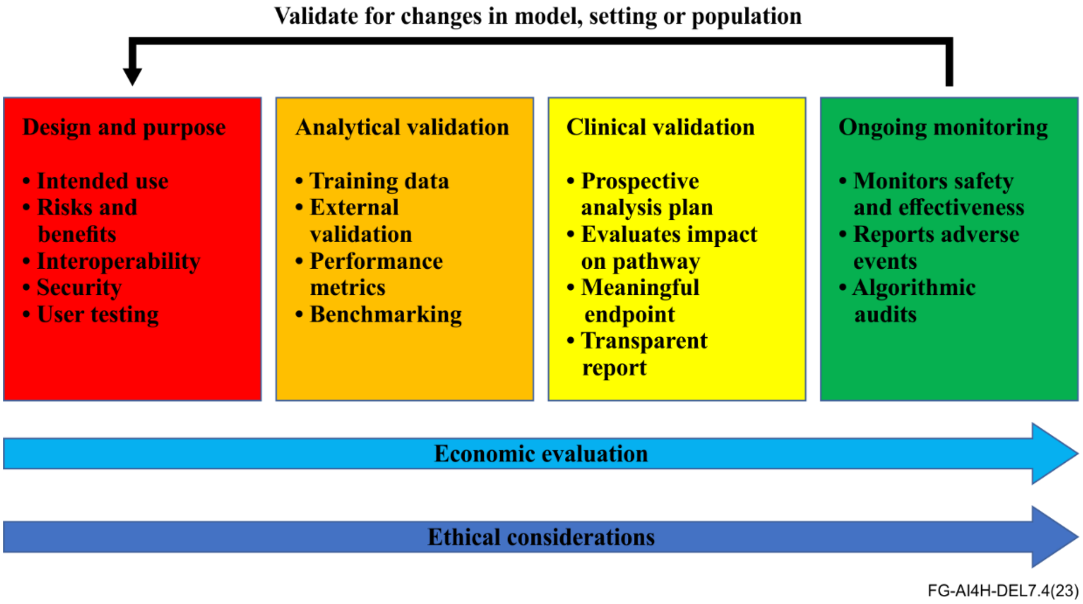

**Figure 1 – Framework for evaluation of AI technologies in health**

## 7 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 from the perspectives of the key stakeholders and a demonstration that AI is the most suitable option to solve it [IDEO design kit], [People + AI guidebook]. 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.org].

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 may have a greater impact. The use of AI technology should be preceded by rigorous analysis and evaluation to ensure it is suitable and 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).

### 7.1 Understanding the problem and intended use

The deployment of AI technologies in healthcare can vary depending on the context. The context might differ in many ways, from established health system infrastructure to completely new clinical pathways. 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:

1) Identification and description of the specific problem to be solved (population, input data required, output data from model, setting). For example, in an AI health technology designed to identify high risk patients with sepsis, the intended use should include target age-groups for which it is suitable and the setting (e.g., intensive care units, ICU versus non-ICU). Additionally, developers should consider the range of clinical information needed for the problem and the intended use.

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

3) A consideration of any special circumstances related to the intended users or context. For example, in paediatric 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.

## 7.2 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 such as an improvement of the patient experience, including reduced waiting times and better clinical outcomes (e.g., improved survival rates, reduced complications compared with current context relevant standard of care), quicker linkage from diagnosis to care, or reduced out-of-pocket expenditure.

– Clinical workflow benefits such as a reduced administrative burden on health care professionals (HCPs), increased time to care, and providing a better HCP experience.

– Health system benefits such as efficiencies found or created in pathways, improved detection of cases, better allocation of resources, cost savings, addressing shortages of skilled HCPs.

## 7.3 Describing potential risks and harms

Consideration of the potential risks of an AI model are an important component of model evaluation. Risks may include, but are not limited to:

– Patient level risks like harmful consequences due to misclassification, misdiagnosis, delayed care, under- or overdiagnosis or unnecessary treatment, or consequences of bias in the AI technology. Clinical workflow risks including removing safeguards, additional time, and administrative or cognitive task burden for HCPs.

– Clinical workflow risks including removing safeguards, additional time, and administrative or cognitive task burden for HCPs.

– 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 by organizations including the national institute for health and care excellence [NICE], US food and drug administration (FDA), and the European Union (EU) medical device regulation (MDR). In general, the higher the risk class, the greater the requirement to demonstrate that a technology's benefits outweigh any potential risks, and greater is 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 the international medical device regulator forum (IMDRF)/FDA definition [IMDRF 2013] and its associated risk classification [IMDRF 2014] that can inform subsequent regulatory, clinical requirements to a certain extent.

## 7.4 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., digital imaging and communications medicine (DICOM), fast healthcare interoperability resources (FHIR)) that support interoperability are essential and compliance with standards where developed offer some assurances, but the full evaluation depends on the local requirements and the local context [DICOM standard], [HL7 FHIR standard]. There remains much work and implementation to be done to create common standards that would bring interoperability by default.

## 7.5 User-testing and stakeholder engagement

AI technologies that have had stakeholders engaged in the design following a user centred 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) [GOV.UK 2020].

## 7.6 Privacy and security

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

# 8 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 to understand this performance in comparison to the current standard for that use case. For the purposes of this report, we use the term 'algorithmic validation' to describe this evaluation of the adequacy of the AI model, *in silico* in contrast to 'clinical validation' in which the whole AI health technology is evaluated in the context of the clinical pathway.

## 8.1 Requirements

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. 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 the area under the receiver operator curve (AUC) and the area under the precision-recall curve (AUPRC) may also be helpful.

At present, the current standard of care is often a health professional doing the same task, but other relevant comparators may include other AI technologies. Over time, the standard of care will likely evolve to humans and AI (augmented intelligence). Beyond this, a world where AI works alone (autonomous AI) in a clinical setting currently appears to be quite a way off but would require a similar comparative approach.

## 8.2 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. A major limitation is that AI technologies may not perform well in populations or contexts that are different from that in which the training data was collected.

External validation is critical for all health AI models to show that the model can be used beyond the data with which it was tested and trained 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:

- A description of the input data type and source including where, when, and how it was collected.
- A description of the demographic spread of the data including gender/sex, age, and race/ethnicity. These data points help indicate how inclusive the data is, and how representative it is of the target population for the intended use of the AI health technology.
- 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.

The 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 that underpins the quality of the 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.

## 8.3 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 with 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 of 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(s) 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 in 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 a systematic under-performance in one or more groups.

## 8.4 Benchmarking against the current standard of care or other AI models

In order to understand the performance of the tool, an 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 for both developers and regulators. Benchmarking against unseen datasets also has a number of potential uses beyond the 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 that 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 an 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 made fully open and reproducible, may be required to have better performance to be attractive to buyers or commissioners 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 national health service [NHS Datasets], in the United Kingdom (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], [Morley et al., 2022].

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 (ITU/WHO FG-AI4H Open code initiative, 2020) [OCI 2020], [aiaudit.org].

## 8.5 Reader study

A reader study evaluates the accuracy and clinical performance of a technology when used by a group of human readers i.e., contextual use by the intended user [Gennaro, 2018]. A reader study, rather than evaluating the 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.

## 8.6 Special considerations

### 8.6.1 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, in some LMICs, this may not only limit the 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 to also compromise the safety and performance of AI tools and is an area which requires ongoing scrutiny and clear targeted action.

Within populations, under-representation of certain groups of people 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 (EHRs)) and instruments such as imaging devices, also the existence of data generation pipelines such as national screening programmes, and adoption of EHRs in well-resourced settings has created an infrastructure that can generate a growing pool of digital data. However, it should be noted that certain information systems operated by HPCs may be predominantly designed for administration, billing and insurance purposes and may not necessarily contain health data of best quality or format for the generation and testing of clinically orientated AI models.

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 [Walter et al., 2020]. The result of this is that these marginalised populations may not benefit from technological advances. Health data poverty – the inability of individuals, groups, or populations to benefit from discovery or innovation due to insufficient data that are adequately representative - can play out at both ends of the economic spectrum [Ibrahim 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, and 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, but are not limited to:

- Concerns with the inclusiveness and representativeness of the data, including systematic underperformance or biases because of under-representation of 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 an additional data collection burden. These biases will emerge during modelling and subsequently diffuse through the resulting algorithm.
- Concerns with the safeguarding of individual privacy. The collection, use, analysis, and sharing of health data have consistently raised broad concerns about individual privacy, and the risk that it may harm an individual or cause a wrong.
- Valid concerns about 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.
- 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.

## 9 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. The performance of a model *in silico* may not translate into the 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, which is 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, in parallel with deployment.

The overarching aim should be to design studies that give confidence in results by minimising bias and, therefore, provide confidence in the reliability and robustness for decision makers. An important aspect of this is reporting transparency of studies, including prospective analysis plans, and ensuring reporting 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 that is robust and meaningful either clinically or for the system. It is important to acknowledge that the 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 that need to be considered when planning the clinical evaluation phase.

The principles of good clinical study design are equally applicable to 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 enhancing the QUAlity and transparency of health research (EQUATOR) network, notably the publication of standard protocol items: recommendations for interventional trials (SPIRIT)-AI [Rivera et al., 2020] (for reporting of study protocols) and consolidated standards of reporting trials (CONSORT)-AI [Liu et al., 2020] (for reporting of trial reports); additional EQUATOR guidelines are currently in development for diagnostic test accuracy studies standards for the reporting of diagnostic accuracy studies (STARD-AI [Sounderajah et al., 2020]) and studies of prediction models [Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)-AI] [Collins and 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 gatekeepers (e.g., health tech assessors, 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) in 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 a retrospective selection of most favourable outcomes or of positive results 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 the 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 the commencement of the study.
- *Reporting of study conduct and results* – open and transparent reporting should align with the registered protocol, include any protocol deviations and full analysis of planned outcomes according to their pre-specified hierarchy. Participant flow (including exclusions at the participant level, exclusions at the 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 a systematic safety reporting may be considered adequate for some AI tools, particularly those that are considered to be low risk to patients. By drawing together good study methodologies, 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 that will enable regulators and other gatekeepers make better decisions regarding AI systems, enabling their populations to benefit from these interventions, whilst also reducing the risk of harm.

## 10 Deployment and ongoing evaluation

### 10.1 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 needs for which the capacity of human resources is inadequate and worsening such as in screening programmes. Second, there 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 an 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 the 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 the 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 centre 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 [Kwong et al., 2022].

The deployment phase also provides greater 'real world' information regarding many of the impacts discussed earlier (*clause 9: 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. In January 2021, the US FDA announced 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 [FDA GMLP] (good machine learning practice for medical device development), 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 new EU artificial intelligence act in April 2021 [Artificial Intelligence Act], which sets out similar principles to the SPS and ACP, albeit less comprehensively described than the US FDA proposals. The UK medicines and healthcare products regulatory agency [MHRA RA] has also provided guidance in their change programme roadmap for software and AI as a medical device [MHRA Roadmap].

*Box 3: Ethical considerations on risks of AI to patient safety*

Patient safety could be at risk from the use of AI which may not be foreseen during regulatory review of the technology. 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 the 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 [Liu et al., 2022]. 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 an outcome (e.g., providing a risk estimate that leads to a patient not receiving a kidney transplant).

There is a 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 [Ibrahim et al. 2021], [Adamson et al., 2018] or for example, people at extremes of age (WHO, *Ageism in artificial intelligence for health, 2022*) [WHO Ageism].

## 10.2 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 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 HCPs, users, patients, and the public also become gatekeepers for discovering and acting upon potential risks.

## 10.3 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 revision 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 [FDA MAUDE] database and the UK MHRA alerts and recall database for medical devices [MHRA Alerts]. 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 and in some cases over extended time periods (for example, where an incorrect diagnosis first results in incorrect treatment, which in turn results in a poor outcome). 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.

## 10.4 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 that enable direct and transparent reporting of adverse events should be in place and users should be supported and encouraged to report openly.

The 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.

### 10.5 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 that 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 that 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 [Liu et al., 2022]. 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 the 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, an 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.

## 11 Economic evaluation

### 11.1 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 - with an intention not only to increase quality or health outcomes, but often driven by a desire for automation and cost reduction of repetitive tasks. As such, the cost of a conventional care pathway versus an AI enabled pathway is often the critical economic information that decision makers are looking for. Conventional health economic analyses often fail to cost an entire pathway of care, as the focus is commonly on specific interventions within an established process. Therefore, while the practice of economic evaluation of AI health technologies will be based on similar principles of economic evaluation for any other health technologies, it will necessarily take a broader view with an aim to represent the process disruption associated with the introduction of AI health technologies.

Economic evaluation is defined as a comparative analysis of two or more interventions in terms of their costs and consequences [Drummond et al., 2015]. It enables the assessment of not only 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. Economic evaluation requires robust evidence relating to health, costs, and resource impacts. Developers are encouraged to invest in generating health economic evidence - context specific for decision makers. An evidence standards framework issued by NICE was recently updated to incorporate adaptive algorithms and align with key regulatory requirements.

The fundamental concept guiding the use of economic evaluation is opportunity cost – the foregone benefits of investing limited resources in 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 programme 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 early development of methods for the economic evaluation of individual health interventions were centred around pharmaceuticals, driven by the governments' need to make evidence-informed and definable decisions. Although the 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.

The conduct of economic evaluation of AI health technologies, in common with digital health interventions (DHI) more broadly, 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 [McNamee et al., 2016]. Many health interventions will differ in marginal cost at scale and clinical effects may alter over time based on the health professional experiences in use in established clinical pathways, however, the dynamics of clinical effects and costs associated with AI health technologies are 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 is 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 generation and 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, a characteristic of AI-health systems is that they will, produce data through routine use in a health system – this enhances the role of real-world evidence (RWE) in the economic evaluation of AI-health technologies, enabling the assessment to be informed by evidence beyond the clinical trial setting and incorporate a growing evidence base on routine clinical practice. While there are multiple guidelines and texts on methods for the economic evaluation of non-digital health interventions, there is relatively limited research on methods for the economic evaluation of AI health technologies. Work particularly related to country specific [Unsworth et al., 2021], [Gomes et al., 2022] and global (WHO, 2016) methods for evaluation of the broader group of digital health technologies (DHT) will be informative to approaches specifically for AI. A framework for the economic evaluation of digital health interventions recently developed by the World Bank includes specific requirements for AI health technologies [Wilkinson et al., 2023], and the body of literature on applied economic evaluation of AI health technologies develops across contexts, more definitive and specific normative guidance on methods and approaches that will be made possible.

## 11.2 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 (HTA) agencies, the objective of an economic evaluation is to inform the 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. Health is a generalised measure such as the quality adjusted life year (QALY) or disability adjusted life year-averted (DALY). In this way a government 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 the aggregation of results into a representation of net benefit and benefit cost ratio. In instances where the health impacts of 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 the 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 clinical pathway costs, resources and impacts are accounted for. This analysis is amenable to DHTs whereby its effects on the clinical care pathway may be multiple. CCA provides decision makers with a full representation of the consequences of the competing courses of action (for example, in case a new diagnostic method is able to reduce the need and/or frequency of other tests and outpatient attendance). 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 the financial implications of investing in a health intervention in a particular context. BIA addresses anticipated expenditure changes (commonly 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], [Budget impact analysis, 2018)]. BIA entails addressing the estimated use and costs of the proposed health technology, estimation of 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 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]. 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 the full net costs of the intervention, aligning with 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 that 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. 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 the ongoing development costs related to the refinement of the algorithm and adjustments to apply the algorithm to additional patient populations.

### 11.3 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]. Patents enable a period of market exclusivity to recoup research and development (R&D) costs by delaying the entry of competition. This may, however, have undesirable impacts including enabling those who hold intellectual property to set 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 is commonly informed by a health care professional to determine what medical interventions they require, influencing their demand for medical technology. Subsequently, in most countries either the government and/or health insurers exercise a degree of influence on the price and utilisation (through coverage/restrictions of indications [Drummond et al., 1997]. Payers often re-evaluate safety and effectiveness evidence as part of the deliberation process; with an objective to reward innovation whilst achieving optimal resource allocation [Barros 2020]. Digital health technologies may come in the form of capital associated with a one-off payment (e.g., MRI/CT), or software as a medical device (SaMD), that may be associated with existing capital (for example CT scanner) and the SaMD product is paid on either a subscription or fee-for-service or fee-per-use basis. Reimbursement of digital health technologies is limited, although an active area of exploration and change is led for example by countries like Germany (made possible by the German digital healthcare act, DVG) and the UK [Gerke et al., 2020].

## 12 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. It enables clinicians, patients, regulators, and other stakeholders to have the evidence they need to assess the safety, effectiveness and likely value of the technology and its performance in their setting. 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 an appropriate algorithmic audit [Liu et al., 2022], [aiaudit.org].

The key components of clinical evaluation have been discussed in detail throughout this working group-clinical evaluation (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.<br>Evaluation of interoperability and security.<br>Description of stakeholder engagement and user testing. |
| Algorithmic validation | Description of data used for development and testing including the size of the dataset, demographics of population (age, gender, sex, race, ethnicity), and setting data was collected (type of facility, date of collection).<br>Type of model used, and performance metrics obtained at internal and independent evaluation with a comparison against the current standard of care.<br>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.<br>Transparent reporting to describe the clinical evidence (using the appropriate EQUATOR guideline and standard performance metrics).<br>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?<br>Clear reporting of audits to describe the post-deployment performance (e.g., Medical algorithmic audit [Liu et al., 2022] and the WHO/ITU AI4H audit template [Verks 2020]. |
| Economic evaluation | A description of the economic evaluation carried out. |

## 13 Conclusions and recommendations for future action

This report provides a framework for current best practice evaluation of AI systems in health, primarily aimed at clinicians, researchers, and policymakers, 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 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 AI health 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 a requirement for high quality datasets to train and evaluate the performance of health AI systems. The working group recommends that clinicians, 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 development of models should evolve from being driven by the availability of data, to prioritising areas of clinical need and addressing health inequalities. Targeting resources towards these priorities is recommended, along with the curation of data required for the evaluation of AI systems at the various stages of development. This will allow decision makers to have better evidence of the extent to which the model generalises, and how it might perform 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 analysis. This is holding back the potential of these tools and affects trust between healthcare professionals and patients. Clinical studies are essential in order to assess whether *in silico* performance translates into a 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 the 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 the 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 what level of evidence is required, it is rare to find the results of these studies reported transparently and openly, often even for widely used digital health tools. This evaluation and the priority setting for digital health tools in all country settings requires a much more active role in health technology assessment, in addition to the role of regulators. Health technology assessment has 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 health technology assessment (HTA) capacity is low, this should be a key focus alongside expanding the use of digital health tools.

In conclusion, the working group provides a framework for the 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 the 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 for AI for health and the role of clinical evaluation. |
| *The adoption of effective, safe, ethical, inclusive, and fair AI systems into health systems is a global concern that requires input from a wide range of stakeholders. Clinical evaluation of AI systems including their underpinning data, performance, safety, and transparent communication of these results 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). |

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

| | |
|---|---|
| | Key findings and considerations:<br>Understanding the problem and intended use<br>– Evaluation of the purpose and demonstration that AI is the most suitable option for a particular clinical problem<br>– 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.)<br>Defining intended benefits<br>– Understanding of the intended benefits to the individual patient, clinical workflow, or health system (or a combination of these)<br>Describing potential risks and harms<br>– Patient level risks (e.g., misclassification, misdiagnosis, automation bias, delayed care, under-or overdiagnosis, unnecessary treatment)<br>– Clinical workflow risks (e.g., additional administrative or cognitive task burden for clinicians)<br>– System level risks (e.g., health economic costs of the clinical impacts)<br>Interoperability and security<br>– Consideration of interoperability requirements of AI technologies with other devices and IT systems (e.g., hardware and software upgrades)<br>– Consideration of security aspects (e.g., regarding data collection, privacy)<br>User-testing and user engagement<br>– 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 report to refer to the evaluation of the AI model *in silico*. |
| | Key findings and considerations:<br>Description of internal and external testing dataset<br>– 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<br>Training/testing data reporting should include<br>– Description of the demographic spread (including gender, sex, age, race, ethnicity)<br>– Performance metrics for the population as a whole and for key groups that might be vulnerable to under-representation in the training data set<br>– Description of data type, source, when, and how it is collected<br>– Quality of the training data, the robustness of the labels, understanding what is used as 'ground truth'<br>Algorithmic validation<br>– Purpose: demonstration of the robustness of the model and the performance on an acceptable level in the intended setting |

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

| | |
|---|---|
| | – 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<br>– External validation: on a representative high-quality dataset with accurate labels; transparent reporting of the performance metrics<br>Benchmarking of performance<br>– Evaluation against an accepted standard should be made<br>– Comparative benchmarking of AI models against an unseen dataset (similar process as the external validation) can improve algorithmic validation<br>– Foster the availability of unseen external datasets by initiatives, local, regional, and governmental bodies for creating / supplying commonly available datasets<br>Building high-quality datasets<br>– Obstacles to the availability of sufficient quality datasets with required labels, and representative of the local population results in a lack of health data for certain people groups or populations ('health data poverty')<br>– May result in poor safety and performance and lead to various harms including exclusion or exposure to under-performance<br>– Major problem: availability of the technology (e.g., EHR, national screening programmes) or even exclusion from the health system especially in non-affluent regions/countries<br>– Issue: low Internet connectivity and IT infrastructure for AI training and implementation. |
| Requirement 3)<br>Clinical validation | The term "clinical validation" in this report refers to the evaluation of AI technologies through interventional or clinical studies. |
| | Key findings and considerations:<br>– 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)<br>– Principles of good clinical study design are equally applicable to AI technologies but are currently not yet established<br>– 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<br>– Several guidance documents help optimise specific study designs when evaluating an AI intervention (through the EQUATOR network)<br>– 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.* |

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

| | |
|---|---|
| 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:<br>Deployment<br>– 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)<br>– Any need for further training or local tuning should be actively sought, as a critical part of the ongoing evaluation for efficacy and safety<br>– Challenge (also for regulators): determine the level of additional evaluation required to appropriately assure version updates of AI products and continuous learning or adaptive algorithms<br>– One approach: "good machine learning principles" (algorithm changes must be transparently labelled in specific protocols for change control plans)<br>Ongoing evaluation<br>– Monitoring of ongoing performance (both safety and effectiveness) is important to determine whether the AI product continues to deliver as expected<br>– AI systems are known to show poor generalisability when encountering new data and unexpected failure in spurious edge cases<br>Regulatory requirements<br>– Manufacturers should consider regulatory requirements: systematically carry out post-deployment monitoring of safety and performance and take corrective action when required<br>– This requires ongoing monitoring of the device's performance, or for an adverse event to be detectable and attributable to the AI device in question<br>– Manufacturer's post market clinical follow-up plan should consider the collection of post-deployment data due to potential harms that might occur within the clinical pathway<br>Relevant stakeholders<br>– Crystallization and definition of the most active stakeholders engaging in post deployment monitoring: users and developers, regulators, etc.<br>– Should be part of the indications for use statement and can inform the level of pot deployment surveillance<br>Algorithmic audits<br>– Purpose: detailed analysis may be performed in an algorithmic audit to determine what, how and why adverse events or algorithmic errors occurred<br>– Variable applicable: local performance monitoring, establish a "baseline performance" by e.g., assessment and prioritization of existing and potential risks. |

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

| | |
|---|---|
| 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:<br>Types of economic evaluation for AI-enabled digital interventions<br>– 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)<br>– Important considerations include the definition of the objective and the users of the economic evaluation<br>– Costs and effects highly dependent on local digital architecture and infrastructure<br>Reimbursement<br>– 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 six 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 explainability; fostering responsibility and accountability; ensuring inclusiveness and equity; promoting AI that is responsive and sustainable. |
| | Key findings and considerations:<br>To meet the appropriate governance, evaluation, and regulation, stakeholders developing AI health technology should:<br>– meet standards of scientific validity, accuracy, and explainability / reproducibility applied to medical technologies (in consideration of the infrastructure and institutional context)<br>– consider the decommissioning of existing services, and replace the previous services with the same or better level<br>– consider both, potential benefits and risks, e.g., with regard to biased (or selective) training data<br>– take full account of the total cost and investment required for its use, including digital infrastructure, training, maintenance, and monitoring costs<br>– 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). |

# Annex A

# Checklist for clinical evaluation of AI for health

This checklist (version 1) for clinical evaluation of AI systems in health is based on the framework for clinical evaluation developed by the FG-AI4H working group on clinical evaluation (WG-CE). It covers important aspects of evaluating an AI system across all relevant phases recommended by the group which are: design and purpose, analytical and clinical validation, ongoing monitoring together with economic evaluation and ethical considerations across all phases. The checklist was created to provide a harmonized and comprehensive approach for developers, implementers, and evaluators of AI systems in healthcare. It has been tested and applied by a research team from the University of Helsinki, Karolinska Institute, and Uppsala University conducting a study on digital microscopy for cervix cancer screening in Republic of (Kenya) and United Republic of (Tanzania) demonstrating its applicability but also showing potential gaps for future work. The checklist can be used as a guidance on the considerations of clinical evaluation by a wide range of stakeholders involved in the development, evaluation, and implementation of AI systems in healthcare, for instance, developers of AI, researchers, clinicians, and regulatory authorities.

Developers and implementors can use this as a checklist to plan their approach for a successful and impactful deployment of their AI system. Decision makers could use this as a framework for evaluating an AI system that is being considered for or undergoing deployment.

## 1 Model design and purpose

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

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

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

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

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

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

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

### *1.2 Define intended benefits*

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

– For example, improvement of the patient experience, including reduced waiting times and better clinical outcomes (e.g., improved survival rates, reduced complications compared with current context relevant standard of care).

– Quicker linkage from diagnosis to care or reduced out-of-pocket expenditure.

*1.2.2 What are the clinical workflow benefits?*

– For example, reduced administrative burden on health care professionals (HCPs).

– Increased time to care for HCP.

– Provision of a better HCP experience.

*1.2.3 What are the health system benefits?*

For example,

– Efficiencies found or created in pathways.

– Improved detection of cases.

– Better allocation of resources.

– Cost savings, addressing shortages of skilled HCPs.

### *1.3 Describe potential risks and harms*

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

*1.3.2 What are the potential <u>clinical workflow risks</u>, including removing safeguards, additional time, and administrative or cognitive task burden for HCPs?*

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

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

### *1.4 Interoperability and security*

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

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

*1.4.3 Does the novel AI technology comply and make use of existing communication standards (e.g., digital imaging and communications medicine (DICOM), fast healthcare interoperability resources (FHIR)?*

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

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

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

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

- For example, user feedback (quantitative or qualitative study).
- Interviews (qualitative study).
- Usability testing (qualitative study).
- Focus groups (qualitative study, delphi studies, quantitative study).
- Ethnographic study (qualitative study) [GOV.UK 2020].

### *1.6 Privacy and security*

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

## 2 Algorithmic validation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

*2.9 How was the "ground truth" established?*

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

## 3 Clinical validation

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

Considerations of specific elements important in clinical studies include:

### *3.1 Describe the study design*

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

### *3.2 Describe the population*

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

### *3.3 Describe the setting*

– Ensure that the study setting reflects the setting (or range of settings) of the intended use, again, diversity of setting is relevant, to provide sufficient confidence in performance outside of ideal scenarios.

### *3.4 Describe intervention(s)*

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

### *3.5 Describe intervention inputs and outputs*

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

### *3.6 Define the comparator*

– The comparator (whether parallel control group or other design) should be a relevant reference. This reference is commonly "standard practice" or "best practice" with a view to informing decision-makers as to whether the intervention reflects an improvement (or not) in current health deliveries.

### *3.7 Describe pre-specified outcomes relevant to all stakeholders*

– Ensure that outcomes are defined in advance and include those that are the most important to patients, and the key stakeholder groups; use of core outcome sets are recommended where they exist for the condition of interest; pre-specification avoids bias through a retrospective selection of the most favourable outcome or of positive results arising through chance and multiple testing.

### *3.8 Process measures*

– Describe relevant impacts on the overall health pathway such as positive or negative changes in time to diagnosis or treatment.

*3.9 Balancing measures*

– Consider upstream, lateral, and downstream consequences including changes in behaviour, changes in resource requirements, and potential ethical implications (such as loss of autonomy).

*3.10 Protocol deviations*

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

*3.11 Define the analysis*

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

*3.12 Describe the reporting of the 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 the commencement of the study.

*3.13 Reporting of study conduct and results*

– Open and transparent reporting should align with the registered protocol, include any protocol deviations, and full analysis of planned outcomes according to their pre-specified hierarchy. Participant flow (including exclusions at the 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].

## 4 Deployment and ongoing monitoring, regulatory requirements, and AI audits

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

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

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

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

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

*4.4 Has the AI technology achieved regulatory approval?*

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

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

4.6 *What stakeholders are monitoring the AI product?*

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

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

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

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

– AI audits can help discover the occurrence of adverse events and also help understand why these happened. Through the AI audit, existing and potential risks can be assessed and prioritised, risk mitigation plans can be put in place, and future audits can monitor whether risk mitigation measures were successful in avoiding harm. Detailed analysis may be performed through a "medical algorithmic audit", such as in [Liu et al., 2022].

## 5 Economic evaluation and reimbursement

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

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

*5.2 Define potential opportunity cost related to the AI model.*

– The foregone benefits of investing limited resources in one course of action rather than another.

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

– These include for example, 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. Health is a generalised measure such as the quality adjusted life year (QALY) or disability adjusted life year-averted (DALY).

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

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

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

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

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

# 6 Communication of results

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

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

# Bibliography

[Adamson et al., 2018] Adamson, A. S., and Smith, A. (2018), *Machine Learning and Health Care Disparities in Dermatology*. JAMA Dermatology, 154(11), 1247. https://doi.org/10.1001/jamadermatol.2018.2348

[Artificial Intelligence Act] The Artificial Intelligence Act. (2021) *The Artificial Intelligence Act*. https://artificialintelligenceact.eu/the-act/

[aiaudit.org] aiaudit.org | *ITU/ WHO AI4H Assessment Platform*. Retrieved from https://aiaudit.org/projects/assessmentplatform/

[Arrow 1962] Arrow, K. J. (1962), *Uncertainty and the Welfare Economics of Medical Care*. PDF, 53(5). https://web.stanford.edu/~jay/health_class/Readings/Lecture01/arrow.pdf

[Barros 2020] Barros, P. P. (2020), *Incentives for R&D: Payment Options and Pricing Challenges*. Office of Health Economics. https://ideas.repec.org/p/ohe/sembri/002289.html

[Budget Impact Analysis] Budget Impact Analysis (2018), YHEC – York Health Economics Consortium. http://yhec.co.uk/glossary/budget-impact-analysis/

[Collins and Moons, 2019] Collins, G. S., and Moons, K. G. M. (2019), *Reporting of artificial intelligence prediction models*. The Lancet, 393(10181), 1577-1579. https://doi.org/10.1016/S0140-6736(19)30037-6

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

[DICOM standard] *dicomstandard*. Retrieved from https://www.dicomstandard.org/

[Djulbegovic et al., 2017] Djulbegovic, B., and Guyatt, G. H. (2017), *Progress in evidence-based medicine: a quarter century on*. The Lancet, 390(10092), 415-423. https://doi.org/10.1016/S0140-6736(16)31592-6

[Drummond et al., 1997] Drummond, M., Jönsson, B., and Rutten, F. (1997), *The role of economic evaluation in the pricing and reimbursement of medicines*. Health Policy, 40(3), 199-215. https://doi.org/10.1016/S0168-8510(97)00901-9

[Drummond et al., 2015] Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L. and Torrance, G. W. (2015), *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press. https://nibmehub.com/opac-service/pdf/read/Methods%20for%20the%20Economic%20Evaluation%20of%20Health%20Care%20Programmes.pdf

[FDA GMLP] FDA. U.S. Food & Drug Administration, *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

[FDA MAUDE] FDA. U.S. Food & Drug Administration. About Manufacturer and User Facility Device Experience (MAUDE). https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/about-manufacturer-and-user-facility-device-experience-maude

[finddx.org] *Access to COVID-19 Tools (ACT) Accelerator*. (2022), Finddx.Org. https://www.finddx.org/covid-19/

[FG-AI4H Del 0.1] ITU/WHO FG-AI4H Deliverable 0.1 (2022), *Common unified terms in artificial intelligence for health*. International Telecommunications Union (ITU). https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/del/DEL0_1-A20220922-%28Prepub%29.pdf

[FG-AI4H D02] ITU/WHO FG-AI4H Deliverable 02 (2022), *Overview of regulatory concepts on artificial intelligence for health*. International Telecommunications Union (ITU). https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/del/DEL02-A20220922-%28Prepub%29.pdf

[Gebru et al., 2018] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021), *Datasheets for Datasets*. arXiv. https://doi.org/10.48550/arXiv.1803.09010

[Gennaro, 2018] Gennaro, G. (2018). *The "perfect" reader study*. European Journal of Radiology, 103, 139–146. https://pubmed.ncbi.nlm.nih.gov/29653758/

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

[GDHP] Global Digital Health Partnership. (2020), *AI for healthcare: Creating an international approach together*. NHS. https://transform.england.nhs.uk/media/documents/GDHP_Creating_an_international_approach_together.pdf

[GOV.UK 2020] GOV.UK (2020), *Evaluating digital health products*. https://www.gov.uk/government/collections/evaluating-digital-health-products

[Gomes et al., 2022] Gomes, M., Murray, E., and Raftery, J. (2022), *Economic Evaluation of Digital Health Interventions: Methodological Issues and Recommendations for Practice*. PharmacoEconomics 40, 367-378. https://doi.org/10.1007/s40273-022-01130-0

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

[HL7 FHIR standard] HL7 FHIR Standard. Retrieved from https://www.hl7.org/fhir/overview.html

[Ingenbleek et al., 2013] 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*. Journal of Product Innovation Management, 30(3), 560-573. https://doi.org/10.1111/jpim.12008

[IMDRF 2013] International Medical Device Regulators Forum (IMDRF) (2013), *Software as a Medical Device (SaMD): Key Definitions*. https://www.imdrf.org/documents/software-medical-device-samd-key-definitions

[IMDRF 2014] International Medical Device Regulators Forum (IMDRF) (2014), *Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations*. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

[Ibrahim et al., 2021] Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D., and Denniston, A. K. (2021), *Health data poverty: an assailable barrier to equitable digital health care*. The Lancet Digital Health, *3*(4), E260-E265. https://doi.org/10.1016/S2589-7500(20)30317-4

[IDEO design kit] IDEO.org Design Kit. Retrieved from https://www.designkit.org/methods.html

[ISPOR] *ISPOR*. Retrieved from https://www.ispor.org/

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

[Kwong et al., 2022] Kwong, J. C. C., Erdman, L., Khondker, A., Skreta, M., Goldenberg, A., McCradden, M. D., Lorenzo, A. J., and Rickard, M. (2022), *The silent trial - the bridge between bench-to-bedside clinical AI applications*. Frontiers in Digital Health. https://doi.org/10.3389/fdgth.2022.929508

[Liu et al., 2019] 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., and 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*. https://doi.org/10.1016/s2589-7500(19)30123-2

[Liu et al., 2020] Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., and the SPIRIT-AI and CONSORT-AI Working Group. (2020), *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. The Lancet Digital Health, 2(10), E537–E548. https://doi.org/10.1016/S2589-7500(20)30218-1

[Larson 2021] Larson, D. B., Harvey, H., Rubin, D. L., Irani, N., Tse, J. R., and 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, 18(3), 413–424. https://doi.org/10.1016/j.jacr.2020.09.060

[Lyell et al., 2021] 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

[Liu et al., 2022] Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., and Oakden-Rayner, L. (2022), *The medical algorithmic audit*. The Lancet Digital Health, 4(5), E384–E397. https://doi.org/10.1016/S2589-7500(22)00003-6

[McNamee et al., 2016] McNamee, P., Murray, E., Kelly, M. P., Bojke, L., Chilcott, J., Fischer, A., West, R., and Yardley, L. (2016). *Designing and Undertaking a Health Economics Study of Digital Health Interventions*. American Journal of Preventive Medicine, 51(5), 852–860. https://pubmed.ncbi.nlm.nih.gov/27745685/

[Macdonald et al., 2021] Macdonald, J., März, M., Oala, L., and Samek, W. (2021), *Interval Neural Networks as Instability Detectors for Image Reconstructions*. Bildverarbeitung für die Medizin 2021 pp 324–329. https://doi.org/10.1007/978-3-658-33198-6_79

[MHRA Alerts] *Alerts, recalls and safety information: drugs and medical devices*. MHRA GOV.UK. Retrieved from https://www.gov.uk/drug-device-alerts

[MHRA RA] MHRA, *Medicines & Healthcare products Regulatory Agency*. GOV.UK. https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency

[MHRA Roadmap] *Software and AI as a Medical Device Change Programme - Roadmap*. (2023), MHRA GOV.UK. https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap

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

[Morley et al., 2022] Morley, J., Murphy, L., Mishra, A., Joshi, I., and Karpathakis, K. (2022), *Governing Data and Artificial Intelligence for Health Care: Developing an International Understanding*. JMIR Formative Research, 6(1):e31623. https://doi.org/10.2196/31623

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

[NHS Datasets] *An Innovator's Guide to the NHS: Navigating the barriers to digital health*. (2020), Boehringer Ingelheim. https://www.boehringer-ingelheim.co.uk/sites/gb/files/documents/innovators_guide.pdf

[NICE] Kaltenthaler, E., Tappenden, P., Paisley, S., and Squires, H. (2011), *NICE DSU Technical Support Document 13: Identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-Effectiveness Models*. National Institute for Health and Care Excellence (NICE). http://www.ncbi.nlm.nih.gov/books/NBK425832/

[OCI 2020] ITU/WHO FG-AI4H Open Code Initiative. *Focus Group on Artificial Intelligence for Health*. GitHub. Retrieved from https://github.com/FG-AI4H

[Open Data Institute 2021] World Health Organization, Open Data Institute. (2021), *Data Governance Maturity and Best-Practices*. Health Data Governance Summit. Pre-read: Health Data as a Global Public Good. 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

[PAHO 2019] PAHO. *Data Governance in Public Health*. | Digital Transformation Toolkit, Knowledge Tools. https://iris.paho.org/bitstream/handle/10665.2/56576/PAHOEIHISdttkt23220027_eng.pdf?sequence=1&isAllowed=y

[People + AI Guidebook] *People + AI Guidebook*. Retrieved from https://pair.withgoogle.com/guidebook

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

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

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

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

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

[SDGs] United Nations *Sustainable Development Goals*, Retrieved from https://sdgs.un.org/goals

[theiet.org] Design and evidence. The Institution of Engineering and Technology. Retrieved from https://www.theiet.org/impact-society/factfiles/healthcare-factfiles/design-and-evidence/

[The Lancet 2018] The Lancet. (2018), *Is digital medicine different?* The Lancet, 392(10142). https://doi.org/10.1016/S0140-6736(18)31562-9

[Unsworth et al., 2021] Unsworth, H., Dillon, B., Collinson, L., Powell, H., Salmon, M., Oladapo, T., Ayiku, L., Shield, G., Holden, J., Patel, N., Campbell, M., Greaves, F., Joshi, I., Powell, J., and Tonnel, A. (2021), *The NICE Evidence Standards Framework for digital health and care technologies – Developing and maintaining an innovative evidence framework with global impact*. DIGITAL HEALTH. https://doi.org/10.1177/20552076211018617

[Verks 2020] Verks, Boris. (2020), ITU/WHO FG-AI4H *DAISAM Audit Reporting Template*. ITU. https://luisoala.net/assets/pdf/standards/FGAI4H-J-048.pdf

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

[Walter et al., 2020] Walter, M., and Carroll, S. R. (2020), *Indigenous Data Sovereignty, governance and the link to Indigenous policy*. Indigenous Data Sovereignty and Policy. 1st Edition. https://doi.org/10.4324/9780429273957-1

[Wiegand et al., 2020] Wiegand, T., Lee, N., Pujari, S., Singh, M., Xu, S., Kuglitsch, M., Lecoultre, M., Riviere-Cinnamond, A., Weicken, E., Wenzel, M., Leite, A. W., Campos, S. and Quast, B. (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

[Wilkinson et al., 2023] Wilkinson, T., Wang, M., Friedman, J., Prestidge, M. (2023), Publication: A Framework for the Economic Evaluation of Digital Health Interventions. Policy Research Working Paper 10407. © World Bank Group. http://hdl.handle.net/10986/39713

[WHO Ageism] World Health Organization (2022), *Ageism in artificial intelligence for health*. https://www.who.int/publications-detail-redirect/9789240040793.

[WHO Governance] World Health Organization. *Ethics and governance of artificial intelligence for health*. (2021), Retrieved from https://www.who.int/publications-detail-redirect/9789240029200

[WHO Framework] World Health Organization. (2021), *Generating Evidence for Artificial Intelligence Based Medical Devices: A Framework for Training Validation and Evaluation*. Retrieved from https://www.who.int/publications-detail-redirect/9789240038462

[WHO Evaluation] World Health Organization. (2016), *Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment*. World Health Organization (WHO). https://apps.who.int/iris/bitstream/handle/10665/252183/9789241511766-eng.pdf?sequence=1&isAllowed=y

[YHEC] *YHEC | Industry Experts in Health Economic Consultancy*. YHEC – York Health Economics Consortium. Retrieved from https://yhec.co.uk/