|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-L-040 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 19 – 21 May 2021 |
| **DOCUMENT** |
| **Source:** | Editors DEL7.4 |
| **Title:** | Updated DEL7.4: Clinical evaluation of AI for health |
| **Purpose:** | Discussion |
| **Contact:** | Co-chairNaomi LeeThe LancetUnited Kingdom | E-Mail:naomi.lee@lancet.com  |
| **Contact:** | Co-chairShubhanan UpadhyayAda HealthGermany | E-Mail:shubs.upadhyay@ada.com |
| **Contact:** | Co-chairEva WeickenFraunhofer Heinrich-Hertz-InstitutGermany | Email:eva.weicken@hhi.fraunhofer.de  |

|  |  |
| --- | --- |
| **Abstract:** | This document provides an overview of the current challenges of *"Clinical Evaluation of AI for Health"*. It is part of the deliverable-series 7.1-7.4 that are outlined by deliverable No.7 "*AI for Health Evaluation considerations".*Although the performance of AI models in health is often measured by their accuracy, establishing confidence among clinicians, patients, researchers and policy makers in the safety, efficacy, and cost-effectiveness of AI solutions in health requires a more comprehensive evaluation.The purpose of the deliverable No.7.4 is to outline the current best practice, the principles and outstanding issues for further considerations related to clinical evaluation of AI models for health. It serves as the output document of the WHO/ITU Focus Group on AI for Health (FG-AI4H) Working group on Clinical Evaluation of AI for Health (WG-CE).This version of the deliverable is issued also as FG-AI4H-L-040 (Online, 19-21 May 2021) |

**Call for participation**

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

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

Change Log

* FG-AI4H meeting L, 19 – 21 May 2021:
	+ Initial formatting clean-up applied by secretariat.
	+ Deliverable 7.4 includes the status update of the current WG-CE outline draft (version 1.1). ***Please* *note that this iteration of the outline draft (version 1.1.) is not finalized yet and corresponds to the document version that was shared with the WG-CE members for review-round 1 in beginning of April 2021. Therefore, the member's comments and feedback (in verbal and written form) have not yet been included***. The work on the implementation of the feedback of review-round 1 is currently in progress.
	+ This iteration of DEL. 7.4 (outline draft version 1.1.) provides an overview of the outcome of the WG-CE activities until the time before the last WG-CE meeting in April 2021(also see chronologic timeline below). Compared to meeting K, the document has been completely updated and restructured following the insights of the WG-CE meetings and discussions.
	+ Next steps:
		- Inclusion of the comments and feedback of review round 1, synchronization of the language within other FG-AI4H working groups and production of outline draft version 1.2
		- circulate outline draft version 1.2 with WC-CE members for review round 2
		- According to need: Follow-up meeting for final comments/review
		- Develop final outline draft > Deliverable 7.4
* The document is based on three earlier versions:
* FG-AI4H meeting K, 27 – 29 January 2021: [K-041](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BBC55470D-455E-4362-AA19-96990C9702A5%7D&file=FGAI4H-K-041.docx&action=default&CT=1621348538619&OR=DocLibClassicUI) – revised version with additions in some sections based on the output of the inaugural WG-CE workshop in October 2020 and the follow-up meetings in subgroups on pre- and post-deployment on clinical evaluation 15&16 December 2020
* FG-AI4H meeting J, 30 September – 2 October 2020: Deliverable draft "*Clinical Evaluation of AI for Health*" [J-053](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7BCBB4539C-36CA-46A4-A015-0A56267FDD77%7D&file=FGAI4H-J-053.docx&action=default), submitted to the secretariat of the ITU/WHO Focus Group on "AI for Health" (tsbfgai4h@itu.int) and presented by the authors.
* Working draft sent as an update on 24 March 2020 to tsbfgai4h@itu.int, titled DEL7.4: "Clinical Evaluation of AI for Health"

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

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

**Contributors**

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

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

All contributors of the outline draft (in alphabetical order):

|  |
| --- |
| Agata Piekut |
| Alexandre Chiavegatto Filho |
| Alixandro Werneck Leite |
| Ally Salim |
| Andrew Bastawrous |
| Andrew Murchison |
| Arnav Mahajan |
| Arne Bathke  |
| Audrey Menezes |
| Daniel Fürstenau |
| Darlington Akogo |
| Dean Ho |
| Ernest Darkoh |
| Falk Schwendicke |
| Farah Magrabi |
| Ferath Kherif |
| Grace Hatton  |
| Harpreet Sood |
| Hussein Ibrahim |
| Inês Sousa |
| Irv Loh |
| Jakir Masud |
| Jana Fehr |
| Jane Elisabeth Carolan |
| Jesse Ehrenfeld |
| Johannes Starlinger |
| Jonghong Jeon |
| Kate Loveys |
| Khondaker Mamun |
| Lina Porras |
| Lisa Murphy |
| Luis Lapao |
| Luis Oala |
| Magen Allen |
| Manish Srivastava |
| Matthew Fenech |
| Melissa McCradden |
| Michal Kurtys |
| Natalie Pankova |
| Nina Linder |
| Oommen John |
| Osama Abou Elkhir |
| Pat Baird |
| Pradeep Balachandran |
| Reza Jarral |
| Rigveda Kadam |
| Robert Gaudin |
| Rose Nakasi |
| Rubeta Matin |
| Saurab Guphta |
| Shariful Islam |
| Stephanie Kuku |
| Stephen Gilbert |
| Tavpritesh Sethi |
| William Wasswa |
| Zdenek Gütter |

**CONTENTS**

|  Page |
| --- |
| [1 Scope 4](#_Toc72257031)[2 Introduction and Background 4](#_Toc72257032)[3 Scoping phase 6](#_Toc72257033)[3.1 Define the problem and intended use 6](#_Toc72257034)[3.2 Intended benefits 7](#_Toc72257035)[3.3 Defining risks 7](#_Toc72257036)[3.4 Interoperability risks 8](#_Toc72257037)[3.5 The importance of clinical stakeholders 8](#_Toc72257038)[3.6 Getting scoping right 8](#_Toc72257039)[4 Design phase 9](#_Toc72257040)[4.1 Taking a user centred approach 9](#_Toc72257041)[4.2 Engaging relevant stakeholders 10](#_Toc72257042)[4.3 Testing a health AI product with end users and stakeholders 10](#_Toc72257043)[5 Development phase 11](#_Toc72257044)[5.1 Training data / building 11](#_Toc72257045)[5.2 Testing and validation 11](#_Toc72257046)[5.3 Clinical studies, safety and efficacy 14](#_Toc72257047)[6 Economic evaluation 17](#_Toc72257048)[6.1 Introduction 17](#_Toc72257049)[6.2 Types of economic evaluation for AI-enabled digital interventions 18](#_Toc72257050)[6.3 How is economic evidence provided? 19](#_Toc72257051)[6.4 Reimbursement 20](#_Toc72257052)[7 Deployment of algorithm 21](#_Toc72257053)[8 Ongoing monitoring 21](#_Toc72257054)[8.1 Introduction 21](#_Toc72257055)[8.2 Regulatory requirements 22](#_Toc72257056)[8.3 Relevant stakeholders for post deployment monitoring 22](#_Toc72257057)[8.4 Algorithmic audits 23](#_Toc72257058)[9 Recommendations 24](#_Toc72257059)[10 Abbreviations 24](#_Toc72257060)[Bibliography 26](#_Toc72257061) |

**List of Tables**

|  Page |
| --- |
| [Table 1: Types of comparative economic evaluation models (extracted from Drummond et al. 2015) 19](#_Toc72257062) |

**List of Figures**

|  Page |
| --- |
| [Figure 1: Overview of datasets involved in a machine learning diagnostic algorithm: model development and evaluation 13](#_Toc72257063) |

FG-AI4H Deliverable DEL7.4

Clinical evaluation of AI for health

Summary

[TBD]

Keywords

[TBD]

# Scope

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

# Introduction and Background

Globally a growing shortage of clinicians, a rapid growth in digital health data, and an expansion in the usage of artificial intelligence (AI) models in other sectors has, along with other factors, contributed to an increasing interest in the use of AI for health and clinical uses.

As a growing number of models become available for use, researchers, patients, clinicians, and policy makers require a framework to understand whether the models are safe, effective, and cost-effective, and also to compare model performance with current standards of care, and between each other. The aim is to produce guidance for current best practice evaluation of AI technologies in health primarily aimed at researchers, clinicians, and policy makers, but that may also be useful for patients, civil society, and developers. A shared understanding of best practice evaluation should facilitate adoption of tools that are safe and effective, and have the potential to improve health outcomes for all. The emphasis throughout this document is on principles of evaluation to ensure that it is generally relevant across all countries, and does not depend on too many assumptions around particular health systems or the agencies involved.

"Evaluation is essential to finding out what works and what does not work, and why […]"[[1]](#footnote-1).

This document will consider how AI models for health should be evaluated along the AI life cycle. It is intended to promote best practice for evaluation of AI systems, rather than be prescriptive whether deployment is permissible or not. The points raised here pay particular consideration to the value proposition to the health ecosystem as a whole. Elements such as cost effectiveness, user perception and healthy equity are vital to the effectiveness of health delivery, but may not be adequately covered by regulatory standards alone. Building a culture of evaluation, best practice and upskilling developers and users of these products is one of the best ways to improve the quality of AI tools, alongside other digital health interventions. This document includes the scoping, design, and development phase of AI-driven technologies, their testing and validation, their assessment in clinical studies, their safety, efficacy, cost-effectiveness, implementation, and ongoing monitoring after deployment. Furthermore, the documentation draws on existing evaluation frameworks and their extensions tailored for special requirements on the evaluation of health AI products (e.g., the EQUATOR network reporting guidelines like CONSORT-AI[[2]](#footnote-2) and SPIRIT-AI[[3]](#footnote-3) for reporting of trial reports and study protocols; the "Model facts label"[[4]](#footnote-4) to ensure a safe use of decision-support tools in clinical practice; or the Machine Learning for Health (ML4H) auditing framework which demonstrates how audits can be carried out in practice)[[5]](#footnote-5). This document identifies potential gaps in the existing evaluation frameworks that may need more work and produces a set of recommendations for future work, as well as highlighting key documents for reference.

Performance measures, often used for AI models, such as accuracy may not translate into clinically meaningful or cost-effective improvements which are safe and cost-effective interventions in a complex clinical environment. While such an evidence-based medical framework exists to evaluate medical innovations such as pharmaceutical and surgical interventions, it has limitations when applied to evaluation of AI based interventions. For pharmaceutical interventions, evaluation is often divided into that required pre- or post-regulatory approval. This may be less appropriate for AI models, even those that require regulatory approval, as most are complex interventions and heavily dependent on context. As such, a more phased evaluation, including a period of mandatory post regulatory approval evaluation could be most appropriate for these tools.

This document is produced by members of the WHO/ITU Focus Group on AI for Health (FG-AI4H) Working Group on Clinical evaluation (WG-CE) in a joint effort with other groups and stakeholders stemming from various fields all over the globe (clinicians, academia, research, commissioning, health-start-ups, NGOs, etc.). WG-CE held its inaugural workshop in October 2020 followed by regular meetings, one of them on special considerations for clinical evaluation in low-and middle-income (LMIC) settings. To ensure that everybody's perspectives and viewpoints from the discussions are covered, the outline document was shared and reviewed by all WG-CE members. FG-AI4H is formed from a collaboration between the United Nations (UN) organizations World Health Organization (WHO) and International Telecommunication Union (ITU). As such, it has a global scope and interest in evaluation that supports the Sustainable Development Goals (SDGs), particularly SDG 3 on good health and well-being for all at all ages. The work of FG-AI4H is closely aligned to the principle that no one should be left behind, and the best practice recommendations look to uphold this principle. In this regard, this document will provide special consideration on clinical evaluation of AI for health implementations in LMIC settings.

When considering the applicability of AI-tools for LMIC-settings, there are a number of potential barriers to equitable access. Availability of representative datasets, with quality annotation is notably lower in LMIC settings. Improving the availability of representative and diverse data including the presentation of underrepresented populations for key medical conditions is a key priority. Complex data sharing arrangements may also be a barrier to validation of tools outside of the setting in which they were developed. Furthermore, poor technical infrastructure and a lack of access to technology (e.g., stable internet provision in rural areas) might be an obstacle, especially in LMIC settings. Transparency of datasets used to train and test is critical to understanding whether the tool may or may not perform in a comparable way in a setting different to where it was developed.

Due to overlapping topics with other FG-AI4H working groups, this guidance will also reference their considerations on, e.g. regulation, data specification, or ethics on AI for health (e.g., regulatory considerations that affect the evaluation of health AI technologies along the AI life cycle, or ethics of research, especially in LMIC settings). Furthermore, this guidance document will provide a consistent nomenclature following FG-AI4H documentation of other working groups to ensure a universal understanding and enable a simple application for users and stakeholders. All these considerations described above will be discussed in detail following the AI life cycle.

# Scoping phase

Innovation in healthcare can have well meaning, technical solutions developed with great optimism, but if poorly developed with only a superficial understanding of the problem being addressed, this can result in failure to implement and achieve the intended outcomes. In evaluating an AI tool, it is important to look for evidence that the developer has identified and understood the clinical problem, and demonstrated that AI is the right technology to solve it (see IDEO's Design Kit[[6]](#footnote-6) for ways to identify and understand a problem and Google's People and AI Guidebook[[7]](#footnote-7) on how to decide if AI is the right solution). In order to do this effectively, it is important for AI developers to engage with end users and clinical stakeholders to ensure the problem space is well understood. Engaging with clinicians and other healthcare professionals to understand clinical workflows and how AI tools could benefit them is crucial to buy-in and success of implementation. Additionally, providing healthcare professionals and/or future end users of an AI technology the opportunity to co-design the solution helps ensure it is fit for purpose, usable and will be adopted.

AI technologies are complex, dependent not only on the constituent code, but also on the training data, clinical setting and user interaction. They are most often deployed into a complex clinical pathway or even introduced into new clinical pathways altogether (for example, into new telemedical pathways or part of the addition of new triage tools). In the latter instance, the use of these new technologies also precipitates a new way of working for users. Therefore, for AI tools', safety and performance are highly context dependent. The description of the use case therefore has a substantial role, both to inform end users where the tool can safely and appropriately be utilised, and for regulated tools (the statement of intended use) to allow regulators to assess if the evidence of the analytical and clinical validation steps taken are appropriate and sufficient for the intended use case.

## Define the problem and intended use

An AI developer should show evidence that in the scoping phase (see AI life cycle), they have considered the following steps:

* Identify and define the specific problem to be solved by exploring the problem especially in LMIC settings. space and context.
* Define the intended users and understand the tasks they are trying to achieve and where exactly an AI solution could fit into that
* Consider if an AI technology is the right solution to the problem (i.e., what other ways could the problem be solved? For example, could the problem be solved by a face-to-face solution or a different technology?)
* Define use case based on intended users for the proposed AI tool in their context and the relevant workflow(s), considering:
* How and where does the tool fit in the clinical workflow?
* How could the interaction with the user affect the performance of the tool?
* How could the interaction with the tool affect the user?
* Define how might the AI tool meet the user needs (i.e., user acceptance criteria)
* In addition, there may be specific considerations related to the intended users or context (for example there may be specific nuance to tools used in paediatric settings that require thought about child protection, unique issues related to particular settings e.g. rural settings with little or no internet provision)
* Another key contextual aspect is around the availability of resources. These are subject to large variability depending on which setting the AI developer might be exploring or scoping. The availability of datasets representative of the target population, data structure and quality, personnel and expertise are good examples. Equally there may be variations of clinical pathways in different regions. There are also socio-cultural variations around data and technology affecting the willingness to design and implement AI tools.

An interesting example of how to consider, and then communicate the intended use and benefits to clinical stakeholders using a model facts label approach: [https://www.nature.com/articles/s41746-020-0253-3,](https://www.nature.com/articles/s41746-020-0253-3) also see https://aiaudit.org/assets/pdf/standards/FGAI4H-J-048.pdf.

## Intended benefits

The developer should demonstrate that they have considered the intended benefits to the individual patient, clinical workflow and system level are important, as this will shape the plan to evaluate them.

What are the clinical or system level benefits that the tool is aiming to have? These may be at patient level, (for example, benefits to patient experience, reduced waiting times). They could also be clinical outcomes (improved survival rates, reduced complications compared with current context relevant standard of care). There may also be advantages to the clinical workflow or system (reduced administrative burden on HCPs, increased time to care, HCP experience, or cost savings).

Understanding the intended benefits and outcomes to influence (and usually improve) through the implementation of an AI technology is important for evaluating its effectiveness. It is recommended that, during the scoping phase, an AI developer creates the first version of a logic model or theory of change. These types of models help define how a product works and how it will lead to the benefits and outcomes it was designed for. A logic model or theory of change is a living tool that can be updated throughout the development of a product, especially as understanding of how it functions within the complex environment deepens. For more information: <https://www.gov.uk/guidance/define-how-your-product-works-evaluating-digital-health-products>

## Defining risks

As with all medical interventions, it is important to weigh up benefits and risks. AI tools can have a myriad of risks (e.g., data security, covered later) attached to them, and this section focuses on risks that could result in clinical harm to patients. Traditional frameworks around risk in medical interventions are focused on those of the individual, but in the case of AI models that might be used on a large scale, population level risks are also pertinent for an AI developer to consider. Depending on the task, the user and the context, the risk profile may vary. There are examples of frameworks to help AI developers to define which risk class their tool might belong to (NICE evidence for effectiveness, FDA, EU MDR). In general, the higher the risk class, the greater the burden on the developer to demonstrate that their tool's benefits outweigh any potential risks, and the greater the responsibility to show how those risks are mitigated.

Developers should show that they have considered potential risks including patient level risks (examples might include misclassification, misdiagnosis, delayed care, under-or overdiagnosis or unnecessary treatment), clinical workflow risks (additional administrative or cognitive task burden for clinicians) and system level risks (the health economic costs of the patient level clinical impacts, plus the potential for these to be replicated at scale across large groups of people), overdiagnosis and overtreatment.

A major system level risk is the exacerbation of bias, data poverty and health inequalities. It must be noted that there is a general risk that AI technologies exacerbate existing well entrenched health inequalities. In the scoping phase, ensuring that the curation of data (both for training and validation) sufficiently represents the population being served is essential to mitigate this. Even for AI technologies targeted for specific patient subpopulations, it is important to ensure this and mitigate for bias. The availability of datasets from electronic health records in richer regions, or even richer areas of LMICs leads to the bias alluded to earlier in this section. It also results in the fact that it will only be these populations that stand to benefit, and those in data and economically poor regions stand to be left behind further. To mitigate against this, AI developers and health systems should, where appropriate, proactively collect data from patients from all regions, especially those furthest away from large urban centres in order to ensure the necessary diversity of data for training AI systems. Care must also be taken to tackle the issues around low internet connectivity and the availability of IT infrastructure for AI training and implementation, particularly in LMICs. Working with local healthcare professionals and adapting to available 'lower tech' IT solutions can be part of ensuring that AI systems are available for the greatest unmet need.

## Interoperability risks

Developers should demonstrate that they have considered interoperability. Interoperability is often overlooked and requires consideration early in the AI development life cycle. In addition to clinical workflow with clinicians, an AI developer should consider the interaction with other devices and other IT systems. It does not suffice to simply be 'connected' - there are nuances and complexities around the coding, structure and potential loss of information between systems that can have clinical as well as economic implications. The work of the IEEE forum summarises this well, "while two devices may be physically connected and one device is providing data to a second device, the most complex interoperability is not occurring unless the second device can 'understand' what it is receiving from the first device and can respond in kind. This understanding requires both a common syntax and common semantics..."

Whilst some groundwork has been done in areas such as imaging, or with the attempts to create communication standards (e.g., FHIR), there remains much work and implementation to be done. One potential solution for AI developers from health systems is an 'interoperability checklist' that could help cover the key considerations.

## The importance of clinical stakeholders

Early and sustained engagement from the clinical community is integral to the eventual implementation of an AI technology, particularly from the perspective of understanding clinical workflow and ensuring that tools work seamlessly and in harmony with clinicians. Equally important are considering how clinician workflow and pathways might be redesigned and recognising that integrating clinicians as a key part of the coding and curation of high- quality datasets will be also be vital in the long-term success of implementation.

## Getting scoping right

The Scoping Phase has the potential to be time well invested if done with a genuine desire to explore the problems and work with clinical stakeholders to understand the problem, outline intended benefits and start thinking about specific and general potential risks. It lays the foundation for a relationship with those clinicians (and a wider group) that extends into collaboration in the design phase and beyond.

# Design phase

When designing, developing and implementing a health AI product it is important to involve end users and stakeholders to ensure the resultant product meets user needs and is solving the problem it was designed to solve (see Scoping Phase for more information). The time spent identifying and understanding the right problem to solve and if AI is the right technology to solve that problem, is arguably the most important (see IDEO's Design Kit[[8]](#footnote-8) for ways to identify and understand a problem and Google's People and AI Guidebook[[9]](#footnote-9) on how to decide if AI is the right solution).

Similarly, it is equally important to involve end users and stakeholders in the design and evaluation of a health AI product to ensure their needs and perspectives are integrated into your assessment. Doing so balances understanding the clinical efficacy of a health AI product with understanding its usability and fit within existing workflows and standards of care.

Core questions when designing a health AI evaluation are:

1. Who are the intended users of the product?
2. Who are the intended beneficiaries of the product?
3. Which user problem(s) or need(s) is AI uniquely positioned to solve?
4. How will intended users and/or beneficiaries be impacted by the delivery of the product?
5. Which stakeholders will be involved in and/or impacted by the delivery of the product and how?
6. What are the evaluation questions (priority order) and how do you plan to answer them?
7. What methods, data and analytics tools will be used?
8. What research ethics protocols and/or guidance do you need to follow? (for example, the UK's Department of Health and Social Care recommends developers of data-driven technologies follow the Guide to Good Practice[[10]](#footnote-10))

## Taking a user centred approach

When designing, developing and evaluating a health AI product, it is recommended to adopt a user-centred approach (see questions 1-4). A human and/or user-centered approach to innovation bases the design, testing, development and evaluation of a health AI product on the needs of the people (users) impacted by the product[[11]](#footnote-11). This approach, commonly known as user-centered design, concerns both the people who will benefit from the service or action of the health AI product and also the people responsible for the delivery and maintenance of the health AI product. Product decisions made through user-centred design are complemented by understanding what is technologically feasible and economically viable[[12]](#footnote-12) for the health AI product's success.

For any digital health evaluation, therefore, you should consider the perspective of different user and stakeholder groups including, but not limited to, patients, healthcare professionals, carers, and hospital administrators or funders. You can also work with users and stakeholders to define evaluation criteria i.e., what is considered the success or failure of the health AI. Please note, the evaluation design needs to ensure there is no discrimination of individuals due to protected characteristics.

## Engaging relevant stakeholders

Engaging relevant stakeholders in the design and execution of your health AI product evaluation is crucial for building support for the evaluation and, subsequently, ensuring access to information and learning from those impacted by the implementation of the health AI product. Unsurprisingly, an evaluation is more likely to fail if stakeholders do not understand why it is important, do not share their insights and/or do not act on the recommendation(s) of the evaluation.

When designing your evaluation, you should clearly define and understand which stakeholders will be involved in and/or impacted by the delivery of the health AI product (see question 4) and why you would engage with them based on what you need to achieve (see the Australian Government's Stakeholder Engagement Framework[[13]](#footnote-13) for guidance on how to identify and plan stakeholder engagement).

These stakeholders can then be grouped based on their influence over the successful implementation and adoption of the health AI product and their level of interest (techniques include power versus interest grid and stakeholder influence diagrams).

There needs to be an ongoing dialogue with stakeholders throughout an evaluation process. Therefore, we recommend adopting an engagement approach that focuses on building trust, confidence and collaborative relationships with stakeholders that enables learning both ways.

## Testing a health AI product with end users and stakeholders

A number of user testing and evaluation methods can be carried out with end users and stakeholders of health AI throughout both the design and development process, and clinical evaluation to understand if/how the health AI product works in practice. These methods are also useful in the scoping phase to help you identify and gain an in-depth understanding of the problem and users' needs.

A mixed methods approach is usually required, especially for different points of enquiry. These methods include, but are not limited to:

* User feedback (descriptive study) [[14]](#footnote-14),[[15]](#footnote-15)
* Interviews (qualitative study) [[16]](#footnote-16)
* Usability testing (qualitative study) [[17]](#footnote-17)
* Focus groups (qualitative study)[[18]](#footnote-18)
* Delphi studies (quantitative study)
* Ethnographic study (qualitative study)[[19]](#footnote-19)

# Development phase

## Training data / building

AI tools are highly dependent on the training data used to develop them. The training data (type of data, location of the data collection, and type of population) determines what the model can analyse and output, and in which situation it could be expected to perform well. It is therefore directly related to the target use case.

In order to facilitate understanding and evaluation of a model, transparency of training data is fundamentally important. A description of the data used in model development (including data used for training, tuning and internal validation) should include the input data type, and source, where and when it was collected, and details of characteristics of the population included.

Quality of the training data, and the robustness of the labels will also affect the tool'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 diagnosis by an expert, understanding the training and experience of these experts, how many experts made a decision and how conflicts were resolved, all provide information which underpins the quality of labelled data.

AI tools are highly dependent on training data and depending on the use case, may not perform well in populations or contexts that are different to that in which the training data was collected. Providing information about the inclusiveness of the data, particularly the gender, age and racial/ethnic demographic information of the population on which the model is trained is important. Where demographic groups are underrepresented in the training population it may be important to specifically understand the tools performance in that group and whether or not the tool can perform adequately. Beyond the population of the training data, context is also critical. External validation enables an understanding of whether a tool can generalise beyond the data in which it was developed. It also may enable tools developed in one setting, for example in a specialist hospital, to be evaluated for use in another setting, for example in primary care. Without external validation however, tools cannot be assumed to generalise beyond the setting and population in which they were developed.

Obtaining datasets that are sufficiently representative, and of sufficient quality can be difficult. Access to representative datasets for validation is a particular issue in many low- and middle-income countries. Where datasets are available in low resource settings, there may also be limitations introduced by the quality of the data. The ability to produce robust datasets with high quality ground truth labels is likely to be affected by limitations elsewhere in the health setting affecting access to diagnosis and treatment. These major challenges have the potential to not only propagate inequality of access, but also to compromise safety and performance of AI tools, and is a potential area of future work (for example the newly launch iDAIR collaborative mentions use of collaborative, distributed, and responsible use of data as one of their main aims. (Advancing Digital Health and Artificial Intelligence Research through Collaboration. I-DAIR. (2020) (<http://i-dair.org/>)

## Testing and validation

For the purposes of this, and other FG AI4H documents, analytical validation refers to the process of validating the AI tool using data, but without performing interventional or clinical studies.

As discussed above, appropriate analytical validation in an independent, quality, external dataset demonstrates that a model is robust and performs to an acceptable level in the intended setting. It also enables the understanding of potential bias and generalisability (and any steps to understand these). 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 [[20]](#footnote-20) possible resulting degradations in predictive confidence or its learned decision heuristics[[21]](#footnote-21) and more [[22]](#footnote-22),[[23]](#footnote-23).

After training, testing should be carried out on an unseen portion of the original dataset, and further tuning may be performed. An AI tool must then be externally validated in a dataset that is independent from that in which it was trained (not merely an unseen portion of the training dataset) in order to demonstrate external validity. Analytical 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 in the intended setting. As with the training dataset, the validation dataset should be of adequate quality, with appropriate robustness of labels, and part of evaluating the AI tool, will require understanding the dataset required for external validation. By understanding the training and testing data, it is important to identify any high-risk cases or cases that may be underrepresented in the external dataset. Failure cases particularly those that are surprising or unusual should also be identified.



Figure 1: Overview of datasets involved in a machine learning diagnostic algorithm: model development and evaluation[[24]](#footnote-24)

The availability of external unseen datasets for analytical validation is a current challenge in many commercial and academic settings, currently requiring collaborations to be established for each tool. Where local, regional, and national bodies are interested in evaluating AI tools, they could hold their own hidden dataset to enable this external validation set (for example, an initiative scheme currently underway by the UK body NHSX to develop, which has 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 GDHP policy document.

In order to understand the performance of the tool, evaluation against an accepted standard should be made. The most appropriate standard for comparison may differ according to the intended use but common examples of standards are human performance in a similar task or other models (for example derived from logistic regression) with - strong evidence-based or mandated standards of accuracy, sensitivity and specificity (for example for screening tools). Depending on the intended use case, the requirement for comparative performance may be more or less stringent (for example when used as a triage or screening tool, a different level of comparative performance may be acceptable compared to a tool used for diagnosis).

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

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

Two of the most important challenges of applying AI in healthcare in LMIC include the diversity of data available for training the algorithms, and the existence of tools for AI deployment in the poorest areas of these countries. The first case arises from the fact that electronic medical records are more frequently available in the richest areas of LMIC, which leads to a higher collection of health data from these patients, improving AI predictions for rich patients but not necessarily for the majority of the country. Thus, there may be an increase in health inequalities as the algorithms will be able to make good decisions for high-income patients, but will provide relatively worse decisions for poorer patients. Increased effort and investments are needed to collect data from patients from all regions, especially those furthest away from large urban centres in order to ensure the necessary diversity of data for training the algorithms.

The second major challenge of using AI in LMIC comes from the low availability of internet and electronic systems necessary to run previously-trained AI algorithms and provide the predictive results to improve healthcare decisions. So even if the first challenge on data diversity is solved, the real-world application of AI algorithms can still be much harder in poorer regions of LMIC. A possible solution in this case is to use the electronic devices that are already routinely available in the care of these patients, such as cell phone applications that allow the insertion of patient data and the presentation of the results of predictive algorithms to healthcare professionals.

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

## Clinical studies, safety and efficacy

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

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

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

The principles of good clinical study design are equally applicable for AI systems. Systematic reviews and meta-analyses have drawn attention to the poor levels of design and reporting in published AI studies, across the whole development pathway[[25]](#footnote-25),[[26]](#footnote-26). Randomised controlled trials (RCT) remain the benchmark of clinical studies, in which key elements help to minimise bias and increase confidence in the findings[[27]](#footnote-27),[[28]](#footnote-28). 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.

As an intervention, AI systems do raise a number of specific challenges and considerations, and this has led to a number of guidance documents to help optimise specific study designs when evaluating an AI intervention. This is being addressed through the publication of AI-specific guidance for different study designs through the EQUATOR network, notably the publication of SPIRIT-AI[[29]](#footnote-29) (for reporting of study protocols) and CONSORT-AI[[30]](#footnote-30) (for reporting of trial reports); additional EQUATOR guidelines are currently in development for diagnostic test accuracy studies (STARD-AI[[31]](#footnote-31)) and studies of prediction models (TRIPOD-AI[[32]](#footnote-32)).

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

* Study design - consider the optimal study design for this intervention that will provide sufficient high-quality evidence across key domains (including efficacy, safety, and cost-effectiveness) to support decision-making by relevant gate-keepers (e.g., regulators, payers, users).
* Population - ensure that the study population (1) reflects the population in which it is intended to be used, and (2) that it is sufficiently diverse to detect under-performance or failure in specific groups;
* Setting - ensure that the study setting reflects the setting (or range of settings) of the intended use; again, diversity of setting is relevant, to provide sufficient confidence of performance outside of ideal scenarios.
* Intervention(s) - ensure that the AI component of any intervention is described sufficiently precisely to ensure results are ascribed to a specific AI system (including version), and would enable replication of the study. This should include product details including version number, supplier and contact details.
* Intervention inputs and outputs- ensure that the following are sufficiently clearly described to enable replication in both trial and clinical deployment contexts (1) the nature of the inputs into the AI system including both human and data elements (such as any data pre-processing); and (2) the nature of the outputs and how this is translated into actions within the healthcare pathway (includes human-computer interaction elements).
* Comparator - the comparator (whether parallel control group or other design) should be a relevant reference. This reference is commonly 'standard practice' or 'best practice' with a view to informing decision-makers as to whether the intervention reflects an improvement (or not) on current health delivery;
* Pre-specified outcomes relevant to all stakeholders - ensure that outcomes include those that are the most important to patients, and the key stakeholder groups; use of core outcome sets are recommended where they exist for the condition of interest; pre-specification avoids bias through retrospective selection of most favourable outcome or of positive result arising through chance and multiple testing.
* Process measures - consider relevant impacts on the overall health pathway such as positive or negative changes in time to diagnosis or treatment.
* Balancing measures - consider upstream, lateral and downstream consequences including changes in behaviour, changes in resource requirements, and potential ethical implications (such as loss of autonomy).
* Protocol deviations - all deviations from study protocol should be recorded and reported. First, such deviations may affect the interpretation of results in relation to pre-specified outcomes. Second, such deviations may provide important information regarding the feasibility and safety of deploying the intervention more widely.
* Analysis - analysis should be pre-specified, and should include sufficient consideration of subgroups to ensure that any deviations of performance and potential risk of harm is detected; errors should be analysed at the individual error level to identify the reasons for failure where possible.
* Reporting of study protocol - the study design should be registered (e.g. on the WHO International Clinical Trials Registry Platform) in advance; additional submission of protocols for publication may enable helpful independent peer review prior to commencement of the study.
* Reporting of study conduct and results - open and transparent reporting should align to the registered protocol include any protocol deviations, and full analysis of planned outcomes according to their pre-specified hierarchy. Participant flow (including exclusions at participant level, exclusions at input data level and losses to follow-up) should be reported according to the CONSORT-AI diagram (adapted from the CONSORT 2010 flow diagram[[33]](#footnote-33)).

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

# Economic evaluation

## Introduction

An important aspect of evaluation for any health technology that is intended to be used at scale is the measurement of the costs (including direct product costs and any associated implementation costs) and wider economic considerations relative to the estimated clinical effect. Economic evaluation is a comparative analysis of two or more interventions in terms of their costs and consequences [REF Drummond 2015], and produces essential evidence for funding decisions about implementing health technologies, whether these investments are made by governments, individuals, companies or donors and development partners. The fundamental concept guiding use of economic evaluation is opportunity cost – the foregone benefits of spending limited resources on one course of action rather than another. By quantifying the costs relative to the consequences through economic evaluation, the best course of action (depending on the objectives of the decision maker), can be informed by the process; whether it be to improve health outcomes or patient experience, generate health systems savings, improve access and equity and a myriad of other potential reasons for investing in a health technology.

Economic evaluation of health interventions examines evidence and uncertainty of an intervention's impact on specific patient populations; comparing differential health benefits and economic costs of more than one alternative (including standard of care and do-nothing scenario if relevant). The intent of cost-effectiveness analysis is to help inform policy so that health interventions that improve patients' lives are recognised, and neither patients nor society overpays for care that does not offer a significant benefit to patients (ICER).

Increasingly, as part of the economic evaluation that enquires into value for money, an estimated extent of health technology uptake and financial implications is required in the form of a budget impact analysis (BIA). The prime purpose of BIA is to assess affordability). BIA addresses anticipated expenditure changes (usually over a 3 to 5 year period) to a specific budget holder that are coupled with a decision to reimburse a new health technology, (York Health Economics Consortium) ass). BIA entails addressing the estimated use and costs of the proposed medical service, an estimation in the changes in use and cost of other 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 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.

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

Inventions in AI supported digital health interventions certainly carry an opportunity cost, but the methods for their economic evaluation require significantly more consideration compared to simple, individually-consumed non-digital health technologies. To address this methodological challenge, the World Bank is engaging in a comprehensive and collaborative effort to develop a framework for economic evaluation of digital health technologies to inform investment, building on existing methodological advancements in the field. The framework is in the stage of consultation and development and is expected for publication and wider application in 2021.

The dynamics of evidence about the clinical effects and costs associated with AI-supported DHIs are central to their economic evaluation. However, the very nature of their evidence is unique. AI-enabled DHIs have a distinct cost profile, where innovation or development costs are substantial and, in some scenarios, at-scale marginal cost that can approach zero. Conversely, the effect is not static and is likely to improve with more use data. Costs and effects are also highly dependent on the local digital architecture and infrastructure, meaning that a generalised approach to economic evaluation (i.e. across a region or grouping of differing contexts) introduces substantial uncertainty. In addition, an AI-enabled DHI will, by definition, produce information through its use – this revolutionizes the real-world evidence (RWE) area of methodological research, enabling economic evaluations to be informed by evidence beyond the clinical trial setting and incorporate evidence from use in clinical practice.

## 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 a national health technology assessment agencies, the objective of an economic evaluation is to inform use of limited resources across the health system, commonly supported by overarching principles including universal health coverage (UHC) and health equity. In this scenario the concept of allocative efficiency is important where the country wishes to distribute resources in a way that maximises outcomes. This has given rise to the common use of cost utility analysis (CUA), where the net incremental costs of an intervention are presented as a ratio to net incremental health's outcomes, and health is a generalised measure such as the quality adjusted life year (QALY) or disability adjusted life year-averted (DALY). In this way a country can assess the likely impact that an intervention can have on "health" - where health is comparable across interventions and diseases, incorporates both positive and negative impacts and represents mortality and morbidity. However, economic evaluation can be used to assess interventions for a variety of different reasons, such as a simple comparison between two or more competing alternative interventions, an analysis within a particular program or institution, or a broader consideration of net benefits under a welfarist framework where cost benefit analysis (CBA) may be employed. Given the variety of methods available and multiple objectives of decision makers who may be assessing AI-enabled digital health interventions, there is no "correct" methodology that should be used, however the analyst and the user of the economic evaluation need to be aware of what the economic evaluation options available and what particular methods mean for central concepts including the opportunity cost of investments, any assumptions that are implicitly incorporated and the uncertainty that this may have on a decision to use, invest in, scale up, or remove an intervention in the health system.

Table 1: Types of comparative economic evaluation models
(extracted from Drummond et al. 2015)

| Type of study | Measurement/‌valuation of costs in both alternatives | Identification of consequences | Measurement/valuation of consequences |
| --- | --- | --- | --- |
| Cost analysis | Monetary units | None | None |
| Cost-effectiveness analysis | Monetary units | Single effect, common to both alternatives, but achieved to different degrees | Natural units (e.g. life years gained, disability days saved etc.) |
| Cost-utility analysis | Monetary units | Single or multiple effects, no necessarily common to both alternatives | Health years (typically measured as quality-adjusted-life-years) |
| Cost-benefit analysis | Monetary units | Single or multiple effects, not necessarily common to both alternatives | Monetary units |

In terms of what has been done to date: Vis et al, 2020 carried out a systematic literature review in 2018 and reported a total of twenty-one distinctive eHTA frameworks are found in literature. A total of 19 of these frameworks address aspects of cost and cost-effectiveness.

## How is economic evidence provided?

Economic evidence can be derived from different methods including: economic evaluations conducted in parallel to clinical trials, literature reviews, registries and dedicated economic evaluation studies that employ health care resource use and quality of life (eg. EQ-5D) instruments to collect the data. Importantly, evidence gathering and assessment should occur both pre and post reimbursement (and ideally be made publicly available) as information 'at the margin' to inform decision making is likely to be improved. The evidence is typically reported via one or more comparative analysis models (listed in Table 1 above), synthesising diverse sources of evidence into a single analytical framework. EE is evidence synthesis by nature – starting with identifying the particular patient population characteristics (including age, gender, and sub-population risk factors), calculating prevalence and incidence rates, identifying the care pathway and appropriate positioning of the technology. With this information modelling and assumptions are a key part. Dealing with uncertainty fundamental: parameter (through probabilistic and/or deterministic sensitive analysis), structural (e.g., local care pathways) and methodological. The approach and methodology for representing the costs of an intervention are a fundamental consideration for any economic evaluation. As the summary output of an economic evaluation is commonly the ratio or aggregation of costs relative to effects, equal importance should be given to demonstrating effect of an intervention and costs of an intervention for a defined population and context. Substantive literature and methodological development have contributed to common understanding of costing methodology for the purposes of economic evaluation in health (ref GHCC, Drummond), however there is little existing guidance on costing approaches tailored to the nuances of an AI enabled DHT.

The basic premise of costing for economic evaluation is that costs should reflect full net costs of the intervention aligning to the specification of the intended decision maker. This requires therefore that when estimating costs, the decision problem, perspective of the decision maker, an understanding of what the intervention would displace in the context of the decision problem should be established. While this principle applies uniformly to costing for any intervention type, some important considerations are required for costing of AI-enabled DHT.

Cost to build the capability:

* build the software
* IP development, or to buy another IP
* regulatory cleared
* Manuals, labelling and infrastructure
* Cost to make the product saleable, e.g. trials that support a US CPT code application, expert consultancy for adoption in different jurisdictions, purchasing of data from different jurisdictions

Cost to maintain the capability

* Software updates for companion technology/technology itself
* Maintaining security

Cost of delivery (dependent on technology)

* Data analyst costs, the time required by them multiplied by their cost, multiplied by some overhead
* Support costs to data analysts.
* Cloud costs
* IP costs for patents licensed

Costs per site:

* Costs to install code and train a site.
* IT costs to initiate connections
* Sales costs to win an adoption site
* There are additional site maintenance costs (management, QC, et)

## Reimbursement

After regulatory approval confirming the safety and effectiveness of a health technology permitting market access, and prior to establishing a significant market penetration, third-party coverage decisions and establishing a reasonable level of reimbursement is required. Pricing of digital health technologies, like other commodities, influences both affordability and access. Initial price setting approaches include price skimming or price penetration (Ingenbleek 2013) and patents enable a period of market exclusivity to recoup R&D costs by delaying the entry of competition. Unlike other typical commodities, there is an imperfect market at play for medical technologies, including DHTs; in both a universal health care system scenario or private health insurance scenario, the consumer (patient) typically does not incur the full cost of the product and further special characteristics of the medical care are described by Arrow (1963). Subsequently, in most countries either the government and/or health insurers exercise a degree of influence to the price and utilisation (through coverage/restrictions of indications (Drummond et al. 1997). Payers often re-evaluate safety and effectiveness evidence as part of the deliberation process; with an objective to reward innovation whilst achieving optimal resource allocation (Barros 2020). Digital health technologies may come in the form of capital associated with a one-off payment (e.g., MRI/CT), or software as a medical device (SaMD), that may be associated with existing capital (for example CT scanner) and the SaMD product is paid on either a subscription or fee-for-service basis. For digital health technologies there are limited national reimbursement opportunities and, within the UK, local negotiations with clinical commissioning groups (CCGs) are required. In the UK, the Innovation and Technology Payment (ITP) is a national reimbursement mechanism for DHTs. To be eligible, products must have NICE support (i.e. MedTech Innovation Briefing or Guidance), demonstrate positive in-year return on investment and be used in at least three NHS organisations (Innovators Guide 2020).

# Deployment of algorithm

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

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

The risk of harm arising from poor generalisability and other performance issues can be considered in terms of a risk matrix of likelihood, and consequence. The likelihood of a fall in performance of the AI system will be increased by the differences between the populations and settings of the deployment phase compared to the test population and setting. Very rapid scaling such as moving to a full nation-wide roll-out based on a successful single 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 including testing *in silico* against external, unseen datasets representing the diversity of the population and setting that the AI system is intended to be deployed into. In most cases this is only partially possible, partly because the data available falls short of the breadth of the population and settings that will be represented within the full deployment. The level of data gap here is a significant risk to the deployment of AI systems, and has been described as a form of *health data poverty* in which the lack of population-appropriate data may prevent safe development, testing, and access to AI systems. Population-specific datasets are therefore a key requisite for enabling countries to benchmark new technologies *in silico* and to provide a degree of assurance prior to exposing their population to them in a care pathway.

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

The deployment phase also provides greater 'real world' information regarding many of the impacts discussed earlier (Clinical studies section), 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.

# Ongoing monitoring

## Introduction

Monitoring of ongoing performance (both safety and effectiveness) is important to determine whether the AI product continues to deliver as expected. AI systems are known to show poor generalisability when encountering new data and unexpected failure in spurious edge cases. Even in the presence of evidence supporting good performance across an aggregate population, it is important to be prepared for unexpected algorithmic outputs and potential adverse outcomes. Additionally, variations in clinical workflow may negatively impact on the overall intended benefits of an AI system.

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

## Regulatory requirements

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

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

## Relevant stakeholders for post deployment monitoring

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

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

## Algorithmic audits

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

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

# Recommendations

1. Procurers of AI tools, should be clear about the economic evaluation that is required for AI tools. Not only is this important to ensure comprehensive evaluation, but it also builds trust in the evaluation system. While it is possible to find examples of where digital strategies explain what kind of evaluation should be done, it is rare to find publicly available examples of where this has been done, and this can undermine public trust in procurement.
2. Priority setting for digital tools in all country settings requires a much more active role for health technology assessment, in addition to the role of regulators. Where fulfilled by two different agencies, close dialogue should be maintained. Where HTA agencies already exist, these need to expand the skillset to include digital interventions. In countries where HTA capacity is low, this should be a key focus alongside expanding digital health tool.
3. Benchmarking of AI tools either by local procurers or by national agencies is likely to important to ensure local performance is acceptable, and to compare performance of tools on particular datasets. The Open Code Initiative is part of the FG-AI4H that is actively developing software that can be used by agencies to do this on their own datasets.
4. Long term analysis of AI tools is required- despite the expanding number of tools, there is a paucity of long-term studies with clinical endpoints and rigorous safety analyses and this is holding back the potential of these tools. There is a lack of evidence on how transferable tools are that are developed in one setting and then used in another. Collaborative studies would accelerate progress and should be considered a priority.
5. Needs based development of tools requires a dedicated effort to collect data in populations that are currently underrepresented and for clinical problems where AI may be effective, but datasets are poor. Not only does this facilitate development but also evaluation. Organisations like I-DAIR have identified this as a key task.
6. All stakeholders must be encouraged to make clinical studies open and accessible. Availability of the clinical evaluation plan is key to building trust and should be made publicly available. Developers should be required to provide comprehensive information about the characteristics of the data underpinning the tool, and details of evaluation, for example as proposed in model facts.

# Abbreviations

[tbc]

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

Bibliography

References for clause 6 (Economic evaluation)

Guidelines

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

References

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

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

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

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

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

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

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

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

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

1. https://www.gov.uk/guidance/get-started-evaluating-digital-health-products [↑](#footnote-ref-1)
2. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. [https://doi.org/10.1038/s41591-020-1034-x](https://www.google.com/url?q=https://doi.org/10.1038/s41591-020-1034-x&sa=D&source=editors&ust=1618395475514000&usg=AOvVaw00JlBdiIlGooxhAHDKsiME). [↑](#footnote-ref-2)
3. Cruz Rivera, S., Liu, X., Chan, AW. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26, 1351–1363 (2020). [https://doi.org/10.1038/s41591-020-1037-7](https://www.google.com/url?q=https://doi.org/10.1038/s41591-020-1037-7&sa=D&source=editors&ust=1618395475516000&usg=AOvVaw3yE7p6aBOS6QbRnj2EQ-ts) [↑](#footnote-ref-3)
4. Sendak, M.P., Gao, M., Brajer, N. *et al.* Presenting machine learning model information to clinical end users with model facts labels. *npj Digit. Med.* 3, 41 (2020). <https://doi.org/10.1038/s41746-020-0253-3> [↑](#footnote-ref-4)
5. Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A. W., Calderon-Ramirez, S., ... & Wiegand, T. (2020, November). ML4H Auditing: From Paper to Practice. In *Machine Learning for Health* (pp. 280-317). PMLR. [↑](#footnote-ref-5)
6. https://www.designkit.org/methods#filter [↑](#footnote-ref-6)
7. https://pair.withgoogle.com/chapter/user-needs/ [↑](#footnote-ref-7)
8. https://www.designkit.org/methods#filter [↑](#footnote-ref-8)
9. https://pair.withgoogle.com/chapter/user-needs/ [↑](#footnote-ref-9)
10. A guide to good practice for digital and data-driven health technologies, <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology> [↑](#footnote-ref-10)
11. Kling R. The Organizational Context of User-Centered Software Designs. MIS Q. 1977;1(4):41-52. doi:10.2307/249021 [↑](#footnote-ref-11)
12. https://www.ideou.com/pages/design-thinking [↑](#footnote-ref-12)
13. Australia Government Department of Health, Stakeholder Engagement Framework, AccountAbility. AA1000SE Stakeholder Engagement Practitioner’s Perspectives. 2005. [↑](#footnote-ref-13)
14. https://www.gov.uk/guidance/user-feedback-study-descriptive-studies [↑](#footnote-ref-14)
15. https://pair.withgoogle.com/chapter/feedback-controls/ [↑](#footnote-ref-15)
16. https://www.gov.uk/guidance/interview-study-qualitative-studies [↑](#footnote-ref-16)
17. https://www.gov.uk/guidance/usability-testing-qualitative-studies [↑](#footnote-ref-17)
18. https://www.gov.uk/guidance/focus-group-study-qualitative-studies [↑](#footnote-ref-18)
19. <https://www.gov.uk/guidance/ethnographic-study-qualitative-studies> [↑](#footnote-ref-19)
20. Macdonald J., März M., Oala L., Samek W. (2021) Interval Neural Networks as Instability Detectors for Image Reconstructions. In: Palm C., Deserno T.M., Handels H., Maier A., Maier-Hein K., Tolxdorff T. (eds) Bildverarbeitung für die Medizin 2021. Informatik aktuell. Springer Vieweg, Wiesbaden. https://doi.org/10.1007/978-3-658-33198-6\_79 [↑](#footnote-ref-20)
21. Hägele, M., Seegerer, P., Lapuschkin, S. *et al.* Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci Rep* **10,** 6423 (2020). https://doi.org/10.1038/s41598-020-62724-2 [↑](#footnote-ref-21)
22. **Data and artificial intelligence assessment methods (DAISAM) reference** Oala, Luis, Balachandran, Pradeep, Cabitza, Federico, Calderon Ramirez, Saul, Chiavegatto Filho, Alexandre, Eitel, Fabian, Extermann, Jérôme, Fehr, Jana, Ghozzi, Stephane, Gilli, Luca, Jaramillo-Gutierrez, Giovanna, Kester, Quist-Aphetsi, Kurapati, Shalini, Konigorski, Stefan, Krois, Joachim, Lippert, Christoph, Martin, Jörg, Merola, Alberto, Murchison, Andrew, Niehaus, Sebastian, Ritter, Kerstin, Samek, Wojciech, Sanguinetti, Bruno, Schwerk, Anne, and Srinivasan, Vignesh

*In Proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting I* 2020 [↑](#footnote-ref-22)
23. **Good practices for health applications of machine learning: Considerations for manufacturers and regulators** Johner, Christian, Balachandran, Pradeep, Oala, Luis, Lee, Aaron .Y., Leite, Alixandro Werneck, Murchison, Andrew, Lin, Anle, Molnar, Christoph, Rumball-Smith, Juliet, Baird, Pat, Goldschmidt, Peter. G., Quartarolo, Pierre, Xu, Shan, Piechottka, Sven, and Hornberger, Zack. In *Proceedings of the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting K* 2021 [↑](#footnote-ref-23)
24. From: A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Faes et al. TVST. Feb 2020, Vol 9,7. [↑](#footnote-ref-24)
25. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019. [https://doi.org/10.1016/S2589-7500(19)30123-2](https://www.google.com/url?q=https://doi.org/10.1016/S2589-7500(19)30123-2&sa=D&source=editors&ust=1618395475511000&usg=AOvVaw0m1syvBkK69z8diafHfLai). [↑](#footnote-ref-25)
26. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. [↑](#footnote-ref-26)
27. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998;316(7126):201. [↑](#footnote-ref-27)
28. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869 [↑](#footnote-ref-28)
29. Cruz Rivera, S., Liu, X., Chan, AW. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26, 1351–1363 (2020). [↑](#footnote-ref-29)
30. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. [https://doi.org/10.1038/s41591-020-1034-x](https://www.google.com/url?q=https://doi.org/10.1038/s41591-020-1034-x&sa=D&source=editors&ust=1618395475514000&usg=AOvVaw00JlBdiIlGooxhAHDKsiME). [↑](#footnote-ref-30)
31. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, Karthikesalingam A, King D, Liu X, Markar SR, McInnes MDF, Panch T, Pearson-Stuttard J, Ting DSW, Golub RM, Moher D, Bossuyt PM, Darzi A. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med. 2020 Jun;26(6):807-808. doi: 10.1038/s41591-020-0941-1. PMID: 32514173. [↑](#footnote-ref-31)
32. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577–9. [↑](#footnote-ref-32)
33. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020. [https://doi.org/10.1038/s41591-020-1034-x](https://www.google.com/url?q=https://doi.org/10.1038/s41591-020-1034-x&sa=D&source=editors&ust=1618395475507000&usg=AOvVaw2cDvol60Z9S9lm1-q_BOcf). [↑](#footnote-ref-33)